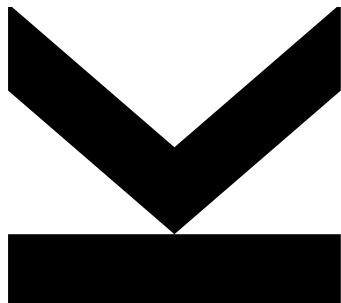


Dr. Mario Ullrich
Institut für Analysis

Version of
February 7, 2023

Mathematics for Artificial Intelligence 1–3



lecture notes (in progress) – winter semester 2022

Preface

These lecture notes belong to the lecture of the same name, and were produced starting in the winter semester 2019. Big part of the work has been done by Julian Hofstadler and Corinna Perchtold, who were employed as undergraduate assistants in this period for the preparation of these notes and corresponding slides.

This is only the second time that this lecture is held at the JKU Linz and therefore, these notes are far from being perfect. However, many parts are taken (or merged) from the repeatedly used lecture notes of Prof. Aicke Hinrichs (“Analysis für Lehramt”, JKU, 2018), Prof. Andreas Neubauer (“Mathematics for Chemistry”, JKU, 2017) and myself (“Klassische Harmonische Analysis”, JKU, 2017). I thank both colleagues for the permission to do so.

Suggestions for improvements and additional comments are appreciated!

Mario Ullrich
(mario.ullrich@jku.at)

October 2022

Contents

Preface	2
1 Sets, numbers and functions	6
1.1 Sets and notation	6
1.2 Relations and functions	11
1.3 Real numbers	23
1.4 Bounded sets, infimum and supremum	26
1.5 Induction and combinatorics	30
1.6 Absolute value	34
1.7 Some elementary functions	36
1.8 Complex numbers	43
1.9 Vectors and norms	48
2 Matrices and systems of linear equations	53
2.1 Matrices	54
2.2 Systems of linear equations	62
2.3 Gaussian elimination	66
2.4 The determinant	78
2.5 Cramer's rule	84
2.6 Inverse matrices	87
3 Sequences and series	93
3.1 Convergence of sequences	94
3.2 Calculation rules for limits	98
3.3 Monotone sequences	103
3.4 Subsequences and accumulation points	107
3.5 Cauchy criterion	112
3.6 Series	114
3.7 Convergence tests	120
3.7.1 Comparison test	121
3.7.2 Root test	122
3.7.3 Ratio test	124
3.7.4 Cauchy's condensation test	126
3.7.5 Leibniz criterion	127
3.8 Power series	128

4 Continuous functions and limits	132
4.1 Calculation rules of continuous functions	136
4.2 Limits of functions	138
4.3 Intermediate and extreme value theorem	148
4.4 Other types of continuity	153
5 Differential calculus	159
5.1 Calculation rules for differentiable functions	163
5.2 Global and local extrema	167
5.3 Mean value theorem and l'Hospital's rule	171
5.4 Monotonicity and convexity	175
5.5 Taylor's theorem	178
5.6 (*) Newton's method	185
6 Basic integration theory	187
6.1 Antiderivatives	188
6.2 Calculation rules for antiderivatives	191
6.3 A first definition of the integral	196
6.4 The fundamental theorem of calculus	202
6.5 Improper integrals	207
6.6 Piecewise continuous functions	210
7 Fourier series	211
7.1 Periodic functions and trigonometric polynomials	212
7.2 Fourier coefficients and Fourier series	215
7.3 First convergence theorems	220
7.4 The theorem of Dirichlet	229
8 Multivariate Calculus	230
8.1 Sequences in \mathbb{R}^d	231
8.2 Continuous functions	235
8.3 Differential calculus	239
8.3.1 Partial derivatives	239
8.3.2 (Total) differentiability	242
8.3.3 Directional derivatives	248
8.3.4 Higher order partial derivatives	249
8.4 Extrema	253

8.4.1	Extrema subject to constraints	263
8.5	Differential calculus for vector-valued functions	268
8.6	Taylor series	274
8.7	Multiple integrals	280
9	Matrices II	290
9.1	Eigenvalues and eigenvectors	290
9.2	Diagonalization	299
9.3	Singular value decomposition	303
10	Basic measure theory and the Lebesgue integral	306
10.1	Measurable functions	312
10.2	The Lebesgue integral	315
10.3	Lebesgue's theorem	324
10.4	Product measures and Fubini's theorem	330
10.5	Connection to probability theory	339
10.5.1	Some special distributions	343
11	Basic functional analysis	346
11.1	Vector spaces	346
11.2	Normed spaces	356
11.2.1	The L_p -spaces	359
11.2.2	Sequences in normed spaces and Banach spaces	367
11.2.3	Metric spaces	373
11.3	Inner products and Hilbert spaces	373
11.3.1	Reproducing kernel Hilbert spaces	380

1 Sets, numbers and functions

In this section we will introduce the most fundamental objects of mathematics – namely sets and its elements, numbers, and relations between them. We then introduce some more involved objects like supremum and infimum, make an excursion to induction and combinatorics, and discuss some special functions that will be important later on. Finally, we treat complex numbers, which are necessary to give solutions to arbitrary polynomial equations.

All these things are the (grammatical) basis on which we build upon. It is therefore essential to understand and memorize every part of this section, **like an alphabet in a foreign language**.

1.1 Sets and notation

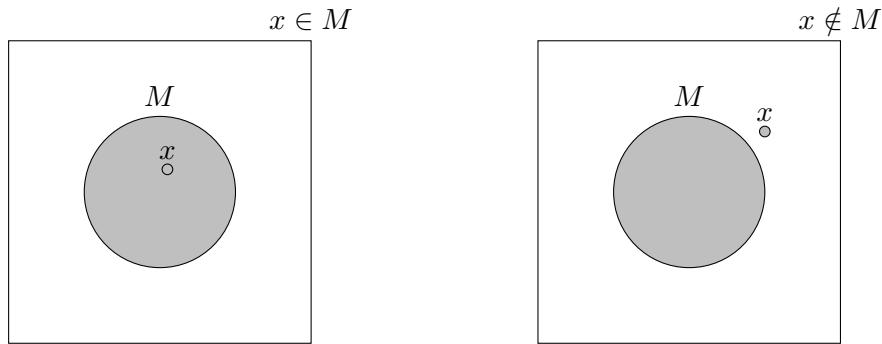
A **set** M is a collection of different 'objects' which we call **elements of M** . This rather intuitive description of a set was first given by Georg Cantor (1845–1918). We use the following notation:

x belongs to M , write $x \in M$,

or

x is not in M , write $x \notin M$.

A typical visualization of a set with an element is:



Some very important (and partly well-known) sets of *numbers* together with their 'symbol' are:

- $\mathbb{N} := \{1, 2, 3, \dots\}$ is the set of **natural numbers**.
- $\mathbb{N}_0 := \{0, 1, 2, 3, \dots\}$ is the set of **natural numbers with zero**.
- $\mathbb{Z} := \{\dots, -3, -2, -1, 0, 1, 2, 3 \dots\}$ is the set of **integer numbers**.
- \mathbb{Q} is the set of **rational numbers**.
- \mathbb{R} is the set of **real numbers**.
- \mathbb{C} is the set of **complex numbers**.

(Note that we write “ $:=$ ” instead of “ $=$ ” if the equation is meant as a **definition**.)

All these sets will be precisely introduced and discussed later in this chapter. First, let us see that there are multiple ways to define sets. The easiest way would be to list all its elements, as for:

- $A := \{0, 1, 2\}$
- $B := \{\text{Artificial Intelligence, Mathematics, Physics, Informatics}\}$

However, if we have sets containing an infinite amount of elements we cannot name all of them. In this case we use dots if it is clear what is contained in the set. For example, we might write

- $G := \{2, 4, 6, \dots\}$ and
- $U := \{1, 3, 5, \dots\}$

for the even and odd natural numbers. However, this may lead to difficulties of interpretation as this is not a unique description. For example, we may define the set of natural numbers by

$$\mathbb{N} := \{1, 2, 3, \dots\} = \{1, 2, 3, 4, 5, 6, 7, \dots\},$$

and the set of all *prime numbers* by

$$\mathbb{P} = \{1, 2, 3, \dots\} = \{1, 2, 3, 5, 7, 11, \dots\},$$

which are not distinguishable, if we only list the first 3 elements.

For a unique definition, it is therefore formally necessary to either *list all elements* of a set or to *precisely specify the properties* of its elements, like in

- $\mathbb{P} := \{n \in \mathbb{N} : n \text{ is prime}\}$
- $G := \{n \in \mathbb{N} : n \text{ is an even number}\}.$

A special but important set is the **empty set**, short \emptyset , which does not contain any element, i.e. $\emptyset = \{\}$.

Two sets M and N can also be related to each other. If for arbitrary $m \in M$ we also have $m \in N$ then we say M is a **subset** of N , and we write $M \subset N$ or $M \subseteq N$. In this case, we also call N a **superset** of M . If we have a look at the sets defined above, we have e.g. $G \subset \mathbb{N}$ and $U \subset \mathbb{N}$. Note that for any set M we have the obvious relations $M \subset M$ and $\emptyset \subset M$.

Sets M and N are called **equal** if they contain the same elements, i.e. $M \subset N$ and $N \subset M$. For example we have

$$A = \{0, 1, 2\} = \{0, 0, 0, 1, 1, 1, 2\} =: \tilde{A}.$$

To verify this equation we have to show that $A \subset \tilde{A}$ and $\tilde{A} \subset A$ and start by showing $A \subset \tilde{A}$. Obviously $0 \in \tilde{A}$, $1 \in \tilde{A}$ and $2 \in \tilde{A}$, hence by definition we have $A \subset \tilde{A}$. The other way around, i.e. $\tilde{A} \subset A$, is left as an exercise. **Note that multiplicities are irrelevant in sets.**

If we have $M \subset N$ and $M \neq N$, then we say that M is a **proper or strict subset** of N and write $M \subsetneq N$. For example $\mathbb{N} \subsetneq \mathbb{N}_0 \subsetneq \mathbb{Z}$.

Remark 1.1. Some authors prefer to use “ \subseteq ” instead of “ \subset ” to indicate that equality is not excluded. (And we also do so sometimes.) The same authors may use “ \subset ” instead of “ \subsetneq ” for proper subsets. So, one might be careful when using different literature.

Sets may contain other sets. An important example is the **power set** $\mathcal{P}(M)$ for a set M , which is the set of all possible subsets of M , i.e.,

$$\mathcal{P}(M) := \{A : A \subset M\}.$$

Consider once more the set $A = \{0, 1, 2\}$, then its power set is given by:

$$\mathcal{P}(A) = \{\emptyset, \{0\}, \{1\}, \{2\}, \{0, 1\}, \{0, 2\}, \{1, 2\}, \{0, 1, 2\}\}.$$

Note that we always have $M \in \mathcal{P}(M)$ and $\emptyset \in \mathcal{P}(M)$. (Important: The statement $M \subset \mathcal{P}(M)$ is usually false! M contains elements, and $\mathcal{P}(M)$ contains sets of elements.)

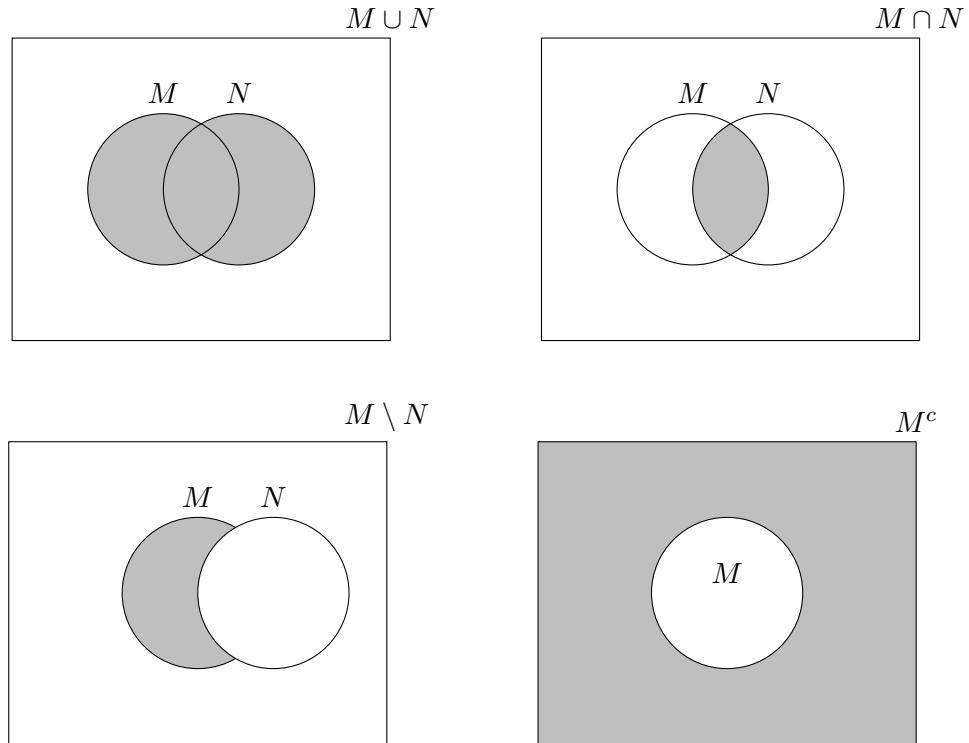
We can also create new sets from given sets, say M and N , by using **set operations**:

The **union** $M \cup N$ contains all elements which belong to the set M and all elements which belong to the set N .

The **intersection** $M \cap N$ consists of all elements which are in both sets M and N .

The **difference** (or relative complement) of M and N , written as $M \setminus N$, is the set of all elements of M which are not contained in N .

If we only work with subsets $M \subset \Omega$ for a fixed set Ω , then we call Ω the **underlying set** or the **universal set**. In this case, we work with the notation $M^c = \Omega \setminus M$ for the **complement** of M (in Ω).



The illustrations above are called **Venn-diagrams** (John Venn, 1834-1923) and are a good tool when working with sets.

However, we often need precise definitions of these set operations in mathematical language.

Definition 1.2. Let M, N be sets in an universal set Ω . We define

- the **union** of M and N by

$$M \cup N := \{x : x \in M \text{ or } x \in N\},$$

- the **intersection** of M and N by

$$M \cap N := \{x : x \in M \text{ and } x \in N\},$$

- the **difference** of M and N

$$M \setminus N := \{x \in M : x \notin N\}$$

- and the **complement** of M (in Ω)

$$M^c := \{x \in \Omega : x \notin M\}.$$

Elements of sets are not ordered since $\{a, b\} = \{b, a\}$. Nevertheless, it is often important to order the objects under consideration.

Definition 1.3 (Tuples and Cartesian product). Let A and B be sets, and $a \in A$ and $b \in B$ arbitrary elements.

- The expression (a, b) , which is sensitive to order, is called a **tuple** or an ordered pair.
- Two tuples (a, b) and (a', b') are equal if and only if $a = a'$ and $b = b'$.
- The set of all tuples

$$A \times B := \{(a, b) : a \in A, b \in B\}$$

is called the **Cartesian product** of the sets A and B .

Note that

- a tuple (a, b) and a set $\{a, b\}$ are completely different objects.
- if we consider more than two sets, say A_1, A_2, \dots, A_d for some $d \in \mathbb{N}$, then we can also define the **(d -fold) Cartesian product**

$$A_1 \times A_2 \times \dots \times A_d := \{(a_1, \dots, a_d) : a_i \in A_i \text{ for all } i = 1, \dots, d\},$$

whose elements (a_1, \dots, a_d) are called **(d -)tuples**.

Example 1.4. Let $A = \{x, y\}$ and $B = \{1, 2, 3\}$, then

$$A \times B = \{(x, 1), (x, 2), (x, 3), (y, 1), (y, 2), (y, 3)\}$$

but

$$B \times A = \{(1, x), (1, y), (2, x), (2, y), (3, x), (3, y)\}.$$

Let us finally fix some other **mathematical language** to make writing mathematical statements more 'elegant'. To this end, we start with the **universal** and the **existential quantifier**, which form the basis of many expressions in mathematical language. First, a **proposition** P is an expression which can either be true or false, like $1 = 1$ or $0 = 1$. If we have that $P(m)$ is a proposition for all elements m of a set M , then we say that $P(\cdot)$ is a **predicate** for M .

Definition 1.5. Let $P(\cdot)$ be a predicate for M , i.e., $P(m)$ is a proposition for every $m \in M$. The **universal quantifier** builds a proposition

$$\forall m \in M: P(m),$$

which is true if and only if *for all* $m \in M$ the proposition $P(m)$ is true.

The **existential quantifier** builds a proposition

$$\exists m \in M: P(m),$$

which is true if and only if there exists *at least one* $m \in M$ such that $P(m)$ is true.

The **uniqueness quantifier** for which

$$\exists! m \in M: P(m)$$

is true if and only if there exists *exactly one* $m \in M$ such that $P(m)$ is true.

Example 1.6. Consider the set $M = \{0, 1, 2\}$ and set $P(m) = (m > 1)$. Inserting all the elements of M into $P(\cdot)$ we get $(0 > 1)$, $(1 > 1)$ and $(2 > 1)$. Clearly, only the last statement is true. Hence, $(\forall m \in M: P(m)) = (\forall m \in M: m > 1)$ is a wrong proposition, while $\exists m \in M: m > 1$ is true. We even have that $\exists! m \in M: m > 1$ is true.

Example 1.7. With these quantifiers we can also give a more mathematical (or 'elegant') definition of a subset. We have that $M \subset N$ if and only if

$$\forall x \in M: x \in N.$$

Moreover, we have $M \subsetneq N$ if and only if $M \subset N$ and $\exists x \in N: x \notin M$.

As you might have already noticed, we will often need the terms "if" or "if and only if", and therefore we define a mathematical symbol for them. Let A and B be two propositions. Then,

- " $A \implies B$ " means " A implies B ." or "If A , then B ."
- " $A \iff B$ " means " A is true if and only if B is true". That is, $A \implies B$ and $B \implies A$.

(We also use "**iff**" as abbreviation of "if and only if".)

With all these notations, we may write certain definitions or statements without any 'usual word' and exclusively with mathematical symbols. For example,

$$M \subset N : \iff (\forall x \in M: x \in N) \iff (\forall x: x \in M \implies x \in N).$$

(Again, we use " $: \iff$ " instead of " \iff " to indicate that this is actually a definition.)

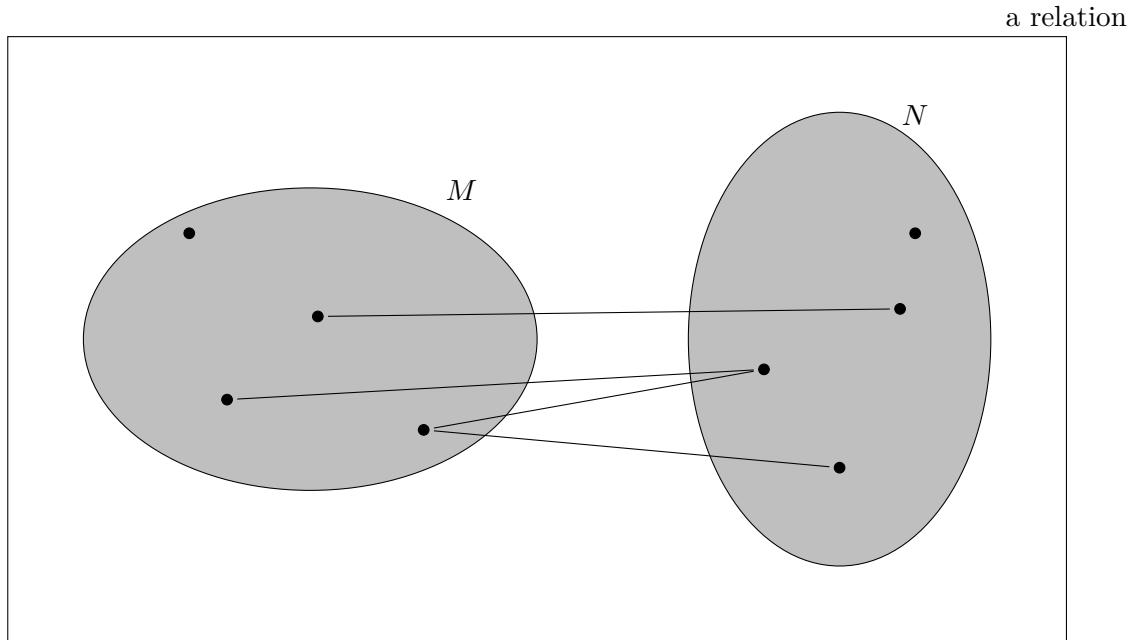
All of these statements are just mathematical language for "Every element of M is also an element of N " or "If an element is in M , then it is also in N ". However, such short (and elegant and exact) notation is often beneficial.

1.2 Relations and functions

Roughly speaking, relations shall describe connections between two objects. Here, we give a formal description and important properties. We then introduce functions, which are used to “map” every element of a set to something different, and discuss special relations that are used to compare, group or order elements of a given set.

Definition 1.8. A **relation** R between two sets M and N is a subset of the cartesian product of M and N , i.e. $R \subseteq M \times N$.

To make things clearer we have a look at the upcoming illustration, which depicts every element of R as a line. As you can see it is possible that $x \in M$ is “connected” to some $y \in N$, which we denote by $(x, y) \in R$. However, this does not have to be the case for every $x \in M$, and, different elements of M may be mapped to the same $y \in N$. Moreover, $x \in M$ can be mapped to more than one element in N .



Example 1.9. Let $M = \{\text{Anna, Philipp, Kevin, Julia}\}$ and $N = \{\text{Corinna, Jakob, Anja}\}$. Now we define a relation $R \subseteq M \times N$, where we have $(x, y) \in R$ if and only if the first letter of x equals the first letter of y . Clearly we have $R = \{(\text{Anna, Anja}), (\text{Julia, Jakob})\} \subsetneq M \times N$.

Now we head to a very important type of relation, i.e., **functions**, which assign to each element of M exactly one element of N .

Definition 1.10. Let $M, N \neq \emptyset$.

We call $f: M \rightarrow N$ a **function** from M to N , if each $x \in M$ is assigned exactly one $f(x) \in N$.

The **mapping rule** is written in the following way $x \mapsto f(x)$.

M is called **domain** (of definition) and N **codomain** of f .

Let $S \subset M$. We define the **image** of S under f as

$$f(S) := \{f(x) : x \in S\} \subseteq N,$$

and the **range** of f as

$$f(M) := \{f(x) : x \in M\} \subseteq N.$$

Moreover, for $T \subset N$ we define the **preimage** of T under f by

$$f^{-1}(T) := \{x : f(x) \in T\} \subseteq M.$$

To show the connection between functions and relations, we define the following.

Definition 1.11. Let $f: M \rightarrow N$ be a function. We define the **graph** of f as

$$G_f := \{(x, f(x)) : x \in M\} \subset M \times N.$$

Note that the graph of a function is a relation. In this sense, all functions induce a relation, but not vice versa.

We can visualize real valued functions by plotting its graph in a usual coordinate system (in \mathbb{R}^2). For $f(x) = x^2$ and $f(x) = x + 1$ this is demonstrated in the next illustration.

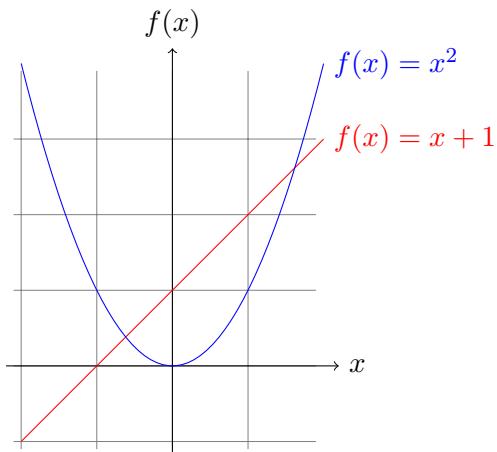


Figure 1: The graph of x^2 and $x + 1$

In what follows, we will define several important properties of relation. We will always demonstrate afterwards what this means for functions.

Definition 1.12. Let $R \subseteq M \times N$ be a relation. R is called

- **injective** if and only if

$$\forall(x_1, y_1), (x_2, y_2) \in R: x_1 \neq x_2 \Rightarrow y_1 \neq y_2,$$

which is equivalent to

$$\forall(x_1, y_1), (x_2, y_2) \in R: y_1 = y_2 \Rightarrow x_1 = x_2.$$

- **surjective** if and only if

$$\forall y \in N \exists x \in M: (x, y) \in R.$$

- **bijective** if and only if it is injective and surjective.

- **functional** if and only if

$$\forall x \in M, y_1, y_2 \in N: (x, y_1), (x, y_2) \in R \Rightarrow y_1 = y_2.$$

Note that the graph of a function is a functional relation, and vice versa. We can therefore rephrase the above definitions for functions. If $f: M \rightarrow N$ is a function we say:

$$f \text{ is injective} \iff \forall x_1, x_2 \in M: x_1 \neq x_2 \Rightarrow f(x_1) \neq f(x_2)$$

$$f \text{ is surjective} \iff \forall y \in N \exists x \in M: f(x) = y$$

$$f \text{ is bijective} \iff \forall y \in N \exists! x \in M: f(x) = y.$$

We call injective, surjective and bijective function also **injections**, **surjections** and **bijections**, respectively.

Let us see some illustrations for a better understanding.

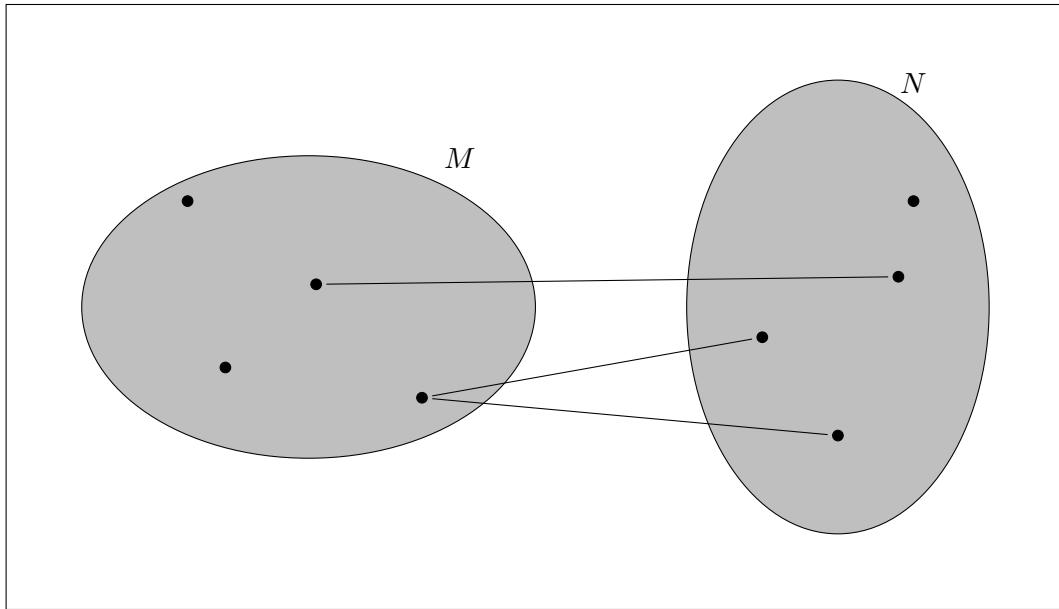


Figure 2: injective relation

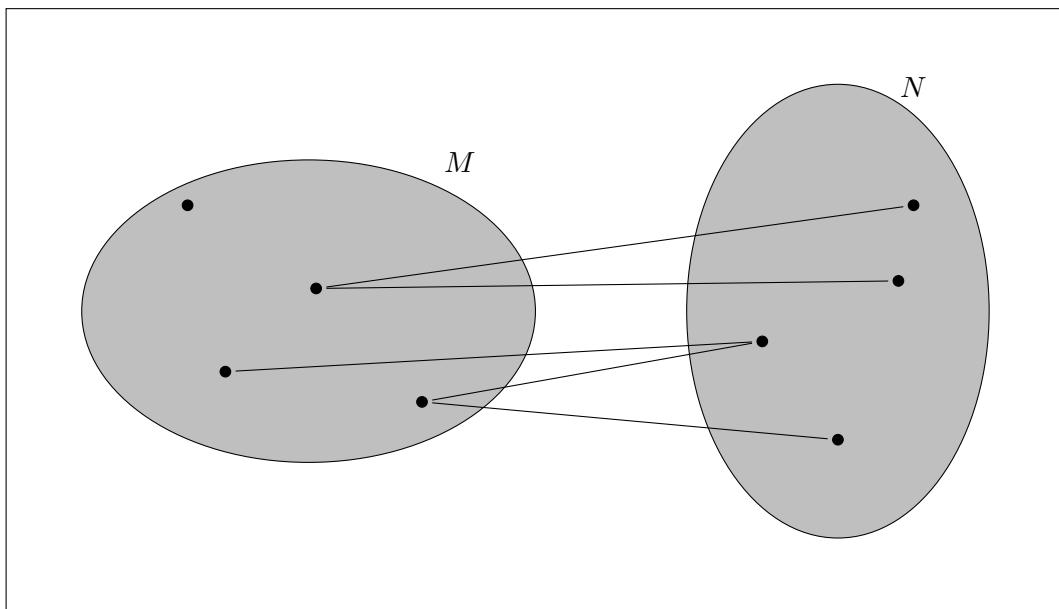


Figure 3: surjective relation

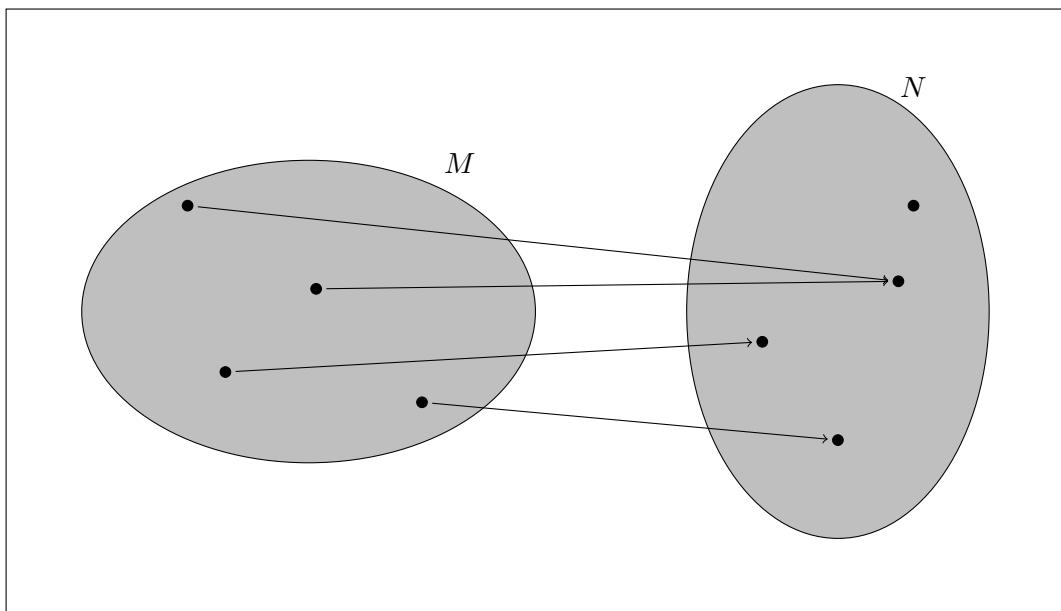


Figure 4: a function (not injective and not surjective)

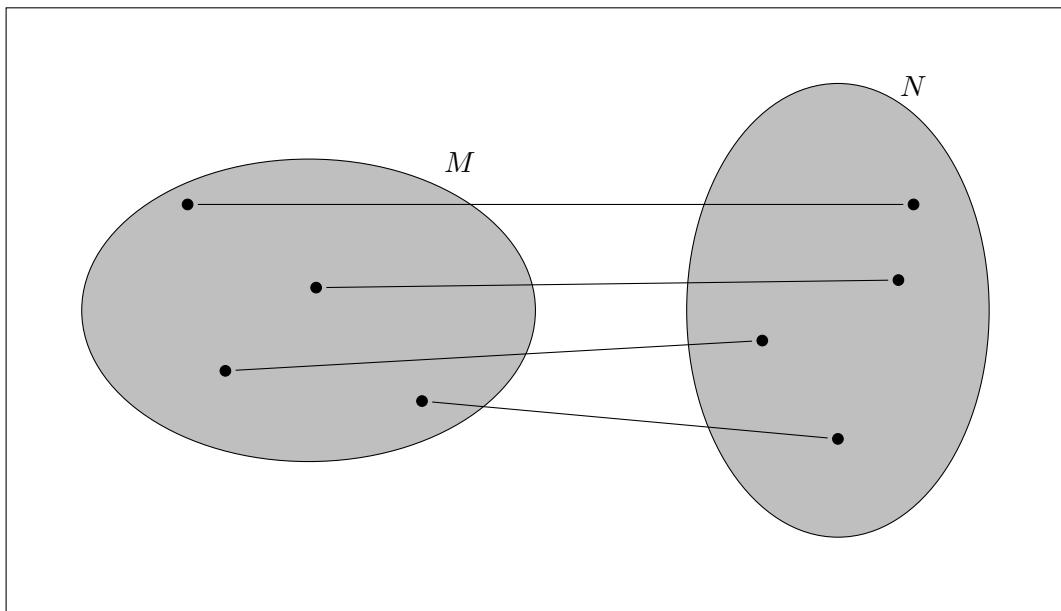


Figure 5: bijective function

We will now define two functions that can in principle be defined on arbitrary sets M . First, we define the **identity function** $Id_M: M \rightarrow M$ which maps each element to itself, i.e., $x \mapsto x$.

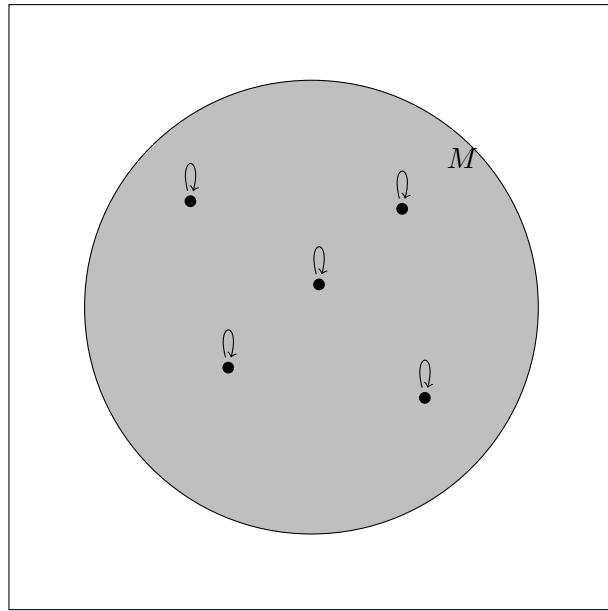


Figure 6: The identity $Id_M: M \rightarrow M$

Let M and N be arbitrary nonempty sets and $c \in N$ be fixed. The function $f: M \rightarrow N$, $x \mapsto c$ is called **constant function**, and is often just denoted by $f = c$.

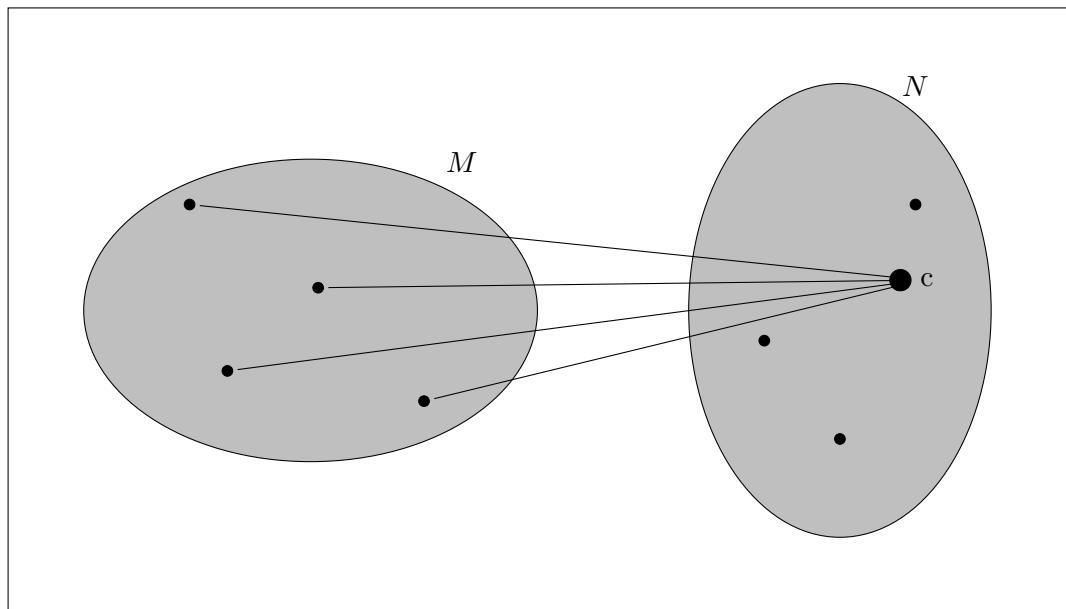


Figure 7: The constant function $f \equiv c$

We now ask ourselves the question: If $f: M \rightarrow N$ is a function and $y \in N$ is given, is there some $x \in M$ such that $f(x) = y$ and is this x unique? This leads us to the concept of an inverse function.

Definition 1.13. Let $f: M \rightarrow N$ and $g: N \rightarrow M$ be functions with the properties

$$\forall x \in M: g(f(x)) = x$$

and

$$\forall y \in N: f(g(y)) = y,$$

then f and g are **inverses** of each other.

In this case we write $f^{-1} := g$ and $g^{-1} := f$ and call f (or g) **invertible**.

Note that we used the notation f^{-1} already for the preimage, see Definition 1.10. There, the input was a set, and the preimage was defined for any function f . For an invertible function $f: M \rightarrow N$ we have, by definition, $f^{-1}(\{f(x)\}) = \{x\}$ and $f(\{f^{-1}(y)\}) = \{y\}$ for any $x \in M$ and $y \in N$. In particular, the preimage of any one-element subset of N has also exactly one element. So, this notation makes sense, if we identify $f^{-1}(y)$ with the unique element in $f^{-1}(\{y\})$.

Remark 1.14. If $f: M \rightarrow N$ is an invertible function and $f^{-1}: N \rightarrow M$ is the inverse function of f , then we have $f^{-1}(f(x)) = Id_M(x)$ and $f(f^{-1}(y)) = Id_N(y)$.

Example 1.15. Let \mathbb{R}^+ be all positive real numbers, and $f: \mathbb{R}^+ \rightarrow \mathbb{R}^+$, $x \mapsto x^2$ and $g: \mathbb{R}^+ \rightarrow \mathbb{R}^+$, $y \mapsto \sqrt{y}$. For fixed $x > 0$ we have $g(f(x)) = \sqrt{x^2} = x$. On the other hand we have $f(g(y)) = (\sqrt{y})^2 = y$, for arbitrary $y > 0$. Therefore, f and g are inverses of each other.

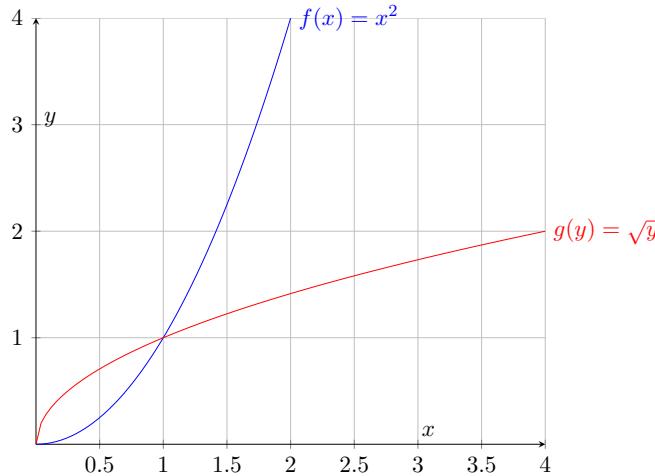


Figure 8: The function $f(x) = x^2$ and its inverse $f^{-1}(x) = \sqrt{x}$

The following theorem provides us with a tool to check if a function has an inverse or not. However, finding a closed formula is often much harder or even impossible. **This is one of the main reasons for using numerical software** and approximations. We will see several examples later during the course.

Theorem 1.16. Let $f: M \rightarrow N$ be a function. Then,

$$f \text{ is invertible} \iff f \text{ is bijective.}$$

Proof. We want to prove an equivalence, and therefore have to prove *both directions*.

First we are going to prove “ \Rightarrow ”:

Since f is invertible, there exists $f^{-1}: N \rightarrow M$ such that

$$f^{-1}(f(x)) = x \quad \forall x \in M$$

and

$$f(f^{-1}(y)) = y \quad \forall y \in N.$$

Recall that bijective means surjective and injective. We show that f is surjective, i.e.,

$$\forall y \in N \exists x \in M : f(x) = y.$$

Let $y \in N$ be arbitrary but fixed. Set $x_0 := f^{-1}(y)$. Since f^{-1} is a function, x_0 is unique. Further $f(x_0) = f(f^{-1}(y)) = y$. Thus for arbitrary $y \in N$ we found a $x \in M$ such that $f(x) = y$. This shows that f is surjective.

Secondly we verify that f is injective. Let $f(x_1) = f(x_2)$ for some $x_1, x_2 \in M$. Since $f(x_1) \in N$ and $f(x_2) \in N$ we may apply f^{-1} and get

$$x_1 = f^{-1}(f(x_1)) = f^{-1}(f(x_2)) = x_2.$$

This shows that f is injective.

We now prove the reverse direction “ \Leftarrow ”: Assume f is bijective and check that f is invertible.

Since f is bijective we have $\forall y \in N \exists! x \in M : f(x) = y$. So for each $y \in N$ we can find a *unique* $x \in M$ such that $f(x) = y$. Now we define $g: N \rightarrow M$ such that it maps each $y \in N$ to this unique $x \in M$, i.e. $g(y) = x$. Moreover we have $f(g(y)) = f(x) = y$ and $g(f(x)) = g(y) = x$.

$\Rightarrow f$ is invertible with $g = f^{-1}$.

□

Invertible (or bijective) functions may be used to **formally define the cardinality of a set**. For this, note that the existence of a bijective function $f: M \rightarrow N$ means that there is a **one-to-one correspondence between M and N** . In particular, both sets must have the same cardinality (aka. size). This means that a set M has cardinality $n \in \mathbb{N}$, i.e., $|M| = n$, if and only if there is a bijective function $f: \{1, \dots, n\} \rightarrow M$. (Clearly, this is equivalent to the existence of a bijective/invertible function $g: M \rightarrow \{1, \dots, n\}$.)

Even more, for two finite sets A, B we have that

- $|A| = |B|$ if there is a bijection $f: A \rightarrow B$,
- $|A| \leq |B|$ if there is a injection $f: A \rightarrow B$,
- $|A| \geq |B|$ if there is a surjection $f: A \rightarrow B$.

(Verify this yourself!)

This notion also allows us (to some extent) to characterize the cardinality of an infinite set.

Definition 1.17. Let A be a set and $n \in \mathbb{N}$. If the elements of A can be labeled by the numbers $\{1, \dots, n\}$, then we say that A has **cardinality** n , and we write

$$|A| = n \quad \text{or} \quad \#A = n.$$

If $|A| < \infty$, i.e., $\exists n \in \mathbb{N}: |A| = n$, then we call A **finite**, or a **finite set**.

If $|A|$ is not finite, then we call A **infinite**, or an **infinite set**, and write $|A| = \infty$.

If A is either finite or there exists a bijection $f: \mathbb{N} \rightarrow A$, then we call A **countable**.

If A is not countable, then we call A **uncountable**.

Note that countability is the precise definition of the “simple” property that the elements of A can be enumerated by the natural numbers $\{1, 2, 3, \dots\}$.

Example 1.18. We have that $f: \mathbb{N} \rightarrow \mathbb{Z}$ with $f(2n) := n$ and $f(2n - 1) := -n + 1$ for $n \in \mathbb{N}$ is a bijection. Verify this yourself! (For example, $f(1) = 0$, $f(2) = 1$, $f(3) = -1$ and so on.) Hence, \mathbb{Z} is a countable set.

Example 1.19. Let M and N be finite sets. Check that $|M \times N| = |M| \cdot |N|$.

We can also consider the **composition of functions**. Let X, Y, Z be non-empty sets, $f: X \rightarrow Y$ and $g: Y \rightarrow Z$ be functions. We then define a function $(g \circ f): X \rightarrow Z$ by **composing** f and g , i.e., $(g \circ f)(x) := g(f(x))$. As an exercise check that $g \circ f$ is indeed a function.

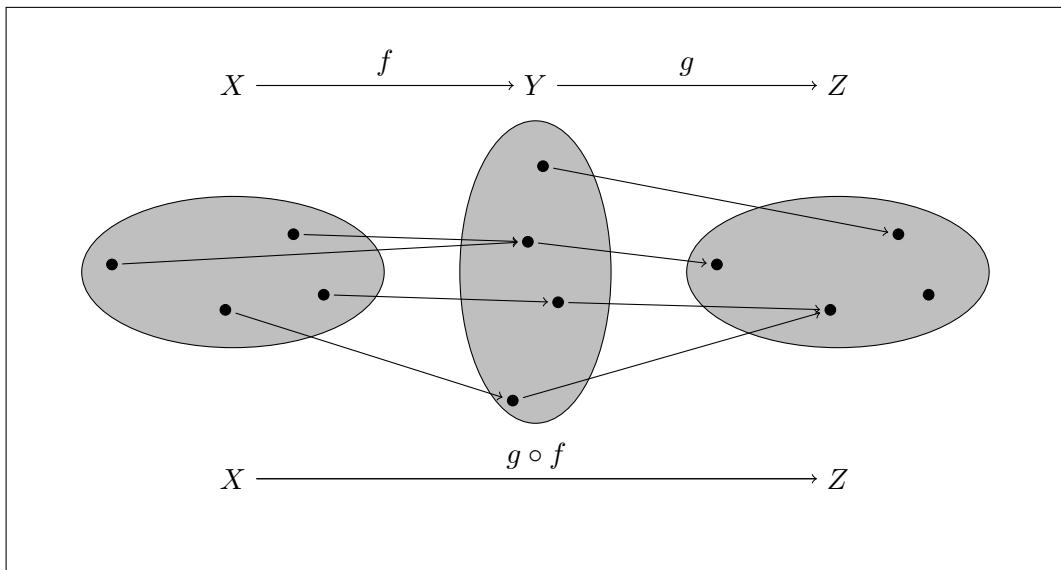


Figure 9: The composition $(g \circ f)(x) := g(f(x))$

Example 1.20. Consider the function $h(x) = \sin(x^2)$. If we set $f(x) = x^2$ and $g(y) = \sin(y)$ we get $(g \circ f)(x) = g(f(x)) = \sin(x^2) = h(x)$. However, since $(f \circ g)(x) = f(g(x)) = (\sin(x))^2$, it is important to mind that, in general, we have $(f \circ g) \neq (g \circ f)$.

Theorem 1.21. Let $f: X \rightarrow Y$ and $g: Y \rightarrow Z$ be invertible functions. Then $g \circ f$ is invertible and

$$(g \circ f)^{-1} = f^{-1} \circ g^{-1}.$$

Proof. Recall that f, g invertible $\iff f, g, f^{-1}, g^{-1}$ bijective. Hence

$$\forall y \in Y \exists! x \in X: f^{-1}(y) = x$$

and

$$\forall z \in Z \exists! y \in Y: g^{-1}(z) = y.$$

Furthermore

$$\forall x \in X \exists! y \in Y: f(x) = y$$

and

$$\forall y \in Y \exists! z \in Z: g(y) = z.$$

This yields

$$(f^{-1} \circ g^{-1})(z) = f^{-1}(g^{-1}(z)) = f^{-1}(y) = x$$

and

$$(g \circ f)(x) = g(f(x)) = g(y) = z.$$

Thus

$$(g \circ f)[(f^{-1} \circ g^{-1})(z)] = (g \circ f)(x) = z$$

and

$$(f^{-1} \circ g^{-1})((g \circ f)(x)) = (f^{-1} \circ g^{-1})(z) = x,$$

which shows that $f^{-1} \circ g^{-1}$ is the inverse of $g \circ f$, and vice versa. □

Finally, let us discuss some relations that are, instead of mappings, used to **compare**, **group** or **order** elements of a set M . For this we consider relations $R \subset M^2 = M \times M$. Such relations have usually nothing to do with functions, but are still very essential. Again, let us give names to some of their important characteristics.

Definition 1.22. Let $R \subseteq M^2$ be a relation for an arbitrary $M \neq \emptyset$. We call R

- **reflexive** if and only if

$$\forall x \in M: (x, x) \in R,$$

- **symmetric** if and only if

$$\forall (x, y) \in R: (y, x) \in R,$$

- **antisymmetric** if and only if

$$\forall x, y \in M: (x, y), (y, x) \in R \implies x = y$$

- **transitive** if and only if

$$\forall x, y, z \in M: (x, y), (y, z) \in R \implies (x, z) \in R$$

- **total** if and only if

$$\forall x, y \in M: (x, y) \in R \text{ or } (y, x) \in R.$$

Some relations which have some of these properties are especially important.

Definition 1.23. A relation is called

- **equivalence relation** if it is reflexive, symmetric and transitive,
- **partial order** if it is reflexive, antisymmetric and transitive,
- **linear order** if it is a partial order and total.

As an example consider the relation $L \subseteq \mathbb{R}^2$, where we define $(x, y) \in L : \iff x \leq y$ for $x, y \in \mathbb{R}$. This is the well-known **smaller or equal** relation in \mathbb{R} and we just write $x \leq y$ for $(x, y) \in L$ in the following. The smaller or equal relation is clearly

- reflexive: $x \leq x$,
- antisymmetric: $x \leq y$ and $y \leq x \implies x = y$,
- transitive: $x \leq y$ and $y \leq z \implies x \leq z$,
- total: $x \leq y$ or $y \leq x$

for all $x, y, z \in \mathbb{R}$, and hence, “ \leq ” is a *linear order*.

Example 1.24. The usual equality relation “ $=$ ” in \mathbb{R} is reflexive, symmetric, antisymmetric and transitive, but not total. It is therefore an equivalence relation and a partial order. As an exercise, show that “ $=$ ” is the only reflexive relation that is symmetric and antisymmetric.

Example 1.25. Define the “strictly less” relation “ $<$ ” by $a < b$ if and only if $a \leq b$ and $a \neq b$. Determine all characteristics of “ $<$ ”, as well as “ \geq ”, and “ $>$ ”. (Relations with the same characteristics as “ $<$ ” are called *strict partial orders*.)

Example 1.26. Let us also mention the *not equal* relation in R , which is defined by $a \neq b$ if and only if $a < b$ or $b < a$. Which characteristics does it have?

Example 1.27. Show that the *divisibility relation* $R \subset \mathbb{N}^2$, with $(a, b) \in R \Leftrightarrow a|b$, i.e., a divides b , is a partial order.

Example 1.28. One can define a partial order on a set of sets by using the subset relation \subset . For example, for $M := \{\emptyset, \{1\}, \{2\}, \{1, 2\}\}$, we have that, e.g., $\emptyset \subset \{1\} \subset \{1, 2\}$. It is easy to see that \subset is reflexive, antisymmetric and transitive. Hence, \subset is a partial order on M . However, since $\{1\} \not\subset \{2\}$ and $\{2\} \not\subset \{1\}$, it is not a linear order on M .

1.3 Real numbers

The commonly known natural numbers or rational numbers (fractions) are not sufficient for a rigorous foundation of mathematical analysis. The historical development shows that for issues concerning analysis, the rational numbers have to be extended to the real numbers. Maybe, you already know a lot about the real numbers. However, you probably do not know that all its properties follow from a few basic ones. So, let us introduce them from scratch.

We begin with the set of **natural numbers**, i.e.,

$$\mathbb{N} := \{1, 2, 3, \dots\}.$$

However, we have seen that such a definition is not precise enough, and we want to define this set solely by its properties.

These properties are called the **Peano axioms**, presented first by *Giuseppe Peano* (1858–1932) around 1889. The only thing we need to assume is that we know what the number “1” is, and what it means to add “1” to a number. (This is very reasonable, when we think about “counting” in real life.)

We then define the natural numbers by the (unique) set \mathbb{N} such that

1. $1 \in \mathbb{N}$,
2. $n \in \mathbb{N} \implies n + 1 \in \mathbb{N}$,
3. $\forall n, m \in \mathbb{N}: n = m \iff n + 1 = m + 1$, and
4. $\forall n \in \mathbb{N}: n + 1 \neq 1$.

In words, 1 is a natural number, for every natural number n , its “successor” $n + 1$ is also a natural number, two natural numbers are equal iff their successors are equal, and no natural number has the successor 1. Although these *axioms* seem to be obvious or redundant, depending on the point of view, they are really the only “axioms” we need to assume to build up most of modern mathematics. (A detailed discussion goes far beyond scope here.)

Next, the set of **natural numbers with zero** (or non-negative integers) is defined by

$$\mathbb{N}_0 := \{0\} \cup \mathbb{N}.$$

The sets \mathbb{N} and \mathbb{N}_0 are closed under addition and multiplication, i.e., for all $n, m \in \mathbb{N}$ we have $n + m \in \mathbb{N}$ and $m \cdot n \in \mathbb{N}$. However, we already get in trouble when we try to work with subtraction, since $21 - 42 = -21 \notin \mathbb{N}$.

The set of **integer numbers** is also closed under subtraction and is defined as

$$\mathbb{Z} := \{0, -1, 1, -2, 2, \dots\} = \{\dots - 2, -1, 0, 1, 2, \dots\} = \mathbb{N}_0 \cup (-\mathbb{N}),$$

where $-\mathbb{N} := \{-n: n \in \mathbb{N}\}$. Clearly, for $a, b \in \mathbb{Z}$ we have $a + b \in \mathbb{Z}$, $a - b \in \mathbb{Z}$ and $a \cdot b \in \mathbb{Z}$. However, division is still a problem if we use integer numbers only.

The set of all fractions of two integers is the set of **rational numbers** which is denoted as

$$\mathbb{Q} := \left\{ \frac{p}{q} : p, q \in \mathbb{Z}, q \neq 0 \right\},$$

where we call p numerator and q denominator.

We call two rational numbers $\frac{k}{n}$ and $\frac{l}{m}$ equal if and only if $km = ln$. Further an integer $k \in \mathbb{Z}$ can be identified with the fraction $\frac{k}{1} \in \mathbb{Q}$. Consequently, the inclusions $\mathbb{N} \subsetneq \mathbb{Z} \subsetneq \mathbb{Q}$ are true.

One main reason to introduce “new” sets of numbers ist that we **want to solve equations** and, in particular, we want to know if a solution exists in a given set.

For example, if we want to solve the equation $a \cdot x + b = 0$, where $a, b \in \mathbb{Q}$ are fixed constants and $a \neq 0$, it is easy to see that

$$x = \frac{-b}{a} \in \mathbb{Q}$$

solves the equation.

However, what about the simple equation $x^2 - 2 = 0$? Is there some $x \in \mathbb{Q}$ such that $x^2 = 2$? The following discussion will show that this is not possible.

Lemma 1.29. *Let $n \in \mathbb{N}$. Then we have n is even $\iff n^2$ is even.*

Proof. We start with showing n is even $\Rightarrow n^2$ is even.

So assume n is even and has the representation $n = 2 \cdot k$ for some $k \in \mathbb{N}$. Then $n^2 = 2 \cdot (2 \cdot k^2)$. Thus n^2 is an even number.

Now we show n^2 is even $\Rightarrow n$ is even.

We prove this by contradiction, i.e., we show that n is odd implies that n^2 is also odd. So, if n has the representation $n = 2 \cdot m + 1$ for some $m \in \mathbb{N}$, we obtain $n^2 = 2m \cdot (2 \cdot m + 1) + 2m + 1$, which is odd. This proves the statement. □

Theorem 1.30. *There is no rational number x , such that $x^2 = 2$.*

Proof. Proof by contradiction, i.e., we assume $\exists x \in \mathbb{Q}: x^2 = 2$ and show that this yields a wrong result. Since $(-x)^2 = x^2$ we may assume that $x > 0$ and that $x = \frac{m}{n}$, where $m, n \in \mathbb{N}$. Furthermore we consider that at least one of the numbers n or m is odd. Otherwise we could cancel by 2. This yields the equation

$$x^2 = \frac{m^2}{n^2} = 2$$

which is equivalent to

$$m^2 = 2 \cdot n^2.$$

Hence m^2 is an even number and consequently (previous lemma) m is an even number and we can write $m = 2 \cdot k$ for some $k \in \mathbb{N}$. So we have

$$m^2 = 4 \cdot k^2 = 2 \cdot n^2,$$

leading to

$$2 \cdot k^2 = n^2.$$

Applying the previous lemma once again, we get the result that n has to be an even number, which contradicts the assumption that either m or n has to be odd. □

Thus the equation $x^2 - 2 = 0$ is not solvable in \mathbb{Q} , but as we all know from school there is a number $\sqrt{2}$ such that $\sqrt{2}^2 = 2$.

Making the so called number line complete, we finally get to the set of **real numbers**. These numbers can have infinitely many decimals, i.e.,

$$\mathbb{R} := \left\{ z + r : z \in \mathbb{Z}, r = \frac{a_1}{10} + \frac{a_2}{100} + \dots; \text{ where } a_1, a_2, \dots \in \{0, 1, \dots, 9\} \right\}.$$

One may think of \mathbb{R} as the set of all points on the number line, i.e., *without holes*. Note that the rational numbers, written as decimal numbers, either have a finite number of digits or the sequence of digits is periodic. This means, that some points on the line are missing. These correspond to numbers which have a non-periodic infinite number of decimals, in other terms which cannot be written as fractions, as various roots, like $\sqrt{2}$, and the numbers π and e . These ones, i.e. numbers in $\mathbb{R} \setminus \mathbb{Q}$ are called **irrational numbers**. We will later see that the set \mathbb{R} is (assumed to be) in some sense **complete**. Such a precise definition was only given in the 19th century, probably by *Karl Weierstraß* (1815–1897).

Although, we will not prove that here, let us note that the **set of rational numbers is countable**, whereas the **set of real numbers are uncountable**. (The proofs can easily be found in the literature and are quite interesting, but they go beyond the scope here.)

In summary the following set relations hold:

$$\mathbb{N} \subsetneq \mathbb{N}_0 \subsetneq \mathbb{Z} \subsetneq \mathbb{Q} \subsetneq \mathbb{R}.$$

Note that the following **calculation rules** are valid for any real numbers:

Axiom 1 (Field axioms). For all $x, y, z \in \mathbb{R}$ we have

Commutativity :	$x + y = y + x,$	$x \cdot y = y \cdot x$
Associativity :	$x + (y + z) = (x + y) + z,$	$x \cdot (y \cdot z) = (x \cdot y) \cdot z$
Distributivity :	$x \cdot (y + z) = (x \cdot y) + (x \cdot z)$	
Identity elements :	$\exists 0 \in \mathbb{R} \forall x \in \mathbb{R}: x + 0 = 0 + x = x,$	
	$\exists 1 \in \mathbb{R} \forall x \in \mathbb{R}: x \cdot 1 = 1 \cdot x = x$	
Inverse elements :	$\forall x \in \mathbb{R} \exists y \in \mathbb{R}: x + y = y + x = 0,$	
	$\forall x \in \mathbb{R} \setminus \{0\} \exists y \in \mathbb{R}: x \cdot y = y \cdot x = 1$	

We call these properties *axioms* because, actually, we somehow *assume* them to be true. (Or how would you prove these statements?) Many of the calculations that follow in the upcoming chapters could also be done with *other fields*. We do not go into details here.

Example 1.31. Another important and well-known field is the *finite field* $\mathbb{Z}_2 := \{0, 1\}$ with the addition $0 + 0 := 0$, $1 + 0 := 1$, $0 + 1 := 1$, $1 + 1 := 0$, and the multiplication $0 \cdot 0 := 0$, $1 \cdot 0 := 0$, $0 \cdot 1 := 0$, $1 \cdot 1 := 1$. (Note that we formally need to define what we mean by addition and multiplication.) These are numbers and operations, computers are working with. Verify yourself that \mathbb{Z}_2 is a field.

1.4 Bounded sets, infimum and supremum

First, recall the following calculation rules for inequalities. Let $a, b, c, d \in \mathbb{R}$, then

- $a < b \implies a \pm c < b \pm c$
- $a < b$ and $c < d \implies a + c < b + d$
- $a < b \implies -a > -b$
- $0 < a < b \implies 0 < \frac{1}{b} < \frac{1}{a}$

These inequalities remain true if we replace $<$ by \leq and $>$ by \geq , respectively.

We now want to **specify the “least” and the “greatest” element** of a set. To define this formally, we first need the definition of bounded set (in \mathbb{R}).

Definition 1.32. Let $A \subseteq \mathbb{R}$. We say A is

- **bounded from above** if and only if

$$\exists C \in \mathbb{R} \forall a \in A: a \leq C$$

and we call C an **upper bound** of A .

- **bounded from below** if and only if

$$\exists c \in \mathbb{R} \forall a \in A: c \leq a$$

and we call c a **lower bound** of A .

- **bounded** if and only if A is bounded from above and from below.

Let us discuss this with the following basic subsets of \mathbb{R} , i.e., **intervals**.

Definition 1.33. Let $a, b \in \mathbb{R}$. Then we define the **closed interval**

$$[a, b] := \{x \in \mathbb{R}: a \leq x \leq b\},$$

half open intervals

$$[a, b) := \{x \in \mathbb{R}: a \leq x < b\}$$

and

$$(a, b] := \{x \in \mathbb{R}: a < x \leq b\},$$

and **open interval**

$$(a, b) := \{x \in \mathbb{R}: a < x < b\}$$

between a and b .

Moreover, we write $[a, \infty) := \{x \in \mathbb{R}: a \leq x\}$ and $(-\infty, b) := \{x \in \mathbb{R}: x < b\}$ etc.

Let $a, b \in \mathbb{R}$ such that $a < b$. Then for the closed interval $[a, b]$ we have that $a, a - 1, a - 42$ are lower bounds and $b, b + 42$ are upper bounds, and the same is true for the corresponding (half) open interval. So, an upper bound is not unique.

To fix a specific upper/lower bound, let us first define the **minimum and maximum** of a set.

Definition 1.34. Let $A \subset \mathbb{R}$ be non-empty and $T \in \mathbb{R}$. Then, T is called

- minimal element or **minimum** of A , denoted by $\min A := T$, if
 - $T \leq A$, i.e., T is a lower bound and
 - $T \in A$, i.e., T is contained in A .
- maximal element or **maximum** of A , denoted by $\max A := T$, if
 - $A \leq T$, i.e., T is an upper bound and
 - $T \in A$, i.e., T is contained in A .

If the maximum/minimum exists, it is clearly unique.

Example 1.35. Let $a, b \in \mathbb{R}$ with $a < b$. Then, $\min[a, b] = a$ and $\max[a, b] = b$.

Example 1.36. The set of the natural numbers $\mathbb{N} \subset \mathbb{R}$ is bounded from below, with $\min \mathbb{N} = 1$.

However, maxima and minima do not have to exist: While b is the least upper bound of both $[a, b]$ and (a, b) , we have that $b \in [a, b]$, but $b \notin (a, b)$. Hence, the set (a, b) does not have a maximum (or minimum), but still we would like to work with the “best possible” bounds for such a set, which are clearly a and b in this example.

For this we define the infimum and the supremum as the **greatest lower bound** and the **least upper bound**, respectively. These objects will be very important in the upcoming analysis.

Definition 1.37. Let $A \subset \mathbb{R}$ be non-empty and $T \in \mathbb{R}$. Then, T is called

- greatest lower bound or **infimum** of A , denoted by $\inf A := T$, if
 - $T \leq A$, i.e., T is a lower bound and
 - $x \leq A \implies x \leq T$, i.e., there is no greater lower bound.
- least upper bound or **supremum** of A , denoted by $\sup A := T$, if
 - $A \leq T$, i.e., T is an upper bound and
 - $A \leq x \implies T \leq x$, i.e., there is no smaller upper bound.

If A is not bounded from above (below) we set $\sup A := \infty$ ($\inf A := -\infty$).

For the empty set we define $\sup \emptyset := -\infty$ and $\inf \emptyset := \infty$.

Clearly, if $\inf A \in A$, then $\inf A = \min A$, and if $\sup A \in A$, then $\sup A = \max A$. In words, if the infimum (or supremum) of a set A is contained in A , then A has a minimum (or maximum) which has the same value.

Moreover, **infimum and supremum are uniquely determined**. To see this, assume that there are two suprema T_1 and T_2 of A . Since $\sup A = T_1$, we have $A \leq T_1$. In addition, since $\sup A = T_2$, we obtain by the second defining property above that $A \leq x \implies T_2 \leq x$. Setting $x = T_1$, we have $T_2 \leq T_1$. In the same way, we get $T_1 \leq T_2$, and hence $T_1 = T_2$.

Example 1.38. Let $a, b \in \mathbb{R}$ with $a < b$. Then,

$$\min[a, b] = \inf[a, b] = \inf[a, b) = \inf(a, b) = \inf(a, b) = a$$

and

$$\max[a, b] = \sup[a, b] = \sup[a, b) = \sup(a, b) = \sup(a, b) = b.$$

However, $\min(a, b)$, $\min(a, b]$, $\max(a, b)$ and $\max[a, b)$ do not exist.

Example 1.39. Let $A = \{x^2 : x \in (-1, 1)\}$, then $\inf A = \min A = 0$ and $\sup A = 1$. Verify yourself!

However, it is not clear at all if **every set has an infimum and supremum**. For example, if we would try the same with \mathbb{Q} instead of \mathbb{R} , i.e., we look for a least upper bound $T \in \mathbb{Q}$ for a set $A \subset \mathbb{Q}$, this would not be true. Consider e.g. the set $A = \{x \in \mathbb{Q} : x^2 \leq 2\}$. If we consider A as a subset of \mathbb{R} , then its supremum is $\sqrt{2} \in \mathbb{R}$. But as a subset of \mathbb{Q} it has no supremum, since $\sqrt{2} \notin \mathbb{Q}$. The reason is, that the rational numbers have “gaps”.

The real numbers \mathbb{R} were precisely defined to have no such “gaps”. However, note that this is actually an assumption and we formalize this by the next axiom of this lecture.

Axiom 2 (Completeness axiom). Every non-empty subset A of \mathbb{R} that is bounded from above has a least upper bound. That is, there always exists $T \in \mathbb{R}$ such that $T = \sup A$.

Let us state an **equivalent definition of supremum and infimum** for bounded sets in \mathbb{R} . Although it looks more complicated at first sight, this formulation is sometimes very helpful.

Let $A \subset \mathbb{R}$ be bounded from below. Then, $T = \inf A$ if and only if

- $T \leq A$, i.e., T is a lower bound and
- $\forall \varepsilon > 0 \exists a \in A : a < T + \varepsilon$, i.e., T comes arbitrarily close to A .

Analogously, let $A \subset \mathbb{R}$ be bounded from above. Then, $T = \sup A$ if and only if

- $A \leq T$, i.e., T is an upper bound and
- $\forall \varepsilon > 0 \exists a \in A : a > T - \varepsilon$, i.e., T comes arbitrarily close to A .

Remark 1.40. (*) Let us add, that all the definitions above (bounded, inf/sup) also make sense, if we replace the “usual” \leq -relation on \mathbb{R} by another partial order on some set M .

For example, consider the subset relation \subset on $M := \{\emptyset, \{1\}, \{2\}, \{1, 2\}\}$, see also Example 1.28. With $A := \{\{1\}, \{2\}\}$, we have $\sup A = \{1, 2\} \in M$. (Verify precisely that this is the least upper bound of A with respect to \subset .) Note that A does not have a maximum. All other suprema are easy to calculate. In particular, every subset of M has a supremum in M , i.e., M with \subset is complete. (What about the same relation on $M' := \{\emptyset, \{1\}, \{2\}\}$?)

Remark 1.41. (*) We call a set Ω with a partial order **complete**, if every $A \subset \Omega$ that is bounded from above has $\sup A \in \Omega$. Thus, very formally, the real numbers \mathbb{R} are assumed to be a *complete field with a linear order*. Note that this would not be true for \mathbb{N} and \mathbb{Z} , as they do not fulfill the field axioms, and also not for \mathbb{Q} because, again, $\{x^2 < 2\}$ has no supremum in \mathbb{Q} .

Let us finally discuss the **Archimedean property**. It is based on the fact that the set of natural numbers is unbounded. Even though this property seems unimpressive, it was of significant importance for real analysis.

Theorem 1.42. *The following assertions hold:*

- a) **Archimedean property:** $\forall x \in \mathbb{R} \exists n \in \mathbb{N}: n > x$, i.e., \mathbb{N} has no upper bound in \mathbb{R} .
- b) $\forall \varepsilon > 0 \exists n \in \mathbb{N}: \frac{1}{n} < \varepsilon$.
- c) For $x, y \in \mathbb{R}$ with $x < y$ there exists a rational number $\frac{m}{n} \in \mathbb{Q}$ with $x \leq \frac{m}{n} \leq y$.

Proof. a) Let us recall that the natural numbers \mathbb{N} are defined solely by the properties $1 \in \mathbb{N}$, and that $n \in \mathbb{N}$ implies that $n + 1 \in \mathbb{N}$.

We first prove by contradiction that \mathbb{N} is unbounded. For this, we assume that \mathbb{N} is bounded. Thus, the supremum $x = \sup \mathbb{N}$ exists by the completeness axiom. As x is the smallest upper bound of \mathbb{N} , $x - 1$ is no upper bound of \mathbb{N} , so there exists a $n \in \mathbb{N}$ with $x - 1 < n$. By addition with 1 this also implies $x < n + 1$. But $n + 1 \in \mathbb{N}$ follows from $n \in \mathbb{N}$, and therefore x is no upper bound of \mathbb{N} , which is a contradiction to the original assumption. So, \mathbb{N} must be unbounded.

In order to show b) and c) one continues as follows:

$$\begin{aligned} \mathbb{N} \text{ has no upper bound} &\iff \forall x \in \mathbb{R} \exists n \in \mathbb{N}: n > x \\ &\iff \forall x > 0 \exists n \in \mathbb{N}: \frac{1}{n} < \frac{1}{x}; \quad \text{set } \varepsilon := \frac{1}{x} \\ &\iff \forall \varepsilon > 0 \exists n \in \mathbb{N}: \frac{1}{n} < \varepsilon. \end{aligned}$$

c) It is sufficient to check the case $x, y > 0$ (Why?). Then we are looking for $m, n \in \mathbb{N}$ such that $nx \leq m \leq ny$. With part a) above, we get a $n \in \mathbb{N}$ with $n(y - x) \geq 1$. Now take the smallest natural number $m \geq nx$ (which clearly exists). Then $m \geq nx$ and $m - 1 < nx$. The last condition implies that $m < nx + 1 \leq ny$, so we get $nx \leq m \leq ny$. □

Based on this, we easily calculate certain infima and suprema of (discrete) sets.

Example 1.43. Let $A = \left\{ \frac{1}{n} : n \in \mathbb{N} \right\}$, then $\inf A = 0$ and $\sup A = \max A = 1$.

To prove that, first note that $0 < \frac{1}{n} \leq 1$ for all $n \in \mathbb{N}$. So, 0 is a lower bound, and 1 is an upper bound. Since $1 \in A$ (for $n = 1$), we therefore obtain that $\sup A = \max A = 1$. For the infimum we need to show that there is no larger lower bound. For this, let $\varepsilon > 0$ be arbitrary. By the Archimedean property there exists an $n \in \mathbb{N}$ with $\frac{1}{n} < \varepsilon$. Hence, ε is not a lower bound. As this works for all $\varepsilon > 0$, we conclude that 0 is the largest lower bound of A , and hence $\inf A = 0$.

Example 1.44. Let $A = \left\{ \frac{1}{n^2 - n - 3} : n \in \mathbb{N} \right\}$, then $\inf A = \min A = -1$ and $\sup A = \max A = \frac{1}{3}$.

1.5 Induction and combinatorics

In the following we are dealing with **mathematical induction** which helps us to prove statements or define objects for all $n \in \mathbb{N}$.

Theorem 1.45 (Mathematical induction). *A predicate $P(n)$ is true for all $n \in \mathbb{N}$ if and only if the following two steps hold:*

a) **Induction basis:** $P(1)$ is true.

b) **Induction step:** if $P(n)$ is true for some $n \in \mathbb{N}$, then $P(n + 1)$ is also true.

Note that, by the Peano axioms, we reach every natural number this way.

The concept of mathematical induction can be used for the **definition** of sequences of objects, which is sometimes quite helpful. If, for instance, $G(n)$ is a quantity that has to be defined for all $n \in \mathbb{N}$, then it is sufficient to define $G(1)$ and, for all $n \in \mathbb{N}$, $G(n + 1)$ in terms of $G(n)$.

One example are the formal definitions of the **sum and product symbols**.

Definition 1.46. Let $a_k \in \mathbb{R}$ for $k \in \mathbb{N}$. Then we define sum and product as follows:

$$\begin{aligned} \sum_{k=1}^1 a_k &:= a_1 & \sum_{k=1}^{n+1} a_k &:= a_{n+1} + \sum_{k=1}^n a_k \\ \prod_{k=1}^1 a_k &:= a_1 & \prod_{k=1}^{n+1} a_k &:= a_{n+1} \cdot \prod_{k=1}^n a_k \end{aligned}$$

k is called the **summation index**.

A special case of products: For $a \in \mathbb{R}$ we have $a^1 := a$ and $a^{n+1} := a \cdot a^n$.

Let us see how to **prove by induction**. As this is an often used and very structural proof, we will use some short notation to highlight the corresponding parts. According to Theorem 1.45, one needs to show that the statement is true for the first element, which we call the induction basis, denoted by **(IB)**. Then, by assuming the induction hypothesis **(IH)**, that the assertion is true for n , we prove that it is also true for $n + 1$, which we call the induction step **(IS)**.

Let us discuss two examples to demonstrate this type of proof.

The first one is (at least without proof) known to many from school. The young Carl Friedrich Gauß (1777–1855) knew it already by the age of nine.

Example 1.47. We prove the formula $\sum_{k=1}^n k = \frac{n(n+1)}{2}$ (*Gauß'sche Summenformel*):

IB ($n = 1$): $\sum_{k=1}^1 k = 1 = \frac{1(1+1)}{2}$ is true.

IH: $\sum_{k=1}^n k = \frac{n(n+1)}{2}$

IS ($n \rightarrow n + 1$) :

$$\sum_{k=1}^{n+1} k = \sum_{k=1}^n k + (n + 1) \stackrel{\text{IH}}{=} \frac{n(n+1)}{2} + \frac{2}{2}(n + 1) = \frac{(n + 1)(n + 2)}{2}.$$

Another important instance of a proof by induction is **Bernoulli's inequality**, which will be essential later.

Theorem 1.48 (Bernoulli's inequality). *Let $x \geq -1$ and $n \in \mathbb{N}$. Then, we have*

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

Proof. Let $x \in \mathbb{R}$ such that $x \geq -1$. We prove the statement for all $n \in \mathbb{N}$ by induction:

IB ($n = 1$): $(1+x) \geq 1+x$ is clearly true.

IH: $(1+x)^n \geq 1+nx$

IS ($n \rightarrow n+1$) :

$$\begin{aligned} (1+x)^{n+1} &= (1+x)^n(1+x) \stackrel{\text{IH}}{\geq} (1+nx)(1+x) \\ &= 1+(n+1)x+nx^2 \geq 1+(n+1)x. \end{aligned}$$

This was what has to be shown. \square

Example 1.49. Try to prove the Bernoulli's inequality with $>$ instead of \geq . Can we use the same assumption on x in this case?

Let us now turn to some combinatorial quantities. These quantities are heavily used when it comes to discrete mathematics or elementary probability theory, as they represent the **number of permutations or subsets** of certain size.

Definition 1.50. The **factorial** $n!$ of a natural number $n \in \mathbb{N}$ is defined inductively by:

$$1! := 1 \quad \text{and} \quad \forall n \in \mathbb{N}: (n+1)! := (n+1) \cdot n!.$$

In addition, we set $0! := 1$.

The **binomial coefficient** $\binom{n}{k}$ (we say “ n choose k ”) for $n, k \in \mathbb{N}_0$ with $n \geq k$ is defined by

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

Clearly, we have $\binom{n}{0} = 1$, $\binom{n}{1} = n$ and $\binom{n}{k} = \binom{n}{n-k}$.

The factorial $n!$ is the number of permutations (orderings) of a set of n numbers. If you pick your first arbitrary element, you have n possibilities for your choice. When it comes to your second decision you have just $n-1$ options left, and so on.

The binomial coefficient $\binom{n}{k}$ is read as n choose k (or “ n over k ”; in German “ n über k ”). It represents the number of ways to choose k unordered outcomes from a set of n possibilities, also known as a combination. E.g., $\binom{n}{2}$ is the number of two-element subsets of $\{1, \dots, n\}$. Clearly, there are $n(n-1)$ ordered pairs $(i, j) \in \{1, \dots, n\}^2$ with $i \neq j$. As the ordering is irrelevant for sets, we have $\frac{n(n-1)}{2}$ two-element subsets which coincides with $\binom{n}{2}$.

The following is a useful formula for binomial coefficients.

Lemma 1.51. *Let $n, k \in \mathbb{N}$ with $k \leq n - 1$. Then we have*

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

Proof.

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

□

The “simple” explanation of this equality is, that subsets of $\{1, \dots, n+1\}$ with $k+1$ elements can be splitted into those that contain the number $n+1$ and those that don’t contain it. To find the number of all sets that contain it, we need to count all k -element subsets of $\{1, \dots, n\}$, and there are $\binom{n}{k}$ of them. The number of all sets that don’t contain it, is the same as the number of all $(k+1)$ -element subsets of $\{1, \dots, n\}$, which is $\binom{n}{k+1}$. So the total number is their sum.

Based on this, we can prove one of the most famous theorems in mathematics.

Theorem 1.52 (Binomial theorem). *Let $x, y \in \mathbb{R}$ and $n \in \mathbb{N}$. Then we have*

$$(x+y)^n = \sum_{k=0}^n \binom{n}{k} x^k y^{n-k}.$$

Proof. If we expand $(x+y)^n$ we get a sum of 2^n expressions of the form $z_1 z_2 \cdots z_n$, where for all $i = 1, 2, \dots, n$ we either have $z_i = x$ or $z_i = y$. Counting all summands where x occurs exactly k times (and y therefore $n-k$ times) gives us $\binom{n}{k}$. Hence there are exactly $\binom{n}{k}$ summands of the form $x^k y^{n-k}$, which proves the theorem.

□

Example 1.53. As a good exercise try to prove the binomial theorem by induction using Lemma 1.51.

Example 1.54. If we set $x = y = 1$ we get

$$2^n = \sum_{k=0}^n \binom{n}{k}.$$

For $x = -1, y = 1$ we get

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

Example 1.55. It is very important to know the following special cases of the binomial theorem. For $x, y \in \mathbb{R}$ we have

$$(x + y)^2 = x^2 + 2xy + y^2.$$

and

$$(x + y)^3 = x^3 + 3x^2y + 3xy^2 + y^3.$$

(Verify these formulas!)

The binomial theorem can also be used to (easily) improve upon Bernoulli's inequality.

Corollary 1.56. *Let $x \geq 0$. Then, for any $k \in \{1, \dots, n\}$, we have*

$$(1 + x)^n \geq 1 + \binom{n}{k} x^k.$$

This lemma is a direct consequence of the binomial theorem, see Theorem 1.52 with $y = 1$. One just leaves out some of the non-negative terms. Note that this is Bernoulli's inequality for $k = 1$. However, note that we need $x \geq 0$ here. (It is false, e.g., for $x = -1$ and $n \geq k = 2$.)

1.6 Absolute value

In this section we want to briefly discuss some essential types of functions, which need to be introduced for the sake of completeness. We start with the absolute value of a number, since this function and its generalizations are heavily used throughout this lecture and beyond.

The **absolute value** of $x \in \mathbb{R}$ is defined by:

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

The graph of the function $x \mapsto |x|$ is shown in Figure 10.

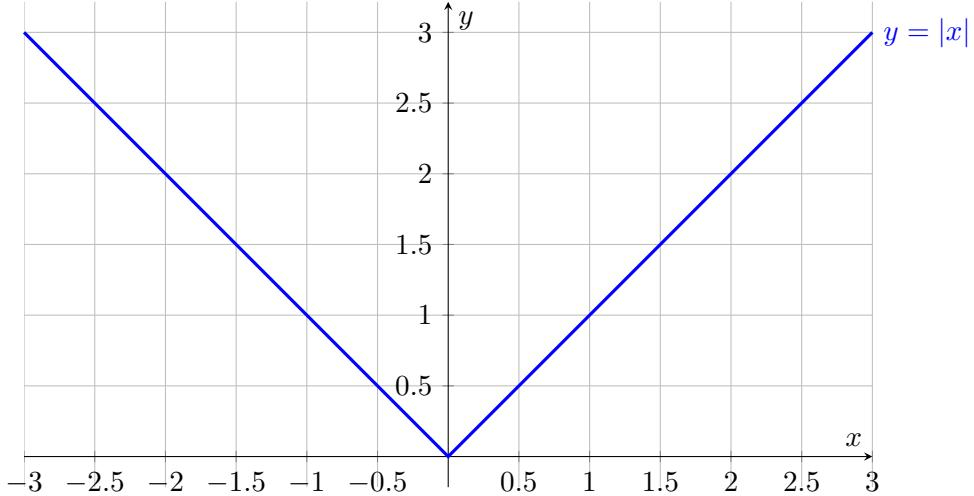


Figure 10: The graph of the function $f(x) = |x|$

Note that this function maps \mathbb{R} onto $\mathbb{R}_{\geq 0} := [0, \infty)$, i.e., the domain and range of $|\cdot|$ are \mathbb{R} and $\mathbb{R}_{\geq 0}$, respectively. As, e.g., $|-1| = |1|$, $|\cdot|$ is not injective on \mathbb{R} , and hence also not bijective. (Try to prove $|-x| = |x|$ formally!)

We list some of the most important properties of the absolute value.

Lemma 1.57. *Let $x, y, z \in \mathbb{R}$ with $z \geq 0$. Then, we have the following properties*

1. $|x| \geq 0$
2. $|x| = 0 \iff x = 0$
3. $|x| \geq x$
4. $|x \cdot y| = |x| \cdot |y|$
5. $|x| \leq z \iff -z \leq x \leq z$

Proof. We prove the lemma by using **case distinction**. In fact, this is a good illustration on how to work with absolute values in general.

We only prove the first statement, the other work similarly, and are left as an exercise.

Case 1: $x \geq 0$, then we have $|x| = x \geq 0$.

Case 2: $x < 0$, then we have $|x| = -x > 0$.

Hence $|x| \geq 0$ for all $x \in \mathbb{R}$. □

Absolute values appear very regularly, and it **essential to know how to work with them**. Mostly, we are interested in the set of all numbers, say x , that satisfy a certain inequality.

So for a given inequality, say $|x - 1| < 2$, we want to find all $x \in \mathbb{R}$ that satisfy it. The set of all these x is called the **solution set** of the inequality, and is often denoted by L .

Example 1.58. We look for the solution set L of the following inequality (in \mathbb{R}):

$$2|x + 3| - 4|x - 1| \geq 8x - 2.$$

We have to distinguish three regions: $x \leq -3$, $-3 < x \leq 1$ and $1 < x$, since the expressions in the absolute values change their signs at these points.

$$\begin{aligned} \text{Case 1 } (x \leq -3): \quad & 2|x + 3| - 4|x - 1| = 2(-x - 3) - 4(-x + 1) = -2x - 6 + 4x - 4 = 2x - 10 \geq 8x - 2 \\ & \iff x \leq -\frac{4}{3} \end{aligned}$$

We therefore obtain: $L_1 := \{x \in \mathbb{R} : x \leq -3, x \leq -\frac{4}{3}\} = (-\infty, -3]$.

$$\begin{aligned} \text{Case 2 } (-3 < x \leq 1): \quad & 2|x + 3| - 4|x - 1| = 2x + 6 + 4x - 4 = 6x + 2 \geq 8x - 2 \\ & \iff x \leq 2 \end{aligned}$$

We therefore obtain: $L_2 := (-3, 1]$.

$$\begin{aligned} \text{Case 3 } (x > 1): \quad & 2|x + 3| - 4|x - 1| = 2x + 6 - 4x + 4 = -2x + 10 \geq 8x - 2 \\ & \iff x \leq \frac{6}{5}, \end{aligned}$$

We therefore obtain: $L_3 := (1, \frac{6}{5}]$.

$$\text{Together: } L = L_1 \cup L_2 \cup L_3 = (-\infty, -3] \cup (-3, 1] \cup (1, \frac{6}{5}] = (-\infty, \frac{6}{5}].$$

The following inequality, which is probably the most well known one, is at the same time the most important (also in more general scenarios). You will not use any other inequality more often than this one.

Theorem 1.59 (Triangle inequality). *Let $x, y \in \mathbb{R}$. Then we have that*

$$|x + y| \leq |x| + |y|.$$

Proof. We already know $x \leq |x|$ and $-x \leq |x|$, which implies

$$x + y \leq |x| + |y| \text{ and } -(x + y) = -x - y \leq |x| + |y|.$$

These are both cases we need to check for proving the triangle inequality. □

The triangle inequality also implies that

$$|x - z| = |x - y + y - z| \leq |x - y| + |y - z|,$$

for all $x, y, z \in \mathbb{R}$. Hence, we can consider the absolute value as a “distance” between numbers.

Moreover, we obtain the following (less obvious) inequality.

Corollary 1.60. *Let $x, y \in \mathbb{R}$. Then we have*

$$|x| - |y| \leq |x - y|.$$

Proof. The triangle inequality yields

$$|a + b| - |b| \leq |a|$$

for arbitrary $a, b \in \mathbb{R}$. We choose $a = x - y$ and $b = y$ to get

$$|x| - |y| \leq |x - y|.$$

For $b = -x$ we get

$$|y| - |x| \leq |x - y|.$$

The result follows from $\max\{|x| - |y|, |y| - |x|\} = \max\{|x| - |y|, |x - y|\}$. □

1.7 Some elementary functions

We now turn to some other elementary functions. Note that it is not necessary to understand all of these definitions completely yet. We just state them here as a reference for later.

- a) Very simple functions are **affine linear functions**, which are defined as

$$f: \mathbb{R} \rightarrow \mathbb{R}$$

$$x \mapsto ax + b$$

with $a, b \in \mathbb{R}$. If $a = 0$, f is called **constant function**. If $a \neq 0$ and $b = 0$ we call f a **linear function**. Note that linear functions satisfy $f(x + y) = f(x) + f(y)$.

(Is the same true for affine linear functions?)

Moreover, they are special cases of **polynomial functions**, which are defined by

$$p: \mathbb{R} \rightarrow \mathbb{R}$$

$$x \mapsto \sum_{i=0}^n a_i x^i$$

with $a_0, \dots, a_n \in \mathbb{R}$.

- b) We may also consider the **power functions**

$$\begin{aligned} f: \mathbb{R}_+ &\rightarrow \mathbb{R} \\ x &\mapsto x^a, \end{aligned}$$

for some fixed $a \in \mathbb{R}$ and $\mathbb{R}_+ := [0, \infty)$, see Figure 11. (In short: $f(x) = x^a$ for $x > 0$.)

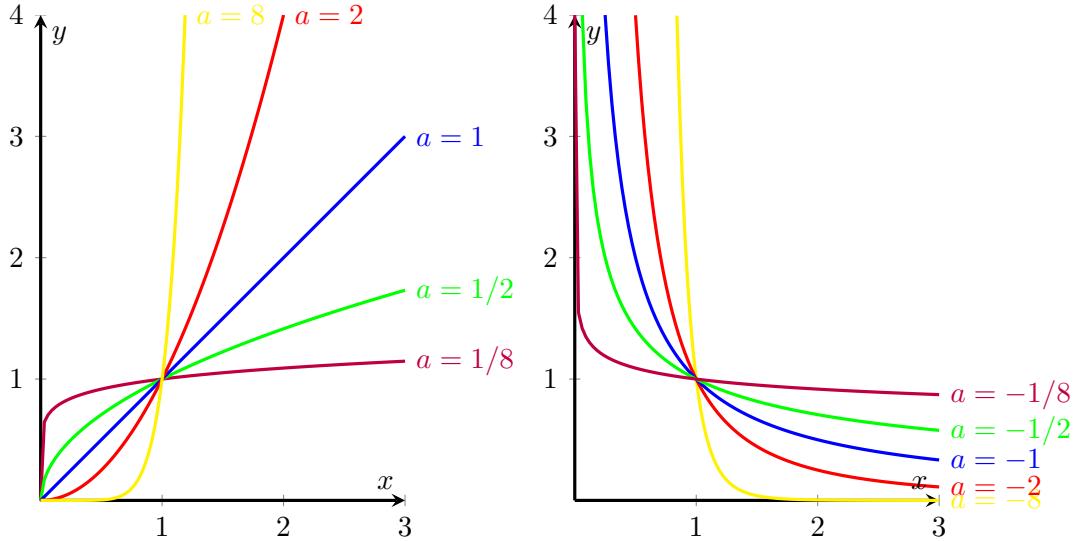


Figure 11: The graph of the function $f(x) = x^a$ for different a

Let us say some words about, how these expressions are precisely defined. Since we know how to multiply two or more numbers, we may clearly define the power functions with non-negative integer exponents, i.e., x^n with $n \in \mathbb{Z}$, even for all $x \in \mathbb{R} \setminus \{0\}$, as in the polynomials above. (We only need to exclude $x = 0$ for negative exponents, because we cannot divide by 0.) Moreover, since we know how a root of a positive number is defined, we can also define, e.g., $\sqrt{2} = 2^{1/2}$ or $3^{1/8}$. ($3^{1/8}$ is just the number z with $z^8 = 3$.) With this, we can define $x^{\frac{n}{m}} = \sqrt[m]{x^n}$ for every $x > 0$, $n \in \mathbb{Z}$ and $m \in \mathbb{N}$. Hence, we know how to define x^q for every $x > 0$ and $q \in \mathbb{Q}$.

But what about $4^{\sqrt{2}}$ or π^π ? (Think a bit how you would calculate/define these numbers!) In this case, i.e., if we consider x^a for some $a \in \mathbb{R} \setminus \mathbb{Q}$, the most natural way is to define them by using supremum or infimum. Taking the monotonicity into account (see Figure 11), we define for $x > 0$ the powers

$$x^a := \begin{cases} \sup\{x^q : q \in \mathbb{Q}, q < a\}, & \text{if } x \geq 1, \\ \inf\{x^q : q \in \mathbb{Q}, q > a\}, & \text{if } 0 < x < 1. \end{cases} \quad (1.1)$$

Note that we know how to compute the expressions inside sup/inf, and that sup/inf always exist as the sets are bounded, due to the Completeness axiom.

With this, we get the well-known calculation rules for arbitrary $x, y \in \mathbb{R}_+$, and $a, b \in \mathbb{R}$ that

$$\begin{array}{lll} x^{-a} = \frac{1}{x^a} & x^a \cdot x^b = x^{a+b} & \frac{x^a}{x^b} = x^{a-b} \\ (x^a)^b = x^{a \cdot b} & (x \cdot y)^a = x^a \cdot y^a & \left(\frac{x}{y}\right)^a = \frac{x^a}{y^a} \end{array}$$

However, note that there is –and there will be– no natural and satisfying way to define arbitrary powers of a negative number, like $(-2)^\pi$. We do not go into detail here.

- c) Other well-known functions are the **exponential functions**, which characterise rapid growth or decay. Exponential functions are defined by

$$f : \mathbb{R} \rightarrow \mathbb{R}_+, \quad x \mapsto a^x,$$

where the basis $a > 0$ is a fixed parameter. That is, in contrast to power functions, the variable $x \in \mathbb{R}$ is the exponent.

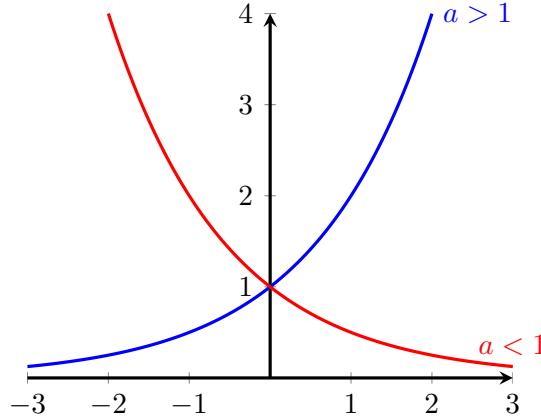


Figure 12: The graph of the function $f(x) = a^x$ for different a

Note that a^x is defined for all $x \in \mathbb{R}$. For a precise definition, we use again supremum and infimum, see (1.1). That is, for $a \geq 1$ we set $a^x := \sup\{a^q : q \in \mathbb{Q}, q < x\}$. Together with $a^x = (\frac{1}{a})^{-x}$ for $0 < a < 1$, this defines a^x for all $a > 0$ and $x \in \mathbb{R}$.

Among all exponential functions, one is particularly important. This is when we choose the basis $a = e \approx 2,7182818284$, where e is **Euler's number**. This special (irrational) number e may be defined by different means:

$$e = \sup \left\{ \left(1 + \frac{1}{n}\right)^n : n \in \mathbb{N} \right\} = \lim_{n \rightarrow \infty} \left(1 + \frac{1}{n}\right)^n = \sum_{k=0}^{\infty} \frac{1}{k!}.$$

(Do not panic yet! We discuss the meaning of these expressions later.)

- d) The exponential functions often appear together with the **logarithmic functions**, which are there inverses:

$$f : \mathbb{R}_+ \rightarrow \mathbb{R}, \quad x \mapsto \log_b(x).$$

Read: logarithm of x to base b . Here, we usually assume that $b > 1$, but in principle one may also consider $b \in (0, 1)$.

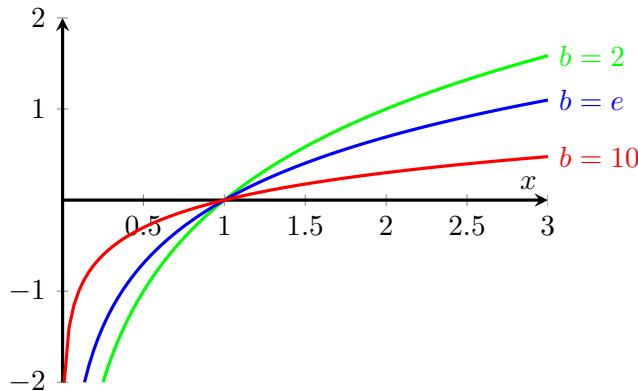


Figure 13: The graph of the function $f(x) = \log_b(x)$ for different b

Some important choices for the base b are

- $b = e \approx 2.71828\dots$, i.e., Euler's number: We write $\ln(x) := \log_e(x) = \log(x)$, and this is called the natural logarithm.
- $b = 2$: We write $\text{lb}(x) := \log_2(x)$, and this is called binary logarithm.
- $b = 10$: We write $\text{lg}(x) := \log_{10}(x)$, and this is called common logarithm.

Remark 1.61. In order to clarify the notation beforehand we agree on the following: For us $\log(x) = \ln(x)$, and in all other cases we are more precise and write $\log_b(x)$.

As inverse function it satisfies the two relations

$$b^{\log_b x} = x \quad \text{and} \quad \log_b(b^x) = x,$$

which define the function values exactly. In particular, this means that $\log_b x$ is the number, say y , such that $b^y = x$.

Moreover, from the rules for exponentiation we obtain the following calculation rules for all $a, b, x, y > 0$:

$$\log_b x^y = y \log_b x$$

$$\log_b(xy) = \log_b x + \log_b y$$

$$\log_b \frac{x}{y} = \log_b x - \log_b y$$

$$\log_b x = \frac{\log_a x}{\log_a b}$$

- e) The **trigonometric functions** play a crucial role in analytic geometry. The most important are:

$\sin(x)$	(sine)
$\cos(x)$	(cosine)
$\tan(x)$	(tangent)
$\cot(x)$	(cotangent)

Figure 14 shows the 'geometric definition' of these functions in the unit circle.

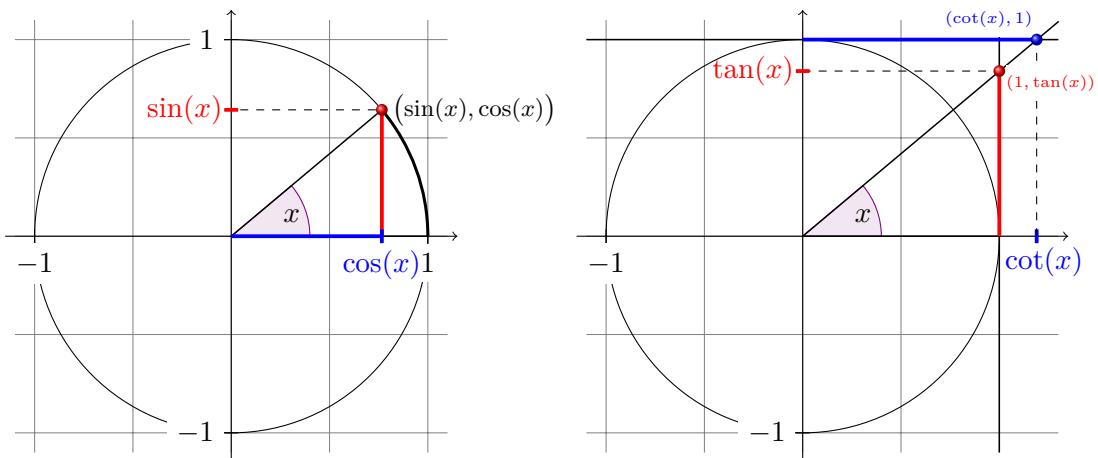


Figure 14: Illustration of trigonometric functions

The variable $x \in \mathbb{R}$ corresponds to the angle that is enclosed between the horizontal axis and the line to the point $(\sin(x), \cos(x))$. It can therefore be given in *degrees* (deg) in the range from 0° and 360° . However, it is more convenient in science to interpret x as the arc length of the part of the unit circle that is contained between the lines. The appropriate unit is called *radians* (rad). Since the length of the unit circle is given by 2π , radians can be obtained from the degrees by:

$$x = \frac{\text{deg}}{180} \pi.$$

The functions $\sin x$ and $\cos x$ are defined for all $x \in \mathbb{R}$, see Figure 15.

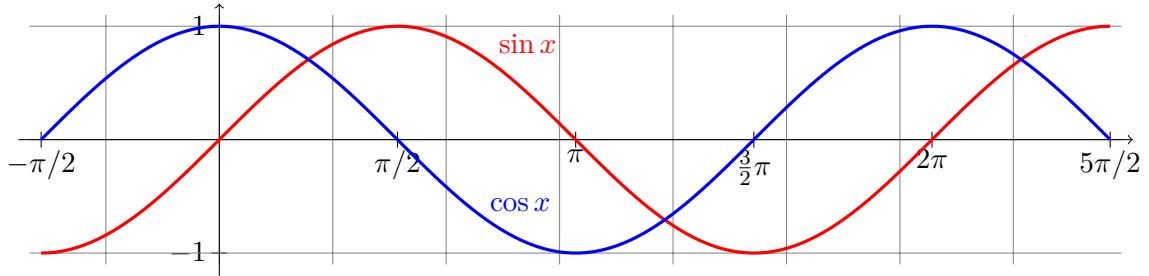


Figure 15: The graphs of \sin and \cos

Using some elementary geometry on the above illustration we obtain the calculation rules:

$$\begin{array}{ll} \sin(-x) = -\sin x, & \cos(-x) = \cos x, \\ \sin\left(x + \frac{\pi}{2}\right) = \cos x, & \cos\left(x + \frac{\pi}{2}\right) = -\sin x, \\ \sin(x + \pi) = -\sin x, & \cos(x + \pi) = -\cos x, \\ \sin(x + 2\pi) = \sin x, & \cos(x + 2\pi) = \cos x. \end{array}$$

Some important values are the following:

φ	0	$\frac{\pi}{6}$	$\frac{\pi}{4}$	$\frac{\pi}{3}$	$\frac{\pi}{2}$	$\frac{2\pi}{3}$	$\frac{3\pi}{4}$	$\frac{5\pi}{6}$	π
$\sin(\varphi)$	$\frac{\sqrt{0}}{2}$	$\frac{\sqrt{1}}{2}$	$\frac{\sqrt{2}}{2}$	$\frac{\sqrt{3}}{2}$	$\frac{\sqrt{4}}{2}$	$\frac{\sqrt{3}}{2}$	$\frac{\sqrt{2}}{2}$	$\frac{\sqrt{1}}{2}$	$\frac{\sqrt{0}}{2}$
$\cos(\varphi)$	$\frac{\sqrt{4}}{2}$	$\frac{\sqrt{3}}{2}$	$\frac{\sqrt{2}}{2}$	$\frac{\sqrt{1}}{2}$	$\frac{\sqrt{0}}{2}$	$-\frac{\sqrt{1}}{2}$	$-\frac{\sqrt{2}}{2}$	$-\frac{\sqrt{3}}{2}$	$-\frac{\sqrt{4}}{2}$

The corresponding values in $(\pi, 2\pi)$ can be obtained from $\sin(x + \pi) = -\sin(x)$ and $\cos(x + \pi) = -\cos(x)$.

Moreover, we have the very important **trigonometric identity**

$$\sin^2 x + \cos^2 x = 1,$$

and the **trigonometric addition formulas**:

$$\sin(x+y) = \sin x \cos y + \cos x \sin y$$

$$\cos(x+y) = \cos x \cos y - \sin x \sin y$$

$$\sin x + \sin y = 2 \sin\left(\frac{x+y}{2}\right) \cos\left(\frac{x-y}{2}\right)$$

$$\sin x - \sin y = 2 \cos\left(\frac{x+y}{2}\right) \sin\left(\frac{x-y}{2}\right)$$

$$\cos x + \cos y = 2 \cos\left(\frac{x+y}{2}\right) \cos\left(\frac{x-y}{2}\right)$$

$$\cos x - \cos y = -2 \sin\left(\frac{x+y}{2}\right) \sin\left(\frac{x-y}{2}\right)$$

The following useful formulas may be deduced:

$$\cos(2x) = \cos^2(x) - \sin^2(x)$$

$$\sin(2x) = 2 \sin(x) \cos(x)$$

$$1 + \cos(2x) = 2 \cos^2(x)$$

$$1 + \sin(2x) = (\sin(x) + \cos(x))^2$$

$$1 - \cos(2x) = 2 \sin^2(x)$$

$$1 - \sin(2x) = (\sin(x) - \cos(x))^2$$

(It's not needed to memorize these formulas, but important to know where to find them.)

Based on sin and cos, we can define $\tan x$ and $\cot x$ by

$$\tan x := \frac{\sin x}{\cos x}, \quad x \neq \frac{\pi}{2} + k\pi, k \in \mathbb{Z},$$

$$\cot x := \frac{\cos x}{\sin x}, \quad x \neq k\pi, k \in \mathbb{Z}.$$

These restrictions are necessary since $\sin(k\pi) = 0 = \cos\left(\frac{\pi}{2} + k\pi\right)$ for all $k \in \mathbb{Z}$, and we are not allowed to divide by zero, see Figure 16.

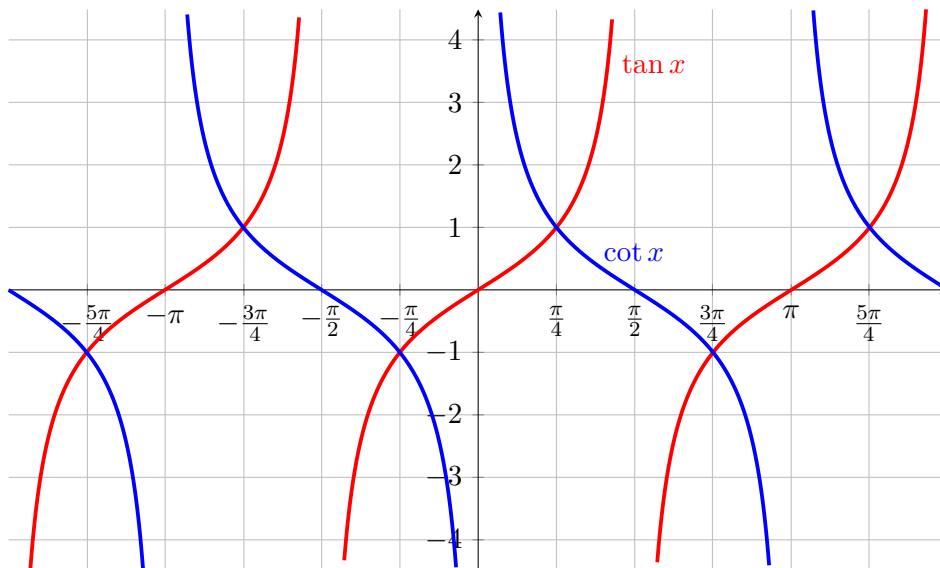


Figure 16: The graphs of tan and cot

Additionally, we can define the **inverse trigonometric functions**. For this, first note that \sin and \tan are increasing and on $[-\frac{\pi}{2}, \frac{\pi}{2}]$ and $(-\frac{\pi}{2}, \frac{\pi}{2})$, respectively, \cos is decreasing on $[0, \pi]$, see Figures 15 and 16. In particular, all three functions are bijective on these intervals (i.e., every value of $\sin(x)$, $\cos(x)$ and $\tan(x)$ is achieved by exactly one x in the interval), and therefore have inverse functions, see Theorem 1.16.

Due to their importance, the inverse trigonometric function have their own name. Namely, **arcsine** $\arcsin := \sin^{-1}: [-1, 1] \rightarrow [-\frac{\pi}{2}, \frac{\pi}{2}]$, the **arccosine** $\arccos := \cos^{-1}: [-1, 1] \rightarrow [0, \pi]$, and the **arctangent** $\arctan := \tan^{-1}: \mathbb{R} \rightarrow [-\pi/2, \pi/2]$. These functions are defined by

$$\begin{aligned} y = \arcsin x &: \iff x = \sin y \text{ and } y \in [-\pi/2, \pi/2], \\ y = \arccos x &: \iff x = \cos y \text{ and } y \in [0, \pi], \\ y = \arctan x &: \iff x = \tan y \text{ and } y \in (-\pi/2, \pi/2), \end{aligned}$$

and it is not hard to see that

- $\arcsin: [-1, 1] \rightarrow [-\frac{\pi}{2}, \frac{\pi}{2}]$ is increasing,
- $\arccos: [-1, 1] \rightarrow [0, \pi]$ is decreasing,
- $\arctan: \mathbb{R} \rightarrow (-\frac{\pi}{2}, \frac{\pi}{2})$ is increasing,

see Figure 17.

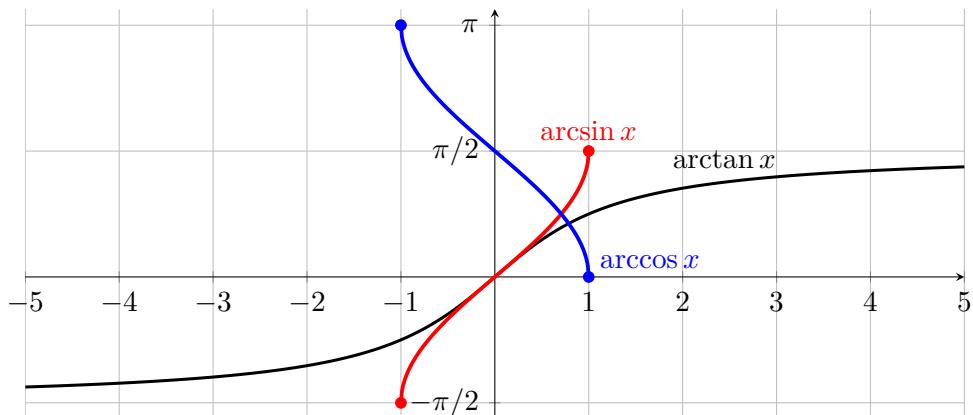


Figure 17: The graphs of \arcsin and \arccos

- f) Finally a special combination of exponential functions yield the **hyperbolic functions**, defined in \mathbb{R} by:

$$\sinh x := \frac{e^x - e^{-x}}{2} \quad (\text{hyperbolic sine})$$

$$\cosh x := \frac{e^x + e^{-x}}{2} \quad (\text{hyperbolic cosine})$$

$$\tanh x := \frac{\sinh x}{\cosh x} \quad (\text{hyperbolic tangent})$$

It obviously holds that $\cosh^2 x - \sinh^2 x = 1$.

1.8 Complex numbers

At a certain point in mathematics we are not able to continue with real numbers anymore. This point is reached, once we want to find a solution of the equation $x^2 + 1 = 0$. If we introduce a new element

$$i := \sqrt{-1},$$

then the equation has the two solutions $\pm i$. This extension leads us to the (field of) complex numbers.

Definition 1.62. The set of all **complex numbers** is defined by

$$\mathbb{C} := \{z = x + iy : x, y \in \mathbb{R}\}.$$

The representation of the term $z = x + iy$ is called the **canonical representation**.

For a complex number $z = x + iy$ we call x the **real part** of z and y the **imaginary part**. We write $\operatorname{Re} z = x$ and $\operatorname{Im} z = y$.

Let $z = x + iy \in \mathbb{C}$. We define the **complex conjugate** of z by

$$\bar{z} := x - iy.$$

In \mathbb{C} we have the following calculation rules for $z = x + iy$ and $w = u + iv$:

- $z + w = (x + u) + i(y + v)$
- $zw = (xu - yv) + i(xv + uy)$

These operations are associative, commutative, distributive and we have

$$z^{-1} = \frac{1}{x + iy} = \frac{1}{x + iy} \frac{x - iy}{x - iy} = \frac{x}{x^2 + y^2} - \frac{iy}{x^2 + y^2}.$$

Hence, \mathbb{C} with these operations is also a field.

Example 1.63. Consider $z = 4 + 2i$, then z is a complex number with $\operatorname{Re} z = 4$, $\operatorname{Im} z = 2$, $\bar{z} = 4 - 2i$ and $z^{-1} = \frac{1}{5} - \frac{1}{10}i$.

Example 1.64. Recall that i is just a symbol for $\sqrt{-1}$. For example, we have

$$\sqrt{-4} = \sqrt{(-1) \cdot 4} = \sqrt{4}\sqrt{-1} = 2i.$$

Note that for $x \in \mathbb{R}$ we have $x \in \mathbb{C}$ with $\operatorname{Im} x = 0$. Thus, $\mathbb{R} \subsetneq \mathbb{C}$.

Formally, one can identify the complex numbers \mathbb{C} with a tuple (x, y) of real numbers. Each complex number $z = x + iy$ can, therefore, be illustrated as a point in the plane, which is called the **complex plane**. The coordinate axes are called real and imaginary axis.

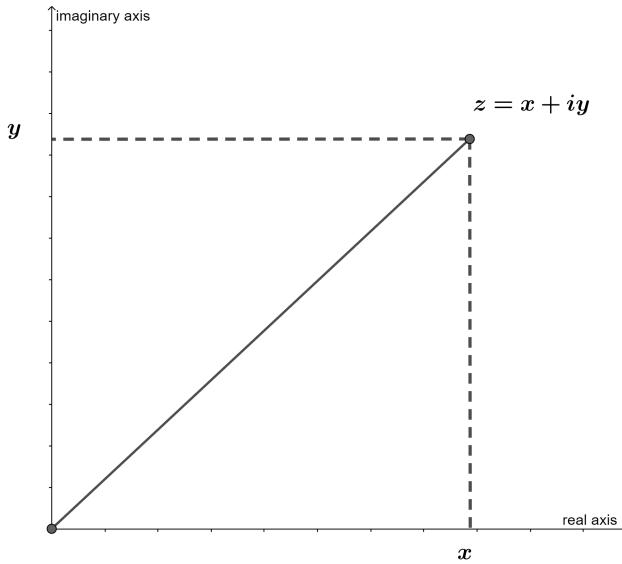


Figure 18: complex plane

(Exercise: Think about where to draw the complex conjugate \bar{z} in the complex plane.)

We define the absolute value of a complex number $z = x + iy$ to be the length of the straight line from $(0,0)$ to (x,y) .

Definition 1.65. Let $z = x + iy \in \mathbb{C}$. We define the **absolute value** of z by

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

Several calculation rules for the absolute value are just the same as for the absolute value on \mathbb{R} . (That's why we use the same name.)

Lemma 1.66. Let $z, w \in \mathbb{C}$. Then the following holds

1. $|z| \geq 0$
2. $|z| = 0 \Leftrightarrow z = 0$
3. $|z| \geq |\operatorname{Re} z|$
4. $|z| \geq |\operatorname{Im} z|$
5. $|zw| = |z||w|$

Proof. 1. - 4. are left as an exercise to the reader. For 5. we see that

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

□

Again, we have the important triangle inequality.

Lemma 1.67 (Triangle inequality). *Let $z, w \in \mathbb{C}$. Then we have*

$$|z + w| \leq |z| + |w|,$$

with equality if and only if $z\bar{w} \geq 0$. (Note that $z\bar{w} \geq 0$ means, in particular, that $z\bar{w} \in \mathbb{R}$.)

Moreover, we have

$$\left| |z| - |w| \right| \leq |z - w|.$$

Proof. For the triangle inequality, observe that

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

Note that the only inequality here is an equality if $z\bar{w} \geq 0$.

For the second inequality, note that

$$|x + y| - |y| \leq |x|.$$

Using $x = z + w$ and $y = -w$ we get

$$|z| - |w| \leq |z + w|$$

and for $y = -z$ we get

$$|w| - |z| \leq |z + w|.$$

Combining these results we get the desired inequality. \square

Let us give a summary about the representation and visualization of complex numbers in the complex plane:

\mathbb{C}	the plane (\mathbb{R}^2)
$z = x + iy$	point with coordinates (x, y)
$\operatorname{Re} z$	x -coordinate
$\operatorname{Im} z$	y -coordinate
real numbers	points on the x -axis
pure imaginary numbers	points on the y -axis
$-z$	z reflected at the origin
\bar{z}	z reflected at the x -axis
$ z $	length, or distance of z to the origin
$ z_1 - z_2 $	distance between z_1 and z_2
$z_1 + z_2$	'vector addition'
$ z_1 + z_2 \leq z_1 + z_2 $	triangle inequality

Instead of using the cartesian coordinates (x, y) for a complex number $z = x + iy$, one can also switch to the so called **polar coordinates** (r, φ) . The following relations hold between them (see Figure 15 and 18):

$$\begin{aligned} r &:= |z| := \sqrt{x^2 + y^2} = \sqrt{(\operatorname{Re} z)^2 + (\operatorname{Im} z)^2} = \sqrt{z\bar{z}} \\ x &:= |z| \cos \varphi \\ y &:= |z| \sin \varphi \end{aligned}$$

Here, we use that every point on the circle $\{(x, y) : x^2 + y^2 = 1\}$ can be written as $(\cos \varphi, \sin \varphi)$ for some $\varphi \in \mathbb{R}$. Actually, one can find such a $\varphi \in [0, 2\pi)$, and this must satisfy $\tan \varphi = \frac{y}{x}$.

Therefore, we can write z in the **trigonometric or polar form**:

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

We call $r = |z|$ the **radius**, and φ the **argument** of the complex number z .

The following question arises: under which condition is the representation (r, φ) unique?
Answer: Yes, if $|z| \neq 0$ and we assume $\varphi \in [0, 2\pi)$ (or any other interval of length 2π).

An easy and useful way of writing complex numbers can be obtained by using **Euler's formula**

$$e^{i\varphi} := \cos \varphi + i \sin \varphi \quad \varphi \in \mathbb{R}.$$

Note that $e^{i\varphi}$ can at the moment only be understood as a symbol for the right hand side above, and it is already useful as such. However, we will see later that this is really an equation, i.e., we will also define e^z for $z \in \mathbb{C}$ and show the above.

To work with this formula, it is essential to mind the following **important values**:

φ	0	$\frac{\pi}{6}$	$\frac{\pi}{4}$	$\frac{\pi}{3}$	$\frac{\pi}{2}$	$\frac{2\pi}{3}$	$\frac{3\pi}{4}$	$\frac{5\pi}{6}$	π
$\sin(\varphi)$	$\frac{\sqrt{0}}{2}$	$\frac{\sqrt{1}}{2}$	$\frac{\sqrt{2}}{2}$	$\frac{\sqrt{3}}{2}$	$\frac{\sqrt{4}}{2}$	$\frac{\sqrt{3}}{2}$	$\frac{\sqrt{2}}{2}$	$\frac{\sqrt{1}}{2}$	$\frac{\sqrt{0}}{2}$
$\cos(\varphi)$	$\frac{\sqrt{4}}{2}$	$\frac{\sqrt{3}}{2}$	$\frac{\sqrt{2}}{2}$	$\frac{\sqrt{1}}{2}$	$\frac{\sqrt{0}}{2}$	$\frac{-\sqrt{1}}{2}$	$\frac{-\sqrt{2}}{2}$	$\frac{-\sqrt{3}}{2}$	$\frac{-\sqrt{4}}{2}$

All remaining important values can be obtained from

$$\sin(x + \pi) = -\sin(x) \quad \text{and} \quad \cos(x + \pi) = -\cos(x)$$

Using these values of the trigonometric functions at the points $\frac{\pi}{2}, \pi, \frac{3\pi}{2}, 2\pi$ we obtain

$$e^{i\frac{\pi}{2}} = i \quad e^{i\pi} = -1 \quad e^{i\frac{3\pi}{2}} = -i \quad e^{i2\pi} = e^{i0} = 1.$$

Additionally, we obtain from these values, together with the periodicity of the trigonometric functions (or standard calculation rules for exponentials), that for $k \in \mathbb{Z}$,

$$e^{ik\pi} = \begin{cases} 1, & \text{if } k \text{ is even,} \\ -1, & \text{if } k \text{ is odd.} \end{cases}$$

(Verify that!)

Using Euler's formula we can write every complex number (in its polar form) as

$$z = re^{i\varphi}.$$

This representation is particularly useful when it comes to multiplication and powers of complex numbers. Given two complex numbers

$$z = re^{i\varphi} \quad \text{and} \quad w = se^{i\psi},$$

we obtain

$$zw = re^{i\varphi} se^{i\psi} = rs e^{i(\varphi+\psi)}.$$

From this formula (with $z = w$) we obtain by induction **de Moivre's formula** for powers z^n of $z = r(\cos \varphi + i \sin \varphi) = re^{i\varphi}$ with $n \in \mathbb{N}$:

$$z^n = r^n (\cos n\varphi + i \sin n\varphi) = r^n e^{in\varphi}.$$

Example 1.68. As an example we calculate $(1+i)^{42}$. We set $z := 1+i$ and bring z in its trigonometric form

$$1+i = \sqrt{2} \left(\cos \frac{\pi}{4} + i \sin \frac{\pi}{4} \right).$$

(Verify this in detail!) By de Moivre's formula we get

$$(1+i)^{42} = (\sqrt{2})^{42} \left(\cos \frac{42\pi}{4} + i \sin \frac{42\pi}{4} \right) = 2^{21} \left(\cos \frac{\pi}{2} + i \sin \frac{\pi}{2} \right) = 2^{21}i.$$

With Euler's formula one may also write this in a more compact way:

$$(1+i)^{42} = (\sqrt{2}e^{\frac{\pi i}{4}})^{42} = 2^{21}e^{\frac{42\pi i}{4}} = 2^{21}e^{10\pi i}e^{\frac{\pi i}{2}} = 2^{21}i.$$

Caution: The above consideration only holds for integer exponents. (Why also for negative integers?) More general exponents need more care and will not be discussed here.

Moreover, the polar form is also very useful for theoretical purposes.

Example 1.69. Show that $|z+w| = |z| + |w|$ for $z, w \in \mathbb{C}$ if and only if z and w have the same argument. (Hint: Consider Lemma 1.67 together with the polar form of z and w .)

It is an important result in mathematics, that complex numbers are really all we need to solve polynomial equations. In fact, the *Fundamental theorem of algebra* (in one of its variants) even gives the precise answer for the number of solutions of polynomial equations. A proof of this result is beyond the scope of this course.

Theorem 1.70 (Fundamental theorem of algebra). *Let $a_k \in \mathbb{C}$ for $k = 0, 1, 2, \dots, n$ such that $a_n \neq 0$. Then, we have the equality*

$$\sum_{k=0}^n a_k z^k = a_n \cdot \prod_{k=1}^n (z - z_k)$$

for some (not necessarily different) $z_1, \dots, z_n \in \mathbb{C}$.

Clearly, each z_k is a **root** of the polynomial $\sum_{k=0}^n a_k z^k$, i.e.,

$$\sum_{k=0}^n a_k \cdot (z_\ell)^k = 0 \quad \text{for all } \ell = 1, \dots, n.$$

Example 1.71. For example, we have

$$z^3 - z^2 + z - 1 = (z^2 + 1)(z - 1) = (z + i)(z - i)(z - 1),$$

and hence, the zeros of the polynomial are 1, i and $-i$.

Note that it is a hard problem in general to find the zeros of a given polynomial of high degree. For such tasks, one usually employs numerical software which, in most cases, can only output *approximations* of the zeros.

1.9 Vectors and norms

Let us finally discuss the important concepts related to **vectors**

$$v = (v_i)_{i=1}^d = \begin{pmatrix} v_1 \\ v_2 \\ \vdots \\ v_d \end{pmatrix} \in \mathbb{C}^d,$$

with d being called the **dimension**. (\mathbb{C}^d denotes the d -fold Cartesian product of \mathbb{C} .) Note that d -dimensional vectors $v \in \mathbb{C}^d$ consist of the d **components** $v_1, \dots, v_d \in \mathbb{C}$.

We define the **addition** and **scalar multiplication** of vectors **component-wise**. That is, for two vectors $u = (u_i)_{i=1}^d, v = (v_i)_{i=1}^d \in \mathbb{C}^d$ and a number $\lambda \in \mathbb{C}$, we define

$$u + v := \begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ u_d \end{pmatrix} + \begin{pmatrix} v_1 \\ v_2 \\ \vdots \\ v_d \end{pmatrix} = \begin{pmatrix} u_1 + v_1 \\ u_2 + v_2 \\ \vdots \\ u_d + v_d \end{pmatrix} \quad \text{and} \quad \lambda \cdot v := \begin{pmatrix} \lambda v_1 \\ \lambda v_2 \\ \vdots \\ \lambda v_d \end{pmatrix}$$

Note that it is important for vector addition that the vectors have the same dimension. If the dimensions of the two vectors do not agree, then their sum is not defined.

(The term “scalar” is used to distinguish this multiplication from the others discussed below.)

Remark 1.72. Here we use the field of complex numbers \mathbb{C} to ‘build’ **complex vectors** $v \in \mathbb{C}^d$. Note that we can easily also consider only **real vectors** $v \in \mathbb{R}^d$, as real vectors are special complex vectors. However, since all definitions here work directly in the complex case, and this will be needed later, we define it in the more general context and comment on necessary changes when needed. Moreover, note that we could define vectors, and the corresponding operations, in a much more general context, as long as the operations for the components are well-defined. This will be discussed much later.

In analogy to the real and complex numbers above, we want to define (and use) a quantity that allows for measuring ‘how large’ a vector is.

Let us begin with the most important choice for such a quantity:

Definition 1.73. Let $v = (v_i)_{i=1}^d \in \mathbb{C}^d$.

Then, we define the **Euclidean norm** (or just **length**) of v by

$$\|v\|_2 := \sqrt{\sum_{i=1}^d |v_i|^2}.$$

Moreover, for $u = (u_i)_{i=1}^d, v = (v_i)_{i=1}^d \in \mathbb{C}^d$ we define the **inner product** (or *dot product*) of u and v by

$$\langle u, v \rangle := \sum_{i=1}^d u_i \bar{v}_i.$$

With this, we have the useful representation

$$\|v\|_2 = \sqrt{\langle v, v \rangle}.$$

For real vectors $v \in \mathbb{R}^d$, these definitions simplify to

$$\|v\|_2 := \sqrt{\sum_{i=1}^d v_i^2} \quad \text{and} \quad \langle u, v \rangle := \sum_{i=1}^d u_i v_i.$$

Remark 1.74. We use the subscript '2' here, because we will study also other norms later. Note that some authors use the notation $|\cdot|$ also for the Euclidean norm, which shows its role as generalization of the absolute value.

From now on, we use x, y, z also as symbols for vectors for notational convenience.

The inner product is (formally) a mapping $\langle \cdot, \cdot \rangle: \mathbb{C}^d \times \mathbb{C}^d \rightarrow \mathbb{C}$, and we have the following properties for all $x, y, z \in \mathbb{C}^d$ and $\lambda \in \mathbb{C}$:

1. Positive definiteness: $x \neq 0 \implies \langle x, x \rangle > 0$ (In particular, $\langle x, x \rangle \in \mathbb{R}$)
2. Linearity in first argument: $\langle x + y, z \rangle = \langle x, z \rangle + \langle y, z \rangle$ and $\langle \lambda x, y \rangle = \lambda \langle x, y \rangle$
3. Conjugate symmetry: $\langle x, y \rangle = \overline{\langle y, x \rangle}$

These properties together imply

$$\langle \lambda x + \mu y, z \rangle = \lambda \langle x, z \rangle + \mu \langle y, z \rangle$$

and

$$\langle x, \lambda y + \mu z \rangle = \bar{\lambda} \langle x, y \rangle + \bar{\mu} \langle x, z \rangle.$$

for all $x, y, z \in \mathbb{C}^d$ and $\lambda, \mu \in \mathbb{C}$. (Verify this!) Note that we need to take the complex conjugate when we take a scalar out of the 'second input'. One says, the inner product is *sesquilinear*.

From these, we directly obtain that the Euclidean norm satisfies

- 1) $\|v\|_2 = 0 \iff v = 0$, where we write $0 := (0)_{i=1}^d$ for the **zero vector**, and
- 2) for any $\lambda \in \mathbb{C}$ and any $v \in \mathbb{C}^d$ we have $\|\lambda v\|_2 = |\lambda| \cdot \|v\|_2$.

The first means that the norm of v is zero if and only if all components of v are zero (which is called *definiteness*), and the second property is called *homogeneity*. Both properties were also clear for the absolute value of a real/complex number, which is also covered by the above in the case $d = 1$. However, there is a third property that is essential for the upcoming considerations. Namely, the triangle inequality, see also Theorem 1.59 and Lemma 1.66.

Theorem 1.75 (Triangle inequality). *For all $x, y \in \mathbb{C}^d$ we have*

$$\|x + y\|_2 \leq \|x\|_2 + \|y\|_2.$$

Moreover, the equality $\|x + y\|_2 = \|x\|_2 + \|y\|_2$ holds if and only if $y = \alpha \cdot x$ for some $\alpha \geq 0$.

Note that in the case that $d = 2$ or $d = 3$ the Euclidean norm coincides with the usual intuition of the 'length' between 0 and the point x , i.e., $\|\cdot\|_2$ is the length of the 'direct way' between 0 and x . This makes the triangle inequality appear to be an obvious statement. However, when it comes to $d > 3$ this might not be that clear, and therefore we present a proof below.

For this, we first need another inequality, which is also of independent interest. The **Cauchy-Schwarz (CS) inequality** gives a relation between the inner product and the Euclidean norm.

Lemma 1.76 (Cauchy-Schwarz inequality). *For $x, y \in \mathbb{C}^d$ we have that*

$$|\langle x, y \rangle| \leq \|x\|_2 \|y\|_2.$$

Moreover, we have the equality $|\langle x, y \rangle| = \|x\|_2 \|y\|_2$ if and only if $y = c \cdot x$ for some $c \in \mathbb{C}$.

Remark 1.77. The choice of the Euclidean norm for the following analysis is quite arbitrary, especially when it comes to non-geometrical applications, and there are sometimes other natural choices. We will discuss some other possibilities soon. However, although the triangle inequality is an essential property also for 'other norms', the (sometimes very helpful) Cauchy-Schwarz inequality is special to the Euclidean norm. Therefore, we usually work with this norm.

Proof. Let us prove the second statement. The first follows by squaring both sides.

First, if either $x = 0$ or $y = 0$, then the statement is clearly true. Otherwise, we use the inequality $ab \leq \frac{a^2+b^2}{2}$, which holds for any real numbers a, b , and follows from $(a - b)^2 \geq 0$. (Verify this!) We define

$$a_i := \frac{|x_i|}{\|x\|_2} \quad \text{and} \quad b_i := \frac{|y_i|}{\|y\|_2}.$$

Observe that, by definition,

$$\sum_{i=1}^d a_i^2 = \sum_{i=1}^d b_i^2 = 1,$$

and therefore also

$$\sum_{i=1}^d \frac{a_i^2 + b_i^2}{2} = 1.$$

This yields that

$$\begin{aligned} \left| \sum_{i=1}^d x_i \bar{y}_i \right| &\leq \sum_{i=1}^d |x_i| \cdot |y_i| \\ &= \|x\|_2 \|y\|_2 \sum_{i=1}^d a_i b_i \\ &\leq \|x\|_2 \|y\|_2 \sum_{i=1}^d \frac{a_i^2 + b_i^2}{2} \\ &= \|x\|_2 \|y\|_2. \end{aligned}$$

(Note that the first inequality, was just the triangle inequality for sums of numbers.)

To obtain equality, both above used inequalities need to be equalities. Equality in the first is equivalent to $x_i \bar{y}_i$ having the same argument for all i , see Exercise 1.69. Equality in the second is equivalent to $a_i = b_i$ for all i (Check that!), which means $|x_i| = r|y_i|$ for all i and some $r > 0$. Since we fixed argument and absolute value, we obtain equality if and only if $x = c \cdot y$ for some $c \in \mathbb{C}$.

□

The CS inequality can now be used to prove the triangle inequality for $\|\cdot\|_2$.

Proof of Theorem 1.75. For $x, y \in \mathbb{C}^d$ we obtain

$$\begin{aligned} \|x + y\|_2^2 &= \langle x + y, x + y \rangle = \langle x, x \rangle + \langle x, y \rangle + \langle y, x \rangle + \langle y, y \rangle \\ &= \|x\|_2^2 + \langle x, y \rangle + \overline{\langle x, y \rangle} + \|y\|_2^2 \\ &= \|x\|_2^2 + 2\operatorname{Re}(\langle x, y \rangle) + \|y\|_2^2 \\ &\leq \|x\|_2^2 + 2|\langle x, y \rangle| + \|y\|_2^2 \\ &\leq \|x\|_2^2 + 2\|x\|_2 \cdot \|y\|_2 + \|y\|_2^2 \\ &= (\|x\|_2 + \|y\|_2)^2. \end{aligned}$$

To obtain equality, note that Lemma 1.76 shows that we need $y = c \cdot x$ for some $c \in \mathbb{C}$ to obtain equality in the second inequality. Using this, we need $\operatorname{Re}(c) = |c|$ to achieve equality in the first inequality above. This holds iff $c \geq 0$. □

We finally introduce some particular important vectors which will appear very often in the upcoming considerations. These are the **unit vectors** $e_1, \dots, e_d \in \mathbb{R}^d$, where e_k is the vector which is zero except for the k -th entry, which is one. That is,

$$e_1 = \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}, \quad e_2 = \begin{pmatrix} 0 \\ 1 \\ \vdots \\ 0 \end{pmatrix}, \quad \dots, \quad e_d = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 1 \end{pmatrix}.$$

Note that they all have norm 1, independent of which norm above one chooses.

One important property of the unit vectors is, that they can be used to **represent arbitrary vectors**. For this, note that $\lambda \cdot e_k$ with $\lambda \in \mathbb{R}$ is the vector with λ in the k -th entry, and zero elsewhere. It is therefore easy to see that

$$\begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_d \end{pmatrix} = \sum_{k=1}^d x_k \cdot e_k \in \mathbb{R}^d.$$

One might guess that working with a representation as given on the right can be quite useful.

Remark 1.78. Although we presented such a representation only with respect to the unit vectors e_1, \dots, e_d , we will see later that this works also with other vectors. A set of vectors that can be used to represent arbitrary vectors in a unique way as above is called a **basis** of \mathbb{R}^d . In this context, the set $\{e_1, \dots, e_d\}$ is called the **standard basis** of \mathbb{R}^d , and the numbers x_k are called **coordinates** of x with respect to $\{e_1, \dots, e_d\}$. It is important to note again that the $x_k \in \mathbb{R}$ are just numbers, and only the $e_k \in \mathbb{R}^d$ are vectors (i.e., elements from \mathbb{R}^d). We will discuss this (and the related concept of a vector space) later in more detail.

2 Matrices and systems of linear equations

In this chapter we want to solve *systems of equations*. That is, we want to find possible values for some *free variables* (or parameters etc.) that fulfill a certain collection of equations. This is one of the most important disciplines in applied mathematics and numerical applications. Here, we start by discussing linear equations, and show how they can be solved explicitly. These equations are called *linear* because they depend on a variable only in a linear way.

For *one free variable*, linear equations are easy to solve:

If we want to solve the equation $ax = b$ for some $a, b \in \mathbb{R}$, then the solution is clearly given by $x = \frac{b}{a}$ if $a \neq 0$. However, if $a = 0$ then this equation can be solved if and only if $b = ax = 0$. That is, for $b \neq 0$ there is no x satisfying the equation, while for $b = 0$ this equation is fulfilled for every $x \in \mathbb{R}$.

For *two free variables*, it is already more involved:

Assume you want to find $x_1, x_2 \in \mathbb{R}$, such that the equations

$$\begin{aligned} a_{11} \cdot x_1 + a_{12} \cdot x_2 &= b_1, \\ a_{21} \cdot x_1 + a_{22} \cdot x_2 &= b_2 \end{aligned}$$

are fulfilled for some given $a_{11}, a_{12}, a_{21}, a_{22}, b_1, b_2 \in \mathbb{R}$. (Note that we use subscripts $a_{k\ell}$ in our notation to keep track 'where' a coefficient appears in the system of equations.)

As before, we see that the first equation is equivalent to

$$x_1 = \frac{b_1 - a_{12} \cdot x_2}{a_{11}} \quad \text{if } a_{11} \neq 0.$$

This can be put into the second equation, which then only depends on the unknown x_2 , and can therefore (potentially) be solved as in the one dimensional case. In this way, we may find a unique solution for x_2 which then implies a solution for x_1 by the equation above. By this procedure, we can find solutions to linear equations, also for larger systems.

However, it is not clear if there really is a (*unique*) *solution* for x_1 and x_2 for every choice of $a_{11}, a_{12}, a_{21}, a_{22}, b_1, b_2 \in \mathbb{R}$. There might also be *no* or *infinitely many* solutions (as for a single equation with $a = 0$).

For demonstration, let us discuss a specific example:

Consider the system of linear equations

$$\begin{aligned} 2x_1 + x_2 &= 1, \\ 6x_1 + 3x_2 &= 2. \end{aligned}$$

The first equation is equivalent to $x_2 = 1 - 2x_1$. If we put this into the second equation, we obtain $6x_1 + 3 - 6x_1 = 2$ which is never fulfilled. That is, there is *no solution* to this system of equations. However, if we change the system to

$$\begin{aligned} 2x_1 + x_2 &= 1, \\ 6x_1 + 3x_2 &= 3, \end{aligned}$$

then the first equation is still equivalent to $x_2 = 1 - 2x_1$. But, after putting this into the second equation, we have $6x_1 + 3 - 6x_1 = 3$ which holds *for every* $x_1 \in \mathbb{R}$. Therefore, the above system is solved by all $(x_1, x_2) \in \mathbb{R}^2$ with $2x_1 + x_2 = 1$, e.g., for $(x_1, x_2) = (0, 1)$ or $(x_1, x_2) = (1, -1)$.

This shows that such systems of equations might be quite sensitive to small changes of the parameters (and this was just a two dimensional example). It is therefore desirable to have

criteria for a given system of equations to be (uniquely) solvable that can be checked more easily and before we start trying to calculate a solution.

Moreover, this procedure is useful only for 'small' systems of equations, say with at most 3 unknowns. It is rather impractical for larger systems, and there are faster methods to solve such systems by hand (and with a computer). This is particularly important since modern applications are usually 'high dimensional'.

The most convenient way to formally work with linear systems is to introduce *matrices*, which is just a way of writing numbers in an array similarly to vectors. For example, we will write the above system of equations by $Ax = b$, where the matrix $A \in \mathbb{R}^{2 \times 2}$ (means that it is a 2×2 array of real numbers) and the vector $b \in \mathbb{R}^2$ are given by

$$A = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} \quad \text{and} \quad b = \begin{pmatrix} b_1 \\ b_2 \end{pmatrix}.$$

We discuss shortly how the *operation* Ax , i.e., *matrix-vector multiplication*, is precisely defined, and that a system of equations may then be written by $Ax = b$. This now looks almost like the one dimensional equation, and one might like to just "divide by A if it is not zero" to obtain a solution. Unfortunately, that's not (always) so easy, and we have to be careful with what it means that " $A \neq 0$ " or to divide by a matrix.

However, we will see that matrices are the 'right' tool to work with such problems, and we will introduce operations and calculation rules for matrices, which will enable us to *transform* systems of equations to others which might be easier to solve. By this, we introduce techniques to solve also large systems of equations in a straight-forward way.

2.1 Matrices

Let us start by introducing *matrices*.

Definition 2.1. Let $n, m \in \mathbb{N}$ and $a_{ij} \in \mathbb{R}$ for $1 \leq i \leq n$ and $1 \leq j \leq m$.

A (real) $m \times n$ -**matrix** (read "m by n matrix") is an array given by

$$A = (a_{ij})_{i,j=1}^{m,n} = \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \dots & a_{mn} \end{pmatrix}.$$

In this case we use the notation $A \in \mathbb{R}^{m \times n}$, and call m and n the **dimensions** of A .

A $m \times 1$ matrix is called a **column vector**, a $1 \times n$ matrix is called a **row vector**, and if $m = n$, then the matrix is called **quadratic**, or a **square matrix**.

Remark 2.2. We mostly consider only matrices of real numbers here. The case of *complex matrices*, i.e., $a_{ij} \in \mathbb{C}$, can be treated analogously, and we write $A \in \mathbb{C}^{m \times n}$. We will comment on differences of the complex case if needed.

Remark 2.3. The index notation at $(a_{ij})_{i,j=1}^{m,n}$ just means that we consider $i = 1, \dots, m$ and $j = 1, \dots, n$. Some authors use $(a_{ij})_{i=1,j=1}^{m,n}$ or even $(a_{ij})_{i=1,\dots,n,j=1,\dots,n}$ for the same.

We now turn to basic operations of matrices. The first two, namely *scalar multiplication* and *matrix addition*, are very easy (and familiar from the corresponding operations for vectors). These operations are **component-wise operations**, meaning that they are performed in each entry of the matrices individually.

The **matrix addition** of two $m \times n$ -matrices $A = (a_{ij})_{i,j=1}^{m,n}$ and $B = (b_{ij})_{i,j=1}^{m,n}$ is defined by *component-wise addition*, i.e.,

$$A + B = (a_{ij} + b_{ij})_{i,j=1}^{m,n} = \begin{pmatrix} a_{11} + b_{11} & a_{12} + b_{12} & \dots & a_{1n} + b_{1n} \\ a_{21} + b_{21} & a_{22} + b_{22} & \dots & a_{2n} + b_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} + b_{m1} & a_{m2} + b_{m2} & \dots & a_{mn} + b_{mn} \end{pmatrix}.$$

Note that it is important here that the matrices have the same dimension. If the dimensions of the two matrices do not agree, then their sum is not defined.

The second operation is **scalar multiplication**, which is the *component-wise product* of a matrix, say $A = (a_{ij})_{i,j=1}^{m,n} \in \mathbb{R}^{m \times n}$, with a (real or complex) number $\lambda \in \mathbb{R}$. That is,

$$\lambda \cdot A = \begin{pmatrix} \lambda a_{11} & \lambda a_{12} & \dots & \lambda a_{1n} \\ \lambda a_{21} & \lambda a_{22} & \dots & \lambda a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \lambda a_{m1} & \lambda a_{m2} & \dots & \lambda a_{mn} \end{pmatrix}$$

(The term “scalar” is used to distinguish this from the matrix product discussed below.)

The third basic operation is the **matrix product**, i.e., the product of two matrices. In contrast to the last operations, the product of two matrices is not defined component-wise:

Given a $m \times n$ -matrix $A = (a_{ij})_{i,j=1}^{m,n}$ and a $n \times p$ -matrix $B = (b_{ij})_{i,j=1}^{n,p}$, we define the $m \times p$ -matrix $C = (c_{ij})_{i,j=1}^{m,p}$ with

$$c_{ij} = \sum_{k=1}^n a_{ik} b_{kj},$$

as the product of A and B , i.e., $C = AB$. This definition shows that the ij -th entry of the product C , i.e., c_{ij} , depends only on the i -th row of A and the j -th column of B . One may mind this rule by “ c_{ij} is the product of the i -th row of A with the j -th column of B ”.

To make this more precise, note that a matrix $A \in \mathbb{R}^{m \times n}$ has m rows of length n , or n columns of length m . So, let us assume that a matrix A has the **rows** $a_1, a_2, \dots, a_m \in \mathbb{R}^{1 \times n}$ and **columns** $c_1, c_2, \dots, c_n \in \mathbb{R}^{m \times 1} = \mathbb{R}^m$. We use the notation

$$A = \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_m \end{pmatrix} = (c_1, c_2, \dots, c_n). \quad (2.1)$$

Note that a_k and c_k are not numbers, but vectors. However, the notation is consistent in the sense that a matrix can be seen as a row vector consisting of column vectors, and vice versa.

Remark 2.4. It does not matter if we put commas in $A = (c_1, c_2, \dots, c_n)$ or not. Both notations are used, and there is actually no room for misunderstanding once we specified clearly what the entries “ c_k ” are.

Moreover, the **matrix-vector product** Ax of a matrix $A = (a_{ij})_{i,j=1}^{m,n}$ and a (column) vector $x = (x_j)_{j=1}^n \in \mathbb{R}^n$ is defined through the matrix product by

$$Ax = \begin{pmatrix} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n \\ a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n \\ \vdots \\ a_{m1}x_1 + a_{m2}x_2 + \cdots + a_{mn}x_n \end{pmatrix} = \begin{pmatrix} \langle a_1, x \rangle \\ \langle a_2, x \rangle \\ \vdots \\ \langle a_m, x \rangle \end{pmatrix} \in \mathbb{R}^m,$$

where $\langle a_i, x \rangle = \sum_{j=1}^n a_{ij}x_j$ is the inner product of a_i and x .

It follows from the definition of the matrix multiplication, that for a matrix $A \in \mathbb{R}^{m \times n}$ of the form (2.1), and another matrix $B \in \mathbb{R}^{n \times p}$ with columns b_1, \dots, b_p , i.e., $B = (b_1, \dots, b_p)$, we have

$$AB = (Ab_1, Ab_2, \dots, Ab_p) = \begin{pmatrix} a_1B \\ a_2B \\ \vdots \\ a_mB \end{pmatrix} = (\langle a_i, b_j \rangle)_{i,j=1}^{m,p} \in \mathbb{R}^{m \times p}.$$

That is, the ij -th entry of AB is the inner product of the i -th row of A with the j -th column of B , as stated above.

(Note that all inner dimensions for the involved matrix(-vector)-products agree.)

Let us now see an example.

Example 2.5. We calculate $C = AB$, where

$$A = \begin{pmatrix} 1 & 6 \\ 2 & 5 \\ 3 & 4 \end{pmatrix} \quad \text{and} \quad B = \begin{pmatrix} 7 & 9 & 1 \\ 8 & 0 & 0 \end{pmatrix}$$

Since $A \in \mathbb{R}^{3 \times 2}$ and $B \in \mathbb{R}^{2 \times 3}$, the product $C = AB \in \mathbb{R}^{3 \times 3}$ will be a 3×3 -matrix.

We calculate the upper left entry $c_{11} = 1 \cdot 7 + 6 \cdot 8 = 55$. In matrix form this looks like

$$\begin{pmatrix} 1 & 6 \\ 2 & 5 \\ 3 & 4 \end{pmatrix} \begin{pmatrix} 7 & 9 & 1 \\ 8 & 0 & 0 \end{pmatrix} = \begin{pmatrix} 55 & * & * \\ * & * & * \\ * & * & * \end{pmatrix},$$

where $*$ is used for entries we do not know yet.

Next we use the first row of A and the second column of B to calculate $c_{12} = 1 \cdot 9 + 6 \cdot 0 = 9$:

$$\begin{pmatrix} 1 & 6 \\ 2 & 5 \\ 3 & 4 \end{pmatrix} \begin{pmatrix} 7 & 9 & 1 \\ 8 & 0 & 0 \end{pmatrix} = \begin{pmatrix} 55 & 9 & * \\ * & * & * \\ * & * & * \end{pmatrix}.$$

To compute the second row of C we have to use the second row of A :

$$\begin{pmatrix} 1 & 6 \\ 2 & 5 \\ 3 & 4 \end{pmatrix} \begin{pmatrix} 7 & 9 & 1 \\ 8 & 0 & 0 \end{pmatrix} = \begin{pmatrix} 55 & 9 & 1 \\ 54 & * & * \\ * & * & * \end{pmatrix},$$

$$\begin{pmatrix} 1 & 6 \\ 2 & 5 \\ 3 & 4 \end{pmatrix} \begin{pmatrix} 7 & 9 & 1 \\ 8 & 0 & 0 \end{pmatrix} = \begin{pmatrix} 55 & 9 & 1 \\ 54 & 18 & * \\ * & * & * \end{pmatrix}.$$

Continuing this procedure we finally get

$$C = AB = \begin{pmatrix} 55 & 9 & 1 \\ 54 & 18 & 2 \\ 53 & 27 & 3 \end{pmatrix}.$$

Note that in this case, also the matrix BA is defined. However, the product BA is a 2×2 -matrix. Namely,

$$BA = \begin{pmatrix} 28 & 91 \\ 8 & 48 \end{pmatrix},$$

and there is no obvious relation between AB and BA .

Note that the matrix product is only defined if the inner dimensions agree. That is, if we want to multiply a $m \times p$ -matrix $A \in \mathbb{R}^{m \times p}$ and a $q \times n$ -matrix $B \in \mathbb{R}^{q \times n}$, then we need that $p = q$. Otherwise, the product is not defined. Note that this implies that we might define the product AB of two matrices A and B , but the “reverse” product BA is not defined. Consider, e.g., the case $A \in \mathbb{R}^{m \times n}$ and $B \in \mathbb{R}^{n \times n}$ with $m \neq n$.

The following rules for calculation, which may remind on the respective rules for numbers, follow easily from the definition:

- $\lambda(A + B) = \lambda A + \lambda B$ for all matrices $A, B \in \mathbb{R}^{m \times n}$,
- $A(B + C) = AB + AC \in \mathbb{R}^{m \times n}$ if $A \in \mathbb{R}^{m \times p}$ and $B, C \in \mathbb{R}^{p \times n}$.

However, even if $A, B \in \mathbb{R}^{n \times n}$, i.e., A and B are quadratic such that AB and BA is defined, we do not have in general that $AB = BA$. That is, matrix multiplication is not commutative.

Moreover, there are **identity elements** for these operations, i.e., there are matrices such that addition/multiplication of them do not change the second matrix. This corresponds to 0 and 1 for real and complex numbers.

For this, let us introduce the following special matrices.

For $m, n \in \mathbb{N}$, we define the **zero matrix** $0_{mn} \in \mathbb{R}^{m \times n}$ by

$$0_{mn} := \begin{pmatrix} 0 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & 0 \end{pmatrix},$$

i.e., $0_{mn} = (a_{ij})_{i,j=1}^{m,n}$ with $a_{ij} = 0$ for all i, j .

If the dimensions of the matrices under consideration are clear, we may just write $0 := 0_{mn}$.

For $n \in \mathbb{N}$, we define the **identity matrix**, or just **identity**, $I_n \in \mathbb{R}^{n \times n}$ by

$$I_n := \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{pmatrix}.$$

This matrix may be written as $I_n = (\delta_{ij})_{i,j=1}^{n,n}$, where δ_{ij} is the **Kronecker delta** defined by

$$\delta_{ij} = \begin{cases} 1, & \text{if } i = j, \\ 0, & \text{otherwise.} \end{cases}$$

Note that the identity is a square matrix, and we may write $I := I_n$ if the dimension is clear.

Let us show formally that the identity matrix is an identity for matrix multiplication, see the *field axioms* (Axiom 1). However, note that we need a different dimension of the identities if the 'other matrix' is not quadratic.

Example 2.6. Let $I_n \in \mathbb{R}^{n \times n}$ and $I_m \in \mathbb{R}^{m \times m}$ be the identity matrices of the given dimensions, and let $A \in \mathbb{R}^{m \times n}$ be an arbitrary $m \times n$ -matrix. Let us check that

$$I_m \cdot A = A \cdot I_n = A.$$

For this, we compute the ij -th entry of $I_m \cdot A$, which we call $(I_m A)_{ij}$. By definition

$$(I_m A)_{ij} = \sum_{k=1}^m \delta_{ik} a_{kj},$$

where the Kronecker delta δ_{ik} is zero if $k \neq i$. Thus the sum reduces to only one term, which is $\delta_{ii} a_{ij} = a_{ij}$. This yields that the ij -th entry of the matrix product $I_m \cdot A$ is a_{ij} , i.e., $I_m \cdot A = A$. A similar calculation yields that $A \cdot I_n = A$.

Note that for a quadratic matrix $A \in \mathbb{R}^{n \times n}$, we have

$$I_n \cdot A = A \cdot I_n = A.$$

□

Recall that the **unit vectors** $e_k = (\delta_{ik})_{i=1}^n \in \mathbb{R}^n$, $k = 1, \dots, n$, are the (column) vectors that contain exactly one 1 and all other entries are zero. Let us also define the **row unit vectors** $e_k^T \in \mathbb{R}^{1 \times n}$, i.e., $e_1^T := (1, 0, \dots, 0)$, $e_2^T := (0, 1, \dots, 0)$ and so on. With them, we can write the identity by

$$I_n = (e_1, e_2, \dots, e_n) = \begin{pmatrix} e_1^T \\ e_2^T \\ \vdots \\ e_n^T \end{pmatrix},$$

i.e., e_k is the k -th column, and e_k^T is the k -th row of I_n .

With the above considerations and $I_n \cdot A = A \cdot I_n = A$, we see that the unit vectors can be used to "extract" the rows and columns from a matrix. That is, given a matrix $A \in \mathbb{R}^{n \times n}$ of the form (2.1), we obtain that

$$Ae_k = c_k \quad \text{and} \quad e_k^T A = a_k, \tag{2.2}$$

i.e., Ae_k gives the k -th column, and $e_k^T A$ gives the k -th row of A . (Verify this!) The same can be done for *rectangular* matrices $A \in \mathbb{R}^{m \times n}$, but one need to consider unit vectors of different length.

The last concept related to matrix multiplication that we will need is the *inverse of a matrix*. That is, for a given matrix $A \in \mathbb{R}^{n \times n}$, the **inverse matrix**, if it exists, is a matrix $A^{-1} \in \mathbb{R}^{n \times n}$ such that

$$AA^{-1} = A^{-1}A = I_n.$$

If an inverse exists, then we call a matrix **invertible** or **regular**, see Section 2.6.

Some matrices are clearly invertible, like the identity with $I_n^{-1} = I_n$. Others are clearly not invertible, like the zero matrix, because 0 cannot be multiplied by any matrix to become “non-zero”. But, in general, it is not easy to see whether a matrix is invertible or not. For example, the matrix $\begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix}$ is invertible, but $\begin{pmatrix} 1 & 2 \\ 3 & 6 \end{pmatrix}$ is not. We will discuss a way to verify if a matrix is invertible, and how to compute an inverse in Section 2.6.

However, let us already add here, that even if we know that a matrix is invertible, it is usually difficult (also computationally) to compute an inverse.

We will come back to this issue later, and present some ways for computing the inverse, at least for ‘small’ matrices. This will be the ultimate tool to solve (certain) systems of linear equations. But we will first discuss some more direct, but less powerful ways to calculate solutions.

Remark 2.7. Another interpretation of an inverse matrix is the following:

Recall that the matrix-vector product Ax of a matrix $A = (a_{ij})_{i=1,j=1}^{m,n}$ and a (column) vector $x \in \mathbb{R}^n$ is a vector with $Ax \in \mathbb{R}^m$. Hence, we can consider a matrix A also as a mapping (aka. function), which maps one vector to another. In this formulation, the equation $Ax = b$ is solvable if b is contained in the *range* of A , see Definition 1.10, and there is a unique solution, if A is *invertible* (\iff bijective), see Definition 1.13. We will see later that this mapping can be bijective only if $m = n$, i.e., if A is a square matrix.

We finally discuss the *transpose* of a matrix. Since the dimensions of a matrix are important, it makes a huge difference if a matrix is $m \times n$ or $n \times m$, and it is quite useful to have a compact notation to somehow ‘switch’ the rows and columns of a matrix. That is, for a given $m \times n$ matrix $A = (a_{ij})$, we define its **transpose** A^T as the $n \times m$ matrix whose rows are the columns of A . To be more precise, the ij -th component of A^T is a_{ji} , i.e.,

$$A^T = (a_{ji})_{j,i=1}^{n,m} = \begin{pmatrix} a_{11} & a_{21} & \dots & a_{m1} \\ a_{12} & a_{22} & \dots & a_{m2} \\ \vdots & \vdots & \ddots & \vdots \\ a_{1n} & a_{2n} & \dots & a_{mn} \end{pmatrix}.$$

With the above notation, we have

$$A^T = \begin{pmatrix} c_1^T \\ c_2^T \\ \vdots \\ c_n^T \end{pmatrix} = (a_1^T, a_2^T, \dots, a_n^T),$$

where the a_k (c_k) are the rows (columns) of A , and note that a transposed row vector is a column vector, and vice versa.

Example 2.8.

$$\begin{pmatrix} 1 & 6 \\ 2 & 5 \\ 3 & 4 \end{pmatrix}^T = \begin{pmatrix} 1 & 2 & 3 \\ 6 & 5 & 4 \end{pmatrix}$$

And

$$\left(\begin{pmatrix} 1 & 6 \\ 2 & 5 \\ 3 & 4 \end{pmatrix}^T \right)^T = \begin{pmatrix} 1 & 2 & 3 \\ 6 & 5 & 4 \end{pmatrix}^T = \begin{pmatrix} 1 & 6 \\ 2 & 5 \\ 3 & 4 \end{pmatrix}$$

In the above example we saw that $(A^T)^T = A$, and this obviously holds in general. (The ij -th component of $(A^T)^T$ is the ji -th component of A^T , which is a_{ij} .)

There is one calculation rule related to the transpose, that is sometimes also very useful for computing the product of matrices. We state this in the following lemma.

Lemma 2.9. *Let $m, n, p \in \mathbb{N}$, $A \in \mathbb{R}^{m \times p}$ and $B \in \mathbb{R}^{p \times n}$. Then,*

$$(AB)^T = B^T A^T.$$

(Note that the order has changed.)

In particular, this lemma shows that $(AA^T)^T = AA^T$ for every $A \in \mathbb{R}^{n \times n}$.

Proof. First of all, note that $B^T \in \mathbb{R}^{n \times p}$ and $A^T \in \mathbb{R}^{p \times m}$. Therefore, the inner dimensions of B^T and A^T agree, and their product is defined.

Now, let us write $(C)_{ij}$ for the ij -th entry of a matrix. By definition, we have that

$$(AB)_{ij} = \sum_{k=1}^p (A)_{ik} (B)_{kj},$$

and

$$(B^T A^T)_{ij} = \sum_{k=1}^p (B^T)_{ik} (A^T)_{kj} = \sum_{k=1}^p (B)_{ki} (A)_{jk} = \sum_{k=1}^p (A)_{jk} (B)_{ki} = (AB)_{ji}.$$

Since $(AB)_{ji} = ((AB)^T)_{ij}$, we see that $B^T A^T = (AB)^T$. □

Some matrices do not change under transposing them.

Definition 2.10. A matrix $A \in \mathbb{R}^{n \times n}$ with $A^T = A$ is called **symmetric**.

Note that symmetric matrices must be quadratic, and we will see later that symmetric matrices have several important properties.

The easiest examples are the identity and the (quadratic) zero matrix. Another important class of symmetric matrices are *diagonal matrices*.

Definition 2.11. A quadratic matrix $A = (a_{i,j}) \in \mathbb{R}^{n \times n}$ is a **diagonal matrix** if there exist numbers d_1, d_2, \dots, d_n such that

$$a_{ij} = \begin{cases} d_i, & \text{if } i = j \\ 0, & \text{otherwise.} \end{cases}$$

The numbers d_i are called diagonal elements of A and we write $A = \text{diag}(d_1, d_2, \dots, d_n)$.

The transpose is often used when working with vectors. Therefore, it is essential to understand especially this case.

Consider a **column vector** $x \in \mathbb{R}^n = \mathbb{R}^{n \times 1}$, then its transposed $x^T \in \mathbb{R}^{1 \times n}$ is a **row vector**. Now, since the inner dimensions in both cases agree, we can define $x^T x$ and xx^T . However, we obtain that $x^T x \in \mathbb{R} = \mathbb{R}^{1 \times 1}$ is a number, but $xx^T \in \mathbb{R}^{n \times n}$ is a quadratic matrix.

Example 2.12. Let $x = (1, 2)^T \in \mathbb{R}^2$. Then, $x^T x = 1^2 + 2^2 = 5$, but

$$xx^T = \begin{pmatrix} 1 & 2 \\ 2 & 4 \end{pmatrix}.$$

The above examples can clearly be generalized to the case of two different vectors $x, y \in \mathbb{R}^n$. That is, we can define the number $x^T y = \sum_{i=1}^n x_i y_i$, which is just the inner product of x and y . In particular, the Euclidean norm of a vector x can be written as $\|x\|_2 = \sqrt{x^T x}$.

Moreover, we can define the matrix $xy^T \in \mathbb{R}^{n \times n}$. One may even define such matrices based on vectors of different dimensions. As these matrices appear rather often in theory and applications, they have been given names.

Definition 2.13. Let $A \in \mathbb{R}^{m \times n}$. If there exist two vectors $x \in \mathbb{R}^m$ and $y \in \mathbb{R}^n$ such that

$$A = xy^T,$$

then we call A a **rank-one matrix**.

These matrices play an important role in the work with high dimensional data.

Remark 2.14. If we consider **complex-valued matrices** $A = (a_{ij}) \in \mathbb{C}^{m \times n}$, then all the definitions above still make sense. However, instead of a transpose A^T , we would speak of the **adjoint matrix** $A^* = (\overline{a_{ji}})$, i.e., the *conjugate transpose*. (That is, we take additionally the complex conjugate in each component.) Clearly, the adjoint equals the transpose if the matrix is real-valued. If $A = A^*$, we call A a **self-adjoint** (or **hermitian**) matrix.

Remark 2.15. The above calculation rules may be compared to the *field axioms* for real numbers, see Axiom 1. Note that many properties are also fulfilled for matrices. However, matrix multiplication is not *commutative*, i.e., we do not have $AB = BA$ in general, and that's why the set of matrices (of same dimensions) is not a field. (One may see that it is a *group* or even a *ring*, but we do not discuss this type of *algebraic structures* here.)

2.2 Systems of linear equations

Let us now consider *systems of linear equations*, which are the most frequently occurring type of *multivariate* (i.e., depending on more than one variable) problems to solve, although they appear to be the easiest. One reason is that many (even “non-linear”) numerical problems can be rewritten as, or approximated by, a (very large) system of linear equations. And although such systems are usually solved by a computer, it is up to the user to transfer the problem under consideration to a *well-defined* linear system. It is therefore indispensable to have a solid understanding of these basic problems.

Throughout this section, there will be the parameters $m, n \in \mathbb{N}$, where

n is the number of unknown variables

and

m is the number of equations that must be fulfilled.

The system of equations we want to solve here, will be of the following form.

Definition 2.16. Let $n, m \in \mathbb{N}$, $a_{ij} \in \mathbb{R}$ and $b_j \in \mathbb{R}$ for $1 \leq i \leq n$ and $1 \leq j \leq m$.

A **system of linear equations** or **linear system** with real coefficients is given by

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 + \dots + a_{1n}x_n &= b_1 \\ a_{21}x_1 + a_{22}x_2 + \dots + a_{2n}x_n &= b_2 \\ \vdots &\quad \vdots & \vdots & \vdots \\ a_{m1}x_1 + a_{m2}x_2 + \dots + a_{mn}x_n &= b_m \end{aligned}$$

The x_i with $1 \leq i \leq n$ are called **variables**, or *unknowns*.

The a_{ij} are called the **coefficients** of the system.

The matrix $A = (a_{ij})_{i=1,j=1}^{m,n}$ is the **matrix of coefficients**.

The tuple $b := (b_1, \dots, b_m)$ is called the **right hand side** (RHS) of the system.

If there exist such numbers $x_1, \dots, x_n \in \mathbb{R}$ that fulfill all the equations, then we call the tuple $x = (x_1, \dots, x_n)$ a **solution** to the linear system.

If there is no solution, then we call the linear system **inconsistent**.

If we recall that the matrix-vector product of the matrix of coefficients $A \in \mathbb{R}^{m \times n}$ and the vector of variables $x \in \mathbb{R}^n$ is defined by

$$Ax = \begin{pmatrix} a_{11}x_1 + a_{12}x_2 + \dots + a_{1n}x_n \\ a_{21}x_1 + a_{22}x_2 + \dots + a_{2n}x_n \\ \vdots \\ a_{m1}x_1 + a_{m2}x_2 + \dots + a_{mn}x_n \end{pmatrix} \in \mathbb{R}^m,$$

we see that the system of linear equations, given in Definition 2.16, can be written in short by

$$Ax = b.$$

Obviously, we are interested in solutions to a linear system. However, as already discussed above, such systems may have *no*, a *unique*, or even *infinitely many* solutions. We will see that

this can be verified by analyzing the matrix of coefficients more detailed. Before we come to this, let us introduce some more notation and discuss some examples.

Definition 2.17. Given a linear system $Ax = b$ with coefficient matrix A and RHS b , then we denote the **set of solutions** by

$$L(A, b) := \{x \in \mathbb{R}^n : Ax = b\}.$$

Example 2.18. Let us consider the examples from the beginning of Section 2. That is, we want to solve the system

$$\begin{aligned} 2x_1 + x_2 &= 1, \\ 6x_1 + 3x_2 &= 3, \end{aligned}$$

and we have already seen that this system is solved by any $(x_1, x_2) \in \mathbb{R}^2$ with $2x_1 + x_2 = 1$. Putting this into our notation, we have that $A = \begin{pmatrix} 2 & 1 \\ 6 & 3 \end{pmatrix}$ and $b = \begin{pmatrix} 1 \\ 3 \end{pmatrix}$, and with this

$$L(A, b) = L\left(\begin{pmatrix} 2 & 1 \\ 6 & 3 \end{pmatrix}, \begin{pmatrix} 1 \\ 3 \end{pmatrix}\right) = \{(x_1, x_2) : 2x_1 + x_2 = 1\}.$$

Recall that changing the RHS to $b = \begin{pmatrix} 1 \\ 2 \end{pmatrix}$ leads to a system without solution, i.e.,

$$L(A, b) = L\left(\begin{pmatrix} 2 & 1 \\ 6 & 3 \end{pmatrix}, \begin{pmatrix} 1 \\ 2 \end{pmatrix}\right) = \emptyset.$$

This example is somehow special because the existence of a solution depends on the RHS. Although this is not a rare case, we will see that there are conditions for a linear system to be *uniquely solvable* for any RHS b . This is of particular interest in applications, where the RHS b usually represents some kind of measurements or requirements, which we may not choose ourselves, and we want to find a solution x to specify some *parameters*.

Before we discuss a systematic way to solve large linear systems, let us see some more examples.

Example 2.19. We want to solve the following problem:

Assume you've ordered 3 pizzas and 5 drinks, but you forgot the individual prices. You only know that you've paid 42 EURO, and that a pizza was 6 EURO more expensive than a drink. To solve this “problem” let x_1 and x_2 be the price of a pizza and a drink, respectively. From our assumption on the overall cost we know that $3x_1 + 5x_2 = 42$, and the second assumption reads $x_1 = x_2 + 6$. This can be written as the linear system

$$\begin{aligned} 3x_1 + 5x_2 &= 42, \\ x_1 - x_2 &= 6. \end{aligned}$$

By substituting $x_2 = x_1 - 6$ in the first equation, we see that a solution must satisfy $8x_1 - 30 = 42$, and we obtain that the price of a pizza is $x_1 = 9$. From $x_2 = x_1 - 6$ we then see that $x_2 = 3$. Therefore,

$$L\left(\begin{pmatrix} 3 & 5 \\ 1 & -1 \end{pmatrix}, \begin{pmatrix} 42 \\ 6 \end{pmatrix}\right) = \left\{\begin{pmatrix} 9 \\ 3 \end{pmatrix}\right\}.$$

Note that $L(A, b)$ is a *set* and therefore, we need to write $L(A, b) = \{x\}$, and not $L(A, b) = x$, if x is the only solution.

Example 2.20. Let us consider the linear system

$$\begin{array}{rcl} x_1 + 2x_2 + 6x_3 & = & 0, \\ 2x_1 + 5x_2 & = & -1, \\ x_2 & = & 3. \end{array}$$

By the third equation we already know that any solution must satisfy $x_2 = 3$. If we plug this into the second equation we see that $x_1 = -8$. Finally, from the first equation, we see that the only possible choice then for x_3 is $\frac{1}{3}$. Therefore, $L(A, b) = \{(-8, 3, \frac{1}{3})\}$, where A and b are the corresponding matrix of coefficients and RHS, respectively.

Remark 2.21. In the above example, we wrote $L(A, b) = \{(x_1, \dots, x_n)\}$. Note that this is a bit inaccurate, because (x_1, \dots, x_n) seems to be a row vector, but we have the convention that elements from \mathbb{R}^n , so in particular the solution $x \in \mathbb{R}^n$, are column vectors. More precisely, we should have written $L(A, b) = \{(x_1, \dots, x_n)^T\}$. However, as it is obvious that the vector is supposed to be a column vector, we usually omit the transpose, for simplicity. (Note that the matrix product Ax would not be defined if x is a row vector.) In the same way, we might define $x = (x_1, \dots, x_n) \in \mathbb{R}^n$ and assume, unless stated otherwise, that x is a column vector.

The next examples are given with solution for your own exercise.

Example 2.22. Consider the linear system

$$\begin{array}{rcl} x_1 & & + x_4 = 0, \\ -4x_2 & & + 16x_4 = 0, \\ 2x_3 & - 6x_4 & = 0. \end{array}$$

Then, the set of solutions is given by

$$L\left(\begin{pmatrix} 1 & 0 & 0 & 1 \\ 0 & -4 & 0 & 16 \\ 0 & 0 & 2 & -6 \end{pmatrix}, \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}\right) = \left\{(-\lambda, 4\lambda, 3\lambda, \lambda) : \lambda \in \mathbb{R}\right\}.$$

In particular, there are infinitely many solutions.

Example 2.23. Consider the linear system

$$\begin{array}{rcl} x_1 + 2x_2 + 6x_3 & = & 0, \\ 2x_1 + 5x_2 & = & -1, \\ x_2 & = & 3, \\ x_1 - x_3 & = & 2. \end{array}$$

Note that we already computed that, if we consider only the first three equations, then there is the unique solution $(x_1, x_2, x_3) = (-8, 3, \frac{1}{3})$, see Example 2.20. However, this solution does not agree with the fourth equation. So there is no solution to this linear system, i.e. $L = \emptyset$.

We now discuss one special case of equations, which will also lead to a first answer to the question if a linear system can have infinitely many solutions.

Definition 2.24. Let $A \in \mathbb{R}^{m \times n}$ and $x \in \mathbb{R}^n$. A linear system of the form

$$Ax = 0,$$

i.e., the RHS is the zero vector $0 := 0_{m1}$, is called a **homogeneous** system (of equations).

To a given linear system $Ax = b$, we call $Ax = 0$ the corresponding homogeneous system.

It is rather easy to see that for any matrix A we have $A \cdot 0 = 0$ (again 0 is a vector here). Just have a look at the definition of the matrix-vector product. This implies that every homogeneous linear system has at least the solution $x = 0$, and this solution is called the **trivial solution**. Written mathematically, we have that $L(A, 0) \supset \{0\}$ for any matrix A .

In some cases, a homogeneous system has more than the trivial solution, see Example 2.22. However, if the trivial solution is also the only solution of a homogeneous system, then the following lemma shows that the solution to a linear system $Ax = b$, if it exists, is unique.

Lemma 2.25. *Let $A \in \mathbb{R}^{m \times n}$, $x, y \in \mathbb{R}^n$ and $b \in \mathbb{R}^m$, such that $Ax = b$ and $Ay = 0$.*

Then, $x + y$ also solves the linear system, i.e., $A(x + y) = b$.

In particular, if there is a solution $y \neq 0$ to $Ay = 0$, and at least one solution x to $Ax = b$, then there are infinitely many solutions.

Moreover, if $Ay = 0$ has only the trivial solution $y = 0$, and there is at least one solution x to $Ax = b$, then this solution x is unique.

Proof. The first statement follows directly from the linearity of matrix multiplication. We obtain

$$A(x + y) = Ax + Ay = b + 0 = b.$$

For the second statement, note that if there is a solution $y \neq 0$ satisfies $Ay = 0$, then also the vector $\lambda \cdot y$ satisfies $A(\lambda y) = \lambda Ay = 0$, and is therefore also a solution. Hence, there are automatically infinitely many solutions to the homogeneous system. Together with the first part, we see that $Ax = b$ has infinitely many solutions.

For the third part, assume that the only solution to $Ay = 0$ is $y = 0$. If there would be two solutions x and z to the linear system, i.e., $Ax = Az = b$, then their difference would satisfy

$$A(x - z) = Ax - Az = b - b = 0.$$

In other words, $x - z$ is a solution to the homogeneous system. As we assumed that the only solution to the homogeneous system is zero, we obtain that $x - z = 0$ or $x = z$. This shows that, if there are two solutions to $Ax = b$, then they are equal, which shows the uniqueness of the solution. □

Remark 2.26. Analogously one can define linear systems also in the complex case, i.e., coefficients and RHS might be complex. Then, we are clearly interested in complex solutions. However, note that every complex equation can be written as two “real” equations by considering the real and complex parts separately. Therefore, every complex linear system can be written as a larger real linear system. We therefore only consider the real case.

2.3 Gaussian elimination

Now we make an important observation which allows us to derive an algorithm for solving linear systems by manipulating matrices. We can do the following operations to a linear system without changing the set of solutions:

- 1) Interchanging any two equations, i.e., changing the order of the equations.
- 2) Multiplying an equation with a scalar $0 \neq \lambda \in \mathbb{R}$.
- 3) Adding a multiple of an equation to another equation.

(Think for a second, why these operations do not change the set of solutions.)

Since every system of linear equations can be written with the help of a matrix, it is clearly of interest how the above operations change the corresponding matrix of coefficients of a linear system. We will see that they indeed allow for successive modifications that lead to “much simpler” matrices, i.e., matrices in *echelon form*. From such a matrix, we will be able to basically *see* if a corresponding linear system is (uniquely) solvable or not.

Let us start by discussing how the above operations to a linear system $Ax = b$ affect the corresponding matrix A . However, note already now that these operations also change the RHS b of a linear system, and this is essential. We will come back to this shortly, but for now we only consider the corresponding matrix of coefficients.

In view of the operations from above that can be used to change a linear system $Ax = b$ without changing the set of solutions, we see that the matrix A is changed in the following way:

- 1) interchanging two rows,
- 2) multiply a row with a scalar $0 \neq \lambda \in \mathbb{R}$, and
- 3) adding a multiple of a row to another row.

For obvious reasons, these operations are called **row operations**.

The goal is now to use these operations in a way that “creates” a lot of zero entries in the matrix. And we make this in a systematic way that “creates” the zero entries always in the “lower left corner” of a matrix until it is of *echelon form*. The resulting matrices will look like

$$\begin{pmatrix} 1 & 5 & -1 \\ 0 & 1 & 2 \\ 0 & 0 & 0 \end{pmatrix} \quad \text{or} \quad \begin{pmatrix} 2 & 3 & -1 & 0 \\ 0 & 0 & 2 & 2 \\ 0 & 0 & 0 & 1 \end{pmatrix}.$$

That is, there are as many as possible 0’s to the left of every row, and they are ordered such that number of zero entries is increasing from top to bottom. In particular, the **leading coefficient** of a row, i.e., the first non-zero entry of a row, is strictly to the right of the leading coefficient of the row above. Before we discuss some examples, let us state the general definition.

Definition 2.27. A matrix $C = (c_{ij}) \in \mathbb{R}^{m \times n}$ of the form

$$C = \begin{pmatrix} 0 & \cdots & 0 & c_{1j_1} & * & \cdots & \cdots & \cdots & * \\ 0 & \cdots & \cdots & 0 & c_{2j_2} & * & \cdots & \cdots & * \\ 0 & \cdots & \cdots & \cdots & 0 & c_{3j_3} & * & \cdots & * \\ \vdots & & & & & & & \ddots & \vdots \\ 0 & \cdots & \cdots & \cdots & \cdots & 0 & c_{kj_k} & * & \cdots & * \\ 0 & \cdots & 0 \\ \vdots & & & & & & & & \vdots \\ 0 & \cdots & 0 \end{pmatrix},$$

where $*$ stands for an arbitrary entry, is in **row echelon form** (ger. 'Treppenform').

That is, there exist numbers $k \leq m$ and $1 \leq j_1 < \cdots < j_k \leq n$ such that for all $1 \leq i \leq k$:

- $c_{ij_i} \neq 0$,
- $c_{ij} = 0$ for all $j < j_i$, i.e. c_{ij_i} is the first non-zero element in the i -th row, and
- $c_{\ell j_i} = 0$ for all $\ell > i$, i.e. c_{ij_i} is the last non-zero element in the j_i -th column.

The number k is called **rank** of the matrix, and we write $\text{rank}(C) := k$.

If, in addition,

- $c_{ij_i} = 1$ for all $i \leq k$, and
- $c_{\ell j_i} = 0$ for all $\ell < i$, i.e. $c_{ij_i} = 1$ is the only non-zero element in the j_i -th column,

then the matrix is in **reduced row echelon form** (ger. 'Treppennormalform').

We do not prove the following statement here formally, but note that it is the basis of the considerations below.

Theorem 2.28. Every matrix can be transformed to (reduced) row echelon form by performing row operations. Moreover, the reduced row echelon form of a matrix is unique.

In contrast, a given matrix A can be transformed by row operations into different matrices in (non-reduced) row echelon form. (For example, multiplying any row by 2 leads to another row echelon form.) But even then, all row echelon forms of a matrix have the same "k", i.e., the same rank. That's why the rank is a characteristic of a matrix, which turns out to be essential.

Definition 2.29. Let $A \in \mathbb{R}^{m \times n}$ be arbitrary. We define the **rank of A** , denoted by $\text{rank}(A)$, as the rank of a corresponding row echelon form C of A .

Note that the definition implies that $\text{rank}(A) \leq \min\{m, n\}$ for any $A \in \mathbb{R}^{m \times n}$, and one may say that the rank is the **number of independent rows** in the matrix. (We will discuss later what this precisely means.)

Let us see some examples, before we show that computing the reduced row echelon form of a matrix is actually solving the corresponding system of linear equations.

Example 2.30 (Diagonal matrices). Diagonal matrices with all diagonal entries non-zero are already in echelon form. By dividing each row by the diagonal entry, we obtain the reduced row echelon form, which equals the identity. In particular, all diagonal matrices with non-zero diagonal entries have the same reduced row echelon form.

Example 2.31 (Rearranging rows). In general, we may easily bring every matrix into echelon form that only contains at most one non-zero entry per row. We just need to rearrange the rows. For example, a row echelon form of

$$\begin{pmatrix} 0 & 0 & 0 & -1 \\ 4 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 3 & 0 & 0 \end{pmatrix} \quad \text{is} \quad \begin{pmatrix} 4 & 0 & 0 & 0 \\ 0 & 3 & 0 & 0 \\ 0 & 0 & 0 & -1 \\ 0 & 0 & 0 & 0 \end{pmatrix}.$$

(In the definition above, we have $(j_1, j_2, j_3) = (1, 2, 4)$, and $c_{11} = 4$, $c_{22} = 3$ and $c_{34} = -1$.)

The rank is 3, and the reduced row echelon form is $\begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$.

Example 2.32 (Removing duplicate rows). Addition of (a multiple of) a row to another is a row operation. In particular, we can subtract a row from another. For example, by subtracting the first row from the second, and twice the first row from the third, we see that the reduced row echelon form of the matrix

$$\begin{pmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 2 & 2 & 2 \end{pmatrix} \quad \text{is} \quad \begin{pmatrix} 1 & 1 & 1 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}.$$

The rank of this matrix is 1. (We have $j_1 = 1$ and $c_{11} = 1$.)

The above example shows that **a rank-one matrix has indeed rank one**.

Given $x \in \mathbb{R}^n$ and $y \in \mathbb{R}^m$, we consider the rank-one matrix $xy^T \in \mathbb{R}^{n \times m}$. By definition of the matrix product, we see that every column of xy^T is a multiple of x , and every row is a multiple of y^T . In particular, the k -th row of xy^T is $x_k \cdot y^T$. Since all rows are multiples of each other, we can proceed as in Example 2.32 to obtain $\text{rank}(xy^T) = 1$.

Example 2.33. For example, let $x = \begin{pmatrix} 0 \\ 1 \\ 2 \end{pmatrix} \in \mathbb{R}^3 = \mathbb{R}^{3 \times 1}$ and $y^T = (2, 3, 4, 5) \in \mathbb{R}^{1 \times 4}$. We obtain

$$\begin{pmatrix} 0 \\ 1 \\ 2 \end{pmatrix} \begin{pmatrix} 2 & 3 & 4 & 5 \end{pmatrix} = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 2 & 3 & 4 & 5 \\ 4 & 6 & 8 & 10 \end{pmatrix} \in \mathbb{R}^{3 \times 4}.$$

Subtracting twice the second row from the third, and interchanging rows, leads to the row echelon form $\begin{pmatrix} 2 & 3 & 4 & 5 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}$. This shows $\text{rank}(xy^T) = 1$.

Let us consider some more examples.

Example 2.34. Let us consider the matrix

$$A = \begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{pmatrix}.$$

We bring this matrix into reduced row echelon form by using only row operations. Systematically, we want to create zeros in the lower left entries. So, we start by subtracting the first row 4-times from the second, which leads to a zero in the first entry of the second row. We indicate this operation by “ $II - 4I$ ”. Afterward, we subtract the first row 7-times from the third (“ $III - 7I$ ”), which leads to a zero in the first entry of the last row, and so on.

$$\begin{array}{c} \left(\begin{array}{ccc} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{array} \right) \\ \xrightarrow{II - 4I} \left(\begin{array}{ccc} 1 & 2 & 3 \\ 0 & -3 & -6 \\ 7 & 8 & 9 \end{array} \right) \\ \xrightarrow{III - 7I} \left(\begin{array}{ccc} 1 & 2 & 3 \\ 0 & -3 & -6 \\ 0 & -6 & -12 \end{array} \right) \\ \xrightarrow{III - 2II} \left(\begin{array}{ccc} 1 & 2 & 3 \\ 0 & -3 & -6 \\ 0 & 0 & 0 \end{array} \right) \\ \xrightarrow{-\frac{1}{3}II} \left(\begin{array}{ccc} 1 & 2 & 3 \\ 0 & 1 & 2 \\ 0 & 0 & 0 \end{array} \right) \\ \xrightarrow{I - 2II} \left(\begin{array}{ccc} 1 & 0 & -1 \\ 0 & 1 & 2 \\ 0 & 0 & 0 \end{array} \right) \end{array}$$

This is the reduced row echelon form of the matrix. Moreover, we can see that the rank is 2.

Note that there are several ways of indicating which row operation is performed. For example, one might be more precise and write “ $II \rightarrow II - 4I$ ” to state that the operation is performed only in the second (II) row. We decided to use this notation with the convention that only the row that appears first will be changed.

Example 2.35. We use “ $I \leftrightarrow II$ ” to indicate that we interchanged the first and the second row. (Note that afterwards, the first row is denoted by II and vice versa.)

$$\begin{array}{c}
 \left(\begin{array}{ccc} 0 & 8 & 0 \\ 3 & 6 & 0 \\ 6 & 0 & 1 \\ 6 & 15 & 0 \end{array} \right) \\
 \xrightarrow{I \leftrightarrow II} \left(\begin{array}{ccc} 3 & 6 & 0 \\ 0 & 8 & 0 \\ 6 & 0 & 1 \\ 6 & 15 & 0 \end{array} \right) \\
 \xrightarrow{III - 2I} \left(\begin{array}{ccc} 3 & 6 & 0 \\ 0 & 8 & 0 \\ 0 & -12 & 1 \\ 6 & 15 & 0 \end{array} \right) \\
 \xrightarrow{IV - 2I} \left(\begin{array}{ccc} 3 & 6 & 0 \\ 0 & 8 & 0 \\ 0 & -12 & 1 \\ 0 & 3 & 0 \end{array} \right) \\
 \xrightarrow{IV - \frac{3}{8}II} \left(\begin{array}{ccc} 3 & 6 & 0 \\ 0 & 8 & 0 \\ 0 & -12 & 1 \\ 0 & 0 & 0 \end{array} \right) \\
 \xrightarrow{III + \frac{3}{2}II} \left(\begin{array}{ccc} 3 & 6 & 0 \\ 0 & 8 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{array} \right) \\
 \xrightarrow{\frac{1}{3}I} \left(\begin{array}{ccc} 1 & 2 & 0 \\ 0 & 8 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{array} \right) \\
 \xrightarrow{\frac{1}{8}II} \left(\begin{array}{ccc} 1 & 2 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{array} \right) \\
 \xrightarrow{I - 2II} \left(\begin{array}{ccc} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{array} \right)
 \end{array}$$

This matrix has therefore rank 3. (We have $(j_1, j_2, j_3) = (1, 2, 3)$ and $c_{11} = c_{22} = c_{33} = 1$.)

As you see it can be rather time consuming to compute the reduced row echelon form even for rather small matrices. However, it is a very **straight-forward method**. Meaning that it is always obvious what to do next, and miscalculation is basically the only source of errors.

Recall that the reduced echelon form of a matrix is unique, but the order of the calculations to find it is not. To make computations generally easier and faster, there are two *rules of thumb*:

- Try to work with rows with fewer non-zero entries.
- If a row contains only one non-zero element, then we can immediately set all other elements in the corresponding column to zero.

The second “shortcut” can clearly be performed by adding multiples of the row with the unique non-zero entry to all other rows. Let’s consider an example.

Example 2.36. We consider the matrix

$$\begin{pmatrix} 2 & 6 & 5 & 0 & 4 \\ 4 & 17 & 8 & 0 & 16 \\ 12 & 42 & 8 & 14 & 0 \\ 0 & 0 & 13 & 0 & 0 \\ 28 & 0 & 3 & 0 & 0 \end{pmatrix}.$$

We see that in the fourth row there is only one non-zero entry, which is in the third column. Hence we can reduce the third column immediately and get

$$\begin{pmatrix} 2 & 6 & 0 & 0 & 4 \\ 4 & 17 & 0 & 0 & 16 \\ 12 & 42 & 0 & 14 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 28 & 0 & 0 & 0 & 0 \end{pmatrix}.$$

Note that we also divided the 4th row by 13 to make things easier.

Next we see that the fifth row is already in the form we wish for the first row. So, we interchange them. Again, we can put all entries in the first column to zero:

$$\begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 17 & 0 & 0 & 16 \\ 0 & 42 & 0 & 14 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 6 & 0 & 0 & 4 \end{pmatrix}.$$

If we now subtract 4-times the last row from the second, we see that the new second row reads $(0, -7, 0, 0, 0)$. Dividing by -7, and putting all other entries in the second column to zero, yields

$$\begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 14 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 4 \end{pmatrix}.$$

Dividing the third row by 14, and interchange III and IV , finally yields the reduced row echelon form

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

We see that the rank is 5, and $j_\ell = \ell$ for $\ell = 1, \dots, 5$, i.e., $(j_1, j_2, j_3, j_4, j_5) = (1, 2, 3, 4, 5)$. Moreover, $c_{\ell\ell} = 1$ for $\ell = 1, \dots, 5$.

Now we discuss how we can **solve linear systems** by calculating (reduced) row echelon forms of matrices. We consider the linear system $Ax = b$ with corresponding matrix $A = (a_{ij}) \in \mathbb{R}^{m \times n}$ and RHS $b = (b_1, \dots, b_m)$, and define the **augmented matrix** $(A|b)$, i.e., we consider the array

$$(A|b) := \left(\begin{array}{cccc|c} a_{11} & a_{12} & \dots & a_{1n} & b_1 \\ a_{21} & a_{22} & \dots & a_{2n} & b_2 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ a_{m1} & a_{m2} & \dots & a_{mn} & b_m \end{array} \right).$$

This means that we add b as new column. We already discussed that by interchanging, multiplying and adding rows we do not change the set of solutions. So, if some $(C|b')$ is obtained from $(A|b)$ only by row operations, then

$$L(A, b) = L(C, b'),$$

where $L(A, b)$ denotes the set of solutions of $Ax = b$, see Definition 2.17.

Now assume that the augmented matrix $(A|b)$ is transformed into an augmented matrix $(C|b')$, where C is in row echelon form. (Here, we consider the vector b as the last column of the matrix, and therefore have to respect it while performing row operations. But we only want to bring A in row echelon form.) From this augmented matrix we can just “see” the solutions of the corresponding linear system. This way of computing solutions is called **Gaussian elimination**.

Before we discuss the general procedure, let us see a minimal example.

Example 2.37. Consider the linear system $Ax = b$ with $A = \begin{pmatrix} 3 & 5 \\ 1 & -1 \end{pmatrix}$ and $b = \begin{pmatrix} 42 \\ 6 \end{pmatrix}$. We have already seen in Example 2.19 that $x = \begin{pmatrix} 9 \\ 3 \end{pmatrix}$ is the unique solution, i.e., $L\left(\begin{pmatrix} 3 & 5 \\ 1 & -1 \end{pmatrix}, \begin{pmatrix} 42 \\ 6 \end{pmatrix}\right) = \left\{\begin{pmatrix} 9 \\ 3 \end{pmatrix}\right\}$. To compute this solution by Gaussian elimination, we bring the augmented matrix $(A|b) = \left(\begin{array}{cc|c} 3 & 5 & 42 \\ 1 & -1 & 6 \end{array} \right)$ into the row echelon form $\left(\begin{array}{cc|c} 1 & -1 & 6 \\ 0 & 8 & 24 \end{array} \right)$ by subtracting 3-times the second row from the first, and then interchanging rows. Dividing the second row by 8 and adding it to the first, we obtain the reduced row echelon form $(C|b') = \left(\begin{array}{cc|c} 1 & 0 & 9 \\ 0 & 1 & 3 \end{array} \right)$, from which we “see” that $x_1 = 9$ and $x_2 = 3$.

Example 2.38. Let us again consider Example 2.20, which is given by $Ax = b$ with $A = \begin{pmatrix} 1 & 2 & 6 \\ 2 & 5 & 0 \\ 0 & 1 & 0 \end{pmatrix}$ and $b = \begin{pmatrix} 0 \\ -1 \\ 3 \end{pmatrix}$. We bring the augmented matrix in reduced row echelon form:

$$\begin{array}{ccc} \left(\begin{array}{ccc|c} 1 & 2 & 6 & 0 \\ 2 & 5 & 0 & -1 \\ 0 & 1 & 0 & 3 \end{array} \right) & \xrightarrow{II - 2I} & \left(\begin{array}{ccc|c} 1 & 2 & 6 & 0 \\ 0 & 1 & -12 & -1 \\ 0 & 1 & 0 & 3 \end{array} \right) & \xrightarrow{\dots} & \left(\begin{array}{ccc|c} 1 & 0 & 6 & -6 \\ 0 & 1 & 0 & 3 \\ 0 & 0 & -12 & -4 \end{array} \right) \\ & \xrightarrow{\dots} & \left(\begin{array}{ccc|c} 1 & 0 & 0 & -8 \\ 0 & 1 & 0 & 3 \\ 0 & 0 & 1 & 1/3 \end{array} \right) & & \end{array}$$

and we see that the unique solution is $x = (-8, 3, 1/3)$.

For the general procedure, assume that A has rank k , i.e., $\text{rank}(A) = k$. We obtain that the augmented matrix $(C|b')$ in row echelon form, which we obtain from $(A|b)$, looks like

$$(C|b') = \left(\begin{array}{cccccc|c} 0 & \dots & 0 & c_{1j_1} & * & \dots & * & b'_1 \\ 0 & \dots & \dots & 0 & c_{2j_2} & * & \dots & b'_2 \\ 0 & \dots & \dots & \dots & 0 & c_{3j_3} & * & b'_3 \\ \vdots & & & & & & \vdots & \vdots \\ 0 & \dots & \dots & \dots & 0 & c_{kj_k} & * & b'_k \\ 0 & \dots & \dots & \dots & \dots & \dots & 0 & b'_{k+1} \\ \vdots & & & & & & \vdots & \vdots \\ 0 & \dots & \dots & \dots & \dots & \dots & 0 & b'_m \end{array} \right),$$

where $*$ stands for an arbitrary entry.

(Note that there are $(n - j_k)$ many $*$'s to the right of c_{kj_k} , and this might be 0 if $j_k = n$.)

Let us first have a closer look at the numbers b'_{k+1}, \dots, b'_m :

If one of them is not equal to zero, then the linear system cannot have a solution. To see this note that the ℓ -th row, with $\ell > k$, corresponds to the equation

$$0 \cdot x_1 + 0 \cdot x_2 + \dots + 0 \cdot x_n = b'_\ell.$$

Since the left hand side is equal to zero for every x , we obtain a contradiction if one of the b'_ℓ 's is not equal to zero. We therefore obtain **the rule**:

If $b'_\ell \neq 0$ for some $\ell = k + 1, \dots, m$, then $L(A, b) = L(C, b') = \emptyset$.

In the other case, i.e., $b'_{k+1} = \dots, b'_m = 0$, the system will always **have a solution**, which we compute **“from the bottom to the top”**:

We first consider the last non-zero equation, i.e., the k -th equation, which reads

$$\sum_{\ell=j_k}^n c_{k\ell} x_\ell = b'_k,$$

and we know that the leading coefficient $c_{kj_k} \neq 0$.

In the case $j_n = n$, we see that this equation reads $c_{kn} x_n = b'_k$, which yields

$$x_n = x_{j_k} = \frac{b'_k}{c_{kj_k}},$$

and we know the value of x_n of any solution $x = (x_1, \dots, x_n)$ of $Ax = b$.

If $j_n < n$, then, by rearranging, the equality is equivalent to

$$x_{j_k} = \frac{1}{c_{kj_k}} \left(b'_k - \sum_{\ell=j_k+1}^n c_{k\ell} x_\ell \right).$$

All possible solutions have to fulfill this identity, otherwise the k -th equation could not be true. Therefore, for any given choice of x_{j_k+1}, \dots, x_n , we have to choose the unique x_{j_k} . However, these x_ℓ with $\ell = j_k + 1, \dots, n$ **can be chosen freely**.

Now assume that we have already fixed the values for x_n, \dots, x_{j_k} . We turn to the next equation, which is the $(k - 1)$ -th:

By the same principle we can give a formula for $x_{j_{k-1}}$ depending only on the x_ℓ with $\ell \geq j_{k-1} + 1$. Since we have only fixed x_n, \dots, x_{j_k} , we can again choose x_ℓ with $\ell = j_{k-1} + 1, \dots, j_k - 1$ freely. (Note that there is no free choice if $j_k = j_{k-1} + 1$.) So, after this step, we have computed the components $x_n, \dots, x_{j_k}, \dots, x_{j_{k-1}}$ of a solution (x_1, \dots, x_n) of $Ax = b$, i.e., the last components.

If we continue this process, we finally see, that there are precisely $k = \text{rank}(A)$ such equalities that “fix” the value of the *unknowns* x_1, \dots, x_{j_k} . These equalities are

$$x_{j_i} = \frac{1}{c_{ij_i}} \left(b'_i - \sum_{\ell=j_i+1}^n c_{i\ell} x_\ell \right)$$

for $i = 1, \dots, k$. But the remaining variables, i.e., the x_ℓ with $\ell \in \{1, 2, \dots, n\} \setminus \{j_1, j_2, \dots, j_k\}$, can all be chosen freely. Hence, when we write down the solution, there are $n - k$ free parameters. This is usually phrased as “a linear system has $n - k$ **degrees of freedom**”.

Let us discuss some examples.

Example 2.39. Consider the linear system $Ax = b$ with the matrix

$$A = \begin{pmatrix} 2 & 1 & -2 \\ 0 & 1 & 2 \\ 0 & 0 & 0 \end{pmatrix}$$

and $b = (5, 1, 0)$. To compute a solution $x = (x_1, x_2, x_3)$, we consider the augmented matrix

$$(A|b) = \left(\begin{array}{ccc|c} 2 & 1 & -2 & 5 \\ 0 & 1 & 2 & 1 \\ 0 & 0 & 0 & 0 \end{array} \right).$$

Since this matrix is already in row echelon form, we do not need to perform row operations. Specifically, we have the row echelon form with $(j_1, j_2) = (1, 2)$. By Gaussian elimination (with $C = A$ and $b' = b$), we can choose the variables with index in $\{1, 2, 3\} \setminus \{j_1, j_2\} = \{3\}$ freely, i.e., we already know that x_3 is a free parameter. Using the formulas established above, we obtain (“from the bottom to the top”)

$$x_2 = \frac{1}{a_{22}} \left(b_2 - \sum_{\ell=3}^3 c_{2\ell} x_\ell \right) = \frac{1}{1} (1 - 2x_3) = 1 - 2x_3$$

and, using this,

$$x_1 = \frac{1}{a_{11}} \left(b_1 - \sum_{\ell=2}^3 c_{1\ell} x_\ell \right) = \frac{1}{2} (5 - 1x_2 + 2x_3) = \frac{1}{2} (5 - (1 - 2x_3) + 2x_3) = 2 + 2x_3.$$

We therefore obtain that the set of solutions is

$$L\left(\left(\begin{smallmatrix} 2 & 1 & -2 \\ 0 & 1 & 2 \\ 0 & 0 & 0 \end{smallmatrix}\right), \left(\begin{smallmatrix} 5 \\ 1 \\ 0 \end{smallmatrix}\right)\right) = \left\{ (x_1, x_2, x_3) \in \mathbb{R}^3 : x_2 = 1 - 2x_3 \text{ and } x_1 = 2 + 2x_3 \right\}.$$

For notational convenience, we choose λ as a name for the free parameter, and write

$$L\left(\left(\begin{smallmatrix} 2 & 1 & -2 \\ 0 & 1 & 2 \\ 0 & 0 & 0 \end{smallmatrix}\right), \left(\begin{smallmatrix} 5 \\ 1 \\ 0 \end{smallmatrix}\right)\right) = \left\{ (2 + 2\lambda, 1 - 2\lambda, \lambda) \in \mathbb{R}^3 : \lambda \in \mathbb{R} \right\}.$$

(Convince yourself that these are the same sets.)

Note that we have infinitely many solutions. For example, $x = (2, 1, 0)$ (for $\lambda = 0$) or $x = (0, 3, -1)$ (for $\lambda = -1$).

However, if we would consider, e.g., the RHS $b' = (0, 0, 1)$, then there would be no solution since the reduced echelon form

$$(A|b') = \left(\begin{array}{ccc|c} 2 & 1 & -2 & 0 \\ 0 & 1 & 2 & 0 \\ 0 & 0 & 0 & 1 \end{array} \right)$$

shows the contradiction in the last equation.

Example 2.40. Note that some formulas get easier when we transfer a linear system into reduced row echelon form. If we consider the example from above, we see that it can be transferred to

$$(A|b) = \left(\begin{array}{ccc|c} 2 & 1 & -2 & 5 \\ 0 & 1 & 2 & 1 \\ 0 & 0 & 0 & 0 \end{array} \right) \rightarrow \left(\begin{array}{ccc|c} 1 & 0 & -2 & 2 \\ 0 & 1 & 2 & 1 \\ 0 & 0 & 0 & 0 \end{array} \right) =: (C|b').$$

Again, the second row is seen to be equivalent to $x_2 = 1 - 2x_3$. However, the first row shows directly that $x_1 = 2 + 2x_3$, and there is no need to plug the already obtained formula for x_2 in, as above. **That's another advantage of the reduced row echelon form.** We obtain again that $L(A, b) = \{(2 + 2\lambda, 1 - 2\lambda, \lambda) \in \mathbb{R}^3 : \lambda \in \mathbb{R}\}$.

Example 2.41. Consider the linear system $Ax = b$ with the matrix

$$A = \begin{pmatrix} 2 & 1 & -2 \\ 0 & 1 & 2 \\ -2 & -2 & 0 \end{pmatrix},$$

with an arbitrary RHS $b = (b_1, b_2, b_3) \in \mathbb{R}^3$. We want to see, depending on b , if there is a solution $x = (x_1, x_2, x_3)$ and compute it. For this, we consider the augmented matrix and its row echelon form

$$(A|b) = \left(\begin{array}{ccc|c} 2 & 1 & -2 & b_1 \\ 0 & 1 & 2 & b_2 \\ -2 & -2 & 0 & b_3 \end{array} \right) \quad \text{and} \quad (C|b') = \left(\begin{array}{ccc|c} 2 & 1 & -2 & b_1 \\ 0 & 1 & 2 & b_2 \\ 0 & 0 & 0 & b_1 + b_2 + b_3 \end{array} \right).$$

Note that we just added the first and the second row to the last. We see that this system has a solution if and only if $b_1 + b_2 + b_3 = 0$. And in this case, repeating the calculations from the last example, we see that x_3 is a free parameter, while $x_2 = b_2 - 2x_3$ and

$$x_1 = \frac{1}{2}(b_1 - x_2 + 2x_3) = \frac{1}{2}(b_1 - b_2 + 4x_3).$$

The set of solutions in the form

$$L\left(\left(\begin{array}{ccc} 2 & 1 & -2 \\ 0 & 1 & 2 \\ -2 & -2 & 0 \end{array} \right), b \right) = \left\{ \left(\frac{b_1 - b_2}{2} + 2\lambda, b_2 - 2\lambda, \lambda \right) \in \mathbb{R}^3 : \lambda \in \mathbb{R} \right\}.$$

Note that a sometimes even faster way of finding the set of solutions is to compute the reduced row echelon form of $(A|b)$ which is

$$\left(\begin{array}{ccc|c} 1 & 0 & -2 & \frac{b_1 - b_2}{2} \\ 0 & 1 & 2 & b_2 \\ 0 & 0 & 0 & b_1 + b_2 + b_3 \end{array} \right).$$

Just subtract the second row from the first in $(C|b')$ and divide the new first row by 2. From this one can see the set of solutions.

Example 2.42. A larger example, that we only present with a short solution is the linear system given by

$$(A|b) = \left(\begin{array}{cccc|c} 9 & -6 & -12 & 30 & -6 \\ 5 & -10 & -20 & 0 & 20 \\ -5 & 2 & 4 & 10 & -10 \\ -2 & -4 & -8 & 40 & -16 \end{array} \right).$$

This has the reduced row echelon form

$$\left(\begin{array}{cccc|c} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 2 & 0 & -2 \\ 0 & 0 & 0 & 1 & -\frac{3}{5} \\ 0 & 0 & 0 & 0 & 0 \end{array} \right).$$

We see that x_3 is a free parameter. From the third equation we get that $x_4 = -\frac{3}{5}$. We deduce from the second equation that $x_2 = -2 - 2x_3$, and the first equation yields that $x_1 = 0$. So we obtain the set of solutions

$$L(A, b) = \left\{ \begin{pmatrix} 0 \\ -2 - 2\lambda \\ \lambda \\ -\frac{3}{5} \end{pmatrix} : \lambda \in \mathbb{R} \right\}.$$

By Gaussian elimination we just found **all solutions** of the linear system $Ax = b$.

Note that a linear system is inconsistent, i.e., has **no solution, if and only if there is a zero row** in C (from the augmented matrix $(C|b')$) and the corresponding $b'_\ell \neq 0$. Since there are exactly m rows, and $k = \text{rank}(A)$ of them are non-zero, we see that a linear system is solvable (independent from the RHS b) if $\text{rank}(A) = m$. In this case, we say that A has **full (row) rank**.

Moreover, since there are exactly n unknowns, and the above procedure only fixes k of them. Hence, there are “ $n - k$ degrees of freedom”. We see that a solution of $Ax = b$, if it exists, is **unique if and only if $\text{rank}(A) = n$** .

We collect these findings in the following lemma, which are particularly meaningful if $m = n$, i.e., if A is a square matrix. This is the case if there are **as many equations as unknowns**.

Lemma 2.43. *Let $A \in \mathbb{R}^{m \times n}$.*

1. *If $\text{rank}(A) < m$, then the linear system $Ax = b$ has no solution for certain $b \in \mathbb{R}^n$.*
2. *If $\text{rank}(A) < n$, then the homogeneous system $Ax = 0$ has infinitely many solutions.*

Hence, if $\text{rank}(A) < \min\{m, n\}$, then the linear system $Ax = b$ has either no or infinitely many solutions, depending on $b \in \mathbb{R}^n$.

Moreover, if $A \in \mathbb{R}^{n \times n}$ is a square matrix, i.e., $m = n$, with $\text{rank}(A) = n$, then the linear system $Ax = b$ has a unique solution for any $b \in \mathbb{R}^n$.

Proof. The lemma follows from the considerations above, together with the fact that the homogeneous system $Ax = 0$ has always at least the solution $x = 0$, see also Lemma 2.25. \square

Let us see an example with a square matrix of full rank.

Example 2.44. We have a look at

$$(A|b) = \left(\begin{array}{ccc|c} 1 & 3 & 0 & 1 \\ 2 & 0 & 2 & 0 \\ 4 & 12 & 2 & 2 \end{array} \right).$$

This matrix is transformed by row operations to the row echelon form

$$\left(\begin{array}{ccc|c} 1 & 3 & 0 & 1 \\ 0 & -6 & 2 & -2 \\ 0 & 0 & 2 & -2 \end{array} \right),$$

which is already enough to see that $\text{rank}(A) = 3$. Therefore, we can compute a unique solution. One way is to transform the augmented matrix further to

$$\left(\begin{array}{ccc|c} 1 & 3 & 0 & 1 \\ 0 & -6 & 0 & 0 \\ 0 & 0 & 2 & -2 \end{array} \right) \quad \text{and then to} \quad \left(\begin{array}{ccc|c} 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & -1 \end{array} \right)$$

We therefore “see” the unique solution $x = (1, 0, -1)$.

It is worth noting that a row echelon form of a square matrix is always an **upper triangular matrix**. That is, the row echelon form is of the form

$$\begin{pmatrix} a_{11} & \dots & a_{1n} \\ & \ddots & \vdots \\ 0 & & a_{nn} \end{pmatrix},$$

i.e., all entries below the diagonal are zero. **The matrix has full rank if and only if all diagonal entries are non-zero.** On the contrary, every upper triangular matrix $A \in \mathbb{R}^{n \times n}$ with non-zero diagonal elements is already in row echelon form.

Matrices with full rank play an important role, in particular, since there is a unique solution to a corresponding linear system independent of the RHS b . However, the technique above only seems to work for a fixed RHS b .

In the following sections we will discuss the inverse of a matrix $A \in \mathbb{R}^{n \times n}$, if it exists, which will lead to a formula for a solution of $Ax = b$ for any b . However, let us add that it is not usual to compute the inverse of a large matrix on a computer, since it is computationally quite hard. Large linear systems are usually solved with (variants of) Gaussian elimination.

Remark 2.45. One straight-forward extension of Gaussian elimination is to solve a linear system with coefficient matrix $A \in \mathbb{R}^{m \times n}$ for different right hand sides, say $b, c \in \mathbb{R}^m$. In this case we consider the augmented matrix

$$(A|b|c)$$

and proceed as above. We then obtain the system $(A'|b'|c')$, which is in row echelon form. (Precisely, we again only want A' to be in row echelon form.)

Then, we can compute solutions x_b by considering $(A'|b')$ and solutions x_c by having a look at $(A'|c')$, since both are clearly in row echelon form. However, if we would now consider another RHS, then we would need to repeat all the computations.

2.4 The determinant

We now introduce the determinant of a matrix, which will be a frequently appearing quantity, and is a good tool to decide if a matrix is invertible or not. Moreover, it is needed to introduce *Cramer's rule* for solving linear systems, and to give an explicit formula for the inverse matrix.

Definition 2.46. Let $A \in \mathbb{R}^{n \times n}$ be a square matrix, and denote the rows of A by $a_1, \dots, a_n \in \mathbb{R}^{1 \times n}$. Moreover, let $\lambda, \mu \in \mathbb{R}$, and $w \in \mathbb{R}^{1 \times n}$.

We define the **determinant** of A by the unique mapping $\det: \mathbb{R}^{n \times n} \rightarrow \mathbb{R}$ such that:

1. For any $1 \leq i \leq n$, the determinant is linear in the i -th row of A , i.e.,

$$\det \begin{pmatrix} a_1 \\ \vdots \\ \lambda a_i + \mu w \\ \vdots \\ a_n \end{pmatrix} = \lambda \det(A) + \mu \det \begin{pmatrix} a_1 \\ \vdots \\ w \\ \vdots \\ a_n \end{pmatrix}$$

2. If there exist $i \neq j$ such that $a_i = a_j$, i.e., two equal rows, then $\det(A) = 0$.
3. The identity matrix $I_n \in \mathbb{R}^{n \times n}$ satisfies

$$\det(I_n) = 1.$$

Note that the determinant is only defined for square matrices.

The definition directly shows the **connection to linear systems**. However, let us show explicitly how the determinant changes under row operations.

Lemma 2.47. Let $A \in \mathbb{R}^{n \times n}$. Then,

1. If the matrix B is obtained by multiplying one row of A by a scalar $\lambda \in \mathbb{R}$, then $\det B = \lambda \det A$. In particular, $\det(\lambda A) = \lambda^n \det(A)$.
2. If the matrix B is obtained by interchanging two rows of A , then $\det B = -\det A$.
3. Adding a multiple of one row to another row does not change the determinant.

Let us show how this follows from the definition.

Proof. The first point follows immediately from Definition 2.46(1) with $\mu = 0$. Moreover, note that λA is the matrix with every row multiplied by λ . Since there are n rows, we have to apply this rule for each row one by one and obtain $\det(\lambda A) = \lambda^n \det(A)$.

For the second point we assume that B is the matrix which is obtained by interchanging the i -th and the j -th row of the matrix A . (The case of column uses again the transpose.) Recall from Definition 2.46(3) that, if a matrix contains a row more than once, then its determinant is

zero. This allows us to “add a zero” in the following way

$$\det A + \det B = \det \begin{pmatrix} a_1 \\ \vdots \\ a_i \\ \vdots \\ a_j \\ \vdots \\ a_n \end{pmatrix} + \det \begin{pmatrix} a_1 \\ \vdots \\ a_j \\ \vdots \\ a_i \\ \vdots \\ a_n \end{pmatrix} = \det \begin{pmatrix} a_1 \\ \vdots \\ a_i \\ \vdots \\ a_j \\ \vdots \\ a_n \end{pmatrix} + \det \begin{pmatrix} a_1 \\ \vdots \\ a_i \\ \vdots \\ a_i \\ \vdots \\ a_n \end{pmatrix} + \det \begin{pmatrix} a_1 \\ \vdots \\ a_j \\ \vdots \\ a_i \\ \vdots \\ a_n \end{pmatrix} + \det \begin{pmatrix} a_1 \\ \vdots \\ a_j \\ \vdots \\ a_j \\ \vdots \\ a_n \end{pmatrix},$$

where the second and the fourth summand on the RHS are just zero.

Now, note that the first and the second, as well as the third and the fourth, do only differ in one row, which allows us to use Definition 2.46(2) to obtain

$$\det A + \det B = \det \begin{pmatrix} a_1 \\ \vdots \\ a_i \\ \vdots \\ a_i + a_j \\ \vdots \\ a_n \end{pmatrix} + \det \begin{pmatrix} a_1 \\ \vdots \\ a_j \\ \vdots \\ a_i + a_j \\ \vdots \\ a_n \end{pmatrix} = \det \begin{pmatrix} a_1 \\ \vdots \\ a_i + a_j \\ \vdots \\ a_i + a_j \\ \vdots \\ a_n \end{pmatrix},$$

where we used again that for the last equality that the corresponding matrices differ only in one row. However, the determinant on the right is just zero because it has two equal rows. (The i -th and the j -th row equals $a_i + a_j$.) We therefore obtain $\det A + \det B = 0$ which gives $\det B = -\det A$, and proves the second point of the Corollary.

For the third point, let B be the matrix, where we added λ -times the j -th row to the i -th row of A . We obtain

$$\det B = \det \begin{pmatrix} a_1 \\ \vdots \\ a_i + \lambda a_j \\ \vdots \\ a_j \\ \vdots \\ a_n \end{pmatrix} = \det \begin{pmatrix} a_1 \\ \vdots \\ a_i \\ \vdots \\ a_j \\ \vdots \\ a_n \end{pmatrix} + \lambda \det \begin{pmatrix} a_1 \\ \vdots \\ a_j \\ \vdots \\ a_j \\ \vdots \\ a_n \end{pmatrix} = \det A,$$

where we use that the last determinant is zero, because of two equal rows. \square

Let us start with the computation of the **determinant of diagonal matrices**.

Example 2.48. We consider the diagonal matrix $D = \text{diag}(d_1, d_2, \dots, d_n)$ with $d_k \in \mathbb{R}$. By observing that the i -th row of D is exactly $d_i \cdot e_i^T$, we obtain

$$\det(D) = \det \begin{pmatrix} d_1 e_1 \\ d_2 e_2 \\ \vdots \\ d_n e_n \end{pmatrix} = d_1 \det \begin{pmatrix} e_1 \\ d_2 e_2 \\ \vdots \\ d_n e_n \end{pmatrix} = \dots = \prod_{i=1}^n d_i \det(I_n) = \prod_{i=1}^n d_i,$$

where we use $\det(I_n) = 1$. In particular, the determinant is zero iff $d_k = 0$ for some k .

It is apparent that the same formula already holds for triangular matrices.

Lemma 2.49. *Let $A \in \mathbb{R}^{n \times n}$ be an **upper triangular matrix**, which is a matrix of the form*

$$A = \begin{pmatrix} a_{11} & \dots & a_{1n} \\ \ddots & \ddots & \vdots \\ 0 & & a_{nn} \end{pmatrix},$$

i.e., all entries below the diagonal are zero. Then,

$$\det A = \prod_{i=1}^n a_{ii}.$$

This formula holds also for lower triangular matrices, which are zero above the diagonal.

Proof. Let us first assume that $\text{rank}(A) < n$. In this case we can produce a zero row in A by using only row operations, see the definition of the rank. Therefore, the determinant is zero in this case.

If $\text{rank}(A) = n$, then we know that A is already in row echelon form and that, in particular, $a_{ii} \neq 0$ for all $i = 1, \dots, n$. Therefore, by adding multiples of rows to other rows (“from bottom to top”), we can transform the matrix to a diagonal matrix without changing its diagonal elements and its determinant. We obtain $\det A = \det(\text{diag}(a_{11}, a_{22}, \dots, a_{nn}))$ which proves the result. \square

The determinant is of special interest, because it can be used to characterize whether a linear system is uniquely solvable, see Lemma 2.43, i.e., the matrix has full rank. We will see soon that is equivalent to the corresponding matrix being regular (aka. invertible).

Theorem 2.50. *Let $A \in \mathbb{R}^{n \times n}$. Then,*

$$\det A \neq 0 \iff \text{rank } A = n$$

Proof. Note that every matrix can be brought into row echelon form by just adding repeatedly rows to other rows, and this does not change the determinant. Since the row echelon form of a square matrix is always a upper triangular matrix, we see that the matrix has full rank, i.e., $\text{rank}(A) = n$, if and only if all diagonal entries of the row echelon form are not zero. By Lemma 2.49 this is equivalent to $\det(A) \neq 0$. \square

Let us see how to calculate determinants, starting with a small example.

Example 2.51. Consider the matrix $A = \begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix}$. We know that adding a multiple of one row to another does not change the determinant. Therefore,

$$\det \begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix} = \det \begin{pmatrix} 1 & 2 \\ 0 & -2 \end{pmatrix} = \det \begin{pmatrix} 1 & 0 \\ 0 & -2 \end{pmatrix} = -2$$

Example 2.52. As above, we see that

$$\det \begin{pmatrix} 2 & 1 & -2 \\ 0 & 1 & 2 \\ -2 & -2 & 0 \end{pmatrix} = \det \begin{pmatrix} 2 & 1 & -2 \\ 0 & 1 & 2 \\ 0 & -1 & -2 \end{pmatrix} = \det \begin{pmatrix} 2 & 1 & -2 \\ 0 & 1 & 2 \\ 0 & 0 & 0 \end{pmatrix} = 0$$

Repeating these computations for **arbitrary matrices**, we obtain rather easy formulas to mind.

Example 2.53 (Determinant of a 2×2 matrix). Consider the matrix

$$A = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix}.$$

Bringing this matrix into upper triangular form, we see that

$$\det(A) = a_{11}a_{22} - a_{12}a_{21}.$$

(Verify yourself!)

Using this formula again to calculate the determinant of the above example, we see

$$\det \begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix} = 1 \cdot 4 - 2 \cdot 3 = -2.$$

Example 2.54 (Determinant of a 3×3 matrix). For an arbitrary 3×3 matrix

$$A = \begin{pmatrix} a & b & c \\ d & e & f \\ g & h & i \end{pmatrix},$$

with $a, \dots, i \in \mathbb{R}$, we can derive a similar formula, which is given by

$$\det A = aei + bfg + cdh - ceg - bdi - afh.$$

This formula is usually called **Rule of Sarrus**, and can be easily minded by noting that one has to multiply only entries on certain diagonals, and add them according to the orientation of these diagonals. (Think for second which entries are multiplied, and how they are summed up.)

Again we want to give an example

$$\det \begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{pmatrix} = 45 + 84 + 96 - 105 - 72 - 48 = 0.$$

Let us present two more calculation rules, without proof. The first shows that transposing does not change the determinant.

Lemma 2.55. *For any square matrix $A \in \mathbb{R}^{n \times n}$ we have*

$$\det(A^T) = \det(A).$$

Note that this shows that Lemma 2.47 also holds if we replace “column” by “row”. That is, **the calculation rules for the determinant also hold for column operations.**

The next rule of calculation shows that the determinant of the product of two matrices is the product of the respective determinants. Recall that the determinant is only defined for square matrices. Therefore, both matrices need to have the same dimensions.

Lemma 2.56. *For any square matrices $A, B \in \mathbb{R}^{n \times n}$ we have*

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

The proofs of the last two lemmas are not hard, but quite long, and therefore we omit them here. Note that the last lemma is mostly of theoretical value as, in general, we do not know if a matrix is a product of *easier* matrices.

Example 2.57. Consider the matrix $A = \begin{pmatrix} 7 & 10 \\ 15 & 22 \end{pmatrix}$. To compute its determinant, we are lucky to know that $\begin{pmatrix} 7 & 10 \\ 15 & 22 \end{pmatrix} = \begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix} \begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix}$, i.e., $A = B^2 = B \cdot B$ with $B := \begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix}$. Since $\det(B) = -2$, see Example 2.53, we obtain

$$\det \begin{pmatrix} 7 & 10 \\ 15 & 22 \end{pmatrix} = \det(B)^2 = 4.$$

Example 2.58. Another important case of Lemma 2.56 is that one of the determinants vanishes, i.e., $\det(A) = 0$ or $\det(B) = 0$. In this case we already know that $\det(AB) = 0$, without actually computing the product AB . For example, since $\det \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} = 0$, we already know that

$$\det \left(\begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix} \right) = 0 \cdot \det \begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix} = 0.$$

Now that we know that the determinant “behaves well” with respect to transposition and multiplication, one might guess that a similar relation also holds for addition. However, and unfortunately, there is **no similar formula for the determinant of the sum of matrices** as the following simple example shows.

Example 2.59. We have a look at

$$A = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} \quad \text{and} \quad B = \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix},$$

such that $A + B = I_2$. We obtain (e.g., by the formula as given in Example 2.53) that

$$\det(A) + \det(B) = 0 + 0 = 0 \neq 1 = \det(A + B).$$

This shows that, in general, we can not extrapolate from the determinant of A and B to the determinant of their sum. (Clearly, there might be exceptions.)

Let us finally introduce the **Laplace expansion** for the determinant, which is also called **co-factor expansion** or **expansion along a row/column**. This formula allows to compute the determinant of large matrices recursively by computing the determinant of smaller matrices. Let us first introduce a bit new notation.

Definition 2.60. Let $A = (a_{ij}) \in \mathbb{R}^{n \times n}$ and $1 \leq i, j \leq n$.

Then, we define the (i, j) -**minor of A** by

$$M_{ij} = \det \begin{pmatrix} a_{1,1} & \dots & a_{1,j-1} & a_{1,j+1} & \dots & a_{1,n} \\ \vdots & & \vdots & \vdots & & \vdots \\ a_{i-1,1} & \dots & a_{i-1,j-1} & a_{i-1,j+1} & \dots & a_{i-1,n} \\ a_{i+1,1} & \dots & a_{i+1,j-1} & a_{i+1,j+1} & \dots & a_{i+1,n} \\ \vdots & & \vdots & \vdots & & \vdots \\ a_{n,1} & \dots & a_{n,j-1} & a_{n,j+1} & \dots & a_{n,n} \end{pmatrix},$$

i.e., M_{ij} is the determinant of the $(n - 1) \times (n - 1)$ -submatrix of A obtained by deleting the i -th row and the j -th column. We call $M = (M_{ij})_{i,j=1}^n$ the **matrix of minors** of A .

Moreover, we define the (i, j) -**cofactor** by $C_{ij} := (-1)^{i+j} M_{ij}$, and call $C = (C_{ij})_{i,j=1}^n$ the **cofactor matrix** of A .

We can now state the following result.

Theorem 2.61 (Laplace expansion). *Let $A = (a_{ij}) \in \mathbb{R}^{n \times n}$. Then, we can compute the determinant of A by expansion along ...*

- the i -th row:

$$\det(A) = \sum_{j=1}^n (-1)^{i+j} a_{ij} M_{ij} = \sum_{j=1}^n a_{ij} C_{ij} \quad (\text{fixed } i)$$

- the j -th column:

$$\det(A) = \sum_{i=1}^n (-1)^{i+j} a_{ij} M_{ij} = \sum_{i=1}^n a_{ij} C_{ij} \quad (\text{fixed } j)$$

As a proof of this result would require a more detailed analysis, we leave it out.

Although this result may look complicated at first sight, it is actually rather simple to apply, and can lead to very fast computations if the matrix under consideration contains many zeros.

Example 2.62. Consider again the matrix $A = \begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix}$. We want to compute the determinant of A using *expansion along the first row*. By Theorem 2.61 (with $i = 1$) we see that

$$\det \begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix} = \sum_{j=1}^2 (-1)^{1+j} a_{1j} M_{1j} = (-1)^{1+1} a_{11} M_{11} + (-1)^{1+2} a_{12} M_{12} = 1 \cdot M_{11} - 2 \cdot M_{12}.$$

Using $M_{11} = \det(4) = 4$ and $M_{12} = \det(3) = 3$, we obtain $\det(A) = 4 - 2 \cdot 3 = -2$, as required.

Example 2.63. Consider the matrix

$$A = \begin{pmatrix} 17 & 8 & 0 & 6 \\ 4 & 7 & 14 & 0 \\ 0 & 13 & 0 & 0 \\ 0 & 3 & 2 & 0 \end{pmatrix}.$$

To compute its determinant we look for a row or column with preferably only one non-zero entry. This makes the Laplace expansion particularly useful, because most of the terms in the sum vanish.

We choose the fourth column, i.e., we take the Laplace expansion with $j = 4$. (Clearly, there are also other *good* choices.) We obtain

$$\det(A) = \sum_{i=1}^4 (-1)^{i+4} a_{i4} M_{i4}.$$

Now, note that $a_{14} = 6$, but $a_{24} = a_{34} = a_{44} = 0$. Therefore,

$$\det(A) = (-1)^5 a_{14} M_{14} = -6 \cdot M_{14}.$$

To compute M_{14} we have to compute the determinant of the matrix that is obtained by deleting the first row and the last column of A . That is,

$$M_{14} = \det \begin{pmatrix} 4 & 7 & 14 \\ 0 & 13 & 0 \\ 0 & 3 & 2 \end{pmatrix} = 104.$$

The last computation could be done directly with the Rule of Sarrus, or by using again Laplace expansion along the second row to see that $M_{14} = 13 \cdot \det \begin{pmatrix} 4 & 14 \\ 0 & 2 \end{pmatrix} = 13 \cdot 8$. We finally obtain $\det(A) = -6 \cdot 104 = -624$.

The examples above show that one may compute determinants very fast by using Laplace expansion. Moreover, it is interesting that some of the entries (like the 17 in the upper left corner) were not even needed in the computation.

2.5 Cramer's rule

In this section we introduce a straight-forward way of computing solutions to linear systems. This also leads to a formula for the inverse matrix, based on certain determinants.

Let us start linear systems with 2 unknowns:

Given a 2×2 -matrix $A = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix}$ and $b = \begin{pmatrix} b_1 \\ b_2 \end{pmatrix}$. The corresponding linear system $Ax = b$ is given by the equations

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 &= b_1, \\ a_{21}x_1 + a_{22}x_2 &= b_2. \end{aligned}$$

Using row operations, this system can be transformed into

$$\begin{aligned} (a_{11}a_{22} - a_{12}a_{21})x_1 &= b_1a_{22} - a_{12}b_2, \\ (a_{11}a_{22} - a_{12}a_{21})x_2 &= a_{11}b_2 - b_1a_{21}. \end{aligned}$$

Now, all the terms appearing in the last system of equations can be written as determinants. Recall that $\det(A) = a_{11}a_{22} - a_{12}a_{21}$. We obtain that the above system can be written by

$$\begin{aligned}\det(A)x_1 &= \det \begin{pmatrix} b_1 & a_{12} \\ b_2 & a_{22} \end{pmatrix}, \\ \det(A)x_2 &= \det \begin{pmatrix} a_{11} & b_1 \\ a_{21} & b_2 \end{pmatrix}.\end{aligned}$$

This shows that, whenever $\det(A) \neq 0$, we can just divide by it to obtain the precise values of x_1 and x_2 , i.e.,

$$x_1 = \frac{1}{\det A} \det \begin{pmatrix} b_1 & a_{12} \\ b_2 & a_{22} \end{pmatrix}$$

and

$$x_2 = \frac{1}{\det A} \det \begin{pmatrix} a_{11} & b_1 \\ a_{21} & b_2 \end{pmatrix}.$$

Note that, to obtain the k -th entry of the (unique) solution x , we just need to replace the k -th column of A by the RHS b and compute the corresponding determinant. After dividing by $\det(A)$ we are done.

We will now see that the computations in the last example also work for more than two equations, i.e., in the case $n > 2$. This is called **Cramer's rule**.

Theorem 2.64 (Cramer's rule). *Let $A \in \mathbb{R}^{n \times n}$ with $\det A \neq 0$, and $b \in \mathbb{R}^n$.*

Then, the linear system $Ax = b$ has the unique solution $x = (x_1, \dots, x_n)^T$ given by

$$x_k = \frac{\det(A_k)}{\det(A)},$$

where A_k is given by

$$A_k := \begin{pmatrix} a_{1,1} & \dots & a_{1,k-1} & b_1 & a_{1,k+1} & \dots & a_{1,n} \\ a_{2,1} & \dots & a_{2,k-1} & b_2 & a_{2,k+1} & \dots & a_{2,n} \\ \vdots & & \vdots & \vdots & \vdots & & \vdots \\ a_{n,1} & \dots & a_{n,k-1} & b_n & a_{n,k+1} & \dots & a_{n,n} \end{pmatrix}.$$

Proof. From Theorem 2.50, we know that $\text{rank } A = n \iff \det A \neq 0$ and so, that there exists a unique solution to the linear system $Ax = b$, see Lemma 2.43. Recall that the vectors $e_k \in \mathbb{R}^n$ with $1 \leq k \leq n$ are the unit vectors, and that $x = (x_1, x_2, \dots, x_n)^T$ is the column vector representing the solution. We now define the matrices

$$X_k = (e_1 \ e_2 \ \dots \ e_{k-1} \ x \ e_{k+1} \ \dots \ e_n).$$

By computing the row echelon form of X_k we see that $\det(X_k) = x_k$ for all $k = 1, \dots, n$. If we denote the columns of A by c_k , i.e., $A = (c_1, \dots, c_n)$, and recall that $Ae_k = c_k$, we obtain

$$\begin{aligned}A \cdot X_k &= (Ae_1 \ Ae_2 \ \dots \ Ae_{k-1} \ Ax \ Ae_{k+1} \ \dots \ Ae_n) \\ &= (c_1 \ c_2 \ \dots \ c_{k-1} \ Ax \ c_{k+1} \ \dots \ c_n).\end{aligned}$$

Since $Ax = b$, we see that $A \cdot X_k = A_k$ with A_k given in the theorem. Using Lemma 2.56, we see that

$$\det(A_k) = \det(A \cdot X_k) = \det(A) \cdot \det(X_k) = x_k \cdot \det A,$$

which proves the result. \square

Let us see some examples.

Example 2.65. We want to solve $Ax = b$, where

$$A = \begin{pmatrix} 1 & 3 \\ 2 & 7 \end{pmatrix} \quad \text{and} \quad b = \begin{pmatrix} 7 \\ 16 \end{pmatrix}.$$

We see that $\det A = 1$, hence Cramer's rule, see Theorem 2.64, implies that

$$x_1 = \det \begin{pmatrix} 7 & 3 \\ 16 & 7 \end{pmatrix} = 49 - 48 = 1.$$

and

$$x_2 = \det \begin{pmatrix} 1 & 7 \\ 2 & 16 \end{pmatrix} = 16 - 14 = 2.$$

The solution to this linear system is therefore $x = (1, 2)^T$.

Example 2.66. Consider the matrix

$$A = \begin{pmatrix} 8 & 1 & 3 \\ 7 & 0 & 11 \\ 5 & 5 & 0 \end{pmatrix},$$

for which we have $\det A = -280$. Then, due to Cramer's rule, we compute the following solution, denoted by $x = (x_1, x_2, x_3)^T$, for the RHS $b = (1, 1, 1)^T$,

$$\begin{aligned} x_1 &= \frac{-1}{280} \det \begin{pmatrix} 1 & 1 & 3 \\ 1 & 0 & 11 \\ 1 & 5 & 0 \end{pmatrix} = \frac{29}{280} \\ x_2 &= \frac{-1}{280} \det \begin{pmatrix} 8 & 1 & 3 \\ 7 & 1 & 11 \\ 5 & 1 & 0 \end{pmatrix} = \frac{27}{280} \\ x_3 &= \frac{-1}{280} \det \begin{pmatrix} 8 & 1 & 1 \\ 7 & 0 & 1 \\ 5 & 5 & 1 \end{pmatrix} = \frac{7}{280}. \end{aligned}$$

2.6 Inverse matrices

We now combine the findings of the last sections to give an explicit formula for the inverse matrix, if it exists. This is particularly useful to compute solutions of a linear system $Ax = b$ if the RHS b is *a priori* not known. Moreover, the inverse is handy when we want to work with a (unique) solution theoretically.

For completeness, let us repeat the definition.

Definition 2.67. Let $A \in \mathbb{R}^{n \times n}$ and assume that there exists some $A' \in \mathbb{R}^{n \times n}$ with the property that

$$A \cdot A' = A' \cdot A = I_n.$$

Then, we say that A is **invertible** or **regular**, and we write $A^{-1} := A'$ to denote the inverse.

Note that a matrix must be a square matrix for both matrix products above being defined. That's why we define the inverse only for $A \in \mathbb{R}^{n \times n}$.

There are two possible interpretations of the inverse:

1. We consider matrices as kind of *numbers* and, for a matrix A , we look for another matrix that is the *inverse element* of A for matrix multiplication, see the field axioms (Axiom 1).
2. We consider matrices as *mappings* $A: \mathbb{R}^n \rightarrow \mathbb{R}^n$ by the matrix-vector product and look for the inverse mapping, see Definition 1.13, which should also be given by a matrix.

Both interpretations are equivalent.

The inverse of a matrix can be used to **solve linear systems** as follows:

Since the inverse matrix A^{-1} satisfies $A^{-1}A = I_n$, we have that

$$Ax = b \iff x = I_n x = A^{-1}Ax = A^{-1}b.$$

Therefore, the solution to the linear system $Ax = b$ is given by $x = A^{-1}b$ (matrix-vector multiplication), whenever A is regular.

From Theorem 2.50 we already know that

$$\det A \neq 0 \iff \text{rank } A = n$$

for a given matrix $A \in \mathbb{R}^{n \times n}$. We now combine that with Lemma 2.43 to show that A is bijective, and hence invertible, in this case.

Let us state that as a lemma.

Lemma 2.68. Let $A \in \mathbb{R}^{n \times n}$ be a square matrix. Then,

$$\det A \neq 0 \iff A \text{ is invertible},$$

and in this case

$$\det(A^{-1}) = \frac{1}{\det A}.$$

Proof. For the equivalence, note that $\text{rank}(A) = n$ if and only if the linear system $Ax = b$ has a unique solution for all $b \in \mathbb{R}^n$, see Lemma 2.43. In other words, the mapping $A: \mathbb{R}^n \rightarrow \mathbb{R}^n$ (i.e., A maps vectors to vectors) is injective (“For every $b \in \mathbb{R}^n$ there is at most one x with $Ax = b$.”) and surjective (“For every $b \in \mathbb{R}^n$ there is some x with $Ax = b$.”). Hence, A is bijective, and therefore invertible (aka. regular), see Theorem 1.16.

From Lemma 2.56 we know that

$$\det(A) \cdot \det(A^{-1}) = \det(A \cdot A^{-1}) = \det(I_n) = 1,$$

whenever A is regular, which proves the claim. \square

Example 2.69. Note that, if A is regular, then A^{-1} exists and is also regular. Hence, the inverse of the inverse exists, and fulfills $(A^{-1})^{-1} = A$. (Verify yourself!)

Note that **the inverse of the product of matrices is the product of the inverses**, but we have to change the order (as for the transpose).

Lemma 2.70. *Let $A, B \in \mathbb{R}^{n \times n}$ be regular matrices. Then,*

$$(A \cdot B)^{-1} = B^{-1} \cdot A^{-1}.$$

In particular, AB is also regular.

Proof. First note that $\det(AB) = \det(A)\det(B) \neq 0$ due to Lemma 2.56 and Lemma 2.68, which shows that AB is regular. If we note that $(AB)(B^{-1}A^{-1}) = A(BB^{-1})A^{-1} = AA^{-1} = I_n$, we see that $(B^{-1}A^{-1})$ is the inverse of AB . \square

For the **computation of the inverse A^{-1}** , denote the columns of A^{-1} by $c_1, \dots, c_n \in \mathbb{R}^n$, i.e.,

$$A^{-1} = (c_1, c_2, \dots, c_n).$$

We then have $A^{-1}e_k = c_k$, where e_k is the k -th unit vector. (We used already earlier that matrix-vector multiplication with a unit vector gives a column of the matrix.) Using the above equivalence, with $x = c_k$ and $b = e_k$, we see that

$$A^{-1}e_k = c_k \iff Ac_k = e_k.$$

That is, we can calculate c_k , i.e., the k -th column of A^{-1} , by solving the linear system $Ax = e_k$.

We can now use Cramer’s rule, together with the Laplace expansion, to compute the inverse of A . Recall that the cofactor matrix of A is defined by $C = (C_{ij}) \in \mathbb{R}^{n \times n}$, where

$$C_{ij} = (-1)^{i+j} M_{ij}$$

and M_{ij} is the (i, j) -minor, i.e., the determinant of the matrix that is obtained by deleting the i -th row and the j -th column, see Definition 2.60.

The following theorem shows that one can compute the inverse of a matrix as the transpose of its cofactor matrix divided by the determinant.

Theorem 2.71. *Let $A \in \mathbb{R}^{n \times n}$ with $\det(A) \neq 0$, and let $C = (C_{ij}) \in \mathbb{R}^{n \times n}$ be the cofactor matrix of A . Then,*

$$A^{-1} = \frac{1}{\det A} C^T,$$

i.e.,

$$(A^{-1})_{ij} = \frac{C_{ji}}{\det A},$$

where $(A^{-1})_{ij}$ denotes the ij -th entry of A^{-1} .

Proof. Fix some $j = 1, \dots, n$. The discussion above shows that the j -th column of A^{-1} can be computed by solving the linear system $Ax = e_j$. Cramer's rule, see Theorem 2.64, yields that the i -th entry of the solution $x = (x_1, \dots, x_n)$ to this linear system is given by

$$x_i = \frac{\det(A_i)}{\det(A)},$$

where

$$A_i = \det \begin{pmatrix} a_{1,1} & \dots & a_{1,i-1} & 0 & a_{1,i+1} & \dots & a_{1,n} \\ \vdots & & \vdots & \vdots & 0 & & \vdots \\ a_{j-1,1} & \dots & a_{j-1,i-1} & 0 & a_{j-1,i+1} & \dots & a_{j-1,n} \\ a_{j,1} & \dots & a_{j,i-1} & 1 & a_{j,i+1} & \dots & a_{j,n} \\ a_{j+1,1} & \dots & a_{j+1,i-1} & 0 & a_{j+1,i+1} & \dots & a_{j+1,n} \\ \vdots & & \vdots & \vdots & & & \vdots \\ a_{n,1} & \dots & a_{n,i-1} & 0 & a_{n,i+1} & \dots & a_{n,n} \end{pmatrix}.$$

We just replaced the i -th column of A by e_j . Now we use Laplace expansion, see Theorem 2.61, with expansion along the i -th column. (Note that in the statement of the Laplace expansion we used the j -th column. Therefore, we need to be careful with the indices.) We see that the only non-zero entry in the i -th column of A_i is the 1 in the j -th row. We obtain (for fixed i)

$$\det(A_i) = (-1)^{i+j} M_{ji} = C_{ji},$$

i.e., the determinant of A_i is the (j, i) -cofactor of A . (Note the reversed indices.) This finally shows that

$$(A^{-1})_{ij} = x_i = \frac{C_{ji}}{\det A}.$$

□

The systematic **procedure to compute the inverse** is:

1. Compute the matrix of minors $M = (M_{ij})_{i,j=1}^n$ of A .
2. Compute the cofactor matrix $C = (C_{ij})_{i,j=1}^n$ with $C_{ij} = (-1)^{i+j} M_{ij}$.
3. Transpose and dividing by determinant to obtain

$$A^{-1} = \frac{1}{\det A} C^T.$$

Let us see some examples.

Example 2.72. Consider the matrix

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

which satisfies $\det A = -2$.

We start by computing the matrix of minors. Note that deleting a row and a column of A makes it to a 1×1 matrix, i.e., a number, and its determinant is just that number. Therefore, we see that the matrix of minors is given by

$$M = \begin{pmatrix} 4 & 3 \\ 2 & 1 \end{pmatrix}.$$

For example, we obtain M_{21} by deleting the second row and the first column of A , and compute the determinant $M_{21} = \det(2) = 2$.

To compute the cofactor matrix C , we have to multiply each entry M_{ij} by $(-1)^{i+j}$, i.e., we multiply with -1 if $i + j$ is odd, and leave all other entries unchanged. We obtain

$$C = \begin{pmatrix} 4 & -3 \\ -2 & 1 \end{pmatrix}.$$

We therefore obtain that

$$A^{-1} = \frac{1}{\det A} C^T = -\frac{1}{2} \begin{pmatrix} 4 & -2 \\ -3 & 1 \end{pmatrix} = \begin{pmatrix} -2 & 1 \\ 3/2 & -1/2 \end{pmatrix}.$$

The result can (and should) be checked by computing

$$AA^{-1} = \begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix} \begin{pmatrix} -2 & 1 \\ 3/2 & -1/2 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} = I_2.$$

The inverse matrix can now be used to calculate the unique solution to $Ax = b$ for any RHS b . We obtain

$$x = A^{-1}b = \begin{pmatrix} -2 & 1 \\ 3/2 & -1/2 \end{pmatrix} \begin{pmatrix} b_1 \\ b_2 \end{pmatrix} = \begin{pmatrix} b_2 - 2b_1 \\ \frac{3b_1 - b_2}{2} \end{pmatrix}.$$

For example, the solution to $Ax = \begin{pmatrix} 3 \\ 1 \end{pmatrix}$, i.e., we have $b = \begin{pmatrix} 3 \\ 1 \end{pmatrix}$, is given by $x = \begin{pmatrix} \frac{3-2}{2} \\ \frac{3 \cdot 3 - 1}{2} \end{pmatrix} = \begin{pmatrix} 1 \\ 4 \end{pmatrix}$.

Example 2.73. Now we compute the inverse of

$$A = \begin{pmatrix} 1 & 0 & 2 \\ 4 & 1 & 8 \\ 0 & 1 & 1 \end{pmatrix},$$

by using Cramer's rule. First, we have that $\det A = 1$. The matrix of minors is found to be

$$M = \begin{pmatrix} -7 & 4 & 4 \\ -2 & 1 & 1 \\ -2 & 0 & 1 \end{pmatrix},$$

where, e.g., $M_{11} = \det\begin{pmatrix} 1 & 8 \\ 1 & 1 \end{pmatrix} = -7$ and $M_{32} = \det\begin{pmatrix} 1 & 2 \\ 4 & 8 \end{pmatrix} = 0$. We obtain the cofactor matrix

$$C = \begin{pmatrix} -7 & -4 & 4 \\ 2 & 1 & -1 \\ -2 & 0 & 1 \end{pmatrix},$$

and therefore the inverse

$$A^{-1} = \frac{1}{\det A} C^T = \begin{pmatrix} -7 & 2 & -2 \\ -4 & 1 & 0 \\ 4 & -1 & 1 \end{pmatrix}.$$

Finally, we want to discuss the **Gauss-Jordan algorithm** to compute the inverse. This method is very similar to the Gaussian elimination, and is sometimes handy, at least for small matrices. (I do not suggest to use this method, as it is prone to error, but others think differently, and so I state it for completeness.)

For this recall that we can apply the Gaussian elimination to more vectors at once, see Remark 2.45, which can be used to solve the linear system for different RHS's simultaneously. From the discussion above, we know that we actually need to solve the linear systems $Ax = e_j$ for all $j = 1, \dots, n$ to obtain all columns of A^{-1} . Hence, we can compute all columns of A^{-1} at once by computing the reduced row echelon form of

$$(A|I_n) = \left(\begin{array}{ccc|cc} a_{11} & \dots & a_{1n} & 1 & 0 \\ \vdots & & \vdots & \ddots & \\ a_{n1} & \dots & a_{nn} & 0 & 1 \end{array} \right).$$

If A is regular, i.e., $\text{rank } A = n$, we know that the reduced row echelon form of A is the identity matrix. Thus, by using Gaussian elimination, we are able to compute

$$(A|I) \longrightarrow (I|A^{-1})$$

by using only row operations. This shows the following result.

Theorem 2.74. *Let $A \in \mathbb{R}^{n \times n}$ be an invertible matrix, i.e., $\det(A) \neq 0$. Then, the reduced row echelon form of*

$$\left(\begin{array}{ccc|cc} a_{11} & \dots & a_{1n} & 1 & 0 \\ \vdots & & \vdots & \ddots & \\ a_{n1} & \dots & a_{nn} & 0 & 1 \end{array} \right)$$

has the form

$$\left(\begin{array}{ccc|cc} 1 & & 0 & a'_{11} & \dots & a'_{1n} \\ & \ddots & & \vdots & & \vdots \\ 0 & & 1 & a'_{n1} & \dots & a'_{nn} \end{array} \right)$$

and it holds that $A^{-1} = (a'_{ij})_{i,j=1}^n$.

Let us consider again the examples from above.

Example 2.75. We want to compute the inverse of

$$A = \begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix}.$$

We apply the Gauss-Jordan algorithm, which means we have to transform the following matrix into its echelon form by

$$\begin{array}{c} \left(\begin{array}{cc|cc} 1 & 2 & 1 & 0 \\ 3 & 4 & 0 & 1 \end{array} \right) \\ \xrightarrow{II - 3I} \left(\begin{array}{cc|cc} 1 & 2 & 1 & 0 \\ 0 & -2 & -3 & 1 \end{array} \right) \\ \xrightarrow{I + II} \left(\begin{array}{cc|cc} 1 & 0 & -2 & 1 \\ 0 & -2 & -3 & 1 \end{array} \right) \\ \xrightarrow{-1/2II} \left(\begin{array}{cc|cc} 1 & 0 & -2 & 1 \\ 0 & 1 & 3/2 & -1/2 \end{array} \right). \end{array}$$

The result clearly agrees with the one from Example 2.72.

Example 2.76. We want to compute again the inverse of

$$A = \begin{pmatrix} 1 & 0 & 2 \\ 4 & 1 & 8 \\ 0 & 1 & 1 \end{pmatrix}.$$

We use the Gauss-Jordan algorithm

$$\begin{array}{c} \left(\begin{array}{ccc|ccc} 1 & 0 & 2 & 1 & 0 & 0 \\ 4 & 1 & 8 & 0 & 1 & 0 \\ 0 & 1 & 1 & 0 & 0 & 1 \end{array} \right) \\ \xrightarrow{II - 4I} \left(\begin{array}{ccc|ccc} 1 & 0 & 2 & 1 & 0 & 0 \\ 0 & 1 & 0 & -4 & 1 & 0 \\ 0 & 1 & 1 & 0 & 0 & 1 \end{array} \right) \\ \xrightarrow{III - II} \left(\begin{array}{ccc|ccc} 1 & 0 & 2 & 1 & 0 & 0 \\ 0 & 1 & 0 & -4 & 1 & 0 \\ 0 & 0 & 1 & 4 & -1 & 1 \end{array} \right) \\ \xrightarrow{I - 2II} \left(\begin{array}{ccc|ccc} 1 & 0 & 0 & -7 & 2 & -2 \\ 0 & 1 & 0 & -4 & 1 & 0 \\ 0 & 0 & 1 & 4 & -1 & 1 \end{array} \right) \end{array}$$

to see $A^{-1} = \begin{pmatrix} -7 & 2 & -2 \\ -4 & 1 & 0 \\ 4 & -1 & 1 \end{pmatrix}$, which is the same result as in Example 2.73.

3 Sequences and series

This section is dedicated to the specification of the idea of *limiting processes*. It forms one of the central ideas of mathematical analysis and defines the basis for essential concepts like continuity, differentiability, integration etc.

A motivational example is the infinite sum of the numbers 2^{-n} , $n = 0, 1, \dots$, i.e.,

$$1 + \frac{1}{2} + \frac{1}{4} + \frac{1}{8} + \frac{1}{16} + \dots$$

We want to give an exact mathematical meaning for an infinite addition of terms which leads us to the definition of a *limit* and the concept of *convergence of sequences*:

That is, we have a sequence of numbers, say a_n , which is given by some rule, e.g., by the recursion $a_{n+1} := f(a_n)$ for some fixed function f , and we want to know what happens if “ n goes to ∞ ”. That is, we want to find out what happens if we repeat such a process infinitely often.

To study such questions, we start again with precise definitions of the objects we are considering.

Definition 3.1 (Sequence). Let $M \neq \emptyset$ be an arbitrary set, and $I \subset \mathbb{Z}$.

An (**infinite**) **sequence** in M is a mapping $a: I \rightarrow M$.

With the notation $a_n := a(n)$, we can write the sequence as $(a_n)_{n \in I}$.

The **range** of a sequence $(a_n)_{n \in I}$ is given by $\{a_n : n \in \mathbb{N}\}$.

The domain I of a sequence is called the **index set** of the sequence.

We write $(a_n)_{n \in I} \subset M$ to say that $\forall n \in I: a_n \in M$,

and we write M^I for the **set of all sequences** in M with index set I .

In most cases, we consider $I = \mathbb{N}$ or $I = \{K, K+1, \dots\}$ for some $K \in \mathbb{Z}$.

In this case, we write $(a_n)_{n \in I} = (a_n)_{n=K}^\infty = (a_n)_{n \geq K} = (a_K, a_{K+1}, a_{K+2}, \dots)$.

If the index set is clear, we may just write (a_n) for $(a_n)_{n \in I}$.

As one considers a sequence as a mapping defined on the index set $I \subset \mathbb{Z}$, one wants to express that we are dealing with a list of elements in a particular order. Thus, we clearly distinguish between the sequence (a_n) and its range $\{a_n : n \in I\}$.

Note that two sequences $(a_n)_{n \in I}$ and $(b_n)_{n \in I}$ are equal if and only if

$$\forall n \in I: a_n = b_n,$$

in this case we write $(a_n)_{n \in I} = (b_n)_{n \in I}$.

In the special cases $M = \mathbb{R}$ or $M = \mathbb{C}$ we say that $(a_n)_{n \in I}$ is a **real-valued or complex-valued sequence**, respectively. We will focus on real-valued sequences in this lecture. However, most statements also hold for complex-valued sequences. We comment on the differences when needed.

To define a sequence, the most common way is to use an **explicit formula**, for instance

$$a_n = 2^n \quad \text{or} \quad b_n = 1 + \frac{1}{n},$$

or by a recursion, i.e., we give one (or more) starting value(s) and a rule how to calculate a new term from previous terms. For the above examples we could write

$$a_1 = 2, \quad a_{n+1} = 2a_n$$

and

$$b_1 = 2, \quad b_{n+1} = b_n \cdot \frac{n(n+2)}{(n+1)^2}.$$

(Verify these formulas!)

Example 3.2 (Fibonacci sequence). One of the most famous sequences, which appears in several areas of natural science, is the so called **Fibonacci sequence**. Here, the recursion depends on more than just the last value. The sequence $(F_n)_{n \in \mathbb{N}}$ is defined by

$$F_1 = 1, F_2 = 1, \text{ and } F_n = F_{n-1} + F_{n-2} \quad \text{for } n \geq 3.$$

The first values of this sequence are $1, 1, 2, 3, 5, 8, 13, 21, 34, 55, \dots$

It is an interesting phenomenon that the quotients F_{n+1}/F_n converge to the **golden ratio** $\frac{1+\sqrt{5}}{2}$ (See e.g. Wikipedia for its importance). We will see later how to prove such statements.

Example 3.3 (Infinite sums aka. series). If we want to add infinitely many numbers, say all the numbers a_n , $n \in \mathbb{N}$, then we can consider the new sequence $s_n = \sum_{k=1}^n a_k$, which can be given recursively by $s_n = s_{n-1} + a_n$, and we would like to know if s_n approaches a certain number when n goes to infinity. These special sequences are called *series*, and we come back to this in Section 3.6.

3.1 Convergence of sequences

As mentioned above, the concept of *convergence* is central to mathematical analysis. Intuitively, it states that the terms of the sequence $(a_n)_{n \in \mathbb{N}}$ approach a *limit* with growing index n . In what follows, we will consider only real- or complex-valued sequences with index set $I = \mathbb{N}$.

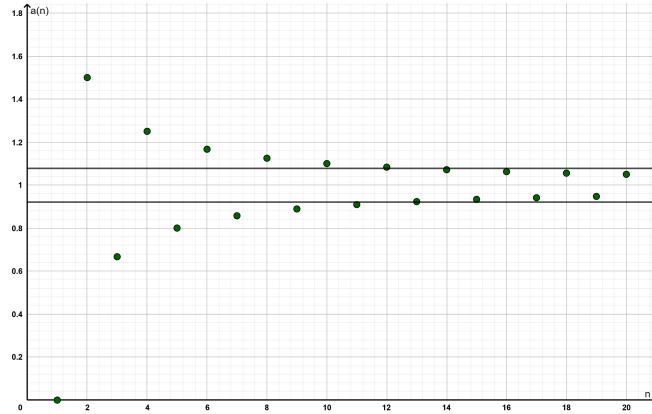


Figure 19: convergence of sequence $1 + \frac{(-1)^n}{n}$

To define what is means to *converge* to something, we need the notion of a neighborhood.

Definition 3.4 (Neighborhood). Let $M = \mathbb{R}$ or $M = \mathbb{C}$, $a \in M$ and $\varepsilon > 0$. We define the **ε -neighborhood** of a in M by

$$U_\varepsilon(a) := \{x \in M : |x - a| < \varepsilon\}.$$

It is mostly clear from the context if we consider real or complex neighborhoods. Note that, for $M = \mathbb{R}$ and $a \in \mathbb{R}$ the ε -neighborhood $U_\varepsilon(a)$ is just the open interval $(a - \varepsilon, a + \varepsilon)$. In the complex case, i.e., $M = \mathbb{C}$ and $a \in \mathbb{C}$, $U_\varepsilon(a)$ is the *disc* of radius ε around a in the complex plane.

Remark 3.5. Note already now, that this definition is quite flexible if we switch to more complex situations. That is, we can define neighborhoods whenever we have a measure for the 'distance' on the set M . We therefore use this notation to get used to it.

We now come to the formal definition of *convergence* to a *limit* a .

Definition 3.6 (Convergence). Let $(a_n)_{n \in \mathbb{N}} \subset \mathbb{C}$ and $a \in \mathbb{C}$. We say that the sequence $(a_n)_{n \in \mathbb{N}}$ **converges to** a if and only if

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

or, equivalently,

$$\forall \varepsilon > 0 \exists n_0 \in \mathbb{N} \forall n \geq n_0: a_n \in U_\varepsilon(a),$$

or, equivalently,

$$\forall \varepsilon > 0 \exists n_0 \in \mathbb{N}: (a_n)_{n=n_0}^\infty \subset U_\varepsilon(a).$$

In this case we call a the **limit** of the sequence and write

$$a = \lim_{n \rightarrow \infty} a_n \quad \text{or} \quad a_n \xrightarrow{n \rightarrow \infty} a \quad \text{or simply} \quad a_n \rightarrow a.$$

$(a_n)_{n \in \mathbb{N}}$ is called **convergent**, or we say that **the limit of** $(a_n)_{n \in \mathbb{N}}$ **exists**, if there exists some $a \in \mathbb{C}$ such that $a_n \rightarrow a$, otherwise $(a_n)_{n \in \mathbb{N}}$ is called **divergent**.

The statement

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

can be equivalently phrased as:

For all $\varepsilon > 0$ we have $|a_n - a| < \varepsilon$ for...

- ... all large enough n .
- ... all but finitely many n .
- ... almost all n .

For the second wording, note that there must be a largest of the finitely many *exceptions* (i.e., the n for which $|a_n - a| \geq \varepsilon$). One may choose n_0 just one larger than this number.

Remark 3.7. Note that the limit does not depend on the first terms of a sequence. So, in particular, if (b_n) is a sequence with $b_n = a_n$ for almost all n , then $a_n \rightarrow a \iff b_n \rightarrow a$.

Let us consider some examples.

Example 3.8. Consider the sequence $(a_n)_{n \in \mathbb{N}}$ with $a_n = \frac{1}{n}$. For each $\varepsilon > 0$ we can find some $n_0 \in \mathbb{N}$ such that $\frac{1}{n_0} < \varepsilon$. This is the Archimedean property.

Since $\frac{1}{n} \leq \frac{1}{n_0}$ for $n \geq n_0$, we obtain $|a_n - 0| = \frac{1}{n} \leq \frac{1}{n_0} < \varepsilon$, and hence $a_n \rightarrow 0$.

Example 3.9. Consider the sequence $(a_n)_{n \in \mathbb{N}}$ with $a_n = (-1)^n$. This sequence is divergent. For a proof we assume the opposite, i.e., that (a_n) converges to some $a \in \mathbb{R}$. Now, by the definition of convergence, we have that there exists some large enough n_0 such that $a_n \in U_{1/2}(a)$ for all $n \geq n_0$. (Note that the $1/2$ is arbitrary here. Every $\varepsilon < 1$ would work.) However, we always have $|a_{n+1} - a_n| > 1$, and therefore, if $a_n \in U_{1/2}(a)$, we have $a_{n+1} \notin U_{1/2}(a)$. In particular, there cannot be an n_0 such that $a_n \in U_{1/2}(a)$ for all $n \geq n_0$: A contradiction. Hence, (a_n) cannot be a convergent sequence.

We now turn to some special properties of sequences or, as one might say, we just give names to sequences with special properties. Afterwards we analyse the relation of these properties.

Definition 3.10 (Null sequence). Let $(a_n)_{n \in \mathbb{N}}$ be a real sequence such that

$$\lim_{n \rightarrow \infty} a_n = 0.$$

Then we call $(a_n)_{n \in \mathbb{N}}$ a **null sequence**.

Example 3.11. The sequences $(\frac{1}{n})_{n \in \mathbb{N}}$ and $(2^{-n})_{n \in \mathbb{N}}$ are null sequences.

We will frequently use the trivial observation that

$$a_n \rightarrow a \iff (a_n - a) \rightarrow 0,$$

i.e., (a_n) converges to a if and only if $(a_n - a)$ is a null sequence. This follows immediately from the definition of the limit.

Example 3.12. The sequence $(a_n)_{n \in \mathbb{N}}$ with $a_n = 1 + \frac{1}{n}$ is not a null sequence.

Since we know that $(1/n)$ is a null sequence, we can set $a = 1$ above, and see that $a_n - a = \frac{1}{n} \rightarrow 0$. Hence, $a_n \rightarrow 1$.

We now turn to a second property of sequences: boundedness.

Definition 3.13. Let $(a_n)_{n \in \mathbb{N}} \subset \mathbb{C}$ be a sequence. We call the sequence **bounded** if

$$\exists R > 0 \forall n \in \mathbb{N}: |a_n| \leq R.$$

Moreover, if $(a_n)_{n \in \mathbb{N}} \subset \mathbb{R}$, we call the sequence **bounded from above** if and only if

$$\exists C \in \mathbb{R} \forall n \in \mathbb{N}: a_n \leq C,$$

and **bounded from below** if and only if

$$\exists c \in \mathbb{R} \forall n \in \mathbb{N}: a_n \geq c.$$

In other words a sequence is bounded (from above/below), if and only if its range is a bounded set (from above/below).

Example 3.14. Clearly, the sequences $((-1)^n)_{n \in \mathbb{N}}$ and $(\frac{42}{n})_{n \in \mathbb{N}}$ are bounded (by 1 and 42, resp.). We also have that $((-1)^n \cdot \frac{42}{n})_{n \in \mathbb{N}}$ is bounded (by 42). Actually, we easily see from the definition that the (term-wise) product of two bounded sequences is bounded.

The triangle inequality shows that the sum of two bounded sequences is also bounded.

Next we observe the following property of convergent sequences.

Theorem 3.15. Let $(a_n)_{n \in \mathbb{N}}$ be a convergent sequence. Then $(a_n)_{n \in \mathbb{N}}$ is bounded.

Proof. Let $a \in \mathbb{C}$ be the limit of $(a_n)_{n \in \mathbb{N}}$, i.e.

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

Fix $\varepsilon_0 = 1$. Then we can find a N such that $\forall n \geq N: |a_n - a| < 1$. The triangle inequality yields

$$|a_n| \leq |a_n - a| + |a| < 1 + |a|,$$

for all $n \geq N$. For the remaining elements $\{a_1, \dots, a_{N-1}\}$ we simply take the maximum, $c_1 = \max\{|a_1|, |a_2|, \dots, |a_{N-1}|\}$. The maximum of a finite amount of real numbers always exists. Hence

$$\forall n \in \mathbb{N}: |a_n| \leq \max\{c_1, |a| + 1\}.$$

□

Example 3.16. The sequence given by $b_n = (-1)^n$ is bounded by 1 but not convergent. So the other direction of the above theorem does not hold in general.

We finally introduce the terminology of *definite divergence* of a real-valued sequence. This concept is used to give a formal definition of the idea that a sequence “**converges to infinity**”.

Definition 3.17. Let $(a_n)_{n \in \mathbb{N}} \subset \mathbb{R}$. The sequence $(a_n)_{n \in \mathbb{N}}$ **tends to ∞** ($= +\infty$) if

$$\forall A > 0 \exists n_0 \in \mathbb{N} \forall n \geq n_0: a_n > A.$$

We write $a_n \rightarrow \infty$ or $\lim_{n \rightarrow \infty} a_n = \infty$, and call ∞ the **improper limit** of (a_n) .

The tendency to $-\infty$ is defined analogously with $a_n < -A$.

If the sequence (a_n) tends to $\pm\infty$, it is called **definitely divergent**.

Note that definitely divergent sequences are necessarily unbounded.

Moreover, we do not have such a concept for complex-valued sequences, as we do not have an order on \mathbb{C} .

Example 3.18. We have $\lim_{n \rightarrow \infty} \sqrt{n} = \infty$ and $\lim_{n \rightarrow \infty} (-n^2) = -\infty$.

3.2 Calculation rules for limits

We now study how to determine the limit of (complicated) sequences. This always follows the same procedure: Either we already know the limit of the sequence under consideration, or one has to split up the sequence into easier parts that can be handled, or splitted again.

To effectively do this, one needs a sufficiently large **collection of known limits**, and we will present the most important below. Together with some rules of calculation, this will allow to compute also the limit of quite complicated sequences.

Let us start with a lemma that shows how to verify that a sequence is a null sequence by **comparison with another null sequence**. The proof is left to the reader.

Lemma 3.19. *If $(a_n)_{n \in \mathbb{N}} \subset \mathbb{C}$ is a null sequence and $(b_n)_{n \in \mathbb{N}} \subset \mathbb{C}$ is a sequence with*

$$|b_n| \leq C \cdot |a_n|^c \quad \text{for some } c, C > 0 \text{ and almost all } n,$$

then (b_n) is also a null sequence.

From this lemma we directly see that the sequences $(\frac{C}{n^c})_{n \in \mathbb{N}}$ for fixed $c, C > 0$ are null sequences.

Let us now consider other important “building blocks”, i.e., limits that may be considered known from now on, together with the corresponding proofs.

The first example is concerned with powers of arbitrary complex bases.

Example 3.20. Let $z \in \mathbb{C}$ with $|z| < 1$. Then

$$\lim_{n \rightarrow \infty} z^n = 0.$$

Proof. For $z = 0$ the result is clear. For $z \neq 0$ we set $x > 0$ such that $|z| = \frac{1}{1+x}$. With the Bernoulli inequality or the binomial formula we get $(1+x)^n \geq 1 + nx$ and obtain

$$|z^n| = \frac{1}{(1+x)^n} \leq \frac{1}{1+nx} \leq \frac{1}{x} \frac{1}{n}.$$

Since $(\frac{1}{n})$ is a null sequence, we get from Lemma 3.19 that z^n is a null sequence as well. \square

As the above sequence is not bounded for $|z| > 1$ (Check yourself!), we obtain from Theorem 3.15 that it cannot be convergent, i.e., (z^n) is divergent if $|z| > 1$.

In the case $|z| = 1$, we cannot say in general if the sequence is convergent or not: Although we have the constant (and therefore clearly convergent sequence) for $z = 1$, the sequence (z^n) does not converge for other z , like $z = -1$ or $z = e^{i\pi/2}$. We leave out the details here.

The next example shows what happens if we replace the n -th power by a n -th root.

Example 3.21. Let $a > 0$. Then, we have

$$\lim_{n \rightarrow \infty} \sqrt[n]{a} = 1.$$

(For $a = 0$ we clearly have $\sqrt[n]{a} \rightarrow 0$.)

Proof. Let us first consider $a \geq 1$. We will show that $x_n := \sqrt[n]{a} - 1$ satisfies $x_n \rightarrow 0$, which proves the statement. Since $a \geq 1$ we have $x_n \geq 0$. By Bernoulli's inequality $(1+x)^n \geq 1+nx$, which holds for $x \geq -1$, we obtain

$$a = (\sqrt[n]{a})^n = (1 + x_n)^n \geq 1 + nx_n.$$

This implies

$$|x_n| = x_n \leq \frac{a-1}{n}.$$

Since $(\frac{1}{n})$ is a null sequence we get that (x_n) is also a null sequence.

For $a < 1$, let $b = 1/a > 1$ and consider $x_n = \sqrt[n]{b} - 1$. From the above we know that (x_n) is a (non-negative) null sequence. Moreover, we have $\sqrt[n]{a} \leq 1$ and therefore

$$|\sqrt[n]{a} - 1| = 1 - \sqrt[n]{a} = \sqrt[n]{a}(\sqrt[n]{b} - 1) \leq 1 \cdot x_n.$$

Again, since (x_n) converges to zero, this proves the statement. \square

The next important limit shows that the constant a in the example above may even be replaced by an unbounded sequence.

Example 3.22.

$$\lim_{n \rightarrow \infty} \sqrt[n]{n} = 1.$$

Proof. Let $x_n := \sqrt[n]{n} - 1$, so we have to show that $x_n \rightarrow 0$. From Lemma 1.56 with $k = 2$ we obtain

$$n = (1 + x_n)^n \geq 1 + \binom{n}{2}x_n^2 = 1 + \frac{n(n-1)}{2}x_n^2.$$

This implies

$$|x_n| = x_n \leq \sqrt{\frac{2}{n}}.$$

Since $(\frac{1}{\sqrt[n]{n}})$ is a null sequence we get that x_n is a null sequence, which proves $\sqrt[n]{n} \rightarrow 1$. \square

We state one last example before we turn to general rules for the calculation with limits. This example could be phrased as “**exponential growth is faster than polynomial growth**”, and is one of the basic arguments when dealing with limits.

Example 3.23. Let $z \in \mathbb{C}$ with $|z| > 1$ and $k \in \mathbb{Z}$. Then,

$$\lim_{n \rightarrow \infty} \frac{n^k}{z^n} = 0.$$

Proof. The proof of this limit combines all the ideas from the preceding examples.

First, note that the limit is already clear from the above if $k \leq 0$. (Why?)

For $k > 0$, set $x := |z| - 1$ with $x > 0$, and assume that $n > 2k$. (This is possible, since we are only interested in large n .) From Lemma 1.56 we obtain

$$|z|^n = (1 + x)^n > \binom{n}{k+1}x^{k+1} = \frac{n \cdot (n-1) \cdots (n-k)}{(k+1)!}x^{k+1}.$$

From $n > 2k$, we obtain that $n - k > n/2$ (or more general $n - k + \ell > n/2$ for all $\ell \in \{0, \dots, k\}$). Therefore,

$$|z|^n = (1+x)^n > \frac{n \cdot (n-1) \cdots (n-k)}{(k+1)!} x^{k+1} > \frac{(n/2)^{k+1}}{(k+1)!} x^{k+1}.$$

It follows

$$\left| \frac{n^k}{z^n} \right| = \frac{n^k}{|z|^n} < \frac{2^{k+1}(k+1)!}{x^{k+1}} \frac{n^k}{n^{k+1}} =: K \cdot \frac{1}{n},$$

where K is a constant (i.e., does not depend on n). As $(\frac{1}{n})$ is a null sequence, we get that $(\frac{n^k}{z^n})$ is also a null sequence. \square

The following calculation rules for convergent sequences and their limits will be very useful if we want to compute the limits of more complicated sequences.

Theorem 3.24. Let $(a_n)_{n \in \mathbb{N}}, (b_n)_{n \in \mathbb{N}}$ be convergent sequences, and let $\lambda \in \mathbb{C}$. Moreover, let $a := \lim_{n \rightarrow \infty} a_n$ and $b := \lim_{n \rightarrow \infty} b_n$. Then, we have

- (i) $\lim_{n \rightarrow \infty} (a_n + b_n) = a + b$
- (ii) $\lim_{n \rightarrow \infty} (\lambda \cdot a_n) = \lambda \cdot a$
- (iii) $\lim_{n \rightarrow \infty} (a_n \cdot b_n) = a \cdot b$
- (iv) If $b \neq 0$, then $\lim_{n \rightarrow \infty} \frac{a_n}{b_n} = \frac{a}{b}$.

Proof. For the first statement we need to show that $\forall \varepsilon > 0 \exists n_0 \in \mathbb{N} : |a_n + b_n - (a + b)| \leq \varepsilon$. Therefore let $\varepsilon > 0$ be arbitrary but fixed. Using the triangle inequality we get that

$$|a_n + b_n - a - b| \leq |a_n - a| + |b_n - b|.$$

Since $(a_n)_{n=1}^\infty$ and $(b_n)_{n=1}^\infty$ are convergent sequences we can find $N \in \mathbb{N}$ such that for all $n \geq N$:

$$|a_n - a| \leq \frac{\varepsilon}{2} \text{ and } |b_n - b| \leq \frac{\varepsilon}{2}.$$

Since this holds for arbitrary ε we get the result.

For the second statement, we use that

$$|\lambda a_n - \lambda a| = |\lambda| \cdot |a_n - a|.$$

As $(a_n)_{n=1}^\infty$ is convergent we get the result.

For the third statement, note that

$$|a_n b_n - ab| \leq |a_n b_n - ab_n| + |ab_n - ab| = |b_n| \cdot |a_n - a| + |a| \cdot |b_n - b|.$$

Since $(b_n)_{n=1}^\infty$ is convergent it is bounded, so $\exists C \in \mathbb{R} \forall n \in \mathbb{N} : |b_n| \leq C$. Hence

$$|a_n b_n - ab| \leq C |a_n - a| + a |b_n - b|.$$

Therefore, the desired statement follows from the convergence of (a_n) and (b_n) .

For the last statement, we only need to prove that $\lim_{n \rightarrow \infty} \frac{1}{b_n} = \frac{1}{b}$, if $b \neq 0$. The more general statement then follows together with part (iii).

Let us assume w.l.o.g. (i.e., *without loss of generality*) that $b > 0$. (Otherwise, consider the sequence $(-b_n)$.) We obtain that $\exists N_0 \in \mathbb{N} \forall n \geq N_0 : b_n > \frac{b}{2}$. (Why?) Hence,

$$\left| \frac{1}{b_n} - \frac{1}{b} \right| = \left| \frac{b - b_n}{b \cdot b_n} \right| < \frac{2}{b^2} |b - b_n|$$

for $n \geq N_0$. We obtain the result from $b_n \rightarrow b$. \square

Remark 3.25. For a complex-valued sequence $(z_n)_{n \in \mathbb{N}}$, convergence to the complex number $z = x + iy \in \mathbb{C}$, i.e., $z_n \rightarrow z$, is equivalent to the convergence of the real and imaginary parts of z_n to x and y , respectively.

The first direction, i.e., that

$$\operatorname{Re} z_n \rightarrow x \quad \text{and} \quad \operatorname{Im} z_n \rightarrow y.$$

implies that $z_n \rightarrow x + iy$, follows from Theorem 3.24(i).

For the reverse statement, choose $\varepsilon > 0$ and n_0 such that $|z_n - z| < \varepsilon$ for $n \geq n_0$ holds. Then, for $n \geq n_0$ it holds that

$$|\operatorname{Re} z_n - \operatorname{Re} z| = |\operatorname{Re}(z_n - z)| \leq |z_n - z| < \varepsilon.$$

As this holds for all $\varepsilon > 0$, this proves the convergence of the real part of z_n to $\operatorname{Re} z = x$. The convergence of the imaginary part can be proven analogously. Consequently, we can say that

$$z_n \rightarrow z \iff \operatorname{Re} z_n \rightarrow \operatorname{Re} z \quad \text{and} \quad \operatorname{Im} z_n \rightarrow \operatorname{Im} z.$$

With the help of calculation rules and the knowledge about limits we can compute more sophisticated limits.

Example 3.26.

$$\lim_{n \rightarrow \infty} \frac{3 - 7\sqrt{n} + 42n}{n} = 3 \cdot \lim_{n \rightarrow \infty} \frac{1}{n} - 7 \lim_{n \rightarrow \infty} \frac{1}{\sqrt{n}} + \lim_{n \rightarrow \infty} 42 = 0 - 0 + 42 = 42.$$

Example 3.27. Let $k \in \mathbb{Z}$. Then, we have

$$\lim_{n \rightarrow \infty} \sqrt[n]{n^k} = 1.$$

Proof. For $k = 0$ this is obvious (and for $k = 1$ we have shown that already in Example 3.22). Moreover, we have for arbitrary $k \in \mathbb{Z}$ that

$$\lim_{n \rightarrow \infty} \sqrt[n]{n^{k+1}} = \lim_{n \rightarrow \infty} \sqrt[n]{n^k n} = \lim_{n \rightarrow \infty} \sqrt[n]{n^k} \lim_{n \rightarrow \infty} \sqrt[n]{n} = \lim_{n \rightarrow \infty} \sqrt[n]{n^k} \cdot 1 = \lim_{n \rightarrow \infty} \sqrt[n]{n^k},$$

as well as

$$\lim_{n \rightarrow \infty} \sqrt[n]{n^{k-1}} = \lim_{n \rightarrow \infty} \sqrt[n]{n^k} \lim_{n \rightarrow \infty} \frac{1}{\sqrt[n]{n}} = \lim_{n \rightarrow \infty} \sqrt[n]{n^k} \cdot \frac{1}{\lim_{n \rightarrow \infty} \sqrt[n]{n}} = \lim_{n \rightarrow \infty} \sqrt[n]{n^k},$$

where we used Theorem 3.24(iv) for the second equality. By induction on k in both directions (with induction basis $k = 0$), we get that the limit is the same for all k , and therefore equals 1. \square

The next result gives another tool for the calculation of difficult limits. This one is helpful when the sequence under consideration can be bounded from above and below by sequences that converge to the same limit.

Theorem 3.28 (Sandwich rule). *Let $(a_n)_{n \in \mathbb{N}}$ and $(c_n)_{n \in \mathbb{N}}$ be convergent real-valued sequences and let $(b_n)_{n \in \mathbb{N}}$ be a sequence such that*

$$a_n \leq b_n \leq c_n \quad \text{for all } n \in \mathbb{N}.$$

If additionally,

$$a := \lim_{n \rightarrow \infty} a_n = \lim_{n \rightarrow \infty} c_n,$$

then $(b_n)_{n \in \mathbb{N}}$ is convergent with

$$\lim_{n \rightarrow \infty} b_n = a.$$

Proof. Let $\varepsilon > 0$ be arbitrary. We have a look at

$$|a - b_n| = \begin{cases} a - b_n, & \text{if } b_n \leq a \\ b_n - a, & \text{otherwise.} \end{cases}$$

Since $(a_n)_{n \in \mathbb{N}}$ and $(c_n)_{n \in \mathbb{N}}$ do both converge to a , we have that there is some n_0 such that $b_n - a \leq c_n - a \leq \varepsilon$ and $a - b_n \leq a - a_n \leq \varepsilon$ for all $n \geq n_0$. Thus, $|b_n - a| < \varepsilon$ for $n \geq n_0$. As this holds for all ε , we obtain $\lim_{n \rightarrow \infty} b_n = a$. \square

Remark 3.29. Note that, as we consider limits, the assumption $a_n \leq b_n \leq c_n$ in the sandwich rule is not essential for the first terms and may be replaced by $\exists N \in \mathbb{N} \ \forall n \geq N: a_n \leq b_n \leq c_n$.

The following example gives a prominent application of the sandwich rule.

Example 3.30. For $a, b \geq 0$ we have

$$\lim_{n \rightarrow \infty} \sqrt[n]{a^n + b^n} = \max\{a, b\}.$$

Proof. W.l.o.g. assume that $b \geq a$. We have that

$$b = \sqrt[n]{b^n} \leq \sqrt[n]{a^n + b^n} \leq \sqrt[n]{2b^n} = \sqrt[n]{2} \cdot b.$$

As $\sqrt[n]{2}$ goes to 1, the sandwich rule (with $a_n = b$ and $c_n = \sqrt[n]{2} \cdot b$) yields the result. \square

In other cases, we may even do not know the precise values of the terms of a sequence since the explicit or recursive formula for them is too complicated. Also in such cases we can possibly apply the sandwich rule to obtain the limit.

Example 3.31. Consider the sequence $b_n = \frac{(1+\sin(n))^n}{n^{2n}}$. Since $|\sin(x)| \leq 1$ for all $x \in \mathbb{R}$, we have $0 \leq \frac{(1+\sin(n))^n}{n^{2n}} \leq \frac{2^n}{n^{2n}} = \frac{1}{n}$ for all $n \in \mathbb{N}$. Using the sandwich rule and that the sequences on both sides are null sequences, we obtain $b_n \rightarrow 0$.

(Note that we did not even need the precise values of $\sin(n)$.)

We end this section with **calculation rules for definitely divergent sequences**:

$$\begin{aligned}
 a_n \rightarrow \infty, b_n \rightarrow \infty &\implies a_n + b_n \rightarrow \infty \text{ and } a_n \cdot b_n \rightarrow \infty \\
 a_n \rightarrow \infty, b_n \rightarrow b &\implies a_n + b_n \rightarrow \infty \\
 a_n \rightarrow \infty, \alpha \in \mathbb{R} &\implies \frac{\alpha}{a_n} \rightarrow 0 \\
 a_n \rightarrow \infty, \alpha > 0 &\implies \alpha \cdot a_n \rightarrow \infty \\
 a_n \rightarrow \infty, \alpha < 0 &\implies \alpha \cdot a_n \rightarrow -\infty
 \end{aligned}$$

(Verify yourself!)

If $a_n \rightarrow \infty$ and $b_n \rightarrow \infty$, no general rule can be given for $(a_n - b_n)$ and $\left(\frac{a_n}{b_n}\right)$. Therefore, also the limit of $(a_n b_n)$ for $a_n \rightarrow 0$ and $b_n \rightarrow \infty$ needs more care, and these limits do not have to exist nor be definitely divergent, consider e.g. $a_n = (-1)^n/n$ and $b_n = n$.

We will come back to the computation of such limits later.

3.3 Monotone sequences

We saw that it is essential to verify that sequences are convergent for applying the rules above. Here, we show that a large class of sequences, namely all monotone and bounded sequences, are convergent. This is an essential insight.

Since we want to assume that the terms of a sequence are (monotonically) *ordered*, we need an order, and therefore only work with real-valued sequences here.

Definition 3.32 (Monotone sequences). A real-valued sequence $(a_n)_{n \in \mathbb{N}}$ is called

- **increasing** if and only if

$$\forall n \in \mathbb{N} : a_{n+1} > a_n,$$

- **non-decreasing** if and only if

$$\forall n \in \mathbb{N} : a_{n+1} \geq a_n,$$

- **decreasing** if and only if

$$\forall n \in \mathbb{N} : a_{n+1} < a_n,$$

- **non-increasing** if and only if

$$\forall n \in \mathbb{N} : a_{n+1} \leq a_n.$$

Moreover, we say that a sequence is **monotone** if it is non-increasing or non-decreasing, and **strictly monotone** if it is either increasing or decreasing.

Note that a sequence that is non-decreasing and non-increasing at the same time, must be a *constant sequence*.

Many of the sequences, that we have discussed so far, were strictly monotone. This holds clearly for the sequences $(n)_{n \in \mathbb{N}}$, $(\frac{1}{n})_{n \in \mathbb{N}}$ or more general $(n^k)_{n \in \mathbb{N}}$ for $k \in \mathbb{Z} \setminus \{0\}$, as well as the sequence $(a^n)_{n \in \mathbb{N}}$ for $a \in (0, 1)$ or $a > 1$. However, for some sequences this is not so easy to see.

Example 3.33. Let us have a look at the sequence given by $a_n := \frac{1}{n^2 - n + 1}$. This is a decreasing sequence, since $a_{n+1} = \frac{1}{(n+1)^2 - (n+1) + 1} = \frac{1}{n^2 + 2n + 1 - n - 1 + 1} = \frac{1}{n^2 + n + 1} < \frac{1}{n^2 - n + 1} = a_n$.

In some cases, a helpful *trick* is to consider the quotients of consecutive terms of a sequence and show that they are bounded (from above or below) by one. That is, e.g., a sequence is **non-decreasing if**

$$\frac{a_{n+1}}{a_n} \geq 1 \quad \text{for all } n \in \mathbb{N},$$

and **non-increasing if**

$$\frac{a_{n+1}}{a_n} \leq 1 \quad \text{for all } n \in \mathbb{N}.$$

Example 3.34. One interesting example, that we will study more detailed soon, is the sequence given by

$$a_n := \left(1 + \frac{1}{n}\right)^n = \left(\frac{n+1}{n}\right)^n$$

We consider quotients of successive terms and observe that

$$\frac{a_{n+1}}{a_n} = \frac{\left(\frac{n+2}{n+1}\right)^{n+1}}{\left(\frac{n+1}{n}\right)^n} = \left(\frac{(n+2)n}{(n+1)^2}\right)^{n+1} \frac{n+1}{n} = \left(1 - \frac{1}{(n+1)^2}\right)^{n+1} \frac{n+1}{n}.$$

The Bernoulli inequality $(1+x)^{n+1} \geq 1 + (n+1)x$, with $x = -\frac{1}{(n+1)^2} \geq -1$, yields

$$\frac{a_{n+1}}{a_n} \geq \left(1 - \frac{1}{n+1}\right) \frac{n+1}{n} = 1.$$

Hence (a_n) is a non-decreasing sequence. (If we note that Bernoulli's inequality is a strict inequality for $x > -1$ with $x \neq 0$, we even obtain that (a_n) is increasing.)

The following result shows that *monotonicity* is a very helpful property as we only have to know if a sequence is bounded to verify whether it is convergent or not. Note that boundedness of a sequence is usually much easier to show.

Theorem 3.35 (Monotonicity principle). *Let $(a_n)_{n \in \mathbb{N}} \subset \mathbb{R}$ be a monotone sequence. Then,*

$$(a_n) \text{ is convergent} \iff (a_n) \text{ is bounded.}$$

Moreover, every monotone and unbounded sequence is definitely divergent.

In particular, if $(a_n)_{n \in \mathbb{N}}$ is non-decreasing (or increasing), then

$$\lim_{n \rightarrow \infty} a_n = \sup\{a_n : n \in \mathbb{N}\} =: \sup(a_n),$$

and if $(a_n)_{n \in \mathbb{N}}$ is non-increasing (or decreasing), then

$$\lim_{n \rightarrow \infty} a_n = \inf\{a_n : n \in \mathbb{N}\} =: \inf(a_n).$$

By the completeness axiom, supremum and infimum exist for every bounded subset of \mathbb{R} .

Proof. We know from Theorem 3.15 that convergent sequences are bounded, which proves the first direction of the statement.

For the second, let us consider the case where $(a_n)_{n \in \mathbb{N}}$ is non-decreasing. The other case, where $(a_n)_{n \in \mathbb{N}}$ is non-increasing, can be treated in the same way (replacing, in particular, sup by inf).

We now assume that $(a_n)_{n \in \mathbb{N}}$ is bounded, and prove that it converges to $a = \sup\{a_n : n \in \mathbb{N}\}$.

Since $(a_n)_{n \in \mathbb{N}}$ is bounded, we get that the range of $(a_n)_{n \in \mathbb{N}}$ is a bounded set, which implies that $a = \sup\{a_n : n \in \mathbb{N}\}$ exists. Let now $\varepsilon > 0$ be arbitrary but fixed. Since a is the supremum of the range of $(a_n)_{n \in \mathbb{N}}$ we get (by definition) that $a - \varepsilon$ is no upper bound for the sequence $(a_n)_{n \in \mathbb{N}}$. Thus, there exists an n_0 with $a_{n_0} > a - \varepsilon$. Since $(a_n)_{n \in \mathbb{N}}$ is non-decreasing, the same then holds for $n \geq n_0$. That is, $\exists n_0 \in \mathbb{N} \forall n \geq n_0 : a - \varepsilon < a_n$. In addition, we clearly have $a_n \leq a < a + \varepsilon$. Hence, $\exists n_0 \in \mathbb{N} \forall n \geq n_0 : |a_n - a| < \varepsilon$. As this holds for all $\varepsilon > 0$, we obtain that $(a_n)_{n \in \mathbb{N}}$ converges to a .

The statement for unbounded sequences can be proven similarly, and is left for the reader. \square

With this theorem we see that there are convergent sequences where we do not have to know the limit to verify that it exists. In some cases, we may even *define* numbers just as limits of specific sequences, because we do not have another (explicit) description. One typical example is **Euler's number**:

Example 3.36 (Euler number). Consider sequences $(a_n)_{n \in \mathbb{N}}, (b_n)_{n \in \mathbb{N}}$ which are defined by

$$a_n := \left(1 + \frac{1}{n}\right)^n = \left(\frac{n+1}{n}\right)^n \quad \text{and} \quad b_n := \left(1 + \frac{1}{n}\right)^{n+1} = \left(\frac{n+1}{n}\right)^{n+1}.$$

Clearly $a_n \leq b_n$. Note that we have considered the sequence (a_n) already in Example 3.34 and showed that it is non-decreasing (which also implies that $a_n \geq a_1 = 2$ for all $n \in \mathbb{N}$). It remains to bound (a_n) from above to show that it is convergent. For this, we show that (b_n) is bounded from above. Together with $a_n \leq b_n$ this implies also the boundedness of (a_n) . In fact, we show that (b_n) is non-increasing, which implies that $b_n \leq b_1 = 4$ for all $n \in \mathbb{N}$. Again we compute quotients of consecutive terms and obtain

$$\begin{aligned} \frac{b_n}{b_{n+1}} &= \frac{\left(\frac{n+1}{n}\right)^{n+1}}{\left(\frac{n+2}{n+1}\right)^{n+2}} = \left(\frac{(n+1)^2}{n(n+2)}\right)^{n+2} \frac{n}{n+1} \\ &= \left(1 + \frac{1}{n(n+2)}\right)^{n+2} \frac{n}{n+1} \geq \left(1 + \frac{1}{n}\right) \frac{n}{n+1} = 1, \end{aligned}$$

where we used Bernoulli's inequality $(1+x)^{n+2} \geq 1 + (n+2)x$, with $x = \frac{1}{n(n+2)} \geq -1$.

From this we have $\frac{b_{n+1}}{b_n} \leq 1$ and therefore that (b_n) is non-increasing. (In fact, it is decreasing.) All in all

$$2 = a_1 \leq a_2 \leq \dots \leq a_n \leq \dots \leq b_n \leq \dots \leq b_2 \leq b_1 = 4.$$

This shows that the limit of $(a_n)_{n \in \mathbb{N}}$ exists and equals $\sup(a_n)$, and we define this limit to be *Euler's number*. Moreover, the limit of (b_n) also exists, equals $\inf(b_n)$, and we show that this is also equal to e . For this consider

$$b_n - a_n = \left(1 + \frac{1}{n}\right)^n \left(1 + \frac{1}{n} - 1\right) = \frac{a_n}{n} \leq \frac{4}{n}.$$

This implies $a_n \leq b_n \leq a_n + \frac{4}{n}$ and the sandwich rule yields $\lim a_n = \lim b_n$. (One may also use that, with $c_n = (n+1)/n$, we have $b_n = a_n \cdot c_n$. Then, $c_n \rightarrow 1$ yields the result.)

Therefore, we have the following **characterizations for Euler's number**

$$e = \lim \left(1 + \frac{1}{n}\right)^n = \sup \left(1 + \frac{1}{n}\right)^n = \lim \left(1 + \frac{1}{n}\right)^{n+1} = \inf \left(1 + \frac{1}{n}\right)^{n+1},$$

and we have $e \approx 2.7182818284590452353\dots$. We will see in the following section, that e may also be defined by the infinite sum $\sum_{k=0}^{\infty} \frac{1}{k!}$.

This example shows, in particular, that it may happen that a sequence is converging but we do not know the precise value of its limit. In some cases, however, we are able to find the limit (or at least some possible candidates) by alternative considerations. For example, if the sequence $(a_n)_{n \in \mathbb{N}}$ is convergent, we have

$$\lim_{n \rightarrow \infty} a_n = \lim_{n \rightarrow \infty} a_{n+1} = \lim_{n \rightarrow \infty} a_{n+2} = \dots,$$

which may just be seen as ignoring the first terms of a sequence. Such equations are particularly useful for recursively defined sequences.

Example 3.37. Let $x > 0$. We consider the following recursively defined sequence:

$$a_1 > 0 \text{ is arbitrary} \quad \text{and} \quad a_{n+1} := \frac{1}{2} \left(a_n + \frac{x}{a_n} \right) \quad \text{for all } n \in \mathbb{N}.$$

It is obvious that (a_n) is a *positive* sequence, i.e., $a_n > 0$ for all $n \in \mathbb{N}$. We now show that (a_n) is decreasing for $n \geq 2$, which implies, by the monotonicity principle, that (a_n) is convergent. (Note that, since we are only interested in limits, it is always ok to prove things only for large n .) We obtain

$$a_{n+1} = \frac{1}{2} \left(a_n + \frac{x}{a_n} \right) \leq a_n \iff a_n + \frac{x}{a_n} \leq 2a_n \iff x \leq a_n^2$$

for all $n \in \mathbb{N}$. This is equivalent to $a_n \geq \sqrt{x}$, because the a_n are positive. We now show that for all $n \geq 2$ it holds that $a_n \geq \sqrt{x}$, or equivalently: $a_{n+1} \geq \sqrt{x}$ for all $n \in \mathbb{N}$:

$$\begin{aligned} a_{n+1} \geq \sqrt{x} &\iff \frac{1}{2} \left(a_n + \frac{x}{a_n} \right) \geq \sqrt{x} \iff a_n^2 + x \geq 2\sqrt{x}a_n \\ &\iff a_n^2 - 2\sqrt{x}a_n + x \geq 0 \\ &\iff (a_n - \sqrt{x})^2 \geq 0. \end{aligned}$$

Since the last statement is clearly true, we finally obtain that $(a_n)_{n \in \mathbb{N}}$ is decreasing for $n \geq 2$, and therefore convergent.

Moreover, we can determine the limit by exploiting the fact that the limits of $(a_n)_{n \in \mathbb{N}}$ and $(a_{n+1})_{n \in \mathbb{N}}$ are the same. Let $a := \lim(a_n) = \lim(a_{n+1})$. Then,

$$a = \lim_{n \rightarrow \infty} a_{n+1} = \lim_{n \rightarrow \infty} \frac{1}{2} \left(a_n + \frac{x}{a_n} \right) = \frac{1}{2} \left(\lim_{n \rightarrow \infty} a_n + \frac{x}{\lim_{n \rightarrow \infty} a_n} \right) = \frac{1}{2} \left(a + \frac{x}{a} \right)$$

With the same calculations as above we see that this equation can only be fulfilled if

$$a = \frac{1}{2} \left(a + \frac{x}{a} \right) \iff a^2 = x \iff a = \pm\sqrt{x}.$$

Since (a_n) is non-negative, $-\sqrt{x}$ cannot be the limit of the sequence and therefore \sqrt{x} is the only possibility, i.e.,

$$\lim_{n \rightarrow \infty} a_n = \sqrt{x}.$$

(Note that the limit would be $-\sqrt{x}$ if the 'starting value' a_1 would be negative, Check yourself!) □

3.4 Subsequences and accumulation points

The concepts of the last sections deal with sequences that converge or, in other words, concentrate around a single point. In some cases, however, also divergent sequences have only some points of interest for very large n . An obvious example is $((-1)^n)_{n \in \mathbb{N}}$.

In this section, we want to formalize the idea of sequences having *more than one limit*, i.e., points where the sequences accumulates for large n . We will show the (to some extent surprising) fact that every bounded sequence has such an accumulation point. We will also specify special convergent subsequences, which leads to the so-called *limit superior* and *limit inferior* of a sequence.

Definition 3.38. Let (n_1, n_2, n_3, \dots) be an increasing sequence of natural numbers and $(a_n)_{n \in \mathbb{N}}$ be a sequence. Then, we call

$$(a_{n_k})_{k \in \mathbb{N}} = (a_{n_k})_{k=1}^{\infty} = (a_{n_1}, a_{n_2}, \dots)$$

a **subsequence** of $(a_n)_{n \in \mathbb{N}}$.

Example 3.39. Consider the sequence given by $b_n = (-1)^n(1 - \frac{1}{n})$. This is not convergent, as it "jumps" between 'close to 1' and 'close to -1'. However, if we take the sequence of even natural numbers, i.e., $(n_1, n_2, n_3, \dots) = (2, 4, 6, \dots)$, then the terms of the subsequence $(b_{n_k})_{k \in \mathbb{N}} = (b_{2n})_{n \in \mathbb{N}} = (b_2, b_4, b_6, \dots)$ are of the form $b_{n_k} = 1 - \frac{1}{2k}$. Hence, $(b_{n_k})_{k \in \mathbb{N}}$ is a convergent sequence, and hence, a convergent subsequence of $(b_n)_{n \in \mathbb{N}}$.

Definition 3.40. Let $(a_n)_{n \in \mathbb{N}} \subset \mathbb{C}$ be a sequence.

We call $a \in \mathbb{C}$ an **accumulation point** of $(a_n)_{n \in \mathbb{N}}$ if there exists a subsequence $(a_{n_k})_{k \in \mathbb{N}}$ of $(a_n)_{n \in \mathbb{N}}$ with

$$\lim_{k \rightarrow \infty} a_{n_k} = a.$$

Equivalently, we may use the definitions

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

or

$$\forall \varepsilon > 0 \ \forall n_0 \in \mathbb{N} \ \exists n \geq n_0 : a_n \in U_\varepsilon(a),$$

or

$$\forall \varepsilon > 0 : \#\{n \in \mathbb{N} : a_n \in U_\varepsilon(a)\} = \infty,$$

i.e., there are infinitely terms of (a_n) in every neighborhood of a .

(Note the interchanged quantifiers compared to the definition of convergence.)

Example 3.41. Considering the example from above, i.e., $b_n = (-1)^n(1 - \frac{1}{n})$, we see that $b_{2n} \rightarrow 1$. Hence, 1 is an accumulation point of (b_n) . Moreover, $b_{2n+1} = -(1 - \frac{1}{2n+1}) \rightarrow -1$ also defines a convergent subsequence, and -1 is therefore also an accumulation point. It is not hard to see that there is no other possible limit of a subsequence.

Next we want to show that each bounded sequence has at least one convergent subsequence. In particular, we show (in the proof) that every sequence contains either a non-increasing or a non-decreasing subsequence (or both), which together with the boundedness implies its convergence, see Theorem 3.35. Note that every subsequence of a bounded sequence is also bounded.

This result bears the names of Bolzano and Weierstrass and is an important technical tool for proofs in many areas of analysis.

Theorem 3.42 (Bolzano-Weierstrass). *Let $(a_n)_{n \in \mathbb{N}} \subset \mathbb{C}$ be a bounded sequence.*

Then, $(a_n)_{n \in \mathbb{N}}$ has at least one convergent subsequence.

That is, $(a_n)_{n \in \mathbb{N}}$ has at least one accumulation point.

Proof. We present a proof only in the real case. The complex case works similar.

We call $m \in \mathbb{N}$ a peak of $(a_n)_{n=1}^{\infty}$ if $\forall n > m$ we have that $a_n < a_m$. If there exist infinitely many peaks of $(a_n)_{n=1}^{\infty}$, denoted by n_1, n_2, n_3, \dots then the sequence $(a_{n_1}, a_{n_2}, a_{n_3}, \dots)$ is decreasing and bounded. Hence it is convergent as we know from before.

If there are at most $l \in \mathbb{N}$ peaks, then we choose n_1 bigger than the largest peak, or, if there are no peaks, then we choose $n_1 = 1$. In both cases n_1 is no peak, hence there exists $n_2 > n_1$ such that $a_{n_2} \geq a_{n_1}$. Furthermore n_2 is no peak, which implies that there exists a $n_3 > n_2$ such that $a_{n_3} \geq a_{n_2}$. If we repeat this process we end up with a non-decreasing subsequence of $(a_n)_{n=1}^{\infty}$, which is also bounded. Therefore this subsequence converges. \square

We can also give another proof of this statement, which is a bit more of geometric flavour.

Alternative proof. Every sequence $(a_n)_{n \in \mathbb{N}}$ has infinitely many (not necessarily different) terms. If infinitely many are equal, we are done, because a list of these terms is a convergent subsequence.

If not, assume w.l.o.g. that $0 \leq a_n \leq 1$ for all n . Every other bounded sequence can be treated the same way. Now, split the interval $[0, 1]$ into the halves $[0, \frac{1}{2}]$ and $[\frac{1}{2}, 1]$. As there are infinitely many points in $[0, 1]$, at least one of the halves must also contain infinitely many points. Now, split up this one, and so on. With this procedure we come arbitrary close to a point a , whose neighborhoods –by construction– all contain infinitely many points. This point is therefore an accumulation point. (We just note here that a is defined as the intersection of infinitely many nested intervals. It follows from the sandwich rule that this intersection is not empty.) \square

Remark 3.43. The Bolzano-Weierstrass theorem is also true for sequences in much more general (e.g. multidimensional) situations.

Example 3.44. Note that every bounded sequence has at least one convergent subsequence but not every sequence with a convergent subsequence is bounded. E.g., consider the sequence (a_n) with $a_{2n} = n$ and $a_{2n-1} = 0$. Clearly, (a_1, a_3, a_5, \dots) is a null sequence, but there is no upper bound for this sequence.

Inspired by the proof of the Bolzano-Weierstrass theorem, we will define two special accumulation points of a sequence, i.e., the *smallest* and the *largest* accumulation point. They can be seen as the *limiting bounds* on the sequence, i.e., every limit of a convergent subsequence must lie between them.

Definition 3.45. Let $(a_n)_{n \in \mathbb{N}} \subset \mathbb{R}$ be a real-valued sequence.

We define the **limes inferior** of $(a_n)_{n \in \mathbb{N}}$ by

$$\liminf_{n \rightarrow \infty} a_n := \lim_{n \rightarrow \infty} \left(\inf_{k \geq n} a_k \right),$$

and the **limes superior** by

$$\limsup_{n \rightarrow \infty} a_n := \lim_{n \rightarrow \infty} \left(\sup_{k \geq n} a_k \right).$$

Note that $(\inf_{k \geq n} a_k)_{n \in \mathbb{N}}$ is a non-decreasing sequence, since we take the infimum over a smaller set if n increases. Hence, together with Theorem 3.35, we can alternatively define the limes inferior by

$$\liminf_{n \rightarrow \infty} a_n = \sup_{n \in \mathbb{N}} \inf_{k \geq n} a_k.$$

Similarly, we obtain for the limes superior that

$$\limsup_{n \rightarrow \infty} a_n = \inf_{n \in \mathbb{N}} \sup_{k \geq n} a_k.$$

Since the infimum and supremum exist of arbitrary bounded sets, we obtain that the limit inferior and limit superior also exist for arbitrary bounded sequences. Moreover, if the sequence is unbounded, then the corresponding 'inner' infimum or supremum (or both) are infinity. So, if we allow limes inferior and limes superior to have also the values $-\infty$ and ∞ , then they **exist for arbitrary sequences**. That is, to every real-valued sequence (a_n) , we may assign the unique values $\liminf(a_n) \in \mathbb{R} \cup \{-\infty, \infty\}$ and $\limsup(a_n) \in \mathbb{R} \cup \{-\infty, \infty\}$.

Example 3.46. Consider the sequence $(a_n)_{n \in \mathbb{N}}$ with $a_n = (-1)^n$.

For all $n \in \mathbb{N}$ we see that $\inf_{k \geq n} a_k = -1$, which shows $\liminf(a_n) = -1$.

Moreover, $\sup_{k \geq n} a_k = 1$ for all n , which gives $\limsup(a_n) = \lim_{n \rightarrow \infty} \sup_{k \geq n} a_k = 1$.

Example 3.47. Consider the sequence $(a_n)_{n \in \mathbb{N}}$ with $a_n = (-1)^n + \frac{1}{n}$.

For $n \in \mathbb{N}$ we see that $\inf_{k \geq n} a_k = \inf\{(-1)^k + \frac{1}{k} : k \geq n\} = -1$ (using the Archimedean property), which shows $\liminf(a_n) = -1$.

Moreover, $\sup_{k \geq n} a_k = 1 + \frac{1}{n}$ for even n and $\sup_{k \geq n} a_k = 1 + \frac{1}{n+1}$ for odd n . (For odd n we have $\sup_{k \geq n} a_k = a_{n+1}$.) This gives $\limsup(a_n) = \lim_{n \rightarrow \infty} \sup_{k \geq n} a_k = 1$.

One may also consider the subsequences (a_{2n}) and (a_{2n+1}) of $(a_n)_{n \in \mathbb{N}}$ with $a_{2n} = 1 + \frac{1}{2n} \rightarrow 1$ and $a_{2n+1} = -1 + \frac{1}{2n+1} \rightarrow -1$, to see that -1 and 1 are accumulation points.

Example 3.48. Consider the sequence $(a_n)_{n \in \mathbb{N}}$ with $a_n = 3 + (-1)^n(1 + \frac{42}{n})$.

We have that

$$\begin{aligned}\sup_{k \geq n} a_k &= \sup \left\{ 3 + (-1)^n \left(1 + \frac{42}{n} \right), 3 + (-1)^{n+1} \left(1 + \frac{42}{n+1} \right), \dots \right\} \\ &= \begin{cases} 4 + \frac{42}{n}, & n \text{ even}, \\ 4 + \frac{42}{n+1}, & n \text{ odd}. \end{cases}\end{aligned}$$

Hence, $\limsup a_n = 4$. In the same way, we obtain $\liminf a_n = 2$.

Example 3.49. Consider the sequence $(a_n)_{n \in \mathbb{N}}$ with $a_n = 2^n + (-2)^n$. The terms of (a_n) equal either 2^{n+1} (for even n) or 0 (for odd n). Therefore, $\liminf(a_n) = 0$ and $\limsup(a_n) = \infty$.

The **limit inferior and limit superior are accumulation points** of (a_n) , if the sequence is bounded. We show this for the limit inferior $b := \liminf(a_n)$.

With $b_n := \inf_{k \geq n} a_k$ we have $b = \lim(b_n)$. (Note that (b_n) is in general no subsequence of (a_n) .) Now let $n_1 \in \mathbb{N}$ be such that $b_1 > a_{n_1} - \frac{1}{2}$. Such an n_1 exists by the definition of the infimum. Next, let $n_2 > n_1$ be such that $b_{n_1+1} > a_{n_2} - \frac{1}{2^2}$, and so on. That is, we obtain an increasing sequence $(n_k)_{k \in \mathbb{N}}$ of natural numbers with $b_{n_k+1} > a_{n_{k+1}} - \frac{1}{2^k}$. In addition, we have by definition that $b_{n_k+1} \leq a_{n_{k+1}}$. We obtain that $|b_{n_k+1} - a_{n_{k+1}}| < \frac{1}{2^k}$, which implies that $(b_{n_k+1} - a_{n_{k+1}})_{k \in \mathbb{N}}$ is a null sequence. Hence, $\lim_{k \rightarrow \infty} b_{n_k+1} = \lim_{k \rightarrow \infty} a_{n_{k+1}}$. Since all subsequences of (b_n) converge to the same limit, we obtain $\lim_{k \rightarrow \infty} a_{n_{k+1}} = \lim(b_n) = b$, i.e., b is an accumulation point.

Moreover, $\liminf(a_n)$ and $\limsup(a_n)$ are **the smallest and largest accumulation point** of (a_n) , respectively. To see this, let $a \in \mathbb{R}$ be any accumulation point of (a_n) , i.e., there exists a subsequence $(a_{n_k})_{k \in \mathbb{N}}$ with $a_{n_k} \rightarrow a$. Then we have the bounds

$$\liminf_{n \rightarrow \infty} a_n \leq a \leq \limsup_{n \rightarrow \infty} a_n.$$

Proof. If $a > \limsup(a_n)$, then there is some $\varepsilon > 0$ and infinitely many terms of (a_n) that are larger than $\limsup(a_n) + \varepsilon$. Hence, $\sup_{k \geq n} a_k \geq \limsup(a_n) + \varepsilon$ for all n : A contradiction to the definition of the limit. \square

Although all accumulation points of (a_n) are bounded from below and above by $\liminf(a_n)$ and $\limsup(a_n)$, respectively, this clearly does not need to hold for all (large enough) terms of the sequence. It may even happen that all elements of a sequence lie outside of the interval $[\liminf(a_n), \limsup(a_n)]$. Consider, e.g., $a_n = (-1)^n(1 + \frac{1}{n}) \notin [-1, 1]$ with $\liminf(a_n) = -1$ and $\limsup(a_n) = 1$.

Moreover, if we are (only) interested in the limiting behavior of the sequence (a_n) , then the trivial bound

$$\inf\{a_n : n \in \mathbb{N}\} \leq a_N \leq \sup\{a_n : n \in \mathbb{N}\},$$

which holds for all $N \in \mathbb{N}$, is useless in general.

We can use the limit inferior and superior to **bound the elements of a sequence for large N** : As \liminf and \limsup are the smallest and largest accumulation point of a sequence, respectively, we obtain have that all elements of (a_n) with large enough n are at least close to the interval $[\liminf(a_n), \limsup(a_n)]$. That is, for all $\varepsilon > 0$ there is some $n_0 \in \mathbb{N}$ such that

$$\liminf(a_n) - \varepsilon \leq a_N \leq \limsup(a_n) + \varepsilon$$

for all $N \geq n_0$. (Verify this!)

We finally show that the limit inferior and limit superior are indeed generalizations of the concept of a limit. Clearly, it is more generally applicable, as \liminf and \limsup are well-defined for arbitrary sequences. The next result shows that, for convergent sequences, all these values are just the same. This follows from the considerations above, and the fact that every subsequence of a convergent sequence converges to the same limit.

Lemma 3.50. *A sequence $(a_n)_{n \in \mathbb{N}} \subset \mathbb{R}$ is convergent (or definitely divergent) if and only if*

$$\liminf_{n \rightarrow \infty} a_n = \limsup_{n \rightarrow \infty} a_n$$

In this case,

$$\lim_{n \rightarrow \infty} a_n = \liminf_{n \rightarrow \infty} a_n \quad \left(= \limsup_{n \rightarrow \infty} a_n\right).$$

This means that **a sequence is convergent if and only if the sequence is bounded and has exactly one accumulation point**.

Remark 3.51 (Complex sequences). Note that the \liminf and \limsup cannot be defined for complex-valued sequences, because we do not have an order on \mathbb{C} , and hence no supremum or infimum. However, it is still true that a complex-valued sequence is convergent if and only if it is bounded and has exactly one accumulation point. For a proof, we may consider \liminf and \limsup of the real and imaginary parts separately.

We finish this section with an extreme example.

Example 3.52. Consider the sequence $(a_n)_{n \in \mathbb{N}}$, which is a list of all rational numbers in $[0, 1]$. We already showed that the rational numbers are dense in \mathbb{R} . Thus every $x \in [0, 1]$ is an accumulation point of $(a_n)_{n \in \mathbb{N}}$. In other words, the set of accumulation points is uncountable and therefore in some sense “larger” than the set of the sequence elements.

3.5 Cauchy criterion

In this section we introduce the *Cauchy criterion* for proving convergence of a sequence. This is, similarly to the monotonicity principle (Theorem 3.35), an important tool to verify that a sequence is convergent, without knowing its limit. The Cauchy criterion will be the dominant technique for proofs of convergence when it comes to higher mathematics, including sequences of more general objects.

The central object is a Cauchy sequence.

Definition 3.53. A sequence $(a_n)_{n \in \mathbb{N}} \subset \mathbb{C}$ is called a **Cauchy sequence** if

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

That is, the terms of a Cauchy sequence are *pairwise* close to each other for large n .

Compare this definition with the definition of convergence in order to gain better understanding.

Example 3.54. The sequence $(a_n)_{n \in \mathbb{N}}$ with $a_n = \frac{1}{n}$ satisfies $|a_n - a_m| = \frac{1}{m} - \frac{1}{n} < \varepsilon$ for $n \geq m > \frac{1}{\varepsilon} =: n_0$. Hence, (a_n) is a Cauchy sequence.

Remark 3.55. For a complex-valued sequence (z_n) with $z_n = x_n + iy_n$ and $(x_n)_{n \in \mathbb{N}}, (y_n)_{n \in \mathbb{N}} \subset \mathbb{R}$ it holds:

$$(z_n) \text{ is a Cauchy sequence} \iff (x_n) \text{ and } (y_n) \text{ are Cauchy sequences.}$$

This follows from

$$\max\{|x_n - x_m|, |y_n - y_m|\} \leq \sqrt{(x_n - x_m)^2 + (y_n - y_m)^2} = |z_n - z_m|.$$

We now prove the important property that every convergent sequence is a Cauchy sequence, and vice versa.

Theorem 3.56 (Cauchy criterion). *Let $(a_n)_{n \in \mathbb{N}} \subset \mathbb{C}$ be a sequence. Then,*

$$(a_n)_{n \in \mathbb{N}} \text{ is convergent} \iff (a_n)_{n \in \mathbb{N}} \text{ is a Cauchy sequence.}$$

Proof. First we show “ (a_n) is convergent \implies (a_n) is Cauchy sequence”:

For this, let $a \in \mathbb{C}$ be the limit of $(a_n)_{n \in \mathbb{N}}$ and $\varepsilon > 0$. By the triangle inequality, we have

$$|a_n - a_m| = |a_n - a + a - a_m| \leq |a_n - a| + |a - a_m|.$$

Since $(a_n)_{n \in \mathbb{N}}$ is convergent we can find $n_0 \in \mathbb{N}$ such that for $n, m \geq n_0$ we have $|a_n - a| \leq \frac{\varepsilon}{2}$ and $|a_m - a| \leq \frac{\varepsilon}{2}$. Hence,

$$|a_n - a_m| \leq |a_n - a| + |a - a_m| \leq \varepsilon.$$

Since this holds for all $\varepsilon > 0$, we get that (a_n) is a Cauchy sequence.

We now show the other direction “ $(a_n)_{n \in \mathbb{N}}$ is Cauchy sequence $\implies (a_n)_{n \in \mathbb{N}}$ is convergent”:

First, we choose $\varepsilon = 1$ in the definition of a Cauchy sequence, and obtain some n_0 such that $|a_n - a_m| < 1$ for all $m, n \geq n_0$. Moreover, the triangle inequality implies

$$|a_n| \leq |a_n - a_{n_0}| + |a_{n_0}| \leq 1 + |a_{n_0}|$$

for all $n \geq n_0$. With $C := \max\{|a_1|, |a_2|, \dots, |a_{n_0-1}|, 1 + |a_{n_0}|\}$ we have

$$\forall n \in \mathbb{N}: |a_n| \leq C$$

which makes $(a_n)_{n \in \mathbb{N}}$ a bounded sequence. The Bolzano-Weierstrass theorem implies that $(a_n)_{n \in \mathbb{N}}$ has at least one accumulation point $a \in \mathbb{C}$. Thus there exists a subsequence $(a_{n_k})_{k \in \mathbb{N}}$ (of $(a_n)_{n \in \mathbb{N}}$) such that

$$\lim_{k \rightarrow \infty} a_{n_k} = a.$$

Using the triangle inequality again we have

$$|a_n - a| \leq |a_n - a_{n_k}| + |a - a_{n_k}|.$$

Let $\varepsilon > 0$. By the convergence of (a_{n_k}) we obtain that $|a - a_{n_k}| < \frac{\varepsilon}{2}$ for large enough k , i.e., for large enough n_k . Moreover, since (a_n) is a Cauchy sequence, $|a_n - a_{n_k}| < \frac{\varepsilon}{2}$ for n and n_k large enough. (Formally, this shows that, for all ε , there exist n_0, k_0 such that for all $n \geq n_0, k \geq k_0$ we have $|a_n - a| < \varepsilon$. But since the statement does not depend on k , we just omit this part.) That is, for arbitrary $\varepsilon > 0$ and n large enough, we have

$$|a_n - a| < \varepsilon,$$

and this shows the convergence of $(a_n)_{n \in \mathbb{N}}$.

□

Remark 3.57. One might show for every convergent sequence discussed so far, that it is a Cauchy sequence. The proof would follow directly the first part of the proof above, and one would not learn much by these computations. (You might still try it on your own!) However, verifying that a sequence is a Cauchy sequence is often much easier, as we will see later on.

Example 3.58. Note that it is not enough that neighboring elements of a sequence become arbitrarily close. For example, the sequence $(a_n)_{n \in \mathbb{N}}$ with $a_n = \sqrt{n}$, which is clearly not convergent, satisfies

$$\begin{aligned} |a_{n+K} - a_n| &= \sqrt{n+K} - \sqrt{n} = \frac{(\sqrt{n+K} - \sqrt{n})(\sqrt{n+K} + \sqrt{n})}{(\sqrt{n+K} + \sqrt{n})} \\ &= \frac{K}{\sqrt{n+K} + \sqrt{n}} < \frac{K}{2\sqrt{n}} \rightarrow 0 \end{aligned}$$

for every fixed $K \in \mathbb{N}$. Hence, ‘terms at fixed distance’ become arbitrarily close together. However, we have, e.g., $|a_{4n} - a_n| = \sqrt{n} \rightarrow \infty$. So, there is no n_0 such that $|a_m - a_n| < 1$ for all $m, n \geq n_0$.

3.6 Series

Now we want to discuss special sequences of the form

$$s_n = \sum_{k=1}^n a_k,$$

where $(a_n)_{n \in \mathbb{N}}$ is a given real- or complex-valued sequence. That is, $s_1 = a_1$, $s_2 = a_1 + a_2$, $s_3 = a_1 + a_2 + a_3$ and so on. The sum of all terms of the sequence $(a_n)_{n \in \mathbb{N}}$, i.e., the limit of the sequence $(s_n)_{n \in \mathbb{N}}$, is one of the main motivations for considering limits at all, and some interesting phenomena appear when it comes to the question if such limits exist.

Let us again start with the necessary definitions.

Definition 3.59. Let $(a_n)_{n \in \mathbb{N}}$ be a sequence and

$$s_n := \sum_{k=1}^n a_k.$$

Then, we call s_n the **n -th partial sum** of the **(infinite) series**

$$\sum_{k=1}^{\infty} a_k \quad \text{or just} \quad \sum a_k.$$

If the sequence of partial sums (s_n) converges to some $s \in \mathbb{C}$, i.e., $s_n \rightarrow s$, then we say that the **series converges** or that the **sequence $(a_n)_{n \in \mathbb{N}}$ is summable**, call s the **sum** of the series, and write

$$\sum_{k=1}^{\infty} a_k := \lim_{n \rightarrow \infty} \sum_{k=1}^n a_k = \lim_{n \rightarrow \infty} s_n = s.$$

If $s_n \rightarrow \pm\infty$ we also write $\sum_{k=1}^{\infty} a_k = \pm\infty$, and say that the series is **definitely divergent**.

Otherwise we call the series $\sum_{k=1}^{\infty} a_k$ **divergent** and the sequence $(a_n)_{n \in \mathbb{N}}$ **not summable**.

Note that *series* is just another word for an infinite sum of elements of a sequence. Moreover, the notation $\sum_{k=1}^{\infty} a_k$ should be understood as a formal symbol for the limit: It might be a number or $\pm\infty$, but it might also not exist (as a number).

One may clearly generalize this concept by considering

$$s_n = \sum_{k=k_0}^n a_k$$

for some $k_0 \in \mathbb{N}$ and $n \geq k_0$, and a sequence $(a_k)_{k=k_0}^{\infty}$. A typical case is e.g. $\sum_{k=0}^{\infty} a_k$. The corresponding definitions should be clear.

Note that we should use the shorter notation $\sum a_k$ only if we exactly know what k_0 is. If different indices appear in a calculation (as almost always), then it is necessary to be more precise to keep track of the summation limits.

The definition above states that a series converges if and only if the sequence $(s_n)_{n=1}^{\infty}$ of partial sums converges. This implies that we can use the results from the previous section to analyze series. Moreover, we will see that there are even more tools for working with series. But first let us see some examples, that will be **essential** for the upcoming considerations.

The first example is one of the most well-known and used infinite sums. Although it is usually considered only for real $q \in (-1, 1)$, we see that it also holds for complex bases.

Example 3.60 (Geometric series). Let $q \in \mathbb{C}$ with $|q| < 1$. Then we have that

$$\sum_{k=0}^{\infty} q^k = \frac{1}{1-q} \quad \text{and} \quad \sum_{k=1}^{\infty} q^k = \frac{q}{1-q}.$$

Both follow from the even more general formula

$$\sum_{k=k_0}^n q^k = \frac{q^{k_0} - q^{n+1}}{1-q},$$

which holds for all $n \geq k_0$ and $q \neq 1$. (Note that the last formula does not contain a limit.)

Proof. We only prove the result for $k_0 = 0$, and leave the rest for the reader.

Let $s_n := \sum_{k=0}^n q^k$ and consider the equation

$$\begin{aligned} (1-q)s_n &= \sum_{k=0}^n q^k - q \cdot \sum_{k=0}^n q^k = \sum_{k=0}^n q^k - \sum_{k=0}^n q^{k+1} = \sum_{k=0}^n q^k - \sum_{k=1}^{n+1} q^k = q^0 - q^{n+1} \\ &= 1 - q^{n+1}. \end{aligned}$$

From the next to the last equation we used the fact that the terms q^k for $k \in \{1, 2, \dots, n\}$ appear in both sums (and the second sum is subtracted from the first). Thus, the only terms that remain are q^0 and q^{n+1} (and the second gets a minus in front). Such arguments, i.e., that many (or all) terms of a series cancel each other out, are called **telescoping tricks** and sums of this form are called **telescoping sums**. We will come back to this kind of series later.

From the above equation we obtain

$$s_n = \frac{1 - q^{n+1}}{1 - q}$$

for all $q \neq 1$. Using this representation of the partial sums we can compute its limit easily. Note that $\frac{1}{1-q}$ is a constant factor and $(q^n)_{n \in \mathbb{N}}$ with $|q| < 1$ is a null sequence, see Example 3.20. Hence

$$\sum_{k=0}^{\infty} q^k = \lim_{n \rightarrow \infty} s_n = \frac{1}{1-q} \lim_{n \rightarrow \infty} (1 - q^{n+1}) = \frac{1}{1-q} (1 - \lim_{n \rightarrow \infty} q^{n+1}) = \frac{1}{1-q}.$$

□

Example 3.61. If we set $q = \frac{1}{2}$ we get e.g. that

$$\sum_{n=0}^{\infty} \frac{1}{2^n} = 2 \quad \text{and} \quad \sum_{n=1}^{\infty} \frac{1}{2^n} = 1$$

Remark 3.62. Note that the *telescoping trick* from above also works for $|q| > 1$ (but not if $q = 1$). It then follows from the explicit formula $s_n = \frac{q^{n+1}-1}{q-1}$, where we just multiplied numerator and denominator by -1 , that (s_n) is unbounded, and therefore divergent. If $q > 1$ (and, in particular, a real number) we see that (s_n) tends to infinity, while the limits simply do not exist for $q \leq -1$. In general, the case $|q| = 1$ needs more care. We will see in Example 3.72 that $\sum_{k=0}^{\infty} q^k$ is divergent for every $|q| \geq 1$, and definitely divergent only for $q \geq 1$.

The next example is quite obvious; still, we need to discuss it.

Example 3.63 (Finite series). Let $(a_n)_{n=1}^{\infty}$ be a *finite sequence*, i.e., $\exists K \in \mathbb{N} \forall k > K: a_k = 0$. Then, we clearly have

$$\sum_{k=1}^{\infty} a_k = \sum_{k=1}^K a_k = a_1 + a_2 + \cdots + a_K.$$

Now we will see that **not all convergent sequences lead to convergent series**. Moreover, this is the most important prototype of a divergent sequence, as it is “almost summable”. We will see later what this means.

Example 3.64 (Harmonic series). Consider the sequence given by $a_n = \frac{1}{n}$. Then, the corresponding series satisfies

$$\sum_{n=1}^{\infty} \frac{1}{n} = \infty,$$

i.e., it is *definitely divergent*. This series is called **harmonic series**.

However, we will see later that $\sum n^{-\alpha}$ is convergent if $\alpha > 1$.

Proof. We show that the sequence of partial sums is bounded from below by a divergent sequence. For this, we successively group the terms in the partial sums, and bound this group by the number of elements times its smallest member:

$$\begin{aligned} s_{2^n} &= 1 + \frac{1}{2} + \left(\frac{1}{3} + \frac{1}{4} \right) + \left(\frac{1}{5} + \frac{1}{6} + \frac{1}{7} + \frac{1}{8} \right) + \cdots + \left(\frac{1}{2^{n-1}+1} + \frac{1}{2^{n-1}+2} + \cdots + \frac{1}{2^n} \right) \\ &\geq 1 + \frac{1}{2} + 2 \cdot \frac{1}{4} + 4 \cdot \frac{1}{8} + \cdots + 2^{n-1} \cdot \frac{1}{2^n} \\ &= 1 + \frac{1}{2} + \frac{1}{2} + \frac{1}{2} + \cdots + \frac{1}{2} \\ &= 1 + \frac{n}{2}. \end{aligned}$$

This means that, with N such that $2^n \leq N < 2^{n+1} \iff n \leq \log_2(N) < n+1$, we obtain $s_N \geq s_{2^n} \geq 1 + n/2 \geq \frac{1+\log_2(N)}{2}$. We obtain $s_n \rightarrow \infty$. □

Example 3.65. We want to discuss the **telescoping trick** once more. This is sometimes a very powerful tool to obtain the precise value of apparently complicated series. Let us therefore prove

$$\sum_{k=1}^{\infty} \frac{1}{k(k+1)} = 1.$$

It is not clear yet that the series on the left hand side converges, and its precise value is also far from obvious. But one might notice that the terms can be expanded to

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

With this, we obtain that

$$s_n = \sum_{k=1}^n \left(\frac{1}{k} - \frac{1}{k+1} \right) = \sum_{k=1}^n \frac{1}{k} - \sum_{k=2}^{n+1} \frac{1}{k} = 1 - \frac{1}{n+1} \rightarrow 1.$$

The above telescoping trick may be generalized in the following way:

Assume that a_k can be written as $a_k = b_k - b_{k+\ell}$ for some $\ell \in \mathbb{N}$. Then,

$$\sum_{k=1}^{\infty} a_k = \sum_{k=1}^{\ell} b_k + \ell \cdot \lim_{k \rightarrow \infty} b_k.$$

Example 3.66. Consider $a_n = \frac{6n+9}{n^2(n+3)^2} = \frac{1}{n^2} - \frac{1}{(n+3)^2} =: b_n - b_{n+3}$. Hence,

$$\sum_{k=1}^n a_k = \sum_{k=1}^n (b_k - b_{k+3}) = \sum_{k=1}^n b_k - \sum_{k=4}^{n+3} b_k = b_1 + b_2 + b_3 + b_{n+1} + b_{n+2} + b_{n+3}.$$

Since $b_n \rightarrow 0$, we obtain $\sum_{k=1}^{\infty} a_k = 1 + \frac{1}{4} + \frac{1}{9} = \frac{49}{36}$.

However, note that **it is rare that we can compute the sum precisely**. Already for the example $\sum_{k=1}^{\infty} \frac{1}{k^2}$, which is very similar to the one above, we need some higher mathematics, to find its precise value $\frac{\pi^2}{6} \approx 1.645$. For many other sums, there is just no closed expression.

Still, we might be interested if the sum exists. For this, we can deduce several techniques from our knowledge about sequences. Let us start with some calculation rules.

Theorem 3.67. Let $\sum_{k=1}^{\infty} a_k$ and $\sum_{k=1}^{\infty} b_k$ be convergent series, and let $c \in \mathbb{C}$. Then we have

$$\sum_{k=1}^{\infty} (a_k + b_k) = \sum_{k=1}^{\infty} a_k + \sum_{k=1}^{\infty} b_k$$

and

$$\sum_{k=1}^{\infty} c \cdot a_k = c \cdot \sum_{k=1}^{\infty} a_k.$$

Proof. For the first equality we observe that

$$\sum_{k=1}^n (a_k + b_k) = \sum_{k=1}^n a_k + \sum_{k=1}^n b_k.$$

For the second equality consider

$$\sum_{k=1}^n c \cdot a_k = c \cdot \sum_{k=1}^n a_k.$$

Since both series are convergent, we obtain the results from Theorem 3.24. □

We now consider two results on the convergence of series that follow directly from the results of the last section. In fact, they are just reformulations of the *monotonicity principle* (Theorem 3.35) and the *Cauchy criterion* (Theorem 3.56).

The first is concerned with series with non-negative terms and bounded partial sums.

Theorem 3.68. *Let $(a_n)_{n=1}^{\infty} \subset \mathbb{R}$ be a non-negative sequence, i.e. $a_k \geq 0$ for all $k \in \mathbb{N}$. Then, the sequence of partial sums $s_n := \sum_{k=1}^n a_k$ is bounded, i.e.,*

$$\exists C \in \mathbb{R} \forall n \in \mathbb{N}: s_n \leq C,$$

if and only if the series $\sum_{k=1}^{\infty} a_k$ converges.

Proof. Since $a_k \geq 0$, we obtain that (s_n) is a non-decreasing, and therefore monotone, sequence. The result follows from the monotonicity principle (Theorem 3.35). \square

This theorem is already enough to show that the aforementioned alternative representation of Euler's number (defined in Section 2) as an infinite sum converges.

Example 3.69 (Euler's number). Consider the series given by the sequence $a_n = \frac{1}{n!}$, starting at 0 in this case. The partial sums are given by

$$s_n = \sum_{k=0}^n \frac{1}{k!}.$$

Now note that $k! = 1 \cdot 2 \cdot 3 \cdots k \geq 1 \cdot 2^{k-1}$ for $k \geq 1$, and therefore

$$s_n \leq 1 + \sum_{k=1}^{\infty} \frac{1}{2^{k-1}} = 1 + 2 \cdot \frac{1/2}{1 - 1/2} = 3.$$

Thus we obtain that the series is convergent.

Let us also compute that the sum of the series $\sum a_n$ is e . The binomial theorem yields

$$\begin{aligned} \left(1 + \frac{1}{n}\right)^n &= \sum_{k=0}^n \binom{n}{k} \frac{1}{n^k} = \sum_{k=0}^n \frac{n!}{k!(n-k)!} \cdot \frac{1}{n^k} \\ &= \sum_{k=0}^n \frac{1}{k!} \cdot \frac{n!}{(n-k)!n^k} \leq s_n, \end{aligned}$$

where the last step follows from $\frac{n!}{(n-k)!} = (n-k+1) \cdot (n-k+2) \cdots (n-1) \cdot n \leq n^k$.

Hence,

$$e := \lim_{n \rightarrow \infty} \left(1 + \frac{1}{n}\right)^n \leq \sum_{k=0}^{\infty} \frac{1}{k!}.$$

On the other hand, for fixed $m \in \mathbb{N}$ and $n \geq m$, we obtain, by similar arguments as above, that

$$\begin{aligned} \lim_{n \rightarrow \infty} \left(1 + \frac{1}{n}\right)^n &= \lim_{n \rightarrow \infty} \sum_{k=0}^n \binom{n}{k} \frac{1}{n^k} \geq \lim_{n \rightarrow \infty} \sum_{k=0}^m \binom{n}{k} \frac{1}{n^k} \\ &= \sum_{k=0}^m \frac{1}{k!} \left(\lim_{n \rightarrow \infty} \frac{n!}{(n-k)!n^k} \right) = \sum_{k=0}^m \frac{1}{k!}, \end{aligned}$$

where the last step follows from $n!/(n-k)! \geq (n-k)^k$.

Note that it was important that the number of elements in the sum was independent of n to

take the limit into the sum. This implies, by taking the limit $m \rightarrow \infty$ on the right hand side, that $\lim_{n \rightarrow \infty} \left(1 + \frac{1}{n}\right)^n \geq \sum_{k=0}^{\infty} \frac{1}{k!}$. Combining both inequalities we see that

$$e = \lim_{n \rightarrow \infty} \left(1 + \frac{1}{n}\right)^n = \sum_{k=0}^{\infty} \frac{1}{k!}.$$

□

The next result is just the Cauchy criterion applied to partial sums.

Theorem 3.70 (Cauchy criterion). *Let $\sum_{k=1}^{\infty} a_k$ be a series. Then we have that $\sum_{k=1}^{\infty} a_k$ is convergent if and only if*

$$\forall \varepsilon > 0 \exists n_0 \in \mathbb{N} \forall m > n \geq n_0: \left| \sum_{k=n+1}^m a_k \right| < \varepsilon.$$

In other words, the series $\sum_{k=1}^{\infty} a_k$ is convergent if and only if the sequence of partial sums is a Cauchy sequence.

Proof. By definition we have that $\sum_{k=1}^{\infty} a_k$ converges if and only if the sequence of partial sums $s_n = \sum_{k=1}^n a_k$ converges. Since, for $m > n$, we have that

$$s_m - s_n = \sum_{k=1}^m a_k - \sum_{k=1}^n a_k = \sum_{k=n+1}^m a_k,$$

we see that the condition in the theorem is equivalent to (s_n) being a Cauchy sequence. This is equivalent to (s_n) being convergent, see Theorem 3.56.

□

This theorem immediately leads to the following simple criterion. In many cases, this is already enough to show that a sequence is *divergent*.

Corollary 3.71. *We have*

$$\sum_{k=1}^{\infty} a_k \text{ is convergent} \implies \lim_{k \rightarrow \infty} a_k = 0,$$

i.e., every summable sequence is a null sequence.

In particular, if (a_n) is not a null sequence, then $\sum a_k$ is divergent.

Proof. We use the Cauchy criterion for series and set $m = n + 1$. Then we get

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

This is the definition of being a null sequence.

□

Example 3.72. From this, we finally obtain that $\sum_{k=0}^{\infty} q^k$ can only be convergent for $|q| < 1$.

Remark 3.73. We have already seen that there are null sequences which do not give a convergent series, e.g. $(\frac{1}{n})_{n=1}^{\infty}$. So the above corollary gives a necessary but not a sufficient condition for a series to be convergent.

Remark 3.74. (*) Indeed the representation of e via a series can be generalized to obtain the known exponential function. This leads to $e^x = \sum_{k=0}^{\infty} \frac{x^k}{k!}$. (We will prove this much later!)

3.7 Convergence tests

We now discuss several criteria to prove that a series is convergent. These *convergence tests* mostly do not lead to the precise sum of a series. However, they are quite generally applicable. The first tests that will be discussed are based on another form of convergence of series, which will turn out to be a stronger criterion.

Definition 3.75. Let $(a_n)_{n \in \mathbb{N}} \subset \mathbb{C}$ be a sequence such that

$$\sum_{k=1}^{\infty} |a_k| < \infty \quad : \iff \quad \exists C \in \mathbb{R} \forall n \in \mathbb{N}: \sum_{k=1}^n |a_k| \leq C,$$

i.e., the series of absolute values of a_n is bounded. Then, we say that the series $\sum_{k=1}^{\infty} a_k$ is **absolutely convergent** or the sequence $(a_n)_{n \in \mathbb{N}}$ is **absolutely summable**.

Note that $\sum |a_k| < \infty$ is really the same as $\sum |a_k|$ being convergent. This follows from Theorem 3.68 and the fact that $|a_k| \geq 0$.

Moreover, note that for non-negative sequences $(a_n)_{n \in \mathbb{N}}$, i.e., $a_k \geq 0$ for all k , absolute summability and summability are just the same.

Let us first consider a known example.

Example 3.76. We have shown already that for $|q| < 1$ the series $\sum_{k=0}^{\infty} q^k$ is convergent. Since

$$\sum_{k=0}^{\infty} |q^k| = \sum_{k=0}^{\infty} |q|^k = \frac{1}{1 - |q|}.$$

can be shown as above, we see that $\sum q^k$ is absolutely convergent in the same range of q .

Example 3.77. The **alternating harmonic series**, i.e. $\sum_{k=1}^{\infty} \frac{(-1)^k}{k}$, is not absolutely convergent, because

$$\sum_{k=1}^{\infty} \left| \frac{(-1)^k}{k} \right| = \sum_{k=1}^{\infty} \frac{1}{k}$$

is the harmonic series, which is divergent. However, we will see later that the alternating harmonic series is convergent.

The next result shows that absolute convergence is indeed a stronger criterion.

Theorem 3.78. *Absolutely convergent series are also convergent.*

Proof. We use the Cauchy criterion and the triangle inequality to prove this result. Let $\sum_{k=1}^{\infty} a_k$ be a absolutely convergent series and $\varepsilon > 0$. By the Cauchy criterion there exists some $n_0 \in \mathbb{N}$ such that for all $m > n \geq n_0$ we have

$$\sum_{k=n}^m |a_k| < \varepsilon.$$

The triangle inequality yields

$$\left| \sum_{k=n}^m a_k \right| \leq \sum_{k=n}^m |a_k| < \varepsilon.$$

Thus the Cauchy criterion implies that $\sum_{k=1}^{\infty} a_k$ is convergent. □

We will now discuss several criteria, called *convergence tests*, that can by used to verify if a series is convergent or not. However, note that these test are sometimes *inconclusive*, i.e., we do not get a definite answer by applying them, and one needs to apply other techniques.

3.7.1 Comparison test

The first test is based on the existence of another series, that is known to be convergent or not. This (quite obvious) test is used in nearly every application, and is clearly based on the numerous examples we are discussing. In particular, some of the other convergence tests are based on a simple application of this one.

Theorem 3.79. *Let $\sum_{k=1}^{\infty} a_k$ and $\sum_{k=1}^{\infty} b_k$ be series.*

- (i) *If $\sum b_k$ is absolutely convergent, and $|a_k| \leq |b_k|$ for all $k \in \mathbb{N}$. Then, $\sum a_k$ is also absolutely convergent.*
- (ii) *If $\sum b_k = \infty$, and $0 \leq b_k \leq a_k$ for all $k \in \mathbb{N}$, then also $\sum a_k = \infty$.*

Proof. Since $\sum |b_k|$ is an upper bound for the partial sums of $\sum |a_k|$, we have that $\sum |a_k|$ is convergent by Theorem 3.68, and therefore finite. This implies that $\sum a_k$ is absolutely convergent. The second point follows from similar argument. We use that the partial sums of $\sum b_k$ are divergent (which is by Theorem 3.68 the same as $\sum b_k = \infty$, since $b_k \geq 0$), and that that partial sums of $\sum a_k$ are just larger. This shows that also $\sum a_k$ is unbounded, and hence divergent. □

Let us consider the series $\sum_{k=1}^{\infty} k^{-c}$ for $c > 0$, with polynomially decaying terms.

Example 3.80. We have

$$\sum_{k=1}^{\infty} \frac{1}{k^2} = \sum_{k=1}^{\infty} \frac{1}{k(k+1)} \frac{k+1}{k} \leq 2 \sum_{k=1}^{\infty} \frac{1}{k(k+1)} = 2 < \infty,$$

and hence that $\sum_{k=1}^{\infty} \frac{1}{k^2}$ is (absolutely) convergent. Similarly, $\sum_{k=1}^{\infty} \frac{1}{\sqrt{k}} \geq \sum_{k=1}^{\infty} \frac{1}{k} = \infty$.

In general, we obtain that

- $\sum_{k=1}^{\infty} \frac{1}{k^c}$ is absolutely convergent for $c \geq 2$, and
- $\sum_{k=1}^{\infty} \frac{1}{k^c} = \infty$ for $c \leq 1$.

(In fact, we will see soon that the series is actually convergent for all $c > 1$.)

One may also consider more complicated series.

Example 3.81. For example, we obtain that $\sum_{k=1}^{\infty} \frac{k^3+4k^2-3}{k^5+k+1}$ is convergent, since

$$\sum_{k=1}^{\infty} \left| \frac{k^3 + 4k^2 - 3}{k^5 + k + 1} \right| \leq \sum_{k=1}^{\infty} \frac{k^3 + 4k^2}{k^5} = 5 \sum_{k=1}^{\infty} \frac{1}{k^2} < \infty,$$

Example 3.82. However, $\sum_{k=1}^{\infty} \frac{k^3+4k^2-3}{k^4-k+1} = \infty$, since

$$\sum_{k=1}^{\infty} \left| \frac{k^3 + 4k^2 - 3}{k^4 - k + 1} \right| \geq \sum_{k=1}^{\infty} \frac{k^3}{k^4} = \sum_{k=1}^{\infty} \frac{1}{k} = \infty.$$

3.7.2 Root test

We now turn to convergence tests that can be applied to the terms of a series, and we do not need precise knowledge about the partial sums. As the proof shows, these tests just follow from a comparison of the series under consideration with a *geometric series*, see Example 3.60.

Theorem 3.83 (root test). *Let $\sum_{k=1}^{\infty} a_k$ be a series.*

(i) *If there exists some $q < 1$ such that*

$$\sqrt[k]{|a_k|} \leq q \quad \text{for almost all } k,$$

then $\sum_{k=1}^{\infty} a_k$ is absolutely convergent.

(ii) *Conversely, if*

$$\sqrt[k]{|a_k|} \geq 1 \quad \text{for infinitely many } k,$$

then $\sum a_k$ is divergent.

Proof. By assumption, there is some k_0 such that $|a_k| \leq q^k$ for $k \geq k_0$. Hence, we get that

$$\sum_{k=1}^{\infty} |a_k| = \sum_{k=1}^{k_0-1} |a_k| + \sum_{k=k_0}^{\infty} |a_k| \leq \sum_{k=1}^{k_0-1} |a_k| + \sum_{k=k_0}^{\infty} q^k \leq \sum_{k=1}^{k_0-1} |a_k| + \sum_{k=0}^{\infty} q^k,$$

where the first inequality comes from Theorem 3.79. Now, $\sum_{k=1}^{k_0-1} |a_k|$ is a finite sum and $\sum_{k=0}^{\infty} q^k$ is a geometric series. As both are finite for $q \in [0, 1)$, we get that $\sum_{k=1}^{\infty} a_k$ converges absolutely. For part (ii), just note that (a_n) fails to converge to 0 under the given assumption. \square

The condition of **the root test can be written equivalently with the help of limits**: For this, consider the limit superior

$$a := \limsup_{k \rightarrow \infty} \sqrt[k]{|a_k|}.$$

Then, the series $\sum a_k$ is

- (i) absolutely convergent, if $a < 1$,
- (ii) divergent, if $a > 1$,
- (iii) and we do not gain any information from the root test, if $a = 1$.

Recall that a not (absolutely) convergent series $\sum a_k$ with $a_k \geq 0$ necessarily satisfies $\sum a_k = \infty$.

Let us discuss some examples.

Example 3.84. Let $a_k = \frac{1}{k^k}$. For $k \geq 2$ we have that $\sqrt[k]{a_k} = \frac{1}{k} \leq \frac{1}{2}$. Thus by the root test the series $\sum_{k=1}^{\infty} \frac{1}{k^k}$ converges absolutely.

Example 3.85. Consider the series $\sum_{k=1}^{\infty} \sin(k) k^{42} 2^{-k}$. Although the terms of the series, say a_k , look complicated and are very large in the beginning (e.g., $a_2 \approx 10^{12}$), we obtain from the comparison and the root test that it converges (absolutely).

For this note that $\sum |\sin(k) k^{42} 2^{-k}| \leq \sum |k^{42} 2^{-k}|$. Now, since $\lim \sqrt[k]{k} = 1$, we obtain that $\lim \sqrt[k]{k^{42} 2^{-k}} = \frac{1}{2}$. It follows that the series is absolutely convergent, and therefore convergent. That is, $\sum_{k=1}^{\infty} \sin(k) k^{42} 2^{-k}$ is just a *number*.

3.7.3 Ratio test

The next test is based on quotients of successive terms of the series.

Theorem 3.86 (ratio test). *Let $\sum_{k=1}^{\infty} a_k$ be a series.*

(i) *If there exists some $q < 1$ such that*

$$a_k \neq 0 \quad \text{and} \quad \left| \frac{a_{k+1}}{a_k} \right| \leq q \quad \text{for almost all } k,$$

then $\sum_{k=1}^{\infty} a_k$ is absolutely convergent.

(ii) *Conversely, if*

$$a_k \neq 0 \quad \text{and} \quad \left| \frac{a_{k+1}}{a_k} \right| \geq 1 \quad \text{for almost all } k,$$

then $\sum a_k$ is divergent.

Proof. As in the proof of the root test we can split the series in $\sum_{k=1}^{\infty} a_k = \sum_{k=1}^{k_0-1} a_k + \sum_{k=k_0}^{\infty} a_k$, where $\sum_{k=1}^{k_0-1} a_k$ is a finite sum and does not change the convergence of the series. We assume w.l.o.g that $k_0 = 1$. By induction we have that $|a_k| \leq q^{k-1}|a_1|$ for all $k \in \mathbb{N}$. An index shift implies

$$\sum_{k=1}^{\infty} |a_k| \leq \sum_{k=1}^{\infty} q^{k-1}|a_1| = \sum_{k=0}^{\infty} q^k|a_1| = |a_1| \sum_{k=0}^{\infty} q^k.$$

Since $q \in [0, 1)$ we get that $\sum_{k=1}^{\infty} a_k$ is absolutely convergent from Theorem 3.79.

For part (ii) we follow similar steps and obtain $|a_k| \geq |a_{k_0}|$ for all $k \geq k_0$ and k_0 large enough. Hence, (a_k) is not a null sequence, and consequently $\sum a_k$ not convergent.

□

The condition of **the ratio test can be written equivalently with the help of limits:**

For this, assume that the limit

$$a := \lim_{k \rightarrow \infty} \left| \frac{a_{k+1}}{a_k} \right|$$

exists. Then, the series $\sum a_k$ is

- (i) absolutely convergent, if $a < 1$,
- (ii) divergent, if $a > 1$,
- (iii) and we do not gain any information from the ratio test, if $a = 1$.

Remark 3.87. One may prove that the ratio test is a bit more special than the root test in the following sense:

Assume that $A := \lim_{k \rightarrow \infty} \left| \frac{a_{k+1}}{a_k} \right|$ exists, then also $B := \lim_{k \rightarrow \infty} \sqrt[k]{|a_k|}$ exists and $A = B$. That is, whenever we successfully applied the ratio test, we may have also applied the root test to come to the same conclusion. (We will not prove this here.) However, the ratio test is sometimes much easier to apply.

Example 3.88. We show that the series $\sum_{k=1}^{\infty} \frac{k^{k/2}}{k!}$ is absolutely convergent. For this, note that

$$\left| \frac{a_{k+1}}{a_k} \right| = \frac{(k+1)^{(k+1)/2} k!}{k^{k/2} (k+1)!} = \frac{(1+1/k)^{k/2}}{\sqrt{k+1}} \leq \frac{2}{\sqrt{k}} \rightarrow 0.$$

The ratio test implies the absolute convergence.

Example 3.89. The root and ratio test have their limitations, e.g., for series whose terms are only polynomially decaying, i.e., $\sum k^{-c}$ for some $c > 0$. Since $\lim \sqrt[k]{k^{-c}} = 1$ independent of c , we cannot distinguish between different c with the root test, although the series is convergent for some c , and divergent for others, see Example 3.80.

3.7.4 Cauchy's condensation test

The convergence test we want to discuss now is only applicable if the terms of the series are a **non-negative and monotone** null sequence. However, this test is very powerful in this case.

Theorem 3.90 (Cauchy's condensation test). *Let $\sum_{k=1}^{\infty} a_k$ be a series with $0 \leq a_{k+1} \leq a_k$ for all k . Then,*

$$\sum_{k=1}^{\infty} a_k \text{ is convergent} \iff \sum_{k=1}^{\infty} 2^k a_{2^k} \text{ is convergent}.$$

Proof. We will bound the series $\sum a_k$ from above and below. This will imply the result by Theorem 3.79. For this, we group the terms of the non-increasing sequence (a_k) into “blocks” with indices $\{2^k, 2^k + 1, \dots, 2^{k+1} - 1\}$, and just bound all elements in such a block by the smallest and the largest one, respectively.

To be precise, note that every natural number can be written uniquely as $2^k + \ell$ for some $k \in \mathbb{N}$ and some $\ell \in \{0, 1, \dots, 2^k - 1\}$ (Check that!), which shows

$$\sum_{k=1}^{\infty} a_k = \sum_{k=1}^{\infty} \sum_{\ell=0}^{2^k-1} a_{2^k+\ell}.$$

Moreover, by simply bounding by maximum or minimum and the number of elements, we obtain

$$2^k a_{2^k+1} \leq \sum_{\ell=0}^{2^k-1} a_{2^k+\ell} \leq 2^k a_{2^k}$$

for all $k \in \mathbb{N}$. This implies the statement. □

We are finally in the position to study the series $\sum_{k=1}^{\infty} k^{-c}$ for $c > 0$.

Lemma 3.91. *We have that*

$$\sum_{k=1}^{\infty} \frac{1}{k^c} < \infty \iff c > 1.$$

Proof. By the condensation test we see that this series is convergent if and only if

$$\sum_{k=1}^{\infty} 2^k (2^k)^{-c} = \sum_{k=1}^{\infty} (2^{1-c})^k < \infty.$$

Now note that this is just a geometric series (with $q = 2^{1-c}$) and we have $2^{1-c} < 1$ if and only if $c > 1$, which proves the result. □

3.7.5 Leibniz criterion

The last convergence test we want to discuss is again only applicable for special series. Again, the terms of the series are based on a **monotone** null sequence. However, we consider their **alternating sum** and show that this is always a convergent series.

Theorem 3.92 (Leibniz criterion). *Let $(a_k)_{k \in \mathbb{N}}$ be monotone with $a_k \rightarrow 0$. Then,*

$$\sum_{k=1}^{\infty} (-1)^k a_k \quad \text{is convergent.}$$

Proof. Assume w.l.o.g. that (a_k) is non-increasing, and therefore non-negative. (Why?) Since the sequence (a_k) is non-increasing, we have that $a_k - a_{k+1} \geq 0$ for all k . Thus

$$s_{2n+2} = s_{2n} + (-1)^{2n+1} a_{2n+1} + (-1)^{2n+2} a_{2n+2} = s_{2n} - (a_{2n+1} - a_{2n+2}) \leq s_{2n}.$$

This means the sequence $(s_{2n})_{n \in \mathbb{N}}$ is non-increasing. The same argument implies that the sequence $(s_{2n-1})_{n \in \mathbb{N}}$ is non-decreasing. Furthermore $s_{2n} - s_{2n-1} = a_{2n} \geq 0$ implies that $s_{2n-1} \leq s_{2n}$. This yields that for all $n \in \mathbb{N}$

$$s_1 \leq s_3 \leq \cdots \leq s_{2n-1} \leq s_{2n} \leq \cdots \leq s_4 \leq s_2.$$

So we have two monotone and bounded sequences, (s_{2n}) and (s_{2n+1}) , which are therefore convergent. We still need to show that their limits are the same. (Otherwise we would have only proven that the series has two accumulation points.) But we clearly have

$$|s_{2n} - s_{2n-1}| = a_{2n+1} \rightarrow 0,$$

which implies that (s_n) is convergent. □

This theorem shows that alternating series are somehow easier to handle than their non-alternating versions. The following example will demonstrate this.

Example 3.93. The **alternating harmonic series** $\sum \frac{(-1)^k}{k}$ converges by the Leibniz criterion, since $a_k = \frac{1}{k}$ is a decreasing null sequence. Later, we will even prove that

$$\sum_{k=1}^{\infty} \frac{(-1)^{k+1}}{k} = \ln(2) \approx 0.693,$$

where $\ln(x)$ is the *natural logarithm*. Recall that the 'normal' harmonic series does not converge.

Example 3.94. Since $a_k = \frac{1}{\sqrt{k+4}}$ is a decreasing null sequence, we have that $\sum (-1)^k a_k$ is convergent. However, note that $b_k = \frac{(-1)^k}{\sqrt{k}}$ is also a null sequence, but not monotone, and $\sum (-1)^k b_k = \sum \frac{1}{\sqrt{k}} = \infty$.

3.8 Power series

We finally consider series that contain a free parameter or, in other words, describe a function whenever they are convergent.

Definition 3.95 (Power series). Let $(a_k)_{k=0}^{\infty} \subset \mathbb{C}$ be a sequence and let $c \in \mathbb{C}$. Then, for $z \in \mathbb{C}$, we define the (formal) **power series** as

$$f(z) := \sum_{k=0}^{\infty} a_k(z - c)^k.$$

We call the a_k the **coefficients** and c the **center** of the power series.

We call

$$R := R(f) := \frac{1}{\limsup_{k \rightarrow \infty} \sqrt[k]{|a_k|}}$$

the **radius of convergence** of the power series f .

(We set $R = \infty$ if $\limsup_{k \rightarrow \infty} \sqrt[k]{|a_k|} = 0$, and $R = 0$ if $\limsup_{k \rightarrow \infty} \sqrt[k]{|a_k|} = \infty$.)

We call $D_f = \{y \in \mathbb{C}: |y - c| < R(f)\}$ the **disc of convergence** of f .

If we consider a power series only for real inputs, i.e., $f(x) = \sum_{k=0}^{\infty} a_k(x - c)^k$ for some $(a_k) \subset \mathbb{R}$, $c, x \in \mathbb{R}$, then we call it a **real power series** and $D_f = \{y \in \mathbb{R}: |y - c| < R(f)\} = (c - R, c + R)$ with $R = R(f)$ is the **interval of convergence**.

The term *formal* in the above definition means that we do not know a priori if the power series converges for a given z . As for series in general: It might be a number or $\pm\infty$, but it might also not exist (as a number).

However, as the name *radius of convergence* already indicates, the number $R(f)$ plays a crucial role for deciding if a series converges.

Theorem 3.96 (Radius of convergence). Let $(a_k)_{k=0}^{\infty} \subset \mathbb{C}$ be a sequence, $c \in \mathbb{C}$, and let f be the corresponding formal power series with radius of convergence $R := R(f)$.

Then, the power series $f(z) = \sum_{k=0}^{\infty} a_k(z - c)^k$ is

- absolutely convergent for $z \in D_f = U_R(c) = \{y \in \mathbb{C}: |y - c| < R\}$, and
- divergent for $z \in \{y \in \mathbb{C}: |y - c| > R\}$.

For z with $|z - c| = R$, we do not get a definitive answer.

Proof. Set $b_k = a_k(z - c)^k$. Set $x = (z - c)$ and consider

$$\limsup_{k \rightarrow \infty} \sqrt[k]{|b_k|} = \limsup_{k \rightarrow \infty} |x| \sqrt[k]{|a_k|} = \frac{|x|}{R}.$$

Thus the root test implies that if $|x| < R$, the series converges. If on the other hand $|x| > R$ we see that (b_k) cannot be a null sequence, so the series cannot be convergent. \square

Example 3.97. Let us consider the power series $f(z) = \sum_{k=0}^{\infty} \frac{z^k}{k}$, i.e., $\sum_{k=0}^{\infty} a_k(z - c)^k$ with $a_k = \frac{1}{k}$ and $c = 0$. We have

$$|\sqrt[k]{a_k}| = \frac{1}{\sqrt[k]{k}} \rightarrow 1.$$

So, the radius of convergence $R(f) = 1$, and we obtain that $\sum \frac{z^k}{k}$ is absolutely convergent if $|z| < 1$.

In some cases, however, it is easier to use the **ratio test** to verify convergence of a power series. Luckily, we have that the radius of convergence can also be given by the corresponding limit, if it exists.

Lemma 3.98. *The radius of convergence of a power series $f(z) := \sum_{k=0}^{\infty} a_k(z - c)^k$ satisfies*

$$R(f) = \lim_{k \rightarrow \infty} \left| \frac{a_k}{a_{k+1}} \right|$$

whenever the limit on the right hand side exists.

(Note the changed order in the quotient in comparison to the ratio test.)

We omit the formal proof.

Example 3.99. Let us consider the real power series $f(x) = \sum_{k=0}^{\infty} \frac{x^k}{k}$. We have

$$\left| \frac{a_k}{a_{k+1}} \right| = \frac{k+1}{k} \rightarrow 1.$$

So, by the ratio test we obtain that $\sum \frac{x^k}{k}$ is absolutely convergent if $|x| < 1$. We will see later that this series describes the *natural logarithm* in this range by $\ln(1-x) = -\sum_{k=0}^{\infty} \frac{x^k}{k}$ for $|x| < 1$.

Example 3.100. We also obtain that the power series

$$\sum_{k=0}^{\infty} \frac{z^k}{k!}$$

is absolutely convergent for every $z \in \mathbb{C}$. To see this, note that

$$\left| \frac{a_k}{a_{k+1}} \right| = k+1 \rightarrow \infty.$$

We obtain by the above ratio test that the radius of convergence is ∞ , and hence, that this power series converges for all $z \in \mathbb{C}$.

We will see later that this series describes the *exponential function*, i.e., $e^z = \sum_{k=0}^{\infty} \frac{z^k}{k!}$.

Note that we can consider

$$\begin{aligned} f: D_f &\rightarrow \mathbb{C} \\ z &\mapsto \sum_{k=0}^{\infty} a_k(z - c)^k \end{aligned}$$

as function, and it would be interesting to **find explicit expressions** as functions, at least for some power series. We already know the most important example:

Example 3.101 (Geometric series). Consider the power series $f(x) := \sum_{k=0}^{\infty} x^k$ for $x \in \mathbb{R}$, i.e., $\sum_{k=0}^{\infty} a_k(z - c)^k$ with $a_k = 1$ and $c = 0$.

We have $\lim \sqrt[k]{a_k} = 1$, and hence $D_f = (-1, 1)$. We know from Example 3.60 that

$$f(x) := \sum_{k=0}^{\infty} x^k = \frac{1}{1-x} \quad \text{for } x \in D_f = (-1, 1).$$

Note that $\frac{1}{1-x}$ is also well-defined for other x , like $x = -2$. However, for $x \notin (-1, 1)$ the equality to the power series does not hold, because f is divergent in this case.

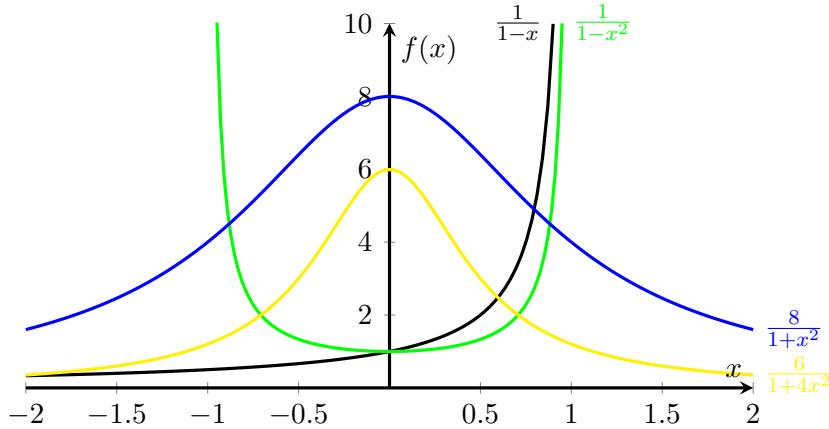


Figure 20: The graph of the functions $\frac{1}{1-x}$, $\frac{1}{1-x^2}$, $\frac{8}{1+x^2}$, $\frac{6}{1+4x^2}$

Many series can be brought into the form of a geometric series, and we can therefore find also explicit expressions for them. However, note this works only if the series converge.

Example 3.102. Consider the power series $f(x) := \sum_{k=0}^{\infty} x^{2k}$ for $x \in \mathbb{R}$.

Using the last example, this can be written as $f(x) = \sum_{k=0}^{\infty} (x^2)^k = \frac{1}{1-x^2}$ for all $x \in (-1, 1)$. Analogously, we find $\sum_{k=0}^{\infty} 8(-1)^k x^{2k} = 8 \sum_{k=0}^{\infty} (-x^2)^k = \frac{8}{1+x^2}$ for $x \in (-1, 1)$, or

$$\sum_{k=0}^{\infty} 6(-4)^k x^{2k} = 6 \sum_{k=0}^{\infty} (-4x^2)^k = \frac{6}{1+4x^2},$$

which only holds for $x \in (-\frac{1}{2}, \frac{1}{2})$. Note that it is also not easy to “see” from the graph of the explicit expression, where it can be written as power series, see Figure 20.

Sometimes it is also helpful to write an explicit function as a series, i.e., a **series expansions**.

Example 3.103. Assume we want to write $f(x) = \frac{1}{x^2}$ as a power series. Then, we can denote $y := 1 - x^2$ such that $1 - y = x^2$. Since we can write $\sum_{k=0}^{\infty} y^k = \frac{1}{1-y}$ for all $|y| < 1$, we see that

$$f(x) = \frac{1}{x^2} = \frac{1}{1-y} = \sum_{k=0}^{\infty} y^k = \sum_{k=0}^{\infty} (1-x^2)^k$$

for all $x \in \mathbb{R}$ with $|1 - x^2| < 1$, i.e., $x \in (-\sqrt{2}, \sqrt{2}) \setminus \{0\}$.

However, rewriting this as a power series $\sum_{k=0}^{\infty} a_k x^k$ for some $(a_k) \subset \mathbb{R}$ would be a rather hard computation. We will see later how to do that in a systematic way using derivatives.

Still we can use the above techniques to **find out where more complicated series converge**.

Example 3.104. Assume we want to find for which $x \in \mathbb{R}$ the series

$$f(x) := \sum_{k=0}^{\infty} \sqrt{k} 2^{-k} (x^2 - 2)^k$$

converges. By setting $y = x^2 - 2$, we see that the power series $g(y) := \sum_{k=0}^{\infty} \sqrt{k} 2^{-k} y^k$ has radius of convergence $R(g) = \left(\limsup \sqrt[k]{\sqrt{k} 2^{-k}} \right)^{-1} = 2$. Hence, g converges, whenever $|y| < 2$. Moreover, we obtain that g diverges for $|y| > 2$. Verifying the cases $y = \pm 2$ separately, we see that g diverges for all $|y| \geq 2$. Using $y = x^2 - 2$, we see that $f(x)$ converges exactly for those $x \in \mathbb{R}$ with $|x^2 - 2| < 2 \iff x^2 \in (0, 4) \iff 0 < |x| < 2$ or, equivalently, $x \in (-2, 0) \cup (0, 2)$.

Example 3.105. If we want to find for which $x \in \mathbb{R}$ the series

$$f(x) := \sum_{k=0}^{\infty} \frac{1}{k+3} 2^k (x^3 - 1)^k$$

converges, we set $y = x^3 - 1$, and consider $g(y) := \sum_{k=0}^{\infty} \frac{1}{k+3} 2^k y^k$. As above, $R(g) = \left(\limsup \sqrt[k]{\frac{1}{k+3} 2^k} \right)^{-1} = 1/2$. Hence, g converges whenever $|y| < 1/2$. Moreover, we obtain that g diverges for $|y| > 1/2$. Verifying the cases $y = \pm \frac{1}{2}$ separately, we see that g diverges for $y = 1/2$ (harmonic series). However, for $y = -1/2$, we have $g(-1/2) = \sum_{k=0}^{\infty} \frac{(-1)^k}{k+3}$, which is convergent by the Leibniz criterion. Using $y = x^3 - 1$, we see that $f(x)$ converges exactly for those $x \in \mathbb{R}$ with $-\frac{1}{2} \leq x^3 - 1 < \frac{1}{2} \iff x^3 \in [\frac{1}{2}, \frac{3}{2}) \iff x \in [\sqrt[3]{\frac{1}{2}}, \sqrt[3]{\frac{3}{2}})$.

Note that, **in general, there is no way to give an explicit form for series as the ones above.**

4 Continuous functions and limits

We now come back to functions and, with the aid sequences and their limits, we will study certain important properties which functions may have. As kind of a warm up we study **continuity**, i.e., we precisely define what we mean by a **continuous functions**. However, on the way we also define the essential concept of *limits of a function* (instead of a sequence), which will be the basis also for the subsequent chapters.

As you probably know from school, a function is continuous if it has '*no jumps*' or '*can be drawn without lifting the pen*', and this is a good intuition. However, this 'definition' does not make sense if it comes to more complex situations. Try, e.g., to draw the function $f(x) = x \cdot \sin(1/x)$ and decide whether it is continuous or not. (Hint: Use your favorite computer algebra software.)

Therefore, we need a formal definition of continuity. Throughout this chapter, we only consider real-valued functions defined on (subsets of) the real numbers.

Definition 4.1. Let $D \subset \mathbb{R}$, $x_0 \in D$ and $f: D \rightarrow \mathbb{R}$. We call f **continuous at x_0** if for all sequences $(x_n)_{n \in \mathbb{N}}$ with $x_n \rightarrow x_0$ we have that $\lim_{n \rightarrow \infty} f(x_n)$ exists and

$$\lim_{n \rightarrow \infty} f(x_n) = f(x_0) = f\left(\lim_{n \rightarrow \infty} x_n\right).$$

If $U \subset D$ and f is continuous at all $x \in U$ we call f continuous on U , and

if f is continuous at all $x \in D$, then we just call f **continuous**.

Note that a function is continuous iff an expression of the form $\lim f(x_n)$ does not depend on the specific sequence (x_n) , but only on its limit $x_0 := \lim(x_n) \in D$.

Roughly speaking, **we can interchange the limit with the function, if it is continuous**.

Here, it is important that $x_0 \in D$ since otherwise $f(x_0)$ may not be defined. Later we will also consider *limits of functions* in the other case.

Let us start with some easy examples.

Example 4.2. Constant functions are clearly continuous.

Example 4.3. The prototype of a **discontinuous function** is the *Heaviside function* which is defined by

$$H(x) = \begin{cases} 0 & \text{if } x < 0, \\ 1 & \text{if } x \geq 0. \end{cases}$$

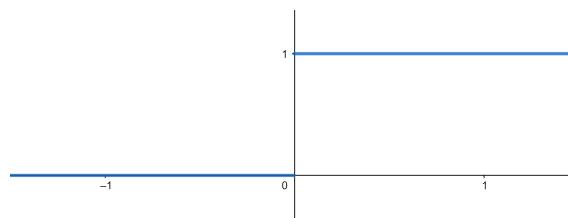


Figure 21: Heaviside function

If we now consider the sequence $x_n = \frac{1}{n}$ and $-x_n = -\frac{1}{n}$ we get that

$$\lim_{n \rightarrow \infty} H(x_n) = \lim_{n \rightarrow \infty} 1 = 1 \neq 0 = \lim_{n \rightarrow \infty} 0 = \lim_{n \rightarrow \infty} H(-x_n)$$

and hence that H is not continuous at 0. However, H is continuous at every other point. (This is because H is then constant in a neighborhood around this point, and constant functions are continuous.) Furthermore, the Heaviside function is not an 'exotic' example of a not continuous function, in fact this function plays an important role in physics.

Example 4.4. The example of the Heaviside function can be extended to 'jump-functions'. Therefore let $I = [a, b]$ be a closed interval, where $a < b$. If there exists some $t \in I$ such that $a \neq t$ and $b \neq t$, then functions of the form

$$f(x) = \begin{cases} c_1 & \text{if } x < t \\ c_2 & \text{if } x \geq t \end{cases}$$

are discontinuous at t , as long as $c_1 \neq c_2$. (For $c_1 = c_2$ this is clearly a constant function on I .)

Often one can find continuity in the following equivalent form, which is called the ε - δ -criterion.

Theorem 4.5 (ε - δ -criterion). *Let $f: D \rightarrow \mathbb{R}$ and $x_0 \in D$. Then, f is continuous at x_0 if and only if*

$$\forall \varepsilon > 0 \exists \delta > 0 \forall x \in D: |x - x_0| < \delta \implies |f(x) - f(x_0)| < \varepsilon.$$

In words: Given $x_0 \in D$. For all (fixed) $\varepsilon > 0$ there exists $\delta > 0$ such that for all $x \in D$ with $|x - x_0| < \delta$ we have that $|f(x) - f(x_0)| < \varepsilon$.

The condition in the above theorem was the first precise definition of a continuous function and is one of the most essential (and frightening) mathematical statements. It may be stated as "a small change in x_0 only allows a small change in $f(x_0)$ ". The precise definition is ultimately due to *Karl Weierstraß* (1815–1897), who is often cited as the "father of modern analysis".

The following figure gives a visualization of the criterion.

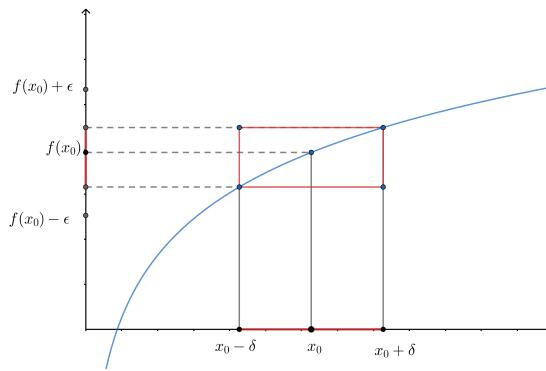


Figure 22: ε - δ -criterion

Proof. First we show f continuous at $x_0 \implies f$ satisfies ε - δ -criterion at x_0 by contradiction. So assume the ε - δ -criterion is not satisfied in x_0 . That is, that there exists some $\varepsilon_0 > 0$ such that

$$\forall \delta > 0 \exists x \in D: |x - x_0| < \delta \text{ and } |f(x) - f(x_0)| \geq \varepsilon_0.$$

Now, for $n \in \mathbb{N}$, let $\delta_n = \frac{1}{n}$. Thus we can find $x_n \in D$ such that $|x_n - x_0| < \frac{1}{n}$ and $|f(x_n) - f(x)| \geq \varepsilon_0$. So we have found a sequence $x_n \rightarrow x$ and $f(x_n) \not\rightarrow f(x)$, which contradicts the continuity of f . Hence, the ε - δ -criterion must be satisfied.

For the other direction assume the ε - δ -criterion holds and let $\varepsilon > 0$ and (x_n) be an arbitrary sequence such that $x_n \rightarrow x_0$. By our assumption we can find $\delta > 0$ such that for all $x \in D$ we have $|x - x_0| < \delta \implies |f(x) - f(x_0)| < \varepsilon$. Since the sequence (x_n) converges to x_0 we have that there exists $n_0 \in \mathbb{N}$ such that

$$n \geq n_0 \implies |x_n - x_0| < \delta \implies |f(x_n) - f(x_0)| < \varepsilon.$$

Hence $x_n \rightarrow x_0 \implies f(x_n) \rightarrow f(x_0)$. □

Example 4.6. The **identity**, i.e. $f(x) := x$, is continuous on \mathbb{R} . We want to prove the statement two times, first by using the definition and then by using the ε - δ -criterion.

Proof by using the definition. Let $x_0 \in \mathbb{R}$ and $(x_n) \subset \mathbb{R}$ be such that $x_n \rightarrow x_0$. Clearly,

$$|f(x_n) - f(x_0)| = |x_n - x_0| \rightarrow 0.$$

This yields that f is continuous at x_0 . As this holds for all $x_0 \in \mathbb{R}$, we have that f is continuous. □

Proof by using ε - δ -criterion. Let $\varepsilon > 0$ and $x_0 \in \mathbb{R}$. We choose $\delta = \varepsilon$ and obtain

$$|x - x_0| < \delta \implies |f(x) - f(x_0)| = |x - x_0| < \delta = \varepsilon.$$

Thus, by the ε - δ -criterion, f is continuous at x_0 . As this holds for all $x_0 \in \mathbb{R}$, we have that f is continuous. □

Example 4.7. Let us consider the **quadratic function** $f(x) = x^2$ on \mathbb{R} . Since for every convergent sequence $(x_n) \subset \mathbb{R}$ with $x_n \rightarrow x$, we have $\lim f(x_n) = \lim x_n^2 = (\lim x_n)^2 = x^2 = f(x)$, and hence that f is continuous on \mathbb{R} . We also give a proof of this fact using the ε - δ -criterion to show the difference to the above example.

Proof by ε - δ -criterion. Let $\varepsilon > 0$ and $x_0 \in \mathbb{R}$. We need to find $\delta > 0$ such that for all x with $|x - x_0| < \delta$ we have that $|f(x) - f(x_0)| = |x^2 - x_0^2| < \varepsilon$. Note that δ may depend on ε and on x_0 but not on x . (This may be justified by the order of the quantifiers in the definition.)

We use the triangle inequality to obtain

$$|x^2 - x_0^2| = |(x - x_0)(x + x_0)| = |x - x_0| |x + x_0| \leq |x - x_0| (|x - x_0| + 2|x_0|).$$

If we now assume $\delta \leq 1$, we obtain that

$$|x - x_0| + 2|x_0| < 1 + 2|x_0|$$

for all x with $|x - x_0| < \delta \leq 1$. If we assume additionally that $\delta \leq \frac{\varepsilon}{1+2|x_0|}$, then

$$|x^2 - x_0^2| < |x - x_0| (1 + 2|x_0|) < \frac{\varepsilon}{1+2|x_0|} (1 + 2|x_0|) = \varepsilon.$$

for all x with $|x - x_0| < \delta \leq \min\{1, \frac{\varepsilon}{1+2|x_0|}\}$. Therefore, we can set $\delta := \min\{1, \frac{\varepsilon}{1+2|x_0|}\}$ and obtain

$$|x - x_0| < \delta \implies |x^2 - x_0^2| < \varepsilon.$$

Note that $\delta > 0$ for all $\varepsilon > 0$, which implies that f is continuous at x_0 . As this also holds for all $x_0 \in \mathbb{R}$, we obtain that f is continuous. \square

Remark 4.8 (*). Note that in the above example, δ depends on ε and x_0 , and it is not hard to see that this dependence is necessary here. It is even no problem that $\delta \rightarrow 0$ if $\varepsilon \rightarrow 0$ and/or $x_0 \rightarrow \pm\infty$, since we only need $\delta > 0$ for all *fixed* ε and x_0 .

Example 4.9. The **root function** $f(x) = \sqrt{x}$ on $[0, \infty)$ is continuous.

Proof. Let $\varepsilon > 0$ and $x, y \in [0, 1]$. The binomial theorem implies

$$|\sqrt{x} - \sqrt{y}|^2 = x + y - 2\sqrt{xy} \leq x + y - 2\min\{x, y\} = |x - y|.$$

This shows that $|\sqrt{x} - \sqrt{y}| \leq |x - y|^{1/2}$. If we now choose $\delta = \varepsilon^2$, we see that $|\sqrt{x} - \sqrt{y}| < \varepsilon$ for all x, y with $|x - y| < \delta$. This shows that f is continuous on $[0, \infty)$. \square

Example 4.10. Let us also show that the **exponential function** $f(x) = a^x$ for fixed $a \in (0, \infty)$ is a continuous function on \mathbb{R} .

Proof. Let (x_n) be convergent with $x_n \rightarrow x_0$. Then, for every $k \in \mathbb{N}$ there is a $n_k \in \mathbb{N}$ such that

$$|x_n - x_0| \leq \frac{1}{k}$$

for all $n \geq n_k$. Using the fact that $a^{1/k} = \sqrt[k]{a} \rightarrow 1$, as $k \rightarrow \infty$, we obtain that $a^{x_n - x_0} \rightarrow 1$, as $n \rightarrow \infty$. This yields

$$|a^{x_n} - a^{x_0}| = |a^{x_0}(1 - a^{x_n - x_0})| \leq a^{x_0}|1 - a^{x_n - x_0}| \rightarrow 0.$$

Hence $\lim_{n \rightarrow \infty} a^{x_n} = a^{x_0}$ for every sequence (x_n) with $x_n \rightarrow x_0$. \square

Example 4.11. The **trigonometric functions sin and cos are continuous on \mathbb{R}** . This can be seen from their 'graphical definition', or also from the inequality $|\sin(x) - \sin(y)| \leq |x - y|$. We do not give a formal proof here, as there will be a very simple one later.

4.1 Calculation rules of continuous functions

Next we want to establish some calculation rules for continuous functions. These rules allow to prove continuity of complicated functions by proving that its (hopefully easier) 'building blocks' are continuous.

Theorem 4.12 (Calculation rules for continuous functions). *Let $f, g: D \rightarrow \mathbb{R}$ be continuous at $x_0 \in D$ and $c \in \mathbb{R}$.*

Then $f + g$, $f \cdot g$ and $c \cdot f$ are continuous at x_0 .

If additionally $g(x_0) \neq 0$, then $\frac{f}{g}$ is also continuous at x_0 .

Proof. The theorem follows immediately from Theorem 3.24 about the calculation rules for limits. We only show that $f + g$ is continuous at x_0 , if f and g are continuous in x_0 . The remaining cases can be shown in the same way. Assume that (x_n) is a sequence in D converging to x_0 . By the calculation rules for limits and the continuity of f and g , we obtain

$$\lim_{n \rightarrow \infty} (f + g)(x_n) = \lim_{n \rightarrow \infty} f(x_n) + \lim_{n \rightarrow \infty} g(x_n) = f(x_0) + g(x_0) = (f + g)(x_0).$$

As this was independent of the specific sequence, $f + g$ is continuous at x_0 . □

Let us discuss some examples.

Example 4.13. Let $p(x) = \sum_{k=0}^n c_k \cdot x^k$, with $c_0, \dots, c_n \in \mathbb{C}$, be a **polynomial of degree n** on \mathbb{R} . Then, $p(x)$ is continuous on \mathbb{R} .

Proof. We have already seen that constant functions and the identity are continuous on \mathbb{R} . Applying the above theorem several times we obtain that also x^k , and therefore $c_k x^k$, are continuous on \mathbb{R} . Adding up these terms and applying the theorem again, we get the result. By Theorem 4.65, we see that p is uniformly continuous on every closed interval $D \subset \mathbb{R}$. If D is an (half-)open interval, say $D = (a, b)$, then note that p is uniformly continuous on $[a, b]$, and therefore uniformly continuous on every subset, e.g., on D . □

Example 4.14. Let $p, q: D \rightarrow \mathbb{R}$ be polynomials, and $S := \{x \in D: q(x) = 0\}$. Then, the **rational function** $\frac{p}{q}$ is continuous on $D \setminus S$.

Proof. We already know that p and q are continuous in D from the last example. Furthermore, from the above theorem, $\frac{p}{q}$ is continuous at x_0 whenever $q(x_0) \neq 0$. Since q has no zeros in D we get the result. □

Example 4.15. A function f of the form

$$f(x) = a_0 + \sum_{k=1}^n (a_k \sin(kx) + b_k \cos(kx)),$$

where all $a_k, b_k \in \mathbb{R}$ and $n \in \mathbb{N}$, is called **(real) trigonometric polynomial**. These functions play an important role in *Fourier analysis* and *signal processing*. In the same way as above, we see that f is continuous on \mathbb{R} .

Example 4.16. Another consequence of the above theorem is the continuity of $\tan x$ at all $x \in \mathbb{R}$ with $\cos x \neq 0$, i.e. $x \neq \frac{\pi}{2} + k\pi$ for all $k \in \mathbb{Z}$. This follows from the representation $\tan x = \frac{\sin x}{\cos x}$ and the continuity of \sin and \cos . Analogously, if $x \neq k\pi$ then $\cot x$ is continuous.

Another type of operation of functions we have to discuss is the **composition of functions**. That is, $(g \circ f)(x) := g(f(x))$, i.e., we first apply one function and then the other function to the output. The following theorem is sometimes particularly helpful in proving continuity of complicated functions.

Theorem 4.17. Let $D, E \subset \mathbb{R}$. Moreover, let $f: D \rightarrow E$ be continuous at $x_0 \in D$, and $g: E \rightarrow \mathbb{R}$ be continuous at $y_0 = f(x_0) \in E$. Then, $g \circ f$ is continuous at x_0 .

Proof. Consider an arbitrary sequence (x_n) in D converging to x_0 . Setting $y_n = f(x_n)$ and using continuity of f and g we obtain

$$\lim_{n \rightarrow \infty} g(y_n) = g\left(\lim_{n \rightarrow \infty} y_n\right) = g\left(\lim_{n \rightarrow \infty} f(x_n)\right) = g\left(f\left(\lim_{n \rightarrow \infty} x_n\right)\right) = g(f(x_0)) = g \circ f(x_0).$$

□

Example 4.18. Let $f: D \rightarrow \mathbb{R}$ be a continuous function (on D). Then, $|f|$ is continuous. This easily follows from the composition $g \circ f$ of f with the continuous function $g(x) = |x|$.

Example 4.19. Let $f, g: D \rightarrow \mathbb{R}$ be continuous functions (on D). Then $\min\{f, g\}$ and $\max\{f, g\}$ are also continuous.

Proof. We know the representations

$$\min\{f, g\} = \frac{f + g - |f - g|}{2}$$

and

$$\max\{f, g\} = \frac{f + g + |f - g|}{2}.$$

(This may be shown by case distinction.) The right hand sides are continuous by the above theorem.

□

We now show that also the **inverse of a continuous function** on intervals is continuous. Recall that the inverse only exists for bijective functions.

Theorem 4.20. Let I be an interval and $f: I \rightarrow D \subset \mathbb{R}$ be a bijective function. If f is continuous on I , then the inverse function f^{-1} is continuous on D .

Note that the interval might be open, closed, bounded or unbounded, but it is important that it is a “connected” set.

Proof. Let $y \in D$ be arbitrary and let (y_n) be a sequence converging to y . We want to show that $f^{-1}(y_n)$ converges to $f^{-1}(y)$. For this, we define $x_n := f^{-1}(y_n)$ and $x := f^{-1}(y)$, thus $y_n = f(x_n)$ and $y = f(x)$. Clearly, the sequence (x_n) is contained in the bounded set $[a, b]$. Hence, the Bolzano-Weierstrass theorem (Theorem 3.42) yields that there exists a convergent subsequence (x_{n_k}) , and we set $z := \lim_{k \rightarrow \infty} x_{n_k}$. Using that the continuity of f and that $z \in [a, b]$, we obtain

$$\lim_{k \rightarrow \infty} y_{n_k} = \lim_{k \rightarrow \infty} f(x_{n_k}) = f(z).$$

Since $y_n \rightarrow y$ also implies $\lim_{k \rightarrow \infty} y_{n_k} = y = f(x)$, we get that $f(z) = f(x)$. Since f is bijective, this implies $z = x$. This gives $\lim_{k \rightarrow \infty} x_{n_k} = x$. As this holds for all convergent subsequences of (x_n) , we see that (x_n) has exactly one accumulation point, and is therefore convergent, i.e., $x_n \rightarrow x$. All in all we have shown that

$$f^{-1}(y_n) = x_n \rightarrow x = f^{-1}(y),$$

which concludes the proof. □

This theorem gives an easy argument for the continuity of several known functions.

Example 4.21. We obtain that **root functions** $f(x) = \sqrt[k]{x}$ are continuous on $[0, \infty)$ for arbitrary $k \in \mathbb{N}$. Just note that f is the inverse function of $f^{-1}(x) = x^k$ on $[0, \infty)$, which is bijective, and continuous because it is a polynomial.

In the same way, we obtain that $f(x) = \frac{1}{\sqrt[k]{x}} = x^{-1/k}$ is continuous on $(0, \infty)$ for every $k \in \mathbb{N}$.

Example 4.22. We also obtain that **logarithmic functions** $f(x) = \log_b(x)$ are continuous on $\mathbb{R}_+ := (0, \infty)$ for every $b > 1$.

Again, just note that f is the inverse function of the exponential function $f^{-1}(x) = b^x$ on \mathbb{R} , which is bijective, and continuous, see Example 4.10. (Note that $f: \mathbb{R}_+ \rightarrow \mathbb{R}$, and hence $f^{-1}: \mathbb{R} \rightarrow \mathbb{R}_+$.)

4.2 Limits of functions

By definition, a function $f: D \rightarrow \mathbb{R}$ is continuous at $x_0 \in D$, if $\lim_{n \rightarrow \infty} f(x_n)$ is the same for every sequence (x_n) with $x_n \rightarrow x_0$. We now want to give a shorter notation for this property, and generalize this concept to $x_0 \notin D$. However, this may only work for x_0 that lie 'close' to D . We specify this in the following definition.

Definition 4.23 (Accumulation point of a set). Let $D \subset \mathbb{R}$ be a non-empty set. We call $x_0 \in \mathbb{R} \cup \{-\infty, +\infty\}$ an **accumulation point** of D if there exists a sequence $(x_n)_{n \in \mathbb{N}}$ in D such that

$$x_0 = \lim_{n \rightarrow \infty} x_n \quad \text{and} \quad x_n \neq x_0 \quad \text{for all } n \in \mathbb{N}.$$

An equivalent definition for $x_0 \in \mathbb{R}$ is

$$\forall \varepsilon > 0: (B_\varepsilon(x_0) \setminus \{x_0\}) \cap D \neq \emptyset,$$

i.e., every ε -neighborhood $B_\varepsilon(x_0)$ around x_0 contains a point of D different from x_0 .

Note that we allow also $\pm\infty$ as accumulation points. Clearly, they are accumulation points if D is not bounded (from below/above).

Moreover, note that accumulation points of a set D may not be contained in D , and that not all points of D are accumulation points.

Example 4.24. The set of all accumulation points of (a, b) (or $[a, b]$) with $a, b \in \mathbb{R}$ is the closed interval $[a, b]$.

The set of all accumulation points of (a, ∞) is $[a, \infty) \cup \{\infty\}$.

Example 4.25. The sets \mathbb{N} and \mathbb{Z} do not contain any accumulation points. However, note that \mathbb{N} has the accumulation point ∞ , and \mathbb{Z} has $\pm\infty$ as accumulation points.

Example 4.26. Consider $M = \left\{ \frac{1}{n} : n \in \mathbb{N} \right\}$. Then, 0 is an accumulation point of M , but 0 is not in M . Moreover, 0 is the only accumulation point, since there is no non-constant sequence in M converging to, e.g., $\frac{1}{42}$. This example shows that M and the set of its accumulation points can even be disjoint.

We now can give a precise definition of what we mean by limits of functions.

Definition 4.27 (Limit of functions). Let $D \subset \mathbb{R}$, $f: D \rightarrow \mathbb{R}$, $y \in \mathbb{R}$ and $x_0 \in \mathbb{R} \cup \{-\infty, \infty\}$ be an accumulation point of D . We call y the **limit of f as $x \rightarrow x_0$** , if for arbitrary sequences (x_n) in D such that $x_n \rightarrow x_0$ and $x_n \neq x_0$ for all $n \in \mathbb{N}$, we have

$$\lim_{n \rightarrow \infty} f(x_n) = y.$$

In this case we use the notation

$$\lim_{x \rightarrow x_0} f(x) = y.$$

In the case $y = \pm\infty$ we say that f **tends to $\pm\infty$ as $x \rightarrow x_0$** .

It is important to note that the existence of the limit $\lim_{x \rightarrow x_0} f(x)$ does not depend on the value $f(x_0)$. However, if f is continuous at $x_0 \in D$ and x_0 is an accumulation point, then this limit must still be equal to $f(x_0)$.

Lemma 4.28. Let $D \subset \mathbb{R}$, $f: D \rightarrow \mathbb{R}$ and $x_0 \in D$ be an accumulation point of D . It holds

$$f \text{ is continuous at } x_0 \iff \lim_{x \rightarrow x_0} f(x) = f(x_0).$$

(Verify this yourself!)

Remark 4.29. We call a point $x_0 \in D$ an **isolated point of D** if it is not an accumulation point. That is, the only sequences that converge to an isolated point x_0 are sequences that are eventually constant, i.e., there is some n_0 such that $x_n = x_0$ for all $n \geq n_0$. (Think e.g. of $D = \mathbb{N}$, which consists only of isolated points.) For such sequences we clearly have $f(x_n) \rightarrow f(x_0)$ and therefore, **a function is always continuous at isolated points of its domain**, but $\lim_{x \rightarrow x_0} f(x)$ is not defined, because there is no sequence with $x_n \rightarrow x_0$ and $x_n \neq x_0$ for all n . However, we will mostly talk about sets D without isolated points, like intervals, here.

Let us discuss some examples.

Example 4.30. Consider the function $f(x) = x^3 - x + 1$ on \mathbb{R} . Since f is continuous (as a polynomial), we obtain $\lim_{x \rightarrow a} f(x) = f(a)$ for every $a \in \mathbb{R}$. So, e.g., $\lim_{x \rightarrow 0} f(x) = f(0) = 1$ and $\lim_{x \rightarrow 2} f(x) = f(2) = 7$.

Example 4.31. Let $f: \mathbb{R} \setminus \{0\} \rightarrow \mathbb{R}$ be defined by $f(x) = \frac{1}{x}$. Since f is continuous at all $x \in \mathbb{R} \setminus \{0\}$, we have, e.g., $\lim_{x \rightarrow 1} f(x) = 1$ or $\lim_{x \rightarrow -2} f(x) = -1/2$. Moreover, we have

$$\lim_{x \rightarrow \infty} f(x) = \lim_{x \rightarrow -\infty} f(x) = 0.$$

However, note that $\lim_{x \rightarrow 0} f(x)$ does not exist. For this, consider $x_n := 1/n$ and $y_n = -1/n$, which satisfy $\lim(x_n) = \lim(y_n) = 0$, but $\lim f(x_n) = \infty$ and $\lim f(y_n) = -\infty$.

Example 4.32. Let $f: \mathbb{R} \setminus \{0\} \rightarrow \mathbb{R}$ be defined by $f(x) = \frac{1}{x^2}$. Since f is continuous at all $x \in \mathbb{R} \setminus \{0\}$, we have, e.g., $\lim_{x \rightarrow 1} f(x) = 1$ or $\lim_{x \rightarrow -2} f(x) = 1/4$. Moreover, we have

$$\lim_{x \rightarrow 0} f(x) = \infty \quad \text{and} \quad \lim_{x \rightarrow \infty} f(x) = \lim_{x \rightarrow -\infty} f(x) = 0.$$

So, $\lim_{x \rightarrow x_0} f(x)$ exists for all $x_0 \in \mathbb{R} \cup \{-\infty, \infty\}$.

Example 4.33 (Euler number). Let us consider an interesting limit that we have considered already for sequences, see Example 3.36. Recall that **Euler's number** was defined by

$$e = \lim_{n \rightarrow \infty} \left(1 + \frac{1}{n}\right)^n = \sup_n \left(1 + \frac{1}{n}\right)^n.$$

It is not hard to see that one obtains the same limit for every $x_n \rightarrow \infty$ in place of $x_n = n$, i.e.,

$$e = \lim_{y \rightarrow \infty} \left(1 + \frac{1}{y}\right)^y = \sup \left\{ \left(1 + \frac{1}{y}\right)^y : y > 0 \right\}.$$

(Verify this precisely!)

With this, we can find a useful representation for the powers e^x for $x \in \mathbb{R}$. First note that by continuity of the function $f(y) := y^x$ on $(0, \infty)$, we obtain

$$e^x = \left(\lim_{y \rightarrow \infty} \left(1 + \frac{1}{y}\right)^y \right)^x = \lim_{y \rightarrow \infty} \left(1 + \frac{1}{y}\right)^{xy}$$

For $x > 0$ we use that for every sequence (y_n) we have $y_n \rightarrow \infty \iff z_n := x \cdot y_n \rightarrow \infty$. Hence,

$$e^x = \lim_{y \rightarrow \infty} \left(1 + \frac{1}{y}\right)^{xy} = \lim_{z \rightarrow \infty} \left(1 + \frac{x}{z}\right)^z = \lim_{n \rightarrow \infty} \left(1 + \frac{x}{n}\right)^n.$$

(We used the substitution $z = x \cdot y \iff y = \frac{z}{x}$. The last equality is just taking the special sequence $z_n = n$.) Using $e^x = \frac{1}{e^{-x}}$, we can prove the same equation for $x < 0$. (Verify this!)

We can now follow exactly the same lines as in Example 3.69, using the binomial theorem, to obtain the representations

$$e^x = \lim_{n \rightarrow \infty} \left(1 + \frac{x}{n}\right)^n = \sum_{k=0}^{\infty} \frac{x^k}{k!},$$

which is valid for all $x \in \mathbb{R}$.

Based on what we know about limits and continuous functions in general, we again obtain the following **rules of calculation**.

Lemma 4.34. *Let $f, g: D \rightarrow \mathbb{R}$, and x_0 an accumulation point of D , such that*

$$A := \lim_{x \rightarrow x_0} f(x) \quad \text{and} \quad B := \lim_{x \rightarrow x_0} g(x)$$

with $A, B \in \mathbb{R}$. Then, we have

- (i) $\lim_{x \rightarrow x_0} f(x) \pm g(x) = A \pm B$,
- (ii) $\lim_{x \rightarrow x_0} f(x)g(x) = A \cdot B$, and
- (iii) if $B \neq 0$, then $\lim_{x \rightarrow x_0} \frac{f(x)}{g(x)} = \frac{A}{B}$.

Moreover, if $g: D \rightarrow E$ and $h: E \rightarrow \mathbb{R}$ are such that $y_0 := \lim_{x \rightarrow x_0} g(x) \in E$ and h is continuous at y_0 , then $\lim_{x \rightarrow x_0} h \circ g(x) = h(y_0)$.

We leave the proof as an exercise. (Compare it to Theorem 3.24.)

Example 4.35. Let $f: \mathbb{R} \rightarrow \mathbb{R}$ with $f(x) = 2^{x^2-x+1}$. Then, we can write $f = h \circ g$ with $h(x) = 2^x$ and $g(x) = x^2 - x + 1$. Both functions are continuous on \mathbb{R} and hence, $\lim_{x \rightarrow x_0} f(x) = 2^{x_0^2-x_0+1}$ for every $x_0 \in \mathbb{R}$.

E.g., we have $\lim_{x \rightarrow 2} f(x) = 8$ and $\lim_{x \rightarrow -2} f(x) = 32$. Moreover, for $x_0 = \pm\infty$, we obtain that $\lim_{x \rightarrow \infty} f(x) = \lim_{x \rightarrow -\infty} f(x) = \infty$. (Here, we use that $x \rightarrow \pm\infty$ implies $g(x) \rightarrow \infty$.)

Example 4.36. Let $f: \mathbb{R} \setminus \{0\} \rightarrow \mathbb{R}$ with $f(x) = \sin(1/x)$. Then, we can write $f = h \circ g$ with $h(x) = \sin(x)$ and $g(x) = \frac{1}{x}$. Since h is continuous on \mathbb{R} and g is continuous at every $x_0 \neq 0$ with $y_0 := \lim_{x \rightarrow x_0} g(x) = g(x_0) = \frac{1}{x_0}$, we obtain $\lim_{x \rightarrow x_0} f(x) = h(y_0) = \sin(\frac{1}{x_0})$ for every $x_0 \neq 0$.

So, e.g., $\lim_{x \rightarrow 2} f(x) = \sin(\frac{1}{2})$ or $\lim_{x \rightarrow 1/\pi} f(x) = \sin(\pi) = 0$. Moreover, we have (for $x_0 = \pm\infty$) that $\lim_{x \rightarrow \infty} f(x) = \lim_{x \rightarrow -\infty} f(x) = \sin(0) = 0$.

However, the limit $\lim_{x \rightarrow 0} f(x)$ does not exist. To see this, define the sequence $x_n = \frac{1}{\pi(n+1/2)}$, and note that $x_n \rightarrow 0$, but $f(x_n) = (-1)^n$, which is not convergent, see Figure 23.

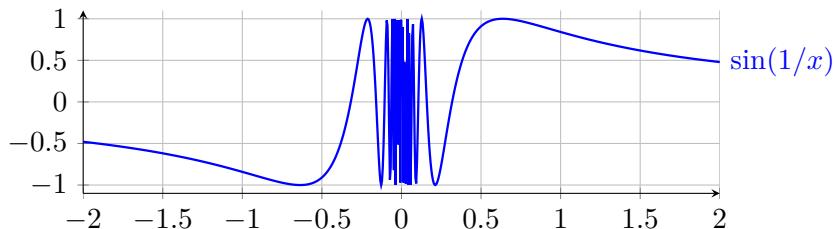


Figure 23: The function $f(x) = \sin(1/x)$

Example 4.37. Let $f: \mathbb{R} \setminus \{0\} \rightarrow \mathbb{R}$ with $f(x) = x^2 \cdot \sin(1/x)$.

Again, all limits at $x_0 \neq 0$ exist. For $x_0 = 0$ the above rules cannot be applied because $\lim_{x \rightarrow 0} \sin(1/x)$ does not exist. However, one might notice that

$$-x^2 \leq x^2 \cdot \sin(1/x) \leq x^2$$

for all $x \neq 0$, since $|\sin(y)| \leq 1$ for all $y \in \mathbb{R}$. Hence, $\lim_{n \rightarrow \infty} |f(x_n)| \leq \lim_{n \rightarrow \infty} x_n^2 = 0$ for every null sequence (x_n) . This implies that $\lim_{x \rightarrow 0} f(x) = 0$, see Figure 24.

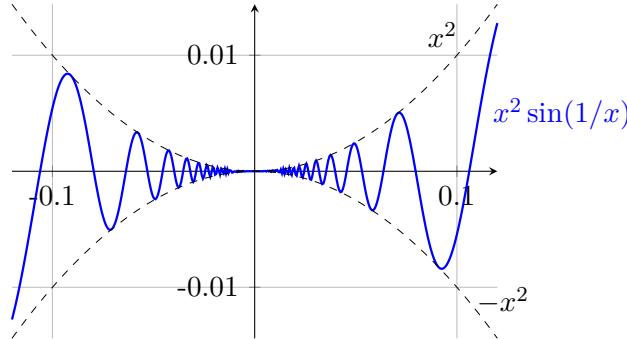


Figure 24: The function $f(x) = x^2 \cdot \sin(1/x)$ with bounds $\pm x^2$

The calculations from the last example should remind you to the application of the *sandwich rule* for limits of sequences, Theorem 3.28. In fact, we have a very similar rule for limits of functions, if two *enclosing functions* have the same limit.

Theorem 4.38 (Sandwich rule for functions). *Let $f, g, h: D \rightarrow \mathbb{R}$ with*

$$f(x) \leq g(x) \leq h(x) \quad \text{for all } x \in D.$$

If x_0 is an accumulation point of D and

$$L := \lim_{x \rightarrow x_0} f(x) = \lim_{x \rightarrow x_0} h(x),$$

then also $\lim_{x \rightarrow x_0} g(x) = L$. (In particular, the limit exists.)

Proof. Consider an arbitrary sequence (x_n) in D converging to x_0 . By assumption, we have $f(x_n) \leq g(x_n) \leq h(x_n)$ for all n . Together with $L := \lim_{n \rightarrow \infty} f(x_n) = \lim_{n \rightarrow \infty} h(x_n)$, Theorem 3.28 implies that $\lim_{n \rightarrow \infty} g(x_n)$ exists, and equals $\lim_{n \rightarrow \infty} g(x_n) = L$. (Note that we used continuity here.) This proves the claim, since $\lim_{n \rightarrow \infty} g(x_n) = L$ holds for every sequence with $x_n \rightarrow x_0$. \square

Most often, we will apply the sandwich rule in the following form.

Corollary 4.39. *Let $f, g: D \rightarrow \mathbb{R}$ with $|f(x)| \leq g(x)$ for all $x \in D$.*

If x_0 is an accumulation point of D and $\lim_{x \rightarrow x_0} g(x) = 0$, then also $\lim_{x \rightarrow x_0} f(x) = 0$. (In particular, the limit exists.)

Another thing that can be seen from the above example is that, sometimes, function are not well-defined at a single point (or even on a set), but we can compute the limit at this point. In such a case, we may “**extend the domain**” of the function using these limits. That is, given a function $f: D \rightarrow \mathbb{R}$ and some accumulation point $x_0 \notin D$ of D (i.e., f is not defined at x_0), such that $y := \lim_{x \rightarrow x_0} f(x)$ exists. Then, we can define the function $g: D \cup \{x_0\} \rightarrow \mathbb{R}$ with

$$g(x) := \begin{cases} f(x) & \text{if } x \neq x_0 \\ y & \text{if } x = x_0. \end{cases}$$

Just by definition, g is continuous at x_0 . That's why g is called a **continuous extension** of f .

Example 4.40. Let $f: \mathbb{R} \setminus \{0\} \rightarrow \mathbb{R}$ with $f(x) = x^2 \cdot \sin(1/x)$.

This function is continuous on $\mathbb{R} \setminus \{0\}$ and we computed in Example 4.37 that $\lim_{x \rightarrow 0} f(x) = 0$. Hence, the function

$$g(x) := \begin{cases} x^2 \cdot \sin(1/x) & \text{if } x \neq 0 \\ 0 & \text{if } x = 0. \end{cases}$$

is continuous on \mathbb{R} .

Example 4.41. Let $D = \mathbb{R} \setminus \{1\}$ and $f: D \rightarrow \mathbb{R}$ be given by

$$f(x) = \frac{x^2 - 1}{x - 1}.$$

Then, f is not well defined in $x_0 = 1$ (since dividing by zero is not allowed). However,

$$\lim_{x \rightarrow 1} \frac{x^2 - 1}{x - 1} = \lim_{x \rightarrow 1} \frac{(x+1)(x-1)}{x-1} = \lim_{x \rightarrow 1} (x+1) = 2.$$

(Note that $\frac{x-1}{x-1} = 1$ can only be used for $x \neq 1$, which holds inside the limit.)

Hence, the function $g: \mathbb{R} \rightarrow \mathbb{R}$ with

$$g(x) = \begin{cases} \frac{x^2 - 1}{x - 1} & \text{if } x \neq 1 \\ 2 & \text{if } x = 1, \end{cases}$$

is continuous on \mathbb{R} . A closer look shows that, in fact, $g(x) = x + 1$ on \mathbb{R} .

Example 4.42. The Heaviside function $H: \mathbb{R} \rightarrow \mathbb{R}$, i.e.,

$$H(x) = \begin{cases} 0 & \text{if } x < 0, \\ 1 & \text{if } x \geq 0. \end{cases}$$

is another example where $\lim_{x \rightarrow 0} H(x)$ does not exist.

However, note that $\lim_{x \rightarrow 0} H(x)$ would exist, if we only allow positive or negative sequences, respectively. This motivates the definition of **one-sided limits** of functions.

Definition 4.43 (One-sided limits). Let $D \subset \mathbb{R}$, $y \in \mathbb{R}$ and $f: D \rightarrow \mathbb{R}$.

- Let x_0 be an accumulation point of $D_+ := D \cap (x_0, \infty)$.

We say that f has a **right-sided limit** y as $x \rightarrow x_0$, if for arbitrary sequences $(x_n) \subset D_+$ with $x_n \rightarrow x_0$ we have that

$$f(x_n) \rightarrow y.$$

We use the notation $\lim_{x \searrow x_0} f(x) = y$ or $\lim_{x \rightarrow x_0^+} f(x) = y$.

- Let x_0 be an accumulation point of $D_- := D \cap (-\infty, x_0)$.

We say that f has a **left-sided limit** y as $x \rightarrow x_0$, if for arbitrary sequences $(x_n) \subset D_-$ with $x_n \rightarrow x_0$ we have that

$$f(x_n) \rightarrow y.$$

We use the notation $\lim_{x \nearrow x_0} f(x) = y$ or $\lim_{x \rightarrow x_0^-} f(x) = y$.

Alternatively, we say that f **approaches** y if $x \rightarrow x_0$ **from the right/left**.

Again, in the case $y = \pm\infty$ we say that f **tends to** $\pm\infty$ as $x \nearrow x_0$ (or $x \searrow x_0$).

Note that the assumption that x_0 is an accumulation point of D_+ just means that there are points in D that are to the right and arbitrarily close to x_0 , i.e., there is a sequence in D_+ converging to x_0 . That x_0 is an accumulation point of D_- means the same 'to the left'.

Example 4.44. Let H be the Heaviside function as above. Then, H is not continuous at 0, but the one-sided limits exist and we have

$$\lim_{x \searrow 0} H(x) = \lim_{x \rightarrow 0^+} H(x) = 1 \quad \text{and} \quad \lim_{x \nearrow 0} H(x) = \lim_{x \rightarrow 0^-} H(x) = 0.$$

Since these limits are different, $\lim_{x \rightarrow 0} H(x)$ does not exist.

Example 4.45. We have a look at the function

$$f(x) = \frac{1}{1 - x^2}$$

on $D = \mathbb{R} \setminus \{\pm 1\}$. We have

$$\lim_{x \searrow 1} \frac{1}{1 - x^2} = -\infty, \quad \lim_{x \nearrow 1} \frac{1}{1 - x^2} = \infty, \quad \lim_{x \rightarrow \infty} \frac{1}{1 - x^2} = \lim_{x \rightarrow -\infty} \frac{1}{1 - x^2} = 0.$$

We see that all one-sided limits of f exist, but the limits at $x_0 = \pm 1$ do not exist.

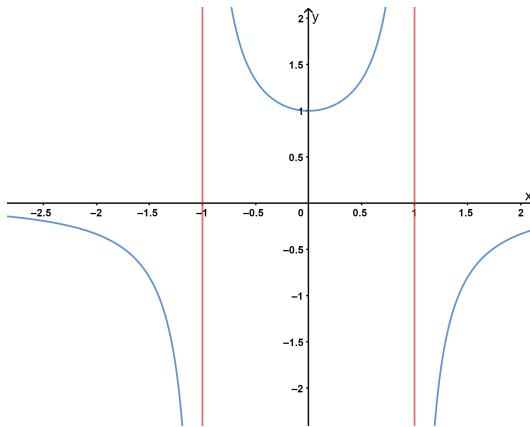


Figure 25: $f(x) = \frac{1}{1-x^2}$

Example 4.46. Let us examine the following function

$$f(x) = \begin{cases} \frac{x-2}{x^2-4} & \text{for } x \neq \pm 2 \\ 0 & \text{for } x = \pm 2 \end{cases}$$

on $D = \mathbb{R}$.

Due to

$$f(x) = \frac{x-2}{x^2-4} = \frac{x-2}{(x-2)(x+2)} = \frac{1}{x+2}$$

for $x \neq -2$ we obtain

$$\lim_{x \rightarrow 2} f(x) = \lim_{x \rightarrow 2} \frac{1}{x+2} = \frac{1}{4}.$$

Hence, f is not continuous at $x = 2$. Moreover, we have $\lim_{x \nearrow -2} f(x) = -\infty$ and $\lim_{x \searrow -2} f(x) = \infty$. So, it is also discontinuous at $x = -2$.

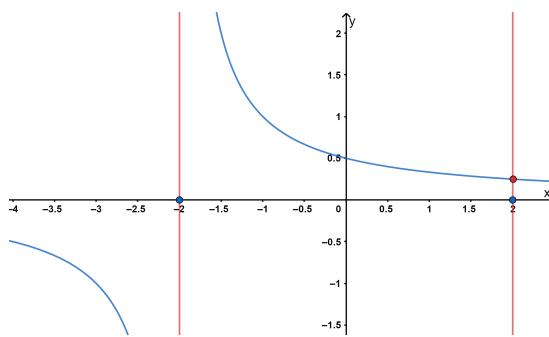


Figure 26: $f(x) = \frac{x-2}{x^2-4}$

We see that one-sided limits can exist although the limit does not exist. In this case, the one-sided limits are still helpful to verify if the (two-sided) limit exists, or even if the function is continuous. This is again an example of a mathematical concept that is just introduced to split a task into two easier ones.

We now prove that $\lim_{x \rightarrow x_0} f(x)$ exists if and only if both one-sided limits exist and are equal.

Theorem 4.47. Let $D \subset \mathbb{R}$, $f: D \rightarrow \mathbb{R}$ and $x_0 \in D$ be an accumulation point of D . Then,

$$\lim_{x \rightarrow x_0} f(x) \text{ exists} \iff \lim_{x \searrow x_0} f(x) \text{ and } \lim_{x \nearrow x_0} f(x) \text{ exist and are equal.}$$

In this case, we have $\lim_{x \rightarrow x_0} f(x) = \lim_{x \nearrow x_0} f(x) = \lim_{x \searrow x_0} f(x)$.

In particular,

$$f \text{ is continuous at } x_0 \iff \lim_{x \nearrow x_0} f(x) = \lim_{x \searrow x_0} f(x) = f(x_0).$$

Proof. First we assume that $\lim_{x \rightarrow x_0} f(x)$ exists. Clearly, right and left handed limits are special cases and therefore exist and $\lim_{x \nearrow x_0} f(x) = \lim_{x \searrow x_0} f(x)$.

For the other direction we assume that right- and left-handed limits in x_0 exist and $\lim_{x \nearrow x_0} f(x) = \lim_{x \searrow x_0} f(x)$. Let (x_n) be an arbitrary sequence such that $x_n \rightarrow x_0$ and $x_n \neq x_0$. We can split (x_n) into two subsequences (y_k^+) and (y_k^-) , where $y_k^+ > x_0$ and $y_k^- < x_0$ for all $k \in \mathbb{N}$. By our assumption we get that

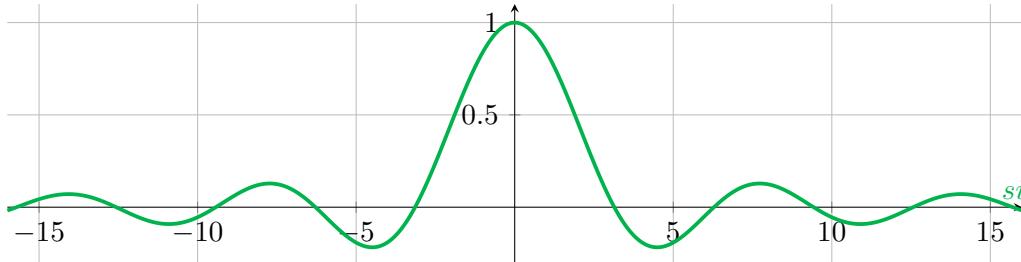
$$\lim_{k \rightarrow \infty} f(y_k^+) = \lim_{k \rightarrow \infty} f(y_k^-).$$

This implies that for arbitrary $\varepsilon > 0$ we can find some $k_0 \in \mathbb{N}$ such that $|f(y_k^+) - f(x_0)| \leq \varepsilon$ and $|f(y_k^-) - f(x_0)| \leq \varepsilon$ for all $k \geq k_0$. Consequently, there is some $n_0 \in \mathbb{N}$ such that $|f(x_n) - f(x_0)| \leq \varepsilon$ for all $n \geq n_0$.

□

Let us finally consider another interesting example, which is one of the most important limits related to trigonometric functions. We consider the function $\text{si}: \mathbb{R} \rightarrow \mathbb{R}$ with

$$\text{si}(x) := \begin{cases} \frac{\sin x}{x}, & \text{if } x \neq 0, \\ 1, & \text{if } x = 0. \end{cases}$$

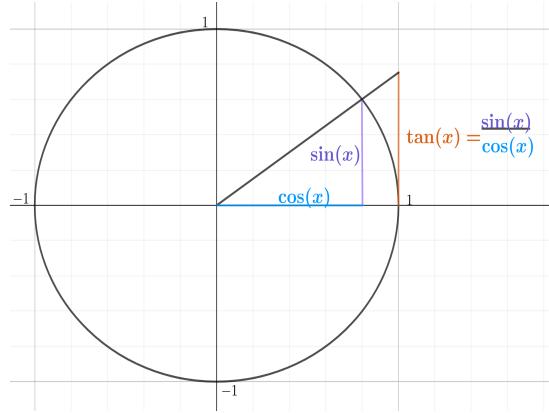


This function is called the **sinus cardinalis** and is clearly well defined for all $x \neq 0$. We will prove now that si is a continuous function on \mathbb{R} .

Example 4.48. It remains to show that

$$\lim_{x \rightarrow 0} \frac{\sin x}{x} = 1.$$

Proof. As \sin is even, i.e. $\sin(-x) = \sin(x)$, it is sufficient to show that $\lim_{x \searrow 0} \frac{\sin x}{x} = 1$. Now, recall the definition of \sin , \cos and \tan on the unit circle, i.e., the circle with radius 1 and center at the origin, see the following figure.



We see that the area of the enclosed circular sector with angle $x \in [0, \frac{\pi}{2})$, which equals $\frac{x}{2}$, is larger than the area of the triangle with legs $\sin x, \cos x$, but smaller than the area of the triangle with legs $\tan x, 1$. Using the corresponding area formulas we obtain

$$\frac{1}{2} \sin x \cos x \leq \frac{1}{2} x \leq \frac{1}{2} \tan x.$$

This is equivalent to

$$\sin x \cos x \leq x \leq \tan x = \frac{\sin x}{\cos x},$$

and therefore to

$$\cos x \leq \frac{\sin x}{x} \leq \frac{1}{\cos x}.$$

(Really try to prove this inequality from the one before!)

We know that $\cos x$ is continuous and $\lim_{x \rightarrow 0} \cos x = \cos(0) = 1$. Using the sandwich rule we obtain

$$1 \leq \lim_{x \rightarrow 0} \frac{\sin x}{x} \leq 1.$$

□

4.3 Intermediate and extreme value theorem

We now discuss two important properties of continuous functions. Both are (or at least look) obvious for 'easy' functions. However, we show that they hold under the weak assumption that the function is continuous and defined on a closed interval.

The first will be the *Intermediate value theorem*, which states that a continuous function on a closed interval attains all values between the function values at the endpoints of the interval.

Theorem 4.49 (Intermediate value theorem). *Let $I = [a, b]$ be a closed interval and $f: I \rightarrow \mathbb{R}$ be a continuous function. Then, for every $y \in \mathbb{R}$ with*

$$\min\{f(a), f(b)\} \leq y \leq \max\{f(a), f(b)\},$$

there exists some $x \in I$ such that

$$f(x) = y.$$

As this theorem holds for every $y \in J := [\min\{f(a), f(b)\}, \max\{f(a), f(b)\}]$, this theorem states that the image of I under f , i.e. $f(I)$, at least contains this interval, i.e. $f(I) \supseteq J$, if I is a closed interval, see Figure 27.

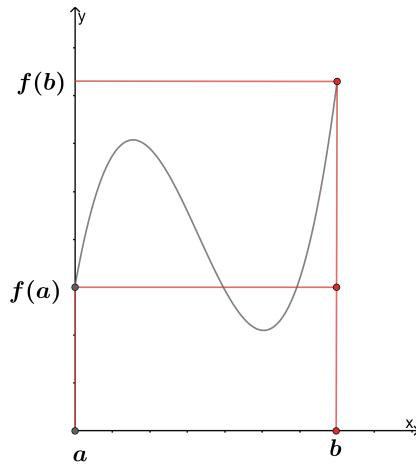


Figure 27: Every value between $f(a)$ and $f(b)$ is attained

Proof. The case $f(a) = f(b)$ is obvious. Now assume w.l.o.g. that $f(a) < f(b)$. The case $f(a) > f(b)$ can be proven in the same way by replacing f by $-f$.

Now let $y \in [f(a), f(b)]$. We will define a sequence (x_n) with $x_n \rightarrow x \in I$ and $f(x) = y$.

First, let $a_1 = a$, $b_1 = b$ and $x_1 = \frac{a_1+b_1}{2}$, i.e., x_1 is the midpoint between a and b .

If $f(x_1) \geq y$, then we set $[a_2, b_2]$ to be the 'left half' of $[a_1, b_1]$. If on the other hand $f(x_1) < y$, we set $[a_2, b_2]$ to be the 'right half' of $[a_1, b_1]$. In both cases we get that $[a_2, b_2] \subset [a_1, b_1]$ and $f(a_2) \leq y \leq f(b_2)$. We iterate this process.

Let's make that more formal: For $n \in \mathbb{N}$ and given a_n, b_n such that $f(a_n) \leq y \leq f(b_n)$, we define

$$x_n := \frac{a_n + b_n}{2},$$

i.e., the midpoint of $[a_n, b_n]$, and set

$$\begin{aligned} a_{n+1} &:= a_n, \\ b_{n+1} &:= x_n, \end{aligned}$$

in the case that $f(x_n) \geq y$. If $f(x_n) < y$, we set

$$\begin{aligned} a_{n+1} &:= x_n, \\ b_{n+1} &:= b_n. \end{aligned}$$

With this, we get two sequences (a_n) and (b_n) with $[a_n, b_n] \subset [a, b]$ and $f(a_n) \leq y \leq f(b_n)$ for all $n \in \mathbb{N}$, and

$$a = a_1 \leq a_2 \leq \cdots \leq a_n \leq \cdots \leq b_n \leq \cdots \leq b_2 \leq b_1 = b.$$

This yields that $(a_n), (b_n)$ are monotone and bounded sequences, which are therefore convergent. Moreover, we have $b_n - a_n = \frac{|a-b|}{2^{n-1}}$, because we always halve the interval, and get that

$$\lim_{n \rightarrow \infty} a_n = \lim_{n \rightarrow \infty} b_n =: x.$$

We clearly have $x \in [a, b]$. (Why?) Using $f(a_n) \leq y \leq f(b_n)$ for all n , we obtain

$$f(x) = \lim_{n \rightarrow \infty} f(a_n) \leq y \leq \lim_{n \rightarrow \infty} f(b_n) = f(x).$$

Hence

$$f(x) = y.$$

□

An important special case is the following corollary, which is sometimes called *Bolzano's theorem*.

Corollary 4.50. Let $f: [a, b] \rightarrow \mathbb{R}$ with $f(a) < 0 < f(b)$.

Then, f has at least one zero in (a, b) , i.e., there is some $t \in (a, b)$ with $f(t) = 0$.

Example 4.51. The intermediate value theorem yields an alternative argument for the existence of arbitrary positive roots: For $a > 0$ and $n \in \mathbb{N}$, let $f: \mathbb{R}_{\geq 0} \rightarrow \mathbb{R}$ be given by $f(x) = x^n - a$. As a polynomial, f is continuous. Furthermore, we have $f(0) = -a < 0$ and $f(1+a) = (1+a)^n - a > 0$. The intermediate value theorem then yields some $x \in [0, 1+a]$ with $f(x) = 0$, i.e. $x^n = a$ or $x = \sqrt[n]{a}$, respectively.

Example 4.52. The intermediate value theorem also yields the existence of real valued zeros of polynomials with an odd degree. Therefore let $n \in \mathbb{N}$ be odd and $p: \mathbb{R} \rightarrow \mathbb{R}$ a polynomial of degree n given by

$$p(x) = a_n x^n + a_{n-1} x^{n-1} + \dots + a_1 x + a_0$$

with $a_n \neq 0$. We assume w.l.o.g. that $a_n > 0$. Clearly we have $\lim_{x \rightarrow \infty} p(x) = +\infty$ and $\lim_{x \rightarrow -\infty} p(x) = -\infty$. Consequently, there exist $a, b \in \mathbb{R}$ with $a < b$ and $p(a) < 0, p(b) > 0$. The intermediate value theorem yields then a $x \in (a, b)$ with $p(x) = 0$.

Example 4.53. The next application refers to **fixed points**, i.e., points that are not changed by a function. Let $f : [0, 1] \rightarrow [0, 1]$ be continuous. Then, there exists a $x \in [0, 1]$ with $f(x) = x$. For the proof we look at the continuous function $g(x) = f(x) - x$. We have $g(0) = f(0) - 0 \geq 0$ and $g(1) = f(1) - 1 \leq 0$, and the intermediate value theorem then yields a $x \in [0, 1]$ with $g(x) = 0$, so we have $f(x) = x$.

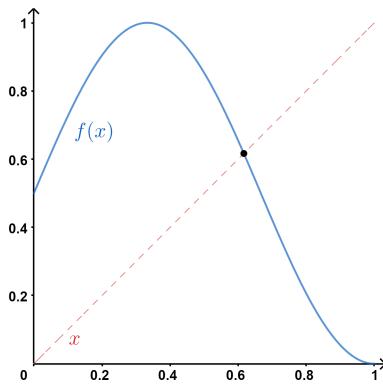


Figure 28: Fixed point of a function

We now want to discuss the *extreme value theorem*, which states that a continuous function on a closed interval attains also its minimal and maximal value.

Let us start with the definition of minimal and maximal points of a function.

We use the term *extremum* if we do not specify if it is a minimum or maximum. Moreover, we call them *global extrema*, as we will later discuss also a *local* variant of this concept.

Definition 4.54. Let D be any set and $f: D \rightarrow \mathbb{R}$.

Then, f has a **(global) minimum at $x_0 \in D$** if

$$f(x) \geq f(x_0) \quad \text{for all } x \in D,$$

and f has a **(global) maximum at $x_0 \in D$** if

$$f(x) \leq f(x_0) \quad \text{for all } x \in D.$$

The point x_0 is called **(global) minimum/maximum point**, or **global extreme point**.

The value $f(x_0)$ is called **minimum/maximum**, or, collectively, **extreme values**.

Note that the minimum/maximum (value) of a function, if it exists, is unique. However, there might still be more than one minimum/maximum point.

Example 4.55. Consider the function $f(x) = x^2$ on $[-1, 1]$.

This function has a unique global minimum point at 0 with minimum value 0, and two global maximum points at -1 and 1 with maximum value 1. Note that the same function on the open interval $(-1, 1)$ does not have (global) maxima.

We now turn to the important *extreme value theorem*.

Theorem 4.56 (Extreme value theorem). *Let $I = [a, b] \subset \mathbb{R}$ be a closed interval and $f: I \rightarrow \mathbb{R}$ be a continuous function. Then there exist $x_{\min}, x_{\max} \in I$ such that*

$$\begin{aligned} f(x_{\min}) &= \inf_{x \in I} f(x) := \inf \{f(x): x \in I\}, \\ f(x_{\max}) &= \sup_{x \in I} f(x) := \sup \{f(x): x \in I\}. \end{aligned}$$

In other words, continuous functions attain their extreme values on closed intervals.

This theorem shows that the infimum $\inf_{x \in I} f(x)$ and supremum $\sup_{x \in I} f(x)$ are attained at some points in I . In fact, **infimum and supremum are actually minimum and maximum**.

Proof. We only show that f attains its maximal value, the other case can be treated similarly. Recall that $f(I) = \{y \in \mathbb{R}: f(x) = y \text{ for some } x \in I\}$. Since I is non-empty, $f(I)$ is non-empty and therefore, $S := \sup f(I)$ exists (the case $S = \infty$ is still allowed here). By the properties of suprema there exists a sequence (y_n) in $f(I)$ such that $y_n \rightarrow S$. Furthermore there exists a sequence (x_n) in I such that $f(x_n) = y_n$. This sequence is bounded, since all $x_n \in I$, i.e., $a \leq x_n \leq b$. By the Bolzano-Weierstrass theorem (Theorem 3.42) there exists a convergent subsequence (x_{n_k}) which converges to some $x_0 \in \mathbb{R}$. Now, by the sandwich rule (Theorem 3.28) and $a \leq x_{n_k} \leq b$, we also obtain that $a \leq x_0 \leq b$, i.e., $x_0 \in I$. The definition of continuity yields

$$S = \lim_{k \rightarrow \infty} y_{n_k} = \lim_{k \rightarrow \infty} f(x_{n_k}) = f(x_0),$$

which proves the claim with $x_{\max} := x_0$ (and implies $S < \infty$). □

Remark 4.57. Note that **it is important that we have a closed interval** in the extreme value theorem. For open intervals the statement is not true in general. For instance, if we consider $f(x) = x$ (we already know that this is a continuous function) and $I = (a, b) = (0, 1)$, then

$$\sup_{x \in I} f(x) = 1 \quad \inf_{x \in I} f(x) = 0,$$

but clearly $0 \notin f(I)$ and $1 \notin f(I)$.

Example 4.58. Another example that shows that the extreme value theorem does not hold in general for open intervals is the function $f(x) = \frac{1}{x}$ on $(0, 1)$. This function is continuous, but unbounded and therefore does not have a maximum.

If we combine both theorems we obtain the following:

- The **extreme value theorem** shows that $m := \inf_{x \in I} f(x)$ and $M := \sup_{x \in I} f(x)$ are attained at some points $x_{\min}, x_{\max} \in I$.
- The **intermediate value theorem** (applied to the interval $[x_{\min}, x_{\max}]$) implies that all intermediate values are attained.
- In short: $f(I) = [\min_{x \in I} f(x), \max_{x \in I} f(x)]$.

- In particular, **continuous functions map closed intervals to closed intervals.**

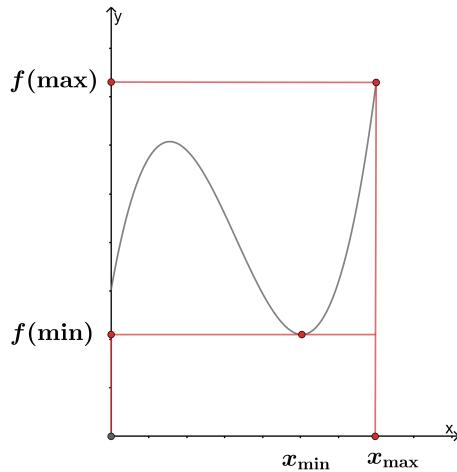


Figure 29: All values between the extreme values $f(x_{\min})$ and $f(x_{\max})$ are attained

One essential corollary (which might appear obvious), is the following.

Corollary 4.59. *Let $I = [a, b] \subset \mathbb{R}$ be a closed interval and $f: I \rightarrow \mathbb{R}$ be a continuous function. Then, there exists some $R \in \mathbb{R}$ such that*

$$|f(x)| \leq R \quad \text{for all } x \in I.$$

In words, continuous functions on closed intervals are bounded, i.e., $\sup_{x \in I} |f(x)| < \infty$.

Try to prove that yourself!

We finally want to discuss shortly how one can actually find a point $x^* \in [a, b]$ such that $f(x^*) = y$ for y with $f(a) \leq y \leq f(b)$. (The intermediate value theorem implies that it exists.) For the sake of simplicity we only discuss the case that $y = 0$. (One might consider $g(x) = f(x) - y$ otherwise.) The proof of the intermediate value theorem leads directly to a (practical) algorithm for the approximation of x^* with $f(x^*) = 0$ that is called the **bisection method**.

Let $f: [a, b] \rightarrow \mathbb{R}$ be continuous with $f(a) < 0 < f(b)$. The bisection method is inductively defined as follows:

1. Start with $k := 1$, $a_1 := a$, and $b_1 := b$.
2. Compute the midpoint $x_k := \frac{1}{2}(a_k + b_k)$.
3. If $f(x_k) > 0$, we set $a_{k+1} := a_k$ and $b_{k+1} := x_k$,
4. otherwise $a_{k+1} := x_k$ and $b_{k+1} := b_k$.
5. Set $k \rightarrow k + 1$ and continue with Step 2.

We always have $a_k, b_k \in [a, b]$ and $f(a_k) < 0 < f(b_k)$, i.e., there is always a zero x^* in the interval $[a_k, b_k]$, and we have

$$|x_k - x^*| \leq \frac{1}{2}(b_k - a_k) = 2^{-k}(b - a).$$

The iteration is stopped if (by chance) $f(x_k) = 0$ holds for some k or if the upper bound is smaller than a prescribed $\varepsilon > 0$. x_k is then used as an “ ε -approximation” for the zero x^* .

Example 4.60. The positive zero of the function $f(x) = x^2 - 2$ is clearly $x^* = \sqrt{2}$. We start the bisection method with $a = 0$ and $b = 2$ and want to achieve an error $|x_k - x^*|$ that is smaller than $\varepsilon = 0.05$.

The requirements of the bisection method are fulfilled, since f is continuous and $f(0) < 0 < f(2)$. To satisfy the prescribed error bound $\varepsilon > 0$, we need that

$$2^{-k}(2 - 0) < \varepsilon \iff k > \log_2\left(\frac{1}{\varepsilon}\right) + 1.$$

For $\varepsilon = 0.05$, we can choose $k = 6$. In fact, for $k = 6$ the bounds shows that the error will be at most $2^{-5} = \frac{1}{32} = 0.03125$. Let us have a look on the first iterations:

k	a_k	b_k	x_k	$f(x_k)$
1	0	2	1	-1
2	1	2	$\frac{3}{2}$	$\frac{1}{4}$
3	1	$\frac{3}{2}$	$\frac{5}{4}$	$-\frac{7}{16}$
4	$\frac{5}{4}$	$\frac{3}{2}$	$\frac{11}{8}$	$-\frac{7}{64}$
5	$\frac{11}{8}$	$\frac{3}{2}$	$\frac{23}{16}$	$\frac{17}{256}$
6	$\frac{11}{8}$	$\frac{23}{16}$	$\frac{45}{32}$	$-\frac{23}{1024}$

The actual error of $x_6 = \frac{45}{32} \approx 1.40625$ to the exact zero $x^* = \sqrt{2} \approx 1.414214$ is, however, only about 0.007964.

4.4 Other types of continuity

Let us introduce two stronger forms of continuity of functions that will be useful later. Note that both are defined for the whole domain, and not at single points.

Definition 4.61. Let $f: D \rightarrow \mathbb{R}$ be a real function. We call f **uniformly continuous on D** if for all sequences $(x_n), (y_n) \subset D$ with $\lim_{n \rightarrow \infty} |x_n - y_n| = 0$ we have that

$$\lim_{n \rightarrow \infty} |f(x_n) - f(y_n)| = 0.$$

By the above considerations, we immediately see that constant functions and the identity $f(x) := x$ are uniformly continuous (Check yourself!), and we will show in Theorem 4.70 that every uniformly continuous function is continuous. To see that uniform continuity is indeed a stronger condition, consider the following example.

Example 4.62. Let $f(x) := x^2$ on \mathbb{R} , which is continuous by Example 4.7. Moreover, consider the sequences defined by $x_n = n + 1/n$ and $y_n = n$, which clearly satisfy $|x_n - y_n| = \frac{1}{n} \rightarrow 0$. However, we have

$$|f(x_n) - f(y_n)| = \left| \left(n + \frac{1}{n} \right)^2 - n^2 \right| = \left| n^2 + 2 + \frac{1}{n^2} - n^2 \right| \rightarrow 2.$$

This shows that f is not uniformly continuous on \mathbb{R} .

As discussed in Example 4.7, the difference in proving continuity of, e.g., the functions x and x^2 , was that we had to choose δ depending on x_0 in the latter case. We will now see that uniformly continuous functions are precisely those, where we may find a δ independent of x_0 in an “ ε - δ -proof” of continuity. (Clearly, such a δ still depends on ε in general.)

Theorem 4.63 (ε - δ -criterion for uniform continuity). *Let $f: D \rightarrow \mathbb{R}$. Then, f is uniformly continuous if and only if*

$$\forall \varepsilon > 0 \exists \delta > 0 \forall x, y \in D: |x - y| < \delta \implies |f(x) - f(y)| < \varepsilon.$$

In words: For all (fixed) $\varepsilon > 0$ there exists $\delta > 0$ such that for all $x, y \in D$ with $|x - y| < \delta$ we have that $|f(x) - f(y)| < \varepsilon$.

From this equivalent definition of uniform continuity, one may already see that it is indeed a stronger condition than continuity. For this, note the additional “*for all*” quantifier in the statement. For a better understanding, try to write the definition of “ f is continuous on D (i.e., for all $x_0 \in D$)” solely with quantifiers, and spot the difference.

Proof. First we prove that the ε - δ -criterion for uniform continuity implies uniform continuity. Therefore assume that

$$\forall \varepsilon > 0 \exists \delta > 0 \forall x, y \in D: |x - y| < \delta \implies |f(x) - f(y)| < \varepsilon,$$

and let $(x_n), (y_n)$ be sequences in D such that $|x_n - y_n| \rightarrow 0$.

Now, fix some $\varepsilon_0 > 0$, so we can find $\delta_0 > 0$ such that

$$\forall x, y \in D: |x - y| < \delta_0 \implies |f(x) - f(y)| < \varepsilon_0.$$

Furthermore, since $|x_n - y_n| \rightarrow 0$, we can find $n_0 \in \mathbb{N}$ such that

$$\forall n \geq n_0: |x_n - y_n| < \delta_0.$$

Hence

$$\forall n \geq n_0: |f(x_n) - f(y_n)| < \varepsilon_0.$$

Since ε_0 was arbitrary, we obtain that $\lim_{n \rightarrow \infty} |f(x_n) - f(y_n)| = 0$. Since also (x_n) and (y_n) were arbitrary, we get that f is uniformly continuous on D from the definition.

The other direction is proved by contradiction. Therefore, we assume

$$\exists \varepsilon_0 > 0 \forall \delta > 0 \exists x, y \in D: |x - y| < \delta \text{ and } |f(x) - f(y)| \geq \varepsilon_0,$$

and we want to show that this implies that f is not uniformly continuous, i.e., that there exist sequences $(x_n), (y_n)$ in D such that $|x_n - y_n| \rightarrow 0$ and $|f(x_n) - f(y_n)| \not\rightarrow 0$. For this, let $\delta_m := \frac{1}{m}$ for all $m \in \mathbb{N}$. From the assumption (with $\delta = \delta_m$), we can find $x_m, y_m \in D$ such that $|x_m - y_m| < \frac{1}{m}$ and $|f(x_m) - f(y_m)| \geq \varepsilon_0$. Thus, we found sequences $(x_m), (y_m)$ such that

$$\forall m \in \mathbb{N}: |x_m - y_m| < \frac{1}{m} \text{ and } |f(x_m) - f(y_m)| \geq \varepsilon_0,$$

i.e., $|x_n - y_n| \rightarrow 0$ and $|f(x_n) - f(y_n)| \not\rightarrow 0$. This finishes the proof. \square

Example 4.64. Try to prove yourself, using the ε - δ -criterion, that the absolute value function, i.e. $f(x) = |x|$, is uniformly continuous on \mathbb{R} .

Uniform continuity is an essential tool for many of the following considerations, in the same way as absolute convergence was essential for series. But uniform continuity is sometimes not so easy to show. The following theorem shows that, however, if a continuous function is defined on a **closed** interval, then it is automatically uniformly continuous.

Theorem 4.65. Let $a, b \in \mathbb{R}$ with $a < b$. A continuous function $f: [a, b] \rightarrow \mathbb{R}$ on a closed interval $[a, b]$ is uniformly continuous.

Together with Theorem 4.70 (see below), this shows that continuity and uniform continuity are just the same for functions defined on closed intervals.

Proof. We prove the result by contradiction, so we assume that there exist sequences (x_n) and (y_n) such that

$$|x_n - y_n| \rightarrow 0 \quad \text{and} \quad \forall n \in \mathbb{N}: |f(x_n) - f(y_n)| \geq \varepsilon_0,$$

for some $\varepsilon_0 > 0$, i.e. that f is not uniformly continuous. Since $[a, b]$ is bounded, and $(x_n) \subset [a, b]$, we have that (x_n) is bounded, and therefore, that there exists a convergent subsequence of (x_n) , say (x_{n_k}) with $x_{n_k} \rightarrow x_0$. This is a consequence of the Bolzano-Weierstrass theorem. Using triangle inequality we see

$$|x_0 - y_{n_k}| \leq |x_0 - x_{n_k}| + |x_{n_k} - y_{n_k}| \rightarrow 0,$$

so (y_{n_k}) also converges to x_0 . The continuity of f and $|\cdot|$ yields

$$\lim_{k \rightarrow \infty} |f(x_{n_k}) - f(y_{n_k})| = \left| f\left(\lim_{k \rightarrow \infty} x_{n_k}\right) - f\left(\lim_{k \rightarrow \infty} y_{n_k}\right) \right| = |f(x_0) - f(x_0)| = 0,$$

which contradicts $|f(x_{n_k}) - f(y_{n_k})| \geq \varepsilon_0$. Hence, f must be uniformly continuous. \square

We come back to a known example.

Example 4.66. Let $f(x) = x^2$ on $D := [a, b]$ for some $a < b$. We know from Example 4.7 that f is continuous on \mathbb{R} , and therefore also on every $D \subset \mathbb{R}$. Moreover, we know from Example 4.62 that f is not uniformly continuous on \mathbb{R} . However, by Theorem 4.65, we see that f is uniformly continuous on every closed interval.

This can also be proven directly by the same considerations as in the proof of Example 4.7. We obtain that, for $\delta = \min\{1, \frac{\varepsilon}{1+2c}\}$ with $c := \max\{|a|, |b|\}$, we have $|f(x) - f(y)| < \varepsilon$ for all $x, y \in [a, b]$ with $|x - y| < \delta$, proving the claim.

The last type of continuity we want to discuss is *Lipschitz continuity*. This one is the strongest, but also the easiest to verify, concept one may consider. Luckily, it is enough to deal with this type for most practical applications.

Definition 4.67. Let $f: D \rightarrow \mathbb{R}$. We call f **Lipschitz continuous** if there exists some $L > 0$ such that

$$\forall x, y \in D: |f(x) - f(y)| \leq L|x - y|.$$

The constant L is called **Lipschitz constant**.

Example 4.68. Again, the constant function $f(x) := c$ and the linear function $f(x) := x$, are Lipschitz continuous. In both cases we can choose the Lipschitz constant $L = 1$. (For the constant function, the Lipschitz constant may be chosen arbitrarily small.)

Example 4.69. It is also not hard to show that $f(x) := x^2$ on D is Lipschitz continuous on arbitrary bounded $D \subset \mathbb{R}$. Check yourself!

Theorem 4.70. Let $f: D \rightarrow \mathbb{R}$. Then,

$$f \text{ is Lipschitz continuous} \implies f \text{ is uniformly continuous} \implies f \text{ is continuous}$$

Proof. First we show that Lipschitz continuous functions are uniformly continuous. For arbitrary $\varepsilon > 0$ we set $\delta = \frac{\varepsilon}{L}$, where L is the Lipschitz constant of f . Using the ε - δ -criterion for uniform continuity we obtain

$$\forall x, y \in D: |x - y| < \delta \implies |f(x) - f(y)| < L \cdot \delta = \varepsilon.$$

As this holds for all $\varepsilon > 0$, the ε - δ -criterion for uniform continuity implies that f is uniformly continuous.

Now assume that f is uniformly continuous on D and let $x_0 \in D$. The uniform continuity yields

$$\forall \varepsilon > 0 \exists \delta > 0 \forall x \in D: |x - x_0| < \delta \implies |f(x) - f(x_0)| < \varepsilon.$$

Hence f is continuous on D . □

This theorem shows that Lipschitz continuity implies the two other forms of continuity we have just discussed. However, one may still ask if some of these concepts are actually the same. We have already seen that the function $f(x) = x^2$ is not uniformly continuous on \mathbb{R} , which shows that uniform continuity is indeed stronger than continuity. The following example shows that a function may be continuous on a closed interval (and therefore uniformly continuous), but not Lipschitz continuous. That is, we do not have the reverse implications in Theorem 4.70, i.e.

$$\text{Lipschitz continuous} \not\iff \text{Uniformly continuous} \not\iff \text{Continuous}.$$

Example 4.71. Let $f: [0, 1] \rightarrow \mathbb{R}$ with $x \mapsto \sqrt{x}$. Then f is uniformly continuous but not Lipschitz continuous.

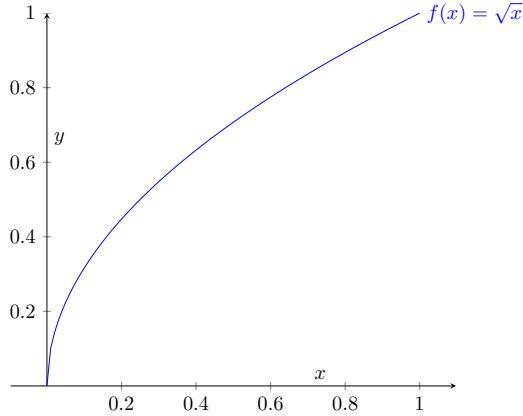


Figure 30: The function $f(x) = \sqrt{x}$

Proof. First we show that f is uniformly continuous on $[0, 1]$. Again, we use the ε - δ -criterion for the purpose of demonstration. Therefore, let $\varepsilon > 0$ and $x, y \in [0, 1]$. The binomial theorem implies

$$|\sqrt{x} - \sqrt{y}|^2 = x + y - 2\sqrt{xy} \leq x + y - 2\min\{x, y\} = |x - y|.$$

This shows that $|\sqrt{x} - \sqrt{y}| \leq |x - y|^{1/2}$. If we now choose $\delta = \varepsilon^2$, we see that $|\sqrt{x} - \sqrt{y}| < \varepsilon$ for all x, y with $|x - y| < \delta$. This shows that f is uniformly continuous on $[0, 1]$.

We now show that, however, f is not Lipschitz continuous. Multiplying and dividing by $|\sqrt{x} + \sqrt{y}|$ (and the binomial theorem) yield

$$|\sqrt{x} - \sqrt{y}| = \frac{|x - y|}{|\sqrt{x} + \sqrt{y}|}.$$

If f would be Lipschitz continuous, then there would exist some $L > 0$ such that

$$|\sqrt{x} - \sqrt{y}| = \frac{|x - y|}{|\sqrt{x} + \sqrt{y}|} < L|x - y|$$

for all $x, y \in [0, 1]$. However, this would mean $L > \frac{1}{|\sqrt{x} + \sqrt{y}|}$ for all $x, y \in [0, 1]$, which is clearly not true, as the right hand side can be made arbitrary large by choosing x and y small enough. Hence f cannot be Lipschitz continuous. □

Example 4.72. Let us finally mention, without formal proof, that the **trigonometric functions sin and cos are Lipschitz continuous on \mathbb{R}** with Lipschitz constant $L = 1$. This can be seen from their 'graphical definition'. See also Figure 31 which shows that all function values lie in the colored cone. The trigonometric functions are therefore also uniformly continuous.

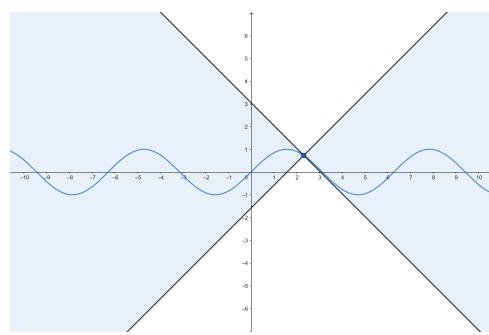


Figure 31: Lipschitz continuity of sin function

5 Differential calculus

In this chapter we want to introduce and study derivatives of real-valued functions which give us a better understanding of how small local changes of the input effect the output of a function. For functions defined on the real line, as will assume throughout this chapter, one may think about the slope (german: *Anstieg*) of the tangent line attached to a point of the graph of the function. As already for continuity, this is a good intuition and, when well understood, makes many of the upcoming results obvious for 'easy' functions. However, we need again a precise definition of a derivative to handle cases where a visualization is not possible or helpful.

Prominent application of the differential calculus are an alternative (and more precise) definition of minima and maxima of a function, and the derivative provides us with information of whether a function is increasing or decreasing in a given point. However, there is much more information 'hidden' in the values of the higher order derivatives at a point. In fact, under certain assumptions on the function, a function can be given approximately in a neighborhood of the point just by knowing some of these values. This will be formalized by means of the *Taylor polynomial*. Finally, we present the very useful *l'Hospital rule*, which is a powerful tool to compute complicated limits.

Let us begin with a precise definition of a derivative of a function, which is clearly only a precise notion (using limits of function) of the slope of a tangent line in a point.

Definition 5.1. Let $I = (a, b)$ and $f: I \rightarrow \mathbb{R}$. We call f **differentiable at $x_0 \in I$** if

$$\lim_{h \rightarrow 0} \frac{f(x_0 + h) - f(x_0)}{h} = \lim_{x \rightarrow x_0} \frac{f(x) - f(x_0)}{x - x_0} \quad \text{exists.}$$

In this case we call this limit **derivative of f at x_0** , and write $f'(x_0)$ or $\frac{d}{dx} f(x_0)$ or $\frac{df}{dx}(x_0)$.

If f is differentiable at every point of I , we call f differentiable (in I) and denote by f' or $\frac{d}{dx} f$ the derivative of f .

From now on we mostly consider functions that are defined on an open interval I . This is because the endpoints of an interval need sometimes more care. (Moreover, the derivative is clearly not defined at isolated points, if the domain of definition contains some.) We comment on the differences when needed.

The expression

$$\frac{f(x_0 + h) - f(x_0)}{h}$$

is called **difference quotient**. Geometrically this is the slope of the secant through the points $(x_0, f(x_0))$ and $(x_0 + h, f(x_0 + h))$. Hence if $f'(x_0)$ exists, it is the slope of the tangent or the function at the point x_0 .

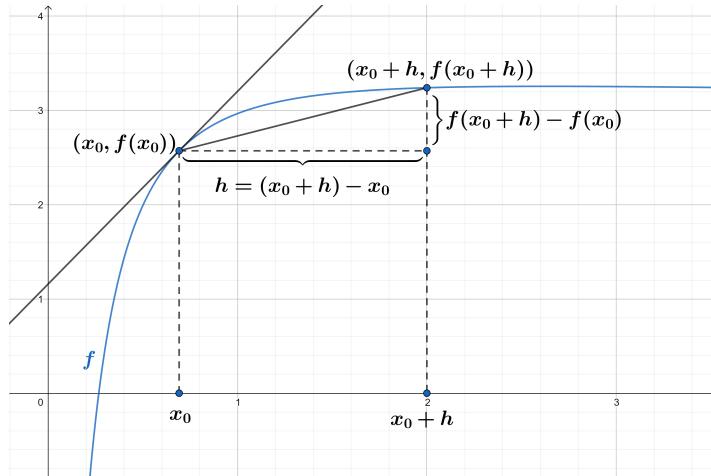


Figure 32: Difference quotient

Obviously, constant functions, i.e., $f(x) = c$ for $c \in \mathbb{R}$, have the derivative $f' = 0$. Let us discuss some more examples which will serve as building blocks for more complicated functions.

Example 5.2. Let $f(x) = x^n$, with $n \in \mathbb{N}$. Then f is differentiable on \mathbb{R} with $f'(x) = nx^{n-1}$.

Proof. We use the binomial theorem to compute

$$\frac{f(x+h) - f(x)}{h} = \frac{(x+h)^n - x^n}{h} = nx^{n-1} + \binom{n}{2}x^{n-2}h + \binom{n}{3}x^{n-3}h^2 + \dots + \binom{n}{n}h^{n-1}.$$

For fixed $x \in \mathbb{R}$, we then compute the limit as $h \rightarrow 0$ and end up with

$$\lim_{h \rightarrow 0} \frac{f(x+h) - f(x)}{h} = nx^{n-1}.$$

□

Example 5.3. Let $f(x) = \frac{1}{x^n}$, $n \in \mathbb{N}$. Then f is differentiable on $\mathbb{R} \setminus \{0\}$ with $f'(x) = -n\frac{1}{x^{n+1}}$. Together with the last example, we have $(1)' = 0$ and $(x^k)' = kx^{k-1}$ for all $k \in \mathbb{Z} \setminus \{0\}$. We will see that this formula in fact holds for all $k \in \mathbb{R} \setminus \{0\}$.

Proof. We want to compute $f'(x)$ for $x \neq 0$. Therefore we use the binomial theorem to obtain

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

Observe that the denominator converges to x^{2n} for $h \rightarrow 0$. To take the limit into the denominator, we have to assume $x \neq 0$. In the numerator, all terms in the latter sum go to zero for $h \rightarrow 0$. Therefore,

$$f'(x) = \lim_{h \rightarrow 0} \frac{h \sum_{k=1}^{n-1} \binom{n}{k+1} x^{n-k} h^k}{x^n \sum_{k=1}^n \binom{n}{k} x^{n-k} h^k} = \frac{-nx^{n-1}}{x^{2n}} = -nx^{-n-1}.$$

□

Example 5.4. Let $f(x) = \sin(x)$. Then f is differentiable on \mathbb{R} and it holds $f'(x) = \cos(x)$, i.e.,

$$\sin'(x) = \cos(x) \quad \text{for all } x \in \mathbb{R}.$$

Proof. With the help of the trigonometric addition formula

$$\sin(x) - \sin(y) = 2 \cos\left(\frac{x+y}{2}\right) \sin\left(\frac{x-y}{2}\right)$$

we can compute that

$$\frac{\sin(x+h) - \sin x}{h} = \frac{2 \cos\left(x + \frac{h}{2}\right) \sin \frac{h}{2}}{h} = \cos\left(x + \frac{h}{2}\right) \frac{\sin \frac{h}{2}}{\frac{h}{2}}.$$

Using $\frac{\sin(x)}{x} \rightarrow 1$ as $x \rightarrow 0$, see Example 4.48 and the continuity of \cos we obtain the result as $h \rightarrow 0$.

□

Example 5.5. In the same way, one can compute that \cos is differentiable on \mathbb{R} and

$$\cos'(x) = -\sin(x),$$

where we us

$$\cos(x) - \cos(y) = -2 \sin\left(\frac{x+y}{2}\right) \sin\left(\frac{x-y}{2}\right).$$

We now turn to the important exponential function $x \mapsto e^x =: \exp(x)$, which is defined by

$$e^x = \lim_{n \rightarrow \infty} \left(1 + \frac{x}{n}\right)^n = \sum_{k=0}^{\infty} \frac{x^k}{k!},$$

To compute its derivative, we need an inequality, which is also of independent interest.

Lemma 5.6. *For all $x < 1$ we have*

$$1 + x \leq e^x \leq \frac{1}{1-x}.$$

The lower bound holds for all $x \in \mathbb{R}$.

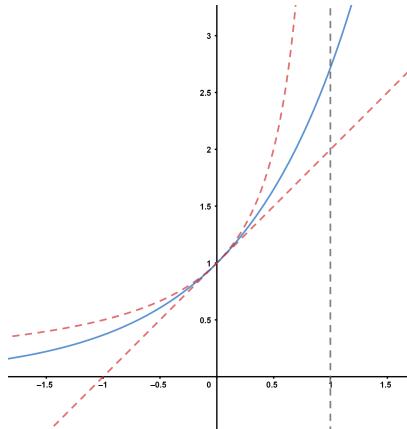


Figure 33: Inequality from Lemma 5.6

Proof. First, recall that *Bernoulli's inequality* (Theorem 1.48) states that

$$(1+y)^n \geq 1+ny \quad \text{for } y \geq -1.$$

With $y = \frac{x}{n}$, we obtain that $(1+\frac{x}{n})^n \geq 1+x$ for all $n \geq |x|$, and hence $e^x \geq 1+x$.

To bound e^x from above, note that the above computation (with $y = -\frac{x}{n}$) shows that $e^{-x} \geq 1-x$. Hence, for $x < 1$, we obtain $e^x \leq \frac{1}{1-x}$. \square

With this, we can compute the derivative of e^x

Example 5.7. Let $f(x) = e^x$, then f is differentiable on \mathbb{R} and $f'(x) = f(x)$, i.e.,

$$(e^x)' = \frac{d}{dx} e^x = e^x.$$

Proof. Again, we compute the difference quotient

$$\frac{f(x+h) - f(x)}{h} = \frac{e^{x+h} - e^x}{h} = e^x \frac{e^h - 1}{h}.$$

From Lemma 5.6 we obtain

$$1+h \leq e^h \leq \frac{1}{1-h}$$

for $h < 1$. This implies

$$\frac{1}{1+|h|} \leq \frac{e^h - 1}{h} \leq \frac{1}{1-|h|}$$

for $|h| < 1$. (Prove this by case distinction.) Hence, $\frac{e^h - 1}{h} \rightarrow 1$ which implies

$$e^x \frac{e^h - 1}{h} \rightarrow e^x.$$

\square

Remark 5.8. Note that calculating a derivative is in fact the same as calculating limits of a function. Therefore, to show that a function is not differentiable at a point x_0 , it is sufficient to find two sequences (h_n) and (\tilde{h}_n) such that $h_n \rightarrow 0$ and $\tilde{h}_n \rightarrow 0$, but

$$\lim_{n \rightarrow \infty} \frac{f(x+h_n) - f(x)}{h_n} \neq \lim_{n \rightarrow \infty} \frac{f(x+\tilde{h}_n) - f(x)}{\tilde{h}_n}.$$

A typical example is the following.

Example 5.9. Let $f(x) = |x|$, then f is not differentiable in 0.

Proof. Set $h_n = \frac{1}{n}$ and $\tilde{h}_n = \frac{-1}{n}$. It follows that

$$1 = \lim_{n \rightarrow \infty} \frac{f(x + h_n) - f(x)}{h_n} \neq \lim_{n \rightarrow \infty} \frac{f(x + \tilde{h}_n) - f(x)}{\tilde{h}_n} = -1.$$

□

This shows that a continuous function is not necessarily differentiable, however, the reverse statement holds.

Theorem 5.10. *Let $f: (a, b) \rightarrow \mathbb{R}$ be differentiable at x_0 . Then, f is continuous at x_0 .*

Proof. By assumption,

$$\lim_{x \rightarrow x_0} \frac{f(x) - f(x_0)}{x - x_0} = f'(x_0).$$

By the calculation rules for limits we get

$$\lim_{x \rightarrow x_0} (f(x) - f(x_0)) = \lim_{x \rightarrow x_0} (x - x_0) \frac{f(x) - f(x_0)}{x - x_0} = \left(\lim_{x \rightarrow x_0} (x - x_0) \right) f'(x_0) = 0.$$

Thus, $\lim_{x \rightarrow x_0} f(x) = f(x_0)$, i.e. f is continuous.

□

5.1 Calculation rules for differentiable functions

As in the previous sections we want to establish some rules of calculation for differentiable functions that will be the main tools to derive the derivatives of complicated functions. In particular, we will establish rules for calculating the derivative of products, quotients and compositions of functions, as well as of the inverse.

If one is not already completely confident with these rules of calculation, it is highly recommended to practice and calculate derivatives of increasingly complicated functions. Note that with the rules that follow, it should be no problem to calculate the derivative of $\sin(x^{42}) e^{\cos(\sqrt{x})}$ step-by-step, although it might be time-consuming.

Theorem 5.11 (Linearity of derivatives). *Let f, g be differentiable at x_0 . Then,*

- $(f + g)'(x_0)$ exists and $(f + g)'(x_0) = f'(x_0) + g'(x_0)$, and
- for any $c \in \mathbb{R}$ we have $(c \cdot f)'(x_0) = c \cdot f'(x_0)$.

Proof. Both follow easily from the calculation rules for limits.

□

Example 5.12. We already know that $(x^n)' = nx^{n-1}$ for all $n \in \mathbb{N}$. By linearity, all polynomials are differentiable on \mathbb{R} and

$$(c_n x^n c_{n-1} x^{n-1} + \dots + c_1 x + c_0)' = c_n n x^{n-1} + c_{n-1} (n-1) x^{n-2} + \dots + c_2 2 x + c_1.$$

Next we have a look at the product of differentiable functions.

Theorem 5.13 (Product rule). *Let f, g be differentiable at x_0 , then $(fg)'(x_0)$ exists and*

$$(fg)'(x_0) = f'(x_0)g(x_0) + g'(x_0)f(x_0).$$

In short, $(fg)' = f'g + g'f$.

Proof. We compute

$$\begin{aligned} (fg)'(x) &= \lim_{h \rightarrow 0} \frac{f(x+h)g(x+h) - f(x)g(x)}{h} \\ &= \lim_{h \rightarrow 0} f(x+h) \frac{g(x+h) - g(x)}{h} + \lim_{h \rightarrow 0} \frac{f(x+h) - f(x)}{h} g(x) \\ &= f(x)g'(x) + f'(x)g(x), \end{aligned}$$

where we use that differentiable functions are continuous. □

The next rule is for the composition of two functions.

Theorem 5.14 (Chain rule). *Let $f: I \rightarrow J$ and $g: J \rightarrow \mathbb{R}$ such that f is differentiable at x_0 and g is differentiable at $f(x_0)$. Then, $(g \circ f)'(x_0)$ exists and*

$$(g \circ f)'(x_0) = g'(f(x_0)) f'(x_0).$$

In short, $(g \circ f)' = (g' \circ f) \cdot f'$.

Proof. We set $y_0 = f(x_0)$ and have a look at the function

$$h(y) := \begin{cases} \frac{g(y) - g(y_0)}{y - y_0} & \text{if } y \neq y_0 \\ g'(y_0) & \text{if } y = y_0. \end{cases}$$

This function is continuous in y_0 and we see that

$$g(y) - g(y_0) = h(y)(y - y_0)$$

for all $y \in J$. This yields, with $y = f(x)$, that

$$\begin{aligned} \lim_{x \rightarrow x_0} \frac{g(f(x)) - g(f(x_0))}{x - x_0} &= \lim_{x \rightarrow x_0} \frac{h(f(x))(f(x) - f(x_0))}{x - x_0} \\ &= \lim_{x \rightarrow x_0} h(f(x)) \lim_{x \rightarrow x_0} \frac{f(x) - f(x_0)}{x - x_0} \\ &= h(y_0)f'(x_0) = g'(y_0)f'(x_0). \end{aligned}$$

Again, we have used that $g \circ f$ is continuous. □

We are now able to consider also quotients of functions.

Theorem 5.15 (Quotient rule). *Let f, g be differentiable at x_0 , and assume that $g(x_0) \neq 0$. Then, $\frac{f}{g}$ is differentiable at x_0 and*

$$\left(\frac{f}{g}\right)'(x_0) = \frac{f'(x_0)g(x_0) - f(x_0)g'(x_0)}{g(x_0)^2}.$$

$$\text{In short, } \left(\frac{f}{g}\right)' = \frac{f'g - fg'}{g^2}.$$

Proof. We want to use the product rule for the functions f and $\frac{1}{g}$. But first we compute $\left(\frac{1}{g}\right)'$. Recall that $\left(\frac{1}{x}\right)' = \frac{-1}{x^2}$ for $x \neq 0$, and that $\frac{1}{g}$ can be written as $h \circ g$, where $h(y) = \frac{1}{y}$. The chain rule yields

$$\left(\frac{1}{g}\right)'(x_0) = -\frac{g'(x_0)}{g(x_0)^2}.$$

Using the product rule, we obtain

$$\left(\frac{f}{g}\right)'(x_0) = \frac{f'(x_0)g(x_0) - f(x_0)g'(x_0)}{g(x_0)^2}.$$

□

These rules allow to compute complicated derivatives by computing several easy derivatives.

Example 5.16. We want to compute the derivative of the tangent function $\tan: (-\frac{\pi}{2}, \frac{\pi}{2}) \rightarrow \mathbb{R}$. By the quotient rule, and $\sin^2 x + \cos^2 x = 1$, we obtain

$$\frac{d}{dx} \tan x = \frac{d}{dx} \left(\frac{\sin x}{\cos x} \right) = \frac{\sin' x \cos x - \sin x \cos' x}{\cos^2 x} = \frac{\cos^2 x + \sin^2 x}{\cos^2 x} = \frac{1}{\cos^2 x} = 1 + \tan^2 x.$$

Example 5.17. The calculation rules for derivatives and the already proven fact that trigonometric functions and (algebraic) polynomials are differentiable, implies that trigonometric polynomials are differentiable on \mathbb{R} .

There is also a formula for differentiating inverse functions. Note that the assumption that a continuous function f is strictly monotone on an interval I is just equivalent to f mapping *bijektivly* to another interval $J \subset \mathbb{R}$.

Theorem 5.18. *Let $f: I \rightarrow \mathbb{R}$ be strictly monotone and continuous. If f is differentiable at $x_0 \in I$, and $f'(x_0) \neq 0$, then f^{-1} is differentiable at $y_0 = f(x_0)$ and*

$$\left(f^{-1}\right)'(y_0) = \frac{1}{f'(x_0)} = \frac{1}{f'(f^{-1}(y_0))}.$$

$$\text{In short, } (f^{-1})' = \frac{1}{f' \circ f^{-1}}.$$

Proof. By Theorem 4.20, we know that f^{-1} is continuous (in y_0). Take an arbitrary sequence $(y_n) \subset f(I)$ with $y_n \rightarrow y_0$ and $y_n \neq y_0$, and define $x_n := f^{-1}(y_n)$. We obtain

$$x_n = f^{-1}(y_n) \rightarrow f^{-1}(y_0) = x_0,$$

as well as $x_n \neq x_0$. Therefore,

$$\lim_{n \rightarrow \infty} \frac{f^{-1}(y_n) - f^{-1}(y_0)}{y_n - y_0} = \lim_{n \rightarrow \infty} \frac{x_n - x_0}{f(x_n) - f(x_0)} = \lim_{n \rightarrow \infty} \frac{1}{\frac{f(x_n) - f(x_0)}{x_n - x_0}} = \frac{1}{f'(x_0)},$$

where we use $f'(x_0) \neq 0$ in the last equality. \square

Example 5.19. We compute the derivative of the natural logarithm $\ln(y)$, $y > 0$. This is the inverse of the function $f(x) = e^x = y$ with derivative $f'(x) = e^x$, see Example 5.7. Thus

$$\frac{d}{dy} \ln(y) = \frac{1}{f'(x)} = \frac{1}{y}.$$

Example 5.20. We use the derivative of $\ln(x)$ and the chain rule to prove that

$$\frac{d}{dx} x^a = ax^{a-1} \quad \text{for arbitrary } a \in \mathbb{R} \text{ and } x > 0.$$

For this, write $x^a = e^{a \ln(x)}$. With $f(x) = a \ln(x)$ and $g(x) = e^x$, we obtain $f'(x) = \frac{a}{x}$ and $g'(x) = e^x$, which yields

$$\frac{d}{dx} x^a = \frac{d}{dx} (g \circ f)(x) = g'(f(x))f'(x) = e^{a \ln(x)} \frac{a}{x} = ax^{a-1}.$$

Example 5.21. Now consider $f(x) = \tan(x)$ for $-\frac{\pi}{2} < x < \frac{\pi}{2}$. We can compute the derivative of the inverse function $\arctan = \tan^{-1}: \mathbb{R} \rightarrow (-\frac{\pi}{2}, \frac{\pi}{2})$ by using the above theorem and obtain

$$\frac{d}{dy} \arctan y = \frac{1}{f'(x)} = \frac{1}{1 + \tan^2 x} = \frac{1}{1 + y^2},$$

where we use Example 5.16 and, again, set $y := f(x)$.

Example 5.22. One can also use the theorem to calculate the derivatives of the inverses of the trigonometric functions, see Section 1.7, to obtain

$$\arcsin'(y) = \frac{1}{\sin'(x)} = \frac{1}{\cos(x)} = \frac{1}{\sqrt{1 - \sin^2(x)}} = \frac{1}{\sqrt{1 - y^2}} \quad \text{with } y = \sin(x),$$

and

$$\arccos'(y) = \frac{1}{\cos'(x)} = \frac{-1}{\sin(x)} = \frac{-1}{\sqrt{1 - \cos^2(x)}} = \frac{-1}{\sqrt{1 - y^2}} \quad \text{with } y = \cos(x),$$

for all $y \in (-1, 1)$. (Note that we used $\sin^2 x + \cos^2 x = 1$.)

5.2 Global and local extrema

In this section we discuss the most important application for derivatives. That is, the calculation and classification of *extreme points*, i.e., points where a function attains a (local) maximum or minimum. Let us first restate the definition of *global extrema*, see Definition 4.54.

Definition 5.23. Let D be any set and $f: D \rightarrow \mathbb{R}$.

Then, f has a **(global) minimum at $x_0 \in D$** if

$$f(x) \geq f(x_0) \quad \text{for all } x \in D,$$

and f has a **(global) maximum at $x_0 \in D$** if

$$f(x) \leq f(x_0) \quad \text{for all } x \in D.$$

The point x_0 is called **(global) minimum/maximum point**, or **global extreme point**.

The value $f(x_0)$ is called **minimum/maximum**, or, collectively, **extreme value**.

Again, note that the extreme values of a function, if they exist, are unique, but there might be several extreme points. (Think about $f(x) = x^2$ on $[-1, 1]$ or $(-1, 1)$, see Example 4.55)

As we want to discuss the connection of extreme points to the derivative of a function, which is a *local* property, we also need a local notion of extreme points.

Definition 5.24. Let $D \subset \mathbb{R}$ and $f: D \rightarrow \mathbb{R}$.

Then, f has a **local minimum at $x_0 \in D$** if there exists $\varepsilon > 0$ such that

$$f(x) \geq f(x_0) \quad \text{for all } x \in D \cap (x_0 - \varepsilon, x_0 + \varepsilon),$$

and a **strict local minimum** if $f(x) > f(x_0)$ for all $x \in D \cap (x_0 - \varepsilon, x_0 + \varepsilon) \setminus \{x_0\}$.

Analogously, we say f has a **local maximum at $x_0 \in D$** if there exists $\varepsilon > 0$ such that

$$f(x) \leq f(x_0) \quad \text{for all } x \in D \cap (x_0 - \varepsilon, x_0 + \varepsilon),$$

and a **strict local maximum** if $f(x) < f(x_0)$ for all $x \in D \cap (x_0 - \varepsilon, x_0 + \varepsilon) \setminus \{x_0\}$.

The point x_0 is called **local maximum/minimum point**, or **local extreme point**.

It follows immediately that a **global extreme point has to be a local extreme point**. However, as we have seen in the last example, there might be already several global extrema and clearly even more local extrema, see Figure 34.

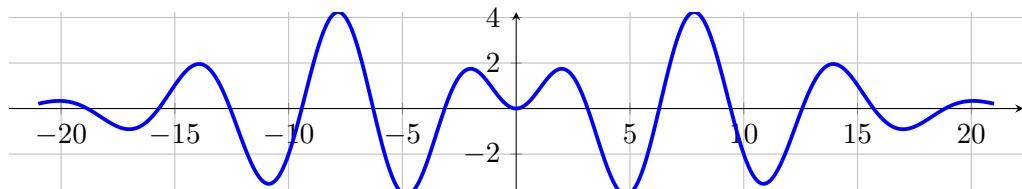


Figure 34: Graph of $x \cdot \sin(x) \cdot e^{-\frac{x^2}{100}}$

We will now discuss how to find (and classify) local extrema by using derivatives. One possible way of **finding a global extremum** of a function $f: [a, b] \rightarrow \mathbb{R}$ then is:

1. Find all local extreme points, say t_1, \dots, t_k .
2. Calculate the function values $f(t_i)$, as well as $f(a)$ and $f(b)$.
3. The largest/smallest value corresponds to the maximum/minimum.

If the function is defined on an open (or unbounded) interval (a, b) with $-\infty \leq a < b \leq \infty$, then calculating $f(a)$ and $f(b)$ in Step 2 should be replaced by calculating the **boundary values** $\lim_{x \rightarrow a} f(x)$ and $\lim_{x \rightarrow b} f(x)$. Clearly, if they lead to the maximal/minimal values, then there exists no global maximum/minimum point, as it is not attained in the domain.

We now turn to step one, i.e., finding all local extrema. For this, note that the above figure shows that the slope of the tangent line attached to an extremum is zero, i.e., the tangent line is horizontal. This means that the derivative at such a point is zero. We now show that this is a necessary condition if the function is differentiable, which means that it is enough to 'check' all points with this condition if you want to find a local extremum.

Theorem 5.25 (Necessary condition for an extreme point). *Let $I = (a, b)$ and $f: I \rightarrow \mathbb{R}$. If $x_0 \in I$ is a local extreme point of f and f is differentiable at x_0 , then*

$$f'(x_0) = 0.$$

We call $x_0 \in I$ with $f'(x_0) = 0$ a **stationary point** of f .

In particular, if a function $f: I \rightarrow \mathbb{R}$ is differentiable, but satisfies $f'(x) \neq 0$ for all $x \in I$, then it cannot have local extreme points in I . Maximum and minimum values can in this case only be attained at the 'boundary' of the interval.

We see that the only possible **candidates for extreme points** inside the domain of a function are the stationary points and the points where f is not differentiable (if they exist). We call these points the **critical points** of f , and the corresponding function values the **critical values**.

We therefore only have to calculate all critical values and the boundary values of a function to determine its supremum and infimum, and to decide if they are (global) maximum and minimum.

Note that, $f'(x_0) = 0$ is not a sufficient condition for having an extreme point at x_0 , i.e., there are functions with $f'(x_0) = 0$, but x_0 is no local maximum/minimum point. Let us consider $f(x) = x^3$ on \mathbb{R} , see Figure 35. Clearly, f has no maximum or minimum at 0, but $f'(0) = 3 \cdot 0^2 = 0$.

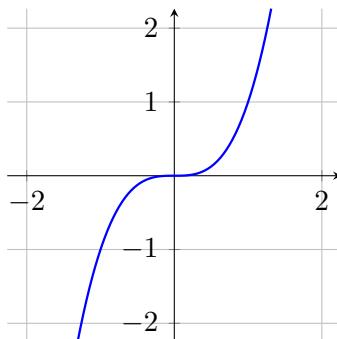


Figure 35: The graph of $f(x) = x^3$

Proof of Theorem 5.25. Let f have a local minimum at x_0 and let $\varepsilon > 0$ be as in the definition required. We now have for $x \in I$ with $x_0 < x < x_0 + \varepsilon$, which implies $x - x_0 > 0$, that

$$\frac{f(x) - f(x_0)}{x - x_0} \geq 0.$$

So, $\lim_{x \searrow x_0} \frac{f(x) - f(x_0)}{x - x_0} \geq 0$. On the other hand for $x_0 - \varepsilon < x < x_0$ we obtain

$$\frac{f(x) - f(x_0)}{x - x_0} \leq 0,$$

thus $\lim_{x \nearrow x_0} \frac{f(x) - f(x_0)}{x - x_0} \leq 0$. Since $f'(x_0)$ exists, we necessarily have $f'(x_0) = 0$. The case of local maxima can be treated similarly. \square

In many cases, it is already enough to determine all stationary points in order to find all (global) extreme points of a function. But only from the function and derivative value at a point, we can not decide if a stationary point is indeed an extreme point and, if it is, if we face a local maximum or a local minimum. This is in particular a problem, if we are not able to draw the function under consideration. However, there is a sufficient condition for having an extrema that involves **higher-order derivatives**.

Let $f: (a, b) \rightarrow \mathbb{R}$ be differentiable and let f' be also differentiable (at x). Then we say that f is **twice differentiable** (at x) and write

$$f''(x) := \frac{d}{dx} \left(\frac{d}{dx} f(x) \right) = \frac{d}{dx} f'(x).$$

This procedure can be repeated as long as derivatives exist and so we can define the **n -th derivative** of f inductively by

$$f^{(n)}(x) := \frac{d^n}{dx^n} f(x) = \frac{d}{dx} f^{(n-1)}(x).$$

In the special case of $n = 2$ or $n = 3$ we write $f''(x)$ or $f'''(x)$, respectively. If the n -th derivative of f at a point exists, then we say that f is **n -times differentiable** at this point. If the n -th derivative of f (at x_0) exists and is a continuous function (at this point), then we say that f is **n -times continuously differentiable** (at x_0).

Example 5.26. Let $f(x) := x^n$ for some $n \in \mathbb{N}$.

By the already discussed rules we obtain $f'(x) = nx^{n-1}$, $f'(x) = n(n-1)x^{n-2}$ or, in general, $f^{(k)}(x) = n(n-1) \cdots (n-k+1) x^{n-k}$ for all $k \leq n$. In particular, $f^{(n)}(x) = n!$ for all $x \in \mathbb{R}$. Since the derivative of the constant function ($n = 0$) is the zero function, we additionally see that $f^{(k)}(x) = 0$ for all $k > n$.

Example 5.27. We now consider $f(x) := x^a$ for arbitrary $a \in \mathbb{R}$, see Example 5.20.

If $a \in \mathbb{N}_0$, we are back in the situation of the last example. That is, the formula for the higher-order derivatives leads to $f^{(k)} \equiv 0$ (i.e., $f^{(k)}(x) = 0$ for all $x \in I$) for all $k > a$.

This does not hold if a is negative or not a natural number, i.e.,

$$\frac{d^k}{dx^k} x^a = a(a-1) \cdots (a-k+1) x^{a-k} \quad \text{for arbitrary } a \in \mathbb{R} \setminus \mathbb{N}_0, x > 0 \text{ and } k \in \mathbb{N}.$$

Example 5.28. Note that all differentiable functions that we discussed so far, were also continuously differentiable, i.e., they possess a continuous derivative (on the whole domain). In fact, it is not easy to find a **differentiable function that is not continuously differentiable**. The classical example of such a function is

$$f(x) := \begin{cases} x^2 \sin(1/x), & \text{if } x \neq 0, \\ 0, & \text{if } x = 0. \end{cases}$$

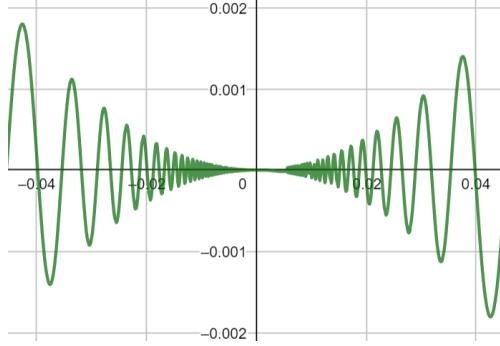


Figure 36: $f(x) = x^2 \sin(\frac{1}{x})$

First of all, by using the bound $|f(x)| \leq |x|^2$, we obtain that f is continuous. By the definition of the derivative at $x_0 = 0$, we obtain that $f'(0) = \lim_{h \rightarrow 0} h \sin(1/h) = 0$, where we again use the boundedness of \sin . For $x_0 \neq 0$ we use the calculation rules for derivatives, and we obtain

$$f'(x) = \begin{cases} 2x \sin(1/x) - \cos(1/x), & \text{if } x \neq 0, \\ 0, & \text{if } x = 0. \end{cases}$$

Hence, the function f is differentiable on \mathbb{R} .

However, f' is not continuous. To see this, note that the limit $\lim_{x \rightarrow 0} f'(x)$ does not exist. (Consider e.g. the sequence $x_n = \frac{1}{n\pi} \rightarrow 0$, which leads to $\cos(1/x_n) = (-1)^n$.) □

We now turn to a sufficient condition that allows to deduce if a stationary point (i.e. $f'(x_0) = 0$) is a local extreme point, and, moreover, if it is a local maximum or a local minimum.

Theorem 5.29 (Second derivative test). *Let $I = (a, b)$, $x_0 \in I$ and $f: I \rightarrow \mathbb{R}$ be twice continuously differentiable at x_0 , i.e., f'' exists and is continuous at x_0 . Moreover, assume that x_0 is a stationary point of f , i.e., $f'(x_0) = 0$. Then,*

$$f''(x_0) > 0 \implies f \text{ has a strict local minimum at } x_0,$$

and

$$f''(x_0) < 0 \implies f \text{ has a strict local maximum at } x_0.$$

If $f''(x_0) = 0$, we do not gain any information about the possible extremum.

Although we could prove this theorem here, using a rather longish reasoning, we will present a very short argument at the end of the following section, see the Proof of Theorem 5.29.

Note that a precise characterization of an extremum would also involve higher-order derivatives and is rather complicated. As the above is usually enough, we do not state the characterization here. However, note that we will learn later that the knowledge of all derivatives $f^{(k)}(x_0)$, $k \in \mathbb{N}_0$, at a given point x_0 , if they exist, allows us to *reconstruct* the function exactly in a neighborhood around x_0 . That is, we can obtain all 'local information' of a function from its higher-order derivative values, if the function can be differentiated infinitely often.

Example 5.30. The function $f(x) = 3x^2 - 6x + 5$ (on \mathbb{R}) satisfies $f'(x) = 6x - 6$ and $f''(x) = 6$, see Figure 37. It therefore has a unique critical point at $x_0 = 1$ which is a local minimum. As it is the only extreme point, and $\lim_{x \rightarrow \pm\infty} f(x) = \infty$, we obtain that f is not bounded from above, i.e., f does not have a maximum, and f has a global minimum at $x_0 = 1$ with minimum value $f(x_0) = 2$.

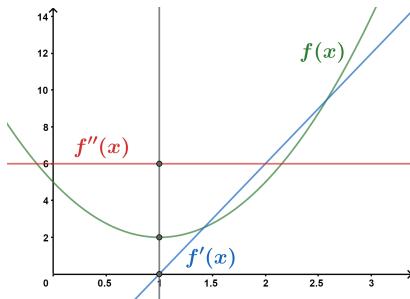


Figure 37: Derivatives of $f(x) = 3x^2 - 6x + 5$

Example 5.31. If $f''(x_0) = 0$, then we can not say if x_0 is an extreme point and, if it is, which one. To see this, consider, e.g., $f(x) = x^3$, $g(x) = x^4$ and $h(x) = -x^4$ on \mathbb{R} , and $x_0 = 0$.

Then, $f'(0) = f''(0) = g'(0) = g''(0) = h'(0) = h''(0) = 0$, but f has no extrema at 0, g has a maximum at 0 and h has a minimum at 0.

5.3 Mean value theorem and l'Hospital's rule

In this section we discuss the mean value theorem, which is a theorem that assures that the derivative of a function attains a certain value, if we only know the function values at the endpoints of the interval. In the same way as the other 'existence theorems', which were ultimately all due to the Bolzano-Weierstrass theorem 3.42, this will be an important tool in the proofs of the upcoming theorems.

In particular, it will imply *l'Hospital's rule* for calculating certain limits, which may be seen as a (fast) way of calculating expressions of the form ' $\frac{0}{0}$ ' or ' $\frac{\infty}{\infty}$ '.

Now we are ready to prove the first mean value theorem, namely the *Theorem of Rolle*. This may be seen as a special case of the upcoming theorem, but it already contains the main idea.

Theorem 5.32 (Rolle). *Let $f: [a, b] \rightarrow \mathbb{R}$ be continuous and differentiable on (a, b) . Furthermore, assume that $f(a) = f(b)$. Then there exists some $\xi \in (a, b)$ such that $f'(\xi) = 0$.*

Proof. For constant f the statement is obvious. So we may assume that f is not constant. Since f is continuous on $[a, b]$ we know from the extreme value theorem (Theorem 4.56) that it has a

maximum and a minimum, which are not equal. Moreover, since $f(a) = f(b)$, one of the (global) extreme points has to be in (a, b) , i.e., not at the boundary points. By Theorem 5.25, this point, say $\xi \in (a, b)$, satisfies $f'(\xi) = 0$. □

Geometrically the above theorem states that the graph of f has at least one point where the tangent is horizontal. Again, if you have a drawing of a function, this result seems to be a trivial statement. However, it will also be necessary in more complex situations, and we will see that it is important to respect all the assumptions.

Example 5.33. It is important the the function is differentiable in the whole interval. To see this, consider the absolute value function $f(x) = |x|$ on $[-1, 1]$, which is continuous and satisfies $f(-1) = f(1) = 1$. However, it is not differentiable at $x_0 = 0$, and at all other points it satisfies $f'(x) = 1$ or $f'(x) = -1$. So, there is no $\xi \in (-1, 1)$ with $f'(\xi) = 0$.

We now use the Theorem of Rolle to prove the *mean value theorem*.

Theorem 5.34 (Mean value theorem). *Let $f: [a, b] \rightarrow \mathbb{R}$ be continuous in $[a, b]$ and differentiable in (a, b) . Then, there exists some $\xi \in (a, b)$ such that*

$$f'(\xi) = \frac{f(b) - f(a)}{b - a}.$$

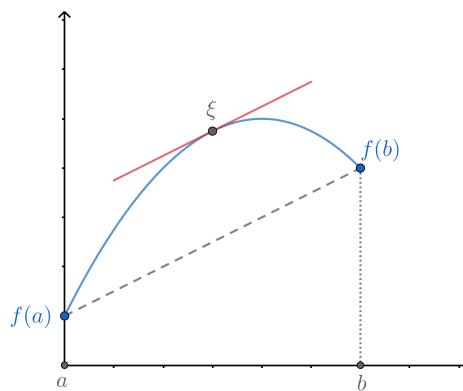


Figure 38: Mean value theorem

Proof. Consider the function

$$h(x) := f(x) - \frac{f(b) - f(a)}{b - a} (x - a).$$

We see that h is continuous and differentiable as f , and satisfies $h(a) = h(b) = f(a)$. Therefore, Rolle's theorem implies that there is some $\xi \in (a, b)$ with $h'(\xi) = 0$. Since

$$h'(x) = f'(x) - \frac{f(b) - f(a)}{b - a},$$

we obtain the result for $x = \xi$. □

Note that the mean value theorem gives information about the derivative of a function, even if we only know the function values at the boundary points. For example, given some $f: [0, 1] \rightarrow \mathbb{R}$ with $f(0) = 0$ and $f(1) = 5$. The mean value theorem states that, if f is differentiable on $(0, 1)$, then its derivative must attain the value 5, i.e., there exists $\xi \in (0, 1)$ with $f'(\xi) = 5$. Note that the function $f(x) = 5x$ is the only linear function with these function values and satisfies $f' \equiv 5$. The theorem then states that *every* function with the same boundary values has also a point with this slope. (Recall that the intermediate value theorem, Theorem 4.49, implies e.g. that there is a $\xi \in (0, 1)$ with $f(\xi) = 2$.)

The following theorem is a slight generalization of the mean value theorem which is the first step in proving l'Hospital's rule, and will be also of interest later.

Theorem 5.35 (General mean value theorem). *Let $f, g: [a, b] \rightarrow \mathbb{R}$ be continuous in $[a, b]$ and differentiable in (a, b) . Then, there exists some $\xi \in (a, b)$ such that*

$$f'(\xi)(g(b) - g(a)) = g'(\xi)(f(b) - f(a)).$$

In particular, if $g' \neq 0$ on (a, b) , then there exists some $\xi \in (a, b)$ such that

$$\frac{f'(\xi)}{g'(\xi)} = \frac{f(b) - f(a)}{g(b) - g(a)}.$$

Proof. We first consider the case that $g(a) = g(b)$. By Rolle's theorem we know that there exists some $\xi \in (a, b)$ with the property that $g'(\xi) = 0$. So we only have to show the first point, as $g' \neq 0$ is clearly not true. To do so we have a look at

$$f'(\xi)(g(b) - g(a)) = 0 = g'(\xi)(f(b) - f(a)),$$

which was to show. For the other case, i.e. $g(a) \neq g(b)$, we define the function

$$h(x) = f(x) - \frac{f(b) - f(a)}{g(b) - g(a)}(g(x) - g(a)).$$

Since f and g are continuous on $[a, b]$ and differentiable on (a, b) , we have that h is continuous on $[a, b]$ and differentiable on (a, b) . Moreover, it is easy to compute that

$$h(a) = h(b).$$

An application of Rolle's theorem yields that there exists some $\xi \in (a, b)$ such that

$$0 = h'(\xi) = f'(\xi) - \frac{f(b) - f(a)}{g(b) - g(a)}g'(\xi).$$

Regrouping this equation leads to

$$f'(\xi)(g(b) - g(a)) = g'(\xi)(f(b) - f(a)),$$

which proves the first point of the theorem. For the second point we assume additionally that $g' \neq 0$ on (a, b) , i.e., $g'(x) \neq 0$ for all $x \in (a, b)$. Thus, we can divide the last but one equation by $g'(\xi)$ and obtain

$$\frac{f'(\xi)}{g'(\xi)} = \frac{f(b) - f(a)}{g(b) - g(a)}.$$

□

We now come back to the computation of limits.

Note that we had problems with examples of the form

$$\lim_{x \rightarrow 4} \frac{x^2 - 16}{x - 4} \quad \text{or} \quad \lim_{x \rightarrow \infty} \frac{4x^2 - 5x}{1 - 3x^2}.$$

If we plug in $x = 4$ in the first limit we get $\frac{0}{0}$, another similar case appears if we would plug in " ∞ " in the second limit as we would get $\frac{\infty}{-\infty}$ (recall that, if x tends to ∞ , then a polynomial behaves like its largest power).

We now introduce **l'Hospital's rule**, which is a method to compute such limits, if they exist.

Theorem 5.36 (l'Hospital). *Let $I = (a, b)$ and $x_0 \in [a, b]$. Let $f, g: I \setminus \{x_0\} \rightarrow \mathbb{R}$ be differentiable on $I \setminus \{x_0\}$ and $g' \neq 0$. Furthermore assume that either $\lim_{x \rightarrow x_0} f(x) = \lim_{x \rightarrow x_0} g(x) = 0$ or $\lim_{x \rightarrow x_0} f(x) = \pm\infty, \lim_{x \rightarrow x_0} g(x) = \pm\infty$ holds. Then we have*

$$\lim_{x \rightarrow x_0} \frac{f(x)}{g(x)} = \lim_{x \rightarrow x_0} \frac{f'(x)}{g'(x)},$$

if the limit on the right hand side exists, or is definitely divergent.

Proof. We only prove the case $\lim_{x \rightarrow x_0} f(x) = \lim_{x \rightarrow x_0} g(x) = 0$. Otherwise replace f by $1/f$, and g by $1/g$. First observe that we can continuously extend f, g to x_0 with $f(x_0) = g(x_0) = 0$. By the general mean value theorem for any $x \in I$ such that $x \neq x_0$ there exists $\xi \in (x_0, x)$ satisfying

$$\frac{f'(\xi)}{g'(\xi)} = \frac{f(x) - f(x_0)}{g(x) - g(x_0)} = \frac{f(x)}{g(x)}.$$

If $x \rightarrow x_0$, it follows that $\xi \rightarrow x_0$ and since $\lim_{\xi \rightarrow x_0} \frac{f'(\xi)}{g'(\xi)}$ exists (this was our assumption) we obtain

$$\lim_{x \rightarrow x_0} \frac{f(x)}{g(x)} = \lim_{x \rightarrow x_0} \frac{f'(x)}{g'(x)}.$$

□

Remark 5.37. This rule of calculation was published in 1696 by *Guillaume Francois Antoine, Marquis de L'Hospital* (1661–1704) in the very first textbook on differential calculus. However, it was actually proven in 1694 by the famous mathematician *Johann Bernoulli* (1667–1748).

As a first application of l'Hospital's rule we prove that the second derivative test is valid.

Proof of Theorem 5.29. For this, let $x_0 \in I$ be such that $f'(x_0) = 0$ and $f''(x_0) > 0$. Since $f'(x_0)$ exists, we have from Theorem 5.10 that f is continuous at x_0 . This implies that $\lim_{x \rightarrow x_0} (f(x) - f(x_0)) = 0$. We obtain, by l'Hospital's rule, $f'(x_0) = 0$ and the definition of the second derivative, that

$$\lim_{x \rightarrow x_0} \frac{f(x) - f(x_0)}{(x - x_0)^2} = \lim_{x \rightarrow x_0} \frac{f'(x)}{2(x - x_0)} = \frac{1}{2} \lim_{x \rightarrow x_0} \frac{f'(x) - f'(x_0)}{x - x_0} = \frac{f''(x_0)}{2} > 0.$$

This shows that the limit on the left hand side exists and is positive. Therefore, the function inside the limit must be positive in a neighborhood of x_0 . In detail: There are some $\varepsilon, \delta > 0$

s.t. $\frac{f(x)-f(x_0)}{(x-x_0)^2} > \delta$ for all x with $0 < |x - x_0| < \varepsilon$. Since $(x - x_0)^2$ is positive for $x \neq x_0$, we obtain that $f(x) - f(x_0) > \delta(x - x_0)^2 > 0$ for x with $0 < |x - x_0| < \varepsilon$. This implies that

$$f(x) > f(x_0) \quad \text{for } x \in I \cap (x_0 - \varepsilon, x_0 + \varepsilon).$$

The case of a local maximum follows by analogous arguments. □

Let us see some more examples where we can use this rule.

Example 5.38.

$$\lim_{x \rightarrow 0} \frac{\sin x}{x} = \lim_{x \rightarrow 0} \frac{\cos x}{1} = 1.$$

Example 5.39.

$$\lim_{x \rightarrow 1} \frac{x^3 - 1}{x - 1} = \lim_{x \rightarrow 1} \frac{3x^2}{1} = 3$$

This rule can also be used several times to calculate a limit as the following examples will show.

Example 5.40.

$$\lim_{x \rightarrow 0} \frac{1 - \cos x}{x^2} = \lim_{x \rightarrow 0} \frac{\sin x}{2x} = \lim_{x \rightarrow 0} \frac{\cos x}{2} = \frac{1}{2}.$$

Example 5.41.

$$\lim_{x \searrow 0} x^x = \exp \left(\lim_{x \searrow 0} x \ln x \right) = \exp \left(\lim_{x \searrow 0} \frac{\ln x}{x^{-1}} \right) = \exp \left(\lim_{x \searrow 0} \frac{x^{-1}}{-x^{-2}} \right) = e^0 = 1.$$

Example 5.42.

$$\lim_{x \rightarrow 0} \frac{1 - \cos \frac{x}{2}}{1 - \cos x} = \lim_{x \rightarrow 0} \frac{\frac{1}{2} \sin \frac{x}{2}}{\sin x} = \lim_{x \rightarrow 0} \frac{\frac{1}{4} \cos \frac{x}{2}}{\cos x} = \frac{1}{4}.$$

5.4 Monotonicity and convexity

A common application of the mean value theorem is the characterization of the monotonicity of a function, which is based on its derivative.

Definition 5.43 (Monotonicity). Let $f: (a, b) \rightarrow \mathbb{R}$.

We call f (**strictly**) **increasing** if

$$\forall x_1, x_2 \in (a, b): x_1 < x_2 \implies f(x_1) < f(x_2),$$

or (**strictly**) **decreasing** if

$$\forall x_1, x_2 \in (a, b): x_1 < x_2 \implies f(x_1) > f(x_2).$$

If we replace ' $<$ ' by ' \leq ', or ' $>$ ' by ' \geq ', then we say f is **non-decreasing** or **non-increasing**, respectively.

A useful condition to check on monotonicity is given in the following theorem.

Theorem 5.44. Let f be differentiable on $I = (a, b)$.

Then,

$$f \text{ non-decreasing} \iff f' \geq 0$$

and

$$f \text{ non-increasing} \iff f' \leq 0.$$

Proof. W.l.o.g. we only prove the first statement.

First, we assume that $f'(x) \geq 0$ for all $x \in I$. By the mean value theorem we get that for $x_1 < x_2 \in I$ there exists a $\xi \in (x_1, x_2)$ such that

$$\frac{f(x_2) - f(x_1)}{x_2 - x_1} = f'(\xi) \geq 0.$$

This implies $f(x_2) \geq f(x_1)$ for arbitrary points $x_1, x_2 \in I$ with $x_1 < x_2$, so f is non-decreasing.

Now assume that f is non-decreasing. This is, $f(x_2) \geq f(x_1)$ for all $x_2 > x_1$. This also implies that $f(x_2) \leq f(x_1)$ for all $x_2 < x_1$. In any case, we get

$$\frac{f(x_2) - f(x_1)}{x_2 - x_1} \geq 0,$$

and therefore, in particular, $f'(x_1) \geq 0$ for all $x_1 \in I$.

□

Remark 5.45. Note that if ' f is increasing' does not imply that $f' > 0$ in general. As an example consider $f(x) = x^3$, which is an increasing function, but $f'(0) = 0$.

Example 5.46. The function $f(x) = e^x$ is increasing on \mathbb{R} , since $\forall x \in \mathbb{R}: f'(x) = e^x > 0$. This implies that f has no stationary points.

Example 5.47. The function $f(x) = -\ln(x)$ is decreasing on \mathbb{R}^+ , since $\forall x > 0: f'(x) = -\frac{1}{x} < 0$. Again, this implies that there are no stationary points.

The second derivative helps us to determine the shape of a function, once plotted in a coordinate system. If a line between two points of a curve is above the graph, we call the function **convex**, if otherwise the line is below the graph, we say the function is **concave**.

Definition 5.48. Let I be an interval and $f: I \rightarrow \mathbb{R}$. We say that f is **convex** in I if $\forall \lambda \in (0, 1)$ and $x_0, x_1 \in I$ there holds

$$f((1 - \lambda)x_0 + \lambda x_1) \leq (1 - \lambda)f(x_0) + \lambda f(x_1).$$

We call f **concave** if $\forall \lambda \in (0, 1)$ and $x_0, x_1 \in I$ we have

$$f((1 - \lambda)x_0 + \lambda x_1) \geq (1 - \lambda)f(x_0) + \lambda f(x_1).$$

Remark 5.49. From the definition we obtain immediately f is concave if and only if $-f$ is convex.

Example 5.50. Have a look at the graphs of $f(x) = x^2$ and $g(x) = \ln(x)$, see Figure 39. Clearly, $f(x)$ is strictly convex and $g(x)$ is strictly concave on \mathbb{R}_+ . (Calculate the second derivatives of both functions. What do you see?) The remark above states that $-f(x)$ is strictly concave.

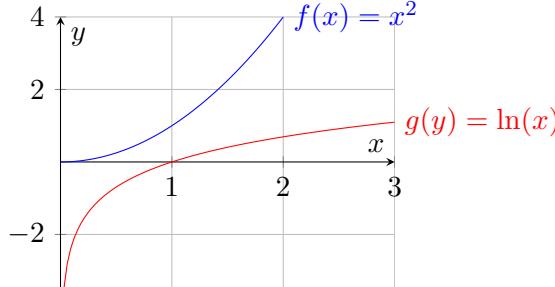


Figure 39: The function $f(x) = x^2$ and $g(x) = \ln(x)$

Remark 5.51 (*). Convexity is a very useful property for certain optimization problems.

For a twice differentiable function it suffices to check the sign of the second derivative.

Theorem 5.52. Let f be twice differentiable on an open interval I . Then f is convex if and only if $f''(x) \geq 0$, or concave if and only if $f''(x) \leq 0$.

Proof. For now we only prove the case of convexity, concavity can be treated analogously.

Assume $f''(x) \geq 0$ for all $x \in I = (a, b)$. Moreover, let $x_0, x_1 \in I$ such that $a < x_0 < x_1 < b$. Then, for any $\lambda \in (0, 1)$, we let $x := (1 - \lambda)x_0 + \lambda x_1 \in (x_0, x_1)$. By the mean value theorem we can find $\xi_0 \in (x_0, x)$ and $\xi_1 \in (x, x_1)$ such that

$$\frac{f(x) - f(x_0)}{x - x_0} = f'(\xi_0) \quad \text{and} \quad \frac{f(x_1) - f(x)}{x_1 - x} = f'(\xi_1).$$

Since $f'' \geq 0$, we get that f' has to be non-decreasing. Hence $f'(\xi_0) \leq f'(\xi_1)$ and we obtain

$$\frac{f(x) - f(x_0)}{\lambda(x_1 - x_0)} = \frac{f(x) - f(x_0)}{x - x_0} \leq \frac{f(x_1) - f(x)}{x_1 - x} = \frac{f(x_1) - f(x)}{(1 - \lambda)(x_1 - x_0)}.$$

The identities $\lambda(x_1 - x_0) = x - x_0$ and $(1 - \lambda)(x_1 - x_0) = x_1 - x$ follow from the definition of x . Using the above inequality and regrouping the terms we obtain

$$f(x) = f((1 - \lambda)x_0 + \lambda x_1) \leq (1 - \lambda)f(x_0) + \lambda f(x_1),$$

which is the definition of convexity.

On the other hand assume that f is convex. Using the inequality in the definition of convex functions and the above calculations we see that it holds

$$\frac{f(x) - f(x_0)}{x - x_0} \leq \frac{f(x_1) - f(x_0)}{x_1 - x_0} \leq \frac{f(x_1) - f(x)}{x_1 - x},$$

for arbitrary $x_0 < x < x_1$. Letting $x \rightarrow x_0$ and $x \rightarrow x_1$ we obtain

$$f'(x_0) \leq \frac{f(x_1) - f(x_0)}{x_1 - x_0} \leq f'(x_1).$$

Thus f' is non-decreasing and therefore $f'' \geq 0$. □

5.5 Taylor's theorem

In the last subsections, we have seen that some (local or global) properties of a function may be characterized by its derivatives. Now, we will show that, under certain assumptions, a function can be characterized exactly (in a neighborhood) just by knowing all higher-order derivatives of a function *at one point*. This shows that the very local behavior of a function can be used to determine it exactly everywhere.

Let us start with a result that shows that we can **approximate a function** of the function in a neighborhood of a point by a polynomial. This is *Taylor's theorem*.

Recall that the **n -th derivative** of $f: (a, b) \rightarrow \mathbb{R}$ (at x) is defined inductively by

$$f^{(n)}(x) := \frac{d^n}{dx^n} f(x) = \frac{d}{dx} f^{(n-1)}(x),$$

if it exists.

Theorem 5.53 (Taylor's theorem). *Let $f: (a, b) \rightarrow \mathbb{R}$ be $(n+1)$ -times differentiable and let $x_0 \in (a, b)$. Then, for all $x \in (a, b)$ there is a ξ between x_0 and x such that*

$$f(x) = \sum_{k=0}^n \frac{f^{(k)}(x_0)}{k!} (x - x_0)^k + \frac{f^{(n+1)}(\xi)}{(n+1)!} (x - x_0)^{n+1}.$$

We call

$$T_n(x) := \sum_{k=0}^n \frac{f^{(k)}(x_0)}{k!} (x - x_0)^k,$$

the **Taylor polynomial of f of order n (at x_0)**.

The term

$$R_n(x) := f(x) - T_n(x) = \frac{f^{(n+1)}(\xi)}{(n+1)!} (x - x_0)^{n+1}$$

is called the **remainder** of the Taylor polynomial (in Lagrange form).

Note that ξ above depends on x . (This may already be clear since it lies between x and x_0 , i.e., $\xi \in (x, x_0)$ or $\xi \in (x_0, x)$.) That's why some authors write $\xi = \xi_x$ to make this explicit. In particular, to obtain the equality above, one needs to find a specific ξ for every x , which is very impractical. However, this formula is helpful, when we want to prove that the error of the Taylor polynomial, which is the difference of $f(x)$ and $T_n(x)$, is 'small' for all $x \in (a, b)$. We just have to bound $f^{n+1}(\xi)$ for all possible $\xi \in (a, b)$.

Proof. Let $x \in (a, b)$ be arbitrary and, w.l.o.g., assume $x > x_0$. Define, for $t \in [x_0, x]$, the function

$$g(t) := f(x) - \sum_{k=0}^n \frac{f^{(k)}(t)}{k!} (x - t)^k - \frac{m}{(n+1)!} (x - t)^{n+1},$$

where we choose m such that $g(x_0) = 0$.

Clearly, $g(x) = 0$. So, together with $g(x_0) = 0$, Rolle's theorem yields the existence of some $\xi \in (x_0, x)$ such that $g'(\xi) = 0$. We compute the first derivative of g (in t) and obtain, using the

product rule, that

$$\begin{aligned} g'(t) &= -\sum_{k=0}^n \frac{f^{(k+1)}(t)}{k!} (x-t)^k + \sum_{k=1}^n \frac{f^{(k)}(t)}{(k-1)!} (x-t)^{k-1} + \frac{m}{n!} (x-t)^n \\ &= -\frac{f^{(n+1)}(t)}{n!} (x-t)^n + m \frac{(x-t)^n}{n!}, \end{aligned}$$

where we have also used a telescoping trick (or just an index shift).

Thus, there exists $\xi \in (x_0, x)$ such that

$$0 = g'(\xi) = -\frac{f^{(n+1)}(\xi)}{n!} (x-\xi)^n + m \frac{(x-\xi)^n}{n!}$$

which holds if and only if $m = f^{(n+1)}(\xi)$. Therefore,

$$0 = g(x_0) = f(x) - \sum_{k=0}^n \frac{f^{(k)}(x_0)}{k!} (x-x_0)^k - \frac{f^{(n+1)}(\xi)}{(n+1)!} (x-x_0)^{n+1},$$

which proves the claim. \square

Remark 5.54. It is easy to check that $T_n^{(k)}(x_0) = f^{(k)}(x_0)$ for all $k = 1, \dots, n$.

One straightforward application is to polynomials. Note that, if $p: \mathbb{R} \rightarrow \mathbb{R}$ is a polynomial of degree n , then all derivatives $p^{(k)}$ of order $k > n$ satisfy $p^{(k)}(x) = 0$ for all $x \in \mathbb{R}$. (If this is not clear to you, prove it!) We therefore obtain that the Taylor polynomial of p of order n is *exact*.

Example 5.55. Let $p: \mathbb{R} \rightarrow \mathbb{R}$ be a polynomial of degree n , then for all $x, x_0 \in \mathbb{R}$ we have

$$p(x) = \sum_{k=0}^n \frac{p^{(k)}(x_0)}{k!} (x-x_0)^k,$$

i.e. the Taylor polynomial of order n is exactly p , independent of x_0 .

Proof. Clearly, p is of the form $p(x) = a_0 + a_1 x + \dots + a_n x^n$. However, the binomial theorem yields

$$x^k = (x - x_0 + x_0)^k = \sum_{\ell=0}^k \binom{k}{\ell} (x - x_0)^\ell x_0^{k-\ell},$$

so we can write p as

$$p(x) = b_0 + b_1(x - x_0) + \dots + b_n(x - x_0)^n,$$

where the b_k clearly depend on x_0 . Differentiating yields

$$\begin{aligned} p(x) &= b_0 + b_1(x - x_0) + b_2(x - x_0)^2 + \dots + b_n(x - x_0)^n \\ p'(x) &= b_1 + 2b_2(x - x_0) + \dots + nb_n(x - x_0)^{n-1} \\ p''(x) &= 2b_2 + 3 \cdot 2b_3(x - x_0) + \dots + n(n-1)b_n(x - x_0)^{n-2} \\ &\vdots \\ p^{(n)}(x) &= n! b_n. \end{aligned}$$

If we now plug in $x = x_0$, we see that most terms vanish and we obtain

$$p^{(k)}(x_0) = k! b_k.$$

\square

The equation in Example 5.55 is sometimes useful to rewrite polynomials. In particular, if one is interested in properties around a specific point.

Example 5.56. Let $p(x) = x^3 - 2x^2 + 3$ and consider $x_0 = -1$. Since p has the derivatives $p'(x) = 3x^2 - 4x$, $p''(x) = 6x - 4$ and $p'''(x) = 6$, we obtain $p(-1) = 0$, $p'(-1) = 7$, $p''(-1) = -10$ and $p'''(-1) = 6$. Example 5.55 implies

$$p(x) = 7(x+1) - 5(x+1)^2 + (x+1)^3.$$

In the same way we may also **expand more general functions around a point** if we know its derivatives, but note that there is an additional error term, which is in general not easy to determine exactly. Under additional assumptions, however, one may give practical bounds.

Corollary 5.57. *In the setting of Theorem 5.53, assume additionally that $f: (a, b) \rightarrow \mathbb{R}$ satisfies $|f^{(n+1)}(x)| \leq M$ for some $M < \infty$ and all $x \in (a, b)$, then*

$$|f(x) - T_n(x)| \leq \frac{M(b-a)^{n+1}}{(n+1)!} \quad \text{for all } x \in (a, b).$$

Proof. The bound follows directly from Theorem 5.53, by noting that $|x - x_0| \leq (b - a)$ for all $x, x_0 \in (a, b)$. □

Although this gives a useful *uniform bound* on the error, it is clearly useless, if we consider functions that are defined on \mathbb{R} . Let us discuss an example.

Example 5.58. Consider the function $f(x) = e^x$ on $I = (-1, 1)$. Assume you want to approximate f by a polynomial with preferably small degree, and you allow an error of at most $\varepsilon = \frac{1}{50}$. We know, see Example 5.7, that $f^{(k)}(x) = e^x$ for all $k \in \mathbb{N}$, and therefore also $f^{(k)}(0) = 1$. Setting $x_0 = 0$, we obtain that

$$T_n(x) = \sum_{k=0}^n \frac{x^k}{k!}.$$

Moreover, note that $|f^{(k)}(x)| \leq e$ for all $x \in (-1, 1)$ and $k \in \mathbb{N}$. It follows from Corollary 5.57 that

$$|e^x - T_n(x)| \leq \frac{e \cdot 2^{n+1}}{(n+1)!}.$$

One may check the first values of n , to see that $n = 7$ is enough. That is, we obtain

$$\left| e^x - \left(1 + x + \frac{x^2}{2} + \frac{x^3}{6} + \frac{x^4}{24} + \frac{x^5}{120} + \frac{x^6}{720} + \frac{x^7}{7!} \right) \right| \leq 0.0173 \leq \frac{1}{50}$$

for all $x \in (-1, 1)$. (Try to visualize this with some computer algebra software!)

One should notice that the upper bound on the error in the last example goes to zero (very fast) with n . This is not only the case for the function e^x , but for all infinitely often differentiable functions that satisfy the assumption of Corollary 5.57 for all $n \in \mathbb{N}$ with the same M , i.e., if

$$\sup \left\{ |f^{(k)}(x)| : k \in \mathbb{N}, x \in (a, b) \right\} \leq M.$$

In this case, we obtain that $\lim_{n \rightarrow \infty} |f(x) - T_n(x)| = 0$, since we always have $\frac{(b-a)^n}{n!} \rightarrow 0$ for $a, b \in \mathbb{R}$, and therefore

$$f(x) = \lim_{n \rightarrow \infty} T_n(x) = \sum_{k=0}^{\infty} \frac{f^{(k)}(x_0)}{k!} (x - x_0)^k$$

for all $x_0 \in (a, b)$, and we call the right hand side the **Taylor series of f** (around x_0).

Remark 5.59. Think for a second about the last formula!

Note that (under the very restrictive conditions needed) the equation holds for all $x, x_0 \in (a, b)$. However, only the right hand side depends on x_0 . Isn't it fascinating that the value of the right hand side does not change, if we change x_0 ?

Note that Taylor series are a special case of the more general **power series**, see Section 3.8. In particular, Taylor's theorem leads to the earlier announced and constructive way to write rather general functions as power series

$$f(x) := \sum_{k=0}^{\infty} a_k (x - x_0)^k$$

for some (real- or complex-valued) sequence $(a_k)_{k \in \mathbb{N}_0}$ and $x_0 \in \mathbb{R}$. See Section 3.8 for some general results on the convergence of the right hand side. In particular, we obtain convergence if $|x - x_0| < R$, where R is the **radius of convergence**

$$R = \liminf_{k \rightarrow \infty} \frac{1}{\sqrt[k]{a_k}}$$

see Theorem 3.96. For Taylor series, we consider $a_k = \frac{f^{(k)}(x_0)}{k!}$.

Remark 5.60. Note that convergence of the series $\sum a_k (x - x_0)^k$ does not mean that this sum equals $f(x)$. For this, we need a bound on the values of the derivatives at all ξ between x_0 and x , see Theorem 5.53. One may get in trouble otherwise. For example, if we write the function $f(x) = e^{|x|}$ as its Taylor series at $x_0 = 1$. Note that at this point $f(x)$ equals e^x , and so would be the Taylor series. This series would converge for all $x \in \mathbb{R}$, but wouldn't be equal to f for $x < 0$. (Note that f is not differentiable at $x_0 = 0$.)

We now bring our conditions in a form that is more useful to decide if a function can be written as Taylor series everywhere. This also allows for an analysis of functions defined on \mathbb{R} . Note that, however, in many cases, the above equality does not hold in the whole domain of definition, but only **in a neighborhood** of the point x_0 .

Theorem 5.61. Let $f: I \rightarrow \mathbb{R}$ (with $I = (a, b)$ or $I = \mathbb{R}$) be an infinitely often differentiable function, and let $x_0 \in I$. If $r > 0$ is such that

$$\lim_{n \rightarrow \infty} \frac{r^n}{n!} \cdot \sup_{\xi \in U_r(x_0)} |f^{(n)}(\xi)| = 0,$$

where $U_r(x_0) := (x_0 - r, x_0 + r) \subset I$, then f can be written by its Taylor series

$$f(x) = \sum_{k=0}^{\infty} \frac{f^{(k)}(x_0)}{k!} (x - x_0)^k \quad \text{for all } x \in U_r(x_0).$$

In particular, if, for every fixed $\xi \in I$, we have

$$\lim_{n \rightarrow \infty} \frac{\sqrt[n]{|f^{(n)}(\xi)|}}{n} = 0,$$

then the above holds for all $x, x_0 \in I$.

Remark 5.62. Note that the latter condition in the theorem is clearly fulfilled if $|f^{(n)}(x)| \leq c \cdot C^n$ for every fixed $x \in I$, where $c, C \geq 1$ may depend on x , but not on n .

Proof. For the first statement, note that $x \in U_r(x_0)$ if and only if $|x - x_0| < r$. We can therefore bound the remainder from Theorem 5.53 by $|R_{n-1}(x)| \leq \frac{|f^{(n)}(\xi)|}{n!} r^n$. Taking into account that ξ is between x_0 and x , and therefore also $\xi \in U_r(x_0)$, we see that the assumption of the theorem implies $|R_{n-1}(x)| \rightarrow 0$.

For the second part, let $x \in I$ be fixed. For $x = x_0$ the statement is obvious. So, w.l.o.g., we assume $x > x_0$. Then, since f is infinitely often differentiable on I , and $[x_0, x]$ is strictly contained in I , we get that $f^{(n+1)}(\xi)$ (i.e., the derivative of $f^{(n)}$ at ξ) exists for all $\xi \in [x_0, x]$. This implies that $f^{(n)}$, and therefore $|f^{(n)}|$, is continuous on $[x_0, x]$, see Theorem 5.10. From Theorem 4.56 we know that a continuous function on a closed interval attains its maximum, say at $\xi^* \in [x_0, x]$, i.e., $\sup_{\xi \in [x_0, x]} |f^{(n)}(\xi)| = |f^{(n)}(\xi^*)|$.

Using our assumption $\lim_{n \rightarrow \infty} \frac{\sqrt[n]{|f^{(n)}(\xi)|}}{n} = 0$ for all $\xi \in I$, we obtain, in particular, that

$$\frac{\sqrt[n]{|f^{(n)}(\xi^*)|}}{n} < \frac{1}{5|x - x_0|} \iff |f^{(n)}(\xi^*)| < \frac{n^n}{5^n |x - x_0|^n}$$

for all large enough n . Moreover, we have that $n^n \leq 4^n n!$ for all $n \in \mathbb{N}$. This can be proven inductively, by using $(1 + \frac{1}{n})^n \leq 4$, see Example 3.36. With R_n from Theorem 5.53, we obtain

$$|R_{n-1}(x)| \leq \frac{|f^{(n)}(\xi^*)|}{n!} |x - x_0|^n \leq \frac{n^n}{n!} \frac{1}{5^n} \leq \left(\frac{4}{5}\right)^n \rightarrow 0,$$

where the second inequality only holds for large enough n . Since $x \in I$ was arbitrary, we get

$$f(x) = \sum_{k=0}^{\infty} \frac{f^{(k)}(x_0)}{k!} (x - x_0)^k \quad \text{for all } x \in I$$

from Theorem 5.53. □

Let us come back to the last example.

Example 5.63. Let $f(x) = e^x$ on \mathbb{R} . As already discussed above, we have $|f^{(n)}(x)| = e^x$, which is independent of n . We can therefore apply the second part of Theorem 5.61 with $I = \mathbb{R}$ and $x_0 = 0$, and get

$$e^x = \sum_{k=0}^{\infty} \frac{x^k}{k!}$$

for all $x \in \mathbb{R}$.

We now consider the **trigonometric functions**. It may come as a surprise, that one can approximate these periodic (wave) functions up to an arbitrary precision by polynomials on the whole real line. However, a visualization of the first Taylor polynomials is very instructive for understanding what's going on.

Example 5.64. Let us have a look at the cosine $\cos: \mathbb{R} \rightarrow \mathbb{R}$.

We clearly know by now that $\cos' = -\sin$, $\cos'' = -\cos$, $\cos^{(3)} = \sin$, $\cos^{(4)} = \cos$, and so on. In any case we have $|\cos^{(k)}(x)| \leq 1$ for all $k \in \mathbb{N}$ and $x \in \mathbb{R}$. Theorem 5.61 now implies that the cosine can be written by its Taylor series for all $x_0, x \in \mathbb{R}$. Choosing the point $x_0 = 0$, we get the function values of the derivatives $\cos(0) = 1$, $\cos'(0) = 0$, $\cos''(0) = -1$, $\cos^{(3)}(0) = 0$, $\cos^{(4)} = 1$, and this “ $1, 0, -1, 0$ ” pattern repeats periodically. In formulas, $\cos^{(2k-1)}(0) = 0$ and $\cos^{(2k)}(0) = (-1)^k$ for all $k \in \mathbb{N}$. We obtain

$$\cos x = \sum_{k=0}^{\infty} (-1)^k \frac{x^{2k}}{(2k)!} = 1 - \frac{x^2}{2!} + \frac{x^4}{4!} - + \dots$$

In the same way, one can calculate the following Taylor series, and prove their convergence:

$$\begin{aligned} \sin x &= \sum_{k=0}^{\infty} (-1)^k \frac{x^{2k+1}}{(2k+1)!} = x - \frac{x^3}{3!} + \frac{x^5}{5!} - + \dots \\ \cosh x &= \sum_{k=0}^{\infty} \frac{x^{2k}}{(2k)!} = 1 + \frac{x^2}{2!} + \frac{x^4}{4!} + \dots \\ \sinh x &= \sum_{k=0}^{\infty} \frac{x^{2k+1}}{(2k+1)!} = x + \frac{x^3}{3!} + \frac{x^5}{5!} + \dots \end{aligned}$$

(Prove this yourself!)

Remark 5.65. In the proof of Theorem 5.61 we have used the bound $n^n \leq 4^n n!$ for all $n \in \mathbb{N}$. Later we will prove **Stirling's formula**, which states that, in fact,

$$\sqrt{2\pi n} \left(\frac{n}{e}\right)^n \leq n! \leq \left(1 + \frac{1}{11n}\right) \sqrt{2\pi n} \left(\frac{n}{e}\right)^n$$

for all $n \in \mathbb{N}$. In particular, it implies $n^n \leq e^n \cdot n!$ for all n , and $\frac{\sqrt[n]{n!}}{n} \rightarrow \frac{1}{e}$.

Example 5.66 (Taylor series of $\ln(x)$). Let us now consider the natural logarithm $f(x) = \ln(x)$ on $\mathbb{R}_+ = (0, \infty)$, which is an example that shows that a Taylor series may converge only for x in a neighborhood of x_0 , i.e. we cannot apply the second part of Theorem 5.61.

For all $x \in \mathbb{R}_+$ we obtain $f'(x) = \frac{1}{x} = x^{-1}$, and therefore

$$f^{(k)}(x) = \frac{(-1)^{k-1}(k-1)!}{x^k}$$

for all $k \in \mathbb{N}$ and $x \in \mathbb{R}_+$, see Example 5.3. With this, the Taylor polynomial of f around $x_0 \in \mathbb{R}_+$ is given by

$$\begin{aligned} T_n(x) &= \ln(x_0) + \sum_{k=1}^n \frac{f^{(k)}(x_0)}{k!} (x - x_0)^k = \ln(x_0) + \sum_{k=1}^n \frac{(-1)^{k-1}}{k} \cdot \frac{(x - x_0)^k}{x_0^k} \\ &= \ln(x_0) + \sum_{k=1}^n \frac{(-1)^{k-1}}{k} \cdot \left(\frac{x}{x_0} - 1 \right)^k. \end{aligned}$$

Clearly, these sums converge absolutely as $n \rightarrow \infty$ if $|\frac{x}{x_0} - 1| < 1$ (use e.g. the root test), which holds if and only if $x \in (0, 2x_0)$. Moreover, for $x = 2x_0$, we have $T_n(x) = \sum_{k=1}^n \frac{(-1)^{k-1}}{k}$, which is the alternating harmonic series and, therefore, convergent, see Example 3.93. For all other x , i.e. $x > 2x_0$, the Taylor series is clearly not convergent, since the terms of the sum are not a null sequence. Hence,

$$\ln(x) = \ln(x_0) - \sum_{k=1}^{\infty} \frac{1}{k} \cdot \left(1 - \frac{x}{x_0} \right)^k$$

holds if and only if $0 < x \leq 2x_0$. Choosing $x_0 = 1$ we obtain the typical series expansion of $\ln(x)$ at $x_0 = 1$:

$$\ln(x) = - \sum_{k=1}^{\infty} \frac{(1-x)^k}{k}$$

for $x \in (0, 2]$. In particular, $\ln(2) = \sum_{k=1}^{\infty} \frac{(-1)^{k-1}}{k}$.

Remark 5.67. Note that one may write functions $f(x)$ as its Taylor series at different points x_0 . This can be used to evaluate, or give a short notation for, certain infinite sums. For example, if we consider the Taylor series of $\ln(x)$ and $\cos(x)$ we see, e.g., that

$$\sum_{k=1}^{\infty} \frac{1}{k} \left(1 - \frac{1}{e} \right)^k = \ln(e) = 1$$

(use $x_0 = e$ and $x = 1$), or

$$\sum_{k=1}^{\infty} \frac{(-\pi)^k}{(2k)!} = \cos(\sqrt{\pi}).$$

Remark 5.68 (Complex arguments). Note that the series in the above expansions of e^x , \sin , \cos etc., are absolutely convergent series for all $x \in \mathbb{R}$. This means, that the series make also sense if we allow x to be a complex number, i.e., $x \in \mathbb{C}$. This is the natural way of extending real valued functions to the complex case.

Example 5.69. Use Taylor series to prove that

$$e^{ix} = \cos x + i \sin x$$

for all $x \in \mathbb{R}$, where $i := \sqrt{-1}$.

5.6 (*) Newton's method

In the last section about differentiability we want to study Newton's method, which is a commonly used method if we want to calculate zeros of a function. Before we prove the convergence of this method we want to discuss the main idea in detail. We are interested in solving

$$f(x) = 0.$$

If f is differentiable then we can use Taylor's theorem to approximate f

$$f(x) \approx f(x_0) + f'(x_0)(x - x_0).$$

Here x_0 is a fixed value, so we can easily solve

$$f(x_0) + f'(x_0)(x - x_0) = 0$$

and obtain $x = x_1 = x_0 - \frac{f(x_0)}{f'(x_0)}$. This implies an algorithm, where for arbitrary x_0 we compute in each step

$$x_{k+1} = x_k - \frac{f(x_k)}{f'(x_k)}.$$

Geometrically, we compute x_{k+1} consecutively as the zero of the tangent of f in the point x_k . Hopefully, this gives a point which is at least very close to a zero of f after enough calculation steps.

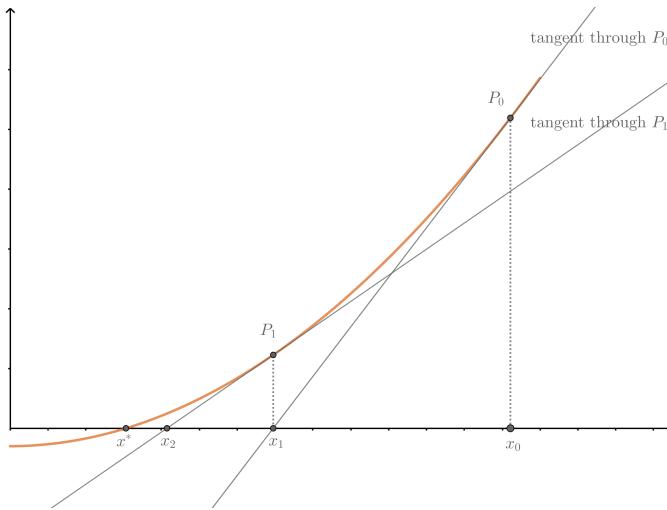


Figure 40: Some steps of Newton's method

Now we consider $f \in C^2(I)$, i.e., twice continuously differentiable functions, such that there exists some $\alpha \in I$ such that $f(\alpha) = 0$. Moreover, we assume that $f'(x) \neq 0$ for all $x \in I$. This ensures that x_{k+1} is always well-defined.

Now we use Taylor's theorem and get for arbitrary $x \in I$ that

$$0 = f(\alpha) = f(x) + f'(x)(\alpha - x) + \frac{f''(\xi)}{2}(\alpha - x)^2,$$

for some $\xi \in (x, \alpha)$ (or (α, x)). We can divide this equation by $f'(x) \neq 0$ and obtain

$$0 = \frac{f(x)}{f'(x)} + (\alpha - x) + \frac{1}{2} \frac{f''(\xi)}{f'(x)}(\alpha - x)^2.$$

Regrouping and plugging in $x = x_n$ yields

$$x_{n+1} - \alpha = x_n - \frac{f(x_n)}{f'(x_n)} - \alpha = \frac{1}{2} \frac{f''(\xi)}{f'(x_n)} (\alpha - x_n)^2.$$

Thus

$$|x_{n+1} - \alpha| = \frac{1}{2} \left| \frac{f''(\xi)}{f'(x_n)} \right| \cdot |x_n - \alpha|^2.$$

So we see that if $\left| \frac{f''(\xi)}{f'(x_n)} \right|$ does not behave too badly, and we choose x_0 not too far from α , the error should decrease quadratically. We will now explain how to choose x_0 , in order to guarantee convergence.

For this, define

$$m_1 := \sup_{x \in I} |f''(x)| \quad \text{and} \quad m_2 := \inf_{x \in I} |f'(x)|$$

and assume $m_2 > 0$. With the above formula we obtain that

$$|x_{n+1} - \alpha| \leq \frac{m_1}{2m_2} \cdot |x_n - \alpha|^2.$$

To guarantee that the method converges, we want that the distance to the solution gets smaller and smaller, i.e., $|x_{n+1} - \alpha| < |x_n - \alpha|$, we need that $\frac{m_1}{2m_2} \cdot |x_n - \alpha| < 1$. That is, we need $|x_0 - \alpha| < \frac{2m_2}{m_1}$, which might be very small, as this then implies the bound for the other n .

This shows the local convergence of **Newton's method**.

Theorem 5.70 (Local convergence of Newton's method). *Let $I = (a, b)$ and $f: C^2(I)$ with $f(\alpha) = 0$ for some $\alpha \in I$. Moreover, assume there is some $\delta > 0$ such that*

$$\delta < \frac{\inf_{x \in U_\delta(\alpha)} |f'(x)|}{\sup_{x \in U_\delta(\alpha)} |f''(x)|}.$$

Then, for all $x_0 \in U_\delta(\alpha)$, the Newton's method converges, and we have $|x_n - \alpha| \leq 2^{-n} |x_0 - \alpha|$.

6 Basic integration theory

In the last section we discussed the differential calculus of functions defined on the real line, or on an open interval $I \subset \mathbb{R}$. With this we discussed certain local properties of functions, and how these properties can be characterized by using derivatives.

Note that the derivative f' of a differentiable function $f: I \rightarrow \mathbb{R}$ is again a function on I . If f is continuously differentiable, then f' is continuous, and so on. This leads to the question, if (at least in some cases) there is also an *inverse operation*, or in other words:

Given some $f: I \rightarrow \mathbb{R}$, is there a differentiable function $F: I \rightarrow \mathbb{R}$ such that $F' = f$?

Such a function will be called a *antiderivative* of f .

In contrary with the definition of the derivative, which is unique if the function is differentiable, there are some problems with this definition. In particular, we will see that, if an antiderivative exists, it is not unique. (One can easily see, e.g., that all constant functions $F(x) = c$, $c \in \mathbb{R}$, are antiderivatives of the zero function $f(x) = 0$.) For other functions, the antiderivative just does not exist. However, we will see that an antiderivative exists for all continuous functions, and even more. This statement is one part of the *fundamental theorem of calculus*.

The other part of the theorem is concerned with the *integral of a function* over an interval:

For a non-negative function $f: [a, b] \rightarrow [0, \infty)$,
the integral $\int_a^b f(x) dx$ is the *area* between the x -axis and the graph of f ,
i.e., the area of the set $\{(x, y) \in \mathbb{R}^2 : a < x < b, 0 < y < f(x)\}$.

Clearly, we have to define precisely what this means. In contrast to derivatives, which were closely connected to the slope at a point, local extrema, and other local properties of a function, the integral is a global quantity. However, we will see later that both concepts are very much related. In particular, if $f: [a, b] \rightarrow \mathbb{R}$ is continuous, then there exists an antiderivative F of f and

$$\int_a^b f(x) dx = F(b) - F(a).$$

This is (the second part of) the *fundamental theorem of calculus*.

In this chapter we will first discuss antiderivatives of functions. Then we introduce a basic definition of an integral, and show the fundamental theorem of calculus, which gives a rather easy way of computing integrals (or areas). Later, we will see the limitations of this (too naive) approach, and turn to the more powerful Lebesgue integral. For this we also need to discuss some basic *measure theory*, i.e., we discuss what we mean by the area of a set.

Let us shortly comment on the existence of “different definitions of an integral”:

Over the centuries, several different approaches have been introduced to define the integral of a function, or equivalently the area of a set, precisely. This may be surprising, since the area is a unique number. And in the ‘simple’ cases that one is usually concerned with (i.e., for functions that *one can draw*), it is indeed the case that all the different approaches, that one can find in the literature, coincide. However, when it comes to theory, and the question which functions are *integrable*, then it is worth to think about clever concepts. In fact, all of the concepts below have some disadvantages, and there are always functions, which are not integrable. This corresponds to the problem that there are sets (which are by the way impossible to draw) to which we cannot assign a meaningful area. We will discuss that shortly later.

Remark 6.1 (History). The first systematic ideas of an *area* (or integral) go back to 500 BC, when people tried to compute the area of simple areas, like land plots. Already in the 3rd century BC, *Archimedes* (c. 287 – c. 212 BC) used these ideas to find an approximation of the area of a circle with radius one, and thereby determined the inequality $3 + \frac{10}{71} < \pi < 3 + \frac{10}{70}$. Only in the 19th century, mathematics was brought to a more formal level, which allowed for very precise statements. The first “correct” definition on an integral was given by *Augustin-Louis Cauchy* (1789–1857) in 1823. This was extended (or improved) by several mathematicians. The most famous approaches are the *Riemann integral*, introduced in 1854 by *Bernhard Riemann* (1826–1866) and by now the classical approach for introducing an integral to students, and the *Lebesgue integral*, introduced in 1902 by *Henri Léon Lebesgue* (1875–1941), which is the one actually used in research. Here, we only shortly comment on the Riemann integral, and focus on the definition of the more powerful Lebesgue integral.

6.1 Antiderivatives

The theory of the previous chapters was mostly done for functions defined on (open) intervals, which was enough to present the most important results. However, all the definitions (and many of the results) would also make sense for functions defined on the union of open intervals, by considering the function separately on each interval, or even on more general domains. But note that for some theorems, like the mean value theorem (Theorem 5.34), it was essential that the domain was an interval.

In order to come closer to more formal (or general) statements of theorems, we will present the following for a larger class of domains. This will also be necessary since we want to define the integral also over general domains.

Definition 6.2. Let $\Omega \subset \mathbb{R}$. Then, we call Ω an **open set**, if

$$\forall x \in \Omega \exists \varepsilon > 0: U_\varepsilon(x) \subset \Omega,$$

where $U_\varepsilon(x) = (x - \varepsilon, x + \varepsilon)$ is the ε -neighborhood of x .

That is, around every point there is a small open interval, that is completely contained in Ω .

Moreover, let $\Omega^c := \mathbb{R} \setminus \Omega$ denote the **complement** of Ω .

Then, we call Ω a **closed set**, if Ω^c is an open set.

Clearly, open intervals (a, b) and $\Omega = \mathbb{R}$ are open. Also sets of the form (a, ∞) , and their unions, are open. Therefore, also $\mathbb{R} \setminus \{0\} = (-\infty, 0) \cup (0, \infty)$ is open. Similarly, we obtain that closed intervals $[a, b]$ are closed since $([a, b])^c = \mathbb{R} \setminus [a, b] = (-\infty, a) \cup (b, \infty)$ is open. Therefore, also sets $\{a\}$, that contain only one element, are closed.

Remark 6.3. By this notation we can present the results in more generality, and without specifying a specific form of the domain. Note that open sets Ω are exactly those sets, where we can define the derivative of a function at every $x \in \Omega$. (We had problems with the boundary points!) However, if we consider in the following an open (or closed) set Ω , you may just think about an open (or closed) interval, or unions of them.

We now turn to antiderivatives.

Definition 6.4. Let $\Omega \subset \mathbb{R}$ be an open set. If $F: \Omega \rightarrow \mathbb{R}$ is a differentiable function such that

$$F'(x) = f(x) \quad \text{for all } x \in \Omega,$$

then we call F an **antiderivative** or **indefinite integral** of f .

We also use the notation

$$F = \int f(x) dx = \int f dx$$

to say that F is a antiderivative of f , and call f the **integrand**.

This definition seems easy to handle since all of us already computed many derivatives. However, if we compute a derivative of a differentiable function, then we always ended up with a (unique) function, and we had a nice point-wise criterion for deciding if a function is differentiable.

This is now different since we want to find a function F , but we only have information about its derivative $F' = f$. This is not enough to end up with a unique antiderivative F . To see this, note that knowing the slope in each point does not give any information about the function values at all. This is because a function with the same derivative might be at any “height”. Let us write this down mathematically. For any function f and F , and $c \in \mathbb{R}$, we have that

$$F \text{ is an antiderivative of } f \iff F + c \text{ is an antiderivative of } f.$$

For this, we only used that the derivative of a constant function equals zero. In particular, if a function has an antiderivative, then it has infinitely many.

Remark 6.5. Note that the notation $\int f(x) dx$, which shall denote a function, might be confusing, because it does not allow for the direct use of function values. E.g., we would never write $\int f(x) dx(2)$ for $F(2)$ or so. Moreover, the correct meaning of $F = \int f(x) dx$ is just “ F is an antiderivative of f ”, which is not an actual equality, but the derivatives of both sides have to coincide everywhere. However, this notation is useful when we want to talk about (properties of) the antiderivative as a function, since we do not have to reserve/waste a new letter.

Let us start with the easy example of the exponential function e^x , which does not change under differentiation.

Example 6.6. For the exponential function, we know that $(e^x)' = e^x$, and therefore that $F(x) = e^x$ is one possible antiderivative, i.e.,

$$\int e^x dx = e^x.$$

However, if one asks for all antiderivatives of e^x , then we have to take $F(x) = e^x + c$ for arbitrary $c \in \mathbb{R}$, i.e.,

$$\int e^x dx = e^x + c.$$

In most applications, it is enough to know just one of the antiderivatives, and therefore we mostly omit the constant c . However, keep in mind that a antiderivative is not unique.

Moreover, note that the two equations above combined do clearly not imply that $e^x = e^x + c$ for every x . The equal signs should be interpreted as “the derivative on the right hand side is e^x ”.

Example 6.7. Now consider $f(x) = \frac{1}{x}$ on $\Omega = \mathbb{R} \setminus \{0\}$, and we show that

$$\int \frac{dx}{x} = \int \frac{1}{x} dx = \ln|x|.$$

First of all, we know from the last chapter that $(\ln(x))' = \frac{1}{x}$ for all $x > 0$. Hence, $F(x) = \ln(x)$ for $x > 0$. But $\ln(x)$ is not defined for $x < 0$ and therefore, it is not obvious how to choose $F(x)$ such that $F'(x) = \frac{1}{x}$. But it is easy to verify that, for $x < 0$, the function $\ln|x| = \ln(-x)$ is well defined and $(\ln(-x))' = \frac{1}{-x} \cdot (-1) = \frac{1}{x}$. This proves the claim.

This is already an example that shows, that it might be hard to find the antiderivative of a given function, but it is easy to verify that a function is an antiderivative.

(Hint: Always double check your antiderivative by calculating its derivative!)

Next we provide a list of antiderivatives which we will use from now on. All of them follow by differentiating the right hand side. (Do this again as an exercise!)

$$\begin{aligned}\int a^x dx &= \frac{a^x}{\ln a}, \quad a > 0, a \neq 1 \\ \int x^a dx &= \frac{x^{a+1}}{a+1}, \quad a \neq -1 \\ \int \frac{dx}{x} &= \ln|x| \\ \int \cos x dx &= \sin x \\ \int \sin x dx &= -\cos x \\ \int \frac{dx}{\cos^2 x} &= \tan x \\ \int \frac{dx}{\sin^2 x} &= -\cot x \\ \int \frac{dx}{1+x^2} &= \arctan x \\ \int \frac{dx}{\sqrt{1-x^2}} &= \arcsin x \\ \int \frac{dx}{\sqrt{x^2-1}} &= \operatorname{arcosh} x \\ \int \frac{dx}{\sqrt{x^2+1}} &= \operatorname{arsinh} x\end{aligned}$$

All the antiderivatives $\int f dx$ above exist on the whole domain where f is defined.

However, not all functions have a antiderivative, as the following example shows.

Example 6.8. If we consider the function $f: \mathbb{R} \rightarrow \mathbb{R}$ with $f(x) = 1$ for $x \geq 0$, and $f(x) = 0$ for $x < 0$, i.e., the Heaviside function, then it is not hard to see that an antiderivative must be constant, say $F(x) = c$, for $x < 0$, and linear, say $F(x) = x + b$ for $x > 0$. Otherwise we would not get $F'(x) = f(x)$ for $x \neq 0$. It remains to consider $x = 0$. Since F has to be differentiable, it has to be continuous, and we obtain that $b = c$. However, it is easy to check that such a function F cannot be differentiable at 0. (F has a kink.)

6.2 Calculation rules for antiderivatives

As for derivatives we will now present some rules that are useful when we want to find the antiderivative of a complicated function, which is composed of some elementary functions, like the ones given above.

However, and unfortunately, although we were able to determine the derivative for nearly any combination of 'easy' functions, it is much harder to find an antiderivative. In fact, a very common strategy is to *guess* an antiderivative, and then to verify it by calculating its derivative. For this, one clearly needs to be well-practiced in calculating derivatives. Moreover, it is sometimes just impossible to determine a closed formula for the antiderivative, even for 'easy looking' functions like e^{-x^2} . We will discuss $\int e^{-x^2} dx$ and similar functions later in more detail.

The first calculation rule for antiderivatives, which directly follows from the corresponding rules for derivatives, is the linearity.

Lemma 6.9 (Linearity). *Let $F = \int f(x) dx$ and $G = \int g(x) dx$. Then, for all $\alpha, \beta \in \mathbb{R}$,*

$$\alpha F + \beta G = \alpha \int f(x) dx + \beta \int g(x) dx = \int \alpha f(x) + \beta g(x) dx.$$

If F and G have different domains, then $F+G$ is understood as a function on the intersection.

Proof. We only have to verify that the derivative of the function on the left equals the one in the integral on the right. By Theorem 5.11, we obtain

$$(\alpha F + \beta G)' = \alpha F' + \beta G' = \alpha f + \beta g,$$

since $F' = f$ and $G' = g$.

□

Now let us see some examples.

Example 6.10. We have

$$\int (x^3 + x^2) dx = \int x^3 dx + \int x^2 dx = \frac{x^4}{4} + \frac{x^3}{3}.$$

In particular, all antiderivatives of $x^3 + x^2$ are of the form $F(x) = \frac{x^4}{4} + \frac{x^3}{3} + c$ for some $c \in \mathbb{R}$.

Example 6.11. We have

$$\begin{aligned} \int (\sqrt{x} + x)^2 dx &= \int (x + 2x^{3/2} + x^2) dx = \int x dx + 2 \int x^{3/2} dx + \int x^2 dx \\ &= \frac{x^2}{2} + \frac{4x^{5/2}}{5} + \frac{x^3}{3}. \end{aligned}$$

In some cases, one may need some modifications of the integrand to bring it to the right form.

Example 6.12.

$$\begin{aligned}\int \frac{x^2 - x^4}{1 - x^4} dx &= \int \frac{1 - x^4 - (1 - x^2)}{1 - x^4} dx = \int 1 - \frac{1 - x^2}{(1 - x^2)(1 + x^2)} dx \\ &= \int 1 dx - \int \frac{1}{1 + x^2} dx = x - \arctan x.\end{aligned}$$

□

In the same way as we used linearity of differentiation above, we can also utilize the other calculation rules from Section 5.1 to deduce rules for antiderivatives.

Recall that the *product rule* states that

$$(fg)' = f'g + fg'$$

whenever the derivatives exist, see Theorem 5.13. Since this equality shows that fg is an antiderivative of $f'g + fg'$, we obtain the rule of **integration by parts** (or **partial integration**) by rearranging the terms.

Lemma 6.13 (Integration by parts). *Let f and g be differentiable functions. Then,*

$$\int f'g dx = fg - \int fg' dx.$$

Let us see an example that shows how this rule is usually applied.

Example 6.14. Assume we want to compute $F = \int \ln(x) dx$, i.e., an antiderivative of $\ln(x)$. Since we know the derivative of \ln , i.e. $(\ln(x))' = \frac{1}{x}$, we may choose $g(x) = \ln(x)$ above. Additionally, we choose $f(x) = x$, which satisfies $f'(x) = 1$. We obtain

$$\begin{aligned}\int \ln(x) dx &= \int 1 \cdot \ln(x) dx = \int f'g dx = fg - \int g'f dx \\ &= x \ln(x) - \int x \frac{1}{x} dx = x \ln(x) - x \\ &= x(\ln(x) - 1).\end{aligned}$$

□

The same 'trick' is very useful for integrating products of polynomials with sine, cosine or exponential functions. However, sometimes one has to integrate several times to obtain an explicit formula for the antiderivative.

Example 6.15. If we want to calculate $\int x \sin x dx$, we set $f'(x) = \sin x$ and $g(x) = x$, which implies that $g'(x) = 1$ and $f(x) = -\cos x$. Partial integration yields

$$\int x \sin x dx = (-\cos x)x - \int (-\cos x) dx = \sin x - x \cos x.$$

Example 6.16. If we want to calculate $\int x^2 e^x dx$, we set $g(x) = x^2$ and $f'(x) = e^x$, which implies that $g'(x) = 2x$ and $f(x) = e^x$. Partial integration yields

$$\int x^2 e^x dx = x^2 e^x - 2 \int x e^x dx.$$

This still involves the integral $\int x e^x dx$, which we calculate by setting $g(x) = x$ and $f'(x) = e^x$, which implies that $g'(x) = 1$ and $f(x) = e^x$. Again, by partial integration

$$\int x e^x dx = x e^x - \int e^x dx = x e^x - e^x,$$

and therefore

$$\int x^2 e^x dx = x^2 e^x - 2x e^x + 2e^x.$$

□

All the examples above involve polynomials (or more precisely, monomials), which vanish after enough differentiation. This is not the case if we consider the product of a trigonometric and a exponential function. In such cases, it may happen that we reach the same integral again after some steps of partial integration. This can be used to calculate the antiderivative.

Example 6.17. We try to calculate $\int \cos(x) 2^x dx$.

We set $f'(x) = \cos(x)$ and $g(x) = 2^x$, which yields $f(x) = \sin(x)$ and $g'(x) = \ln(2) \cdot 2^x$.

Therefore,

$$\int \cos(x) 2^x dx = \sin(x) 2^x - \ln(2) \cdot \int \sin(x) 2^x dx.$$

Partial integration also gives

$$\int \sin(x) 2^x dx = -\cos(x) 2^x + \ln(2) \cdot \int \cos(x) 2^x dx.$$

(Verify yourself!) Both equations together imply

$$\int \cos(x) 2^x dx = \sin(x) 2^x + \ln(2) \cos(x) 2^x - (\ln(2))^2 \cdot \int \cos(x) 2^x dx.$$

Now, the same indefinite integral appears on both sides. Rearranging, and dividing by $(1+\ln(2)^2)$ leads to

$$\int \cos(x) 2^x dx = 2^x \frac{\sin(x) + \ln(2) \cos(x)}{1 + \ln(2)^2}.$$

Example 6.18. Show that $\int \sin(x) e^x dx = e^x \frac{\sin(x)-\cos(x)}{2}$.

The next rule we want to employ is the chain rule, see Theorem 5.14. For this recall that for two differentiable functions F and g , we have

$$(F \circ g)'(x) = \frac{d}{dx} F(g(x)) = F'(g(x)) \cdot g'(x).$$

If F is a antiderivative of f , then this shows that $F \circ g$ is an antiderivative of $(f \circ g) \cdot g'$. This is called the **substitution rule**.

Lemma 6.19 (Substitution rule). *Let $F = \int f(x) dx$ and g be a differentiable function. Then,*

$$F(g(x)) = \int f(g(x)) g'(x) dx.$$

Let us again discuss some examples to understand this rule.

Example 6.20. Assume we want to calculate $\int x^6 \cos(x^7 + 1) dx$.

(This may also be done by partial integration, but would take ages.)

We see that the difficult part is the “ $x^7 + 1$ ” in the cosine. Let’s write $g(x) = x^7 + 1$. Then, we have $g'(x) = 7x^6$. If we now write $f(x) = \cos(x)$, we obtain

$$\int x^6 \cos(x^7 + 1) dx = \frac{1}{7} \int g'(x) f(g(x)) dx.$$

The integral on the right hand side equals $F(g(x))$ by the substitution rule, where $F(x) = \sin(x)$ is the antiderivative of f . We obtain

$$\int x^6 \cos(x^7 + 1) dx = \frac{1}{7} F(g(x)) = \frac{\sin(x^7 + 1)}{7}.$$

The substitution rule is sometimes easier to handle, if we introduce the **substitution** $t = g(x)$. If we then use the very informal reasoning that

$$\frac{dt}{dx} = \frac{d}{dx} t = \frac{d}{dx} g(x) = g'(x),$$

and that this “implies” $dt = g'(x) dx$ ($\iff dx = \frac{dt}{g'(x)}$), then we can write

$$\int f(g(x)) g'(x) dx = \int f(t) dt = F(t).$$

Note that $\int f(t) dt$ is now a function in t , and we have to replace $t = g(x)$ at the end. Although the above arguments were rather informal, we know that this formula is correct from the substitution rule.

Example 6.21. Consider the integral

$$\int (x^3 - 5)^6 \cdot x^2 dx.$$

(Again, one could expand the brackets and integrate all the easy terms, which is rather lengthy.) With the substitution $t = x^3 - 5$, we have $\frac{dt}{dx} = 3x^2$, and therefore $x^2 dx = \frac{dt}{3}$. We obtain

$$\int (x^3 - 5)^6 \cdot x^2 dx = \int t^6 \frac{dt}{3} = \frac{t^7}{21}.$$

If we finally substitute $t = x^3 - 5$, we obtain

$$\int (x^3 - 5)^6 \cdot x^2 dx = \frac{(x^3 - 5)^7}{21}.$$

Example 6.22. There are also quite tricky examples, which can be solved by substitution. Consider for example the function $f(x) = \sqrt{1-x^2}$ for $x \in [0, 1]$, and we want to compute its antiderivative $\int \sqrt{1-x^2} dx$. Here, we use the substitution the other way around and define $x = \sin(t)$ (is equivalent to $t = \arcsin(x)$) for $t \in [0, \frac{\pi}{2}]$. We obtain $\frac{dx}{dt} = \cos(t)$, and therefore

$$\int \sqrt{1-x^2} dx = \int \sqrt{1-\sin^2(t)} \cdot \cos(t) dt.$$

Since $\cos^2(t) + \sin^2(t) = 1$, and $\sqrt{\cos^2(t)} = |\cos(t)| = \cos(t)$ for $t \in [0, \frac{\pi}{2}]$, we get

$$\int \sqrt{1-x^2} dx = \int \cos^2(t) dt \quad \text{for } x = \sin(t).$$

By partial integration, with $f'(t) = g(t) = \cos(t)$, yields

$$\begin{aligned} \int \cos^2(t) dt &= \sin(t) \cos(t) + \int \sin^2(t) dt \\ &= \sin(t) \cos(t) + \int 1 - \cos^2(t) dt \\ &= \sin(t) \cos(t) + t - \int \cos^2(t) dt \end{aligned}$$

From this equation we get $\int \cos^2(t) dt = \frac{1}{2}(\sin(t) \cos(t) + t)$. Putting this in the equation above, re-substituting $x = \sin(t)$ (or $t = \arcsin(x)$), and noting that $\cos(\arcsin(x)) = \sqrt{1-x^2}$ (Try to prove this!), we obtain

$$\int \sqrt{1-x^2} dx = \frac{1}{2} \left(x \sqrt{1-x^2} + \arcsin(x) \right)$$

Note that, although the left hand side looks elementary, its antiderivative could involve functions that one would not expect there, e.g., \arcsin . We will see later how this integral is related to the area of a circle.

□

There is one particularly useful rule that follows directly from the substitution rule. One may even deduce it directly from the product rule by noting that, for a differentiable function f , we have

$$(\ln(|f(x)|))' = \frac{f'(x)}{f(x)},$$

whenever $f(x) \neq 0$. We obtain the following corollary.

Corollary 6.23. Let $f: \Omega \rightarrow \mathbb{R} \setminus \{0\}$ be a differentiable function. Then,

$$\int \frac{f'(x)}{f(x)} dx = \ln |f(x)|.$$

To see that this result follows from the substitution rule, consider the substitution $t = f(x)$.

Example 6.24. The last formula is always useful, when one wants to find the antiderivative of a rational function, such that the numerator is the derivative of the denominator.

An easy application is

$$\int \frac{x}{1+x^2} dx = \frac{1}{2} \int \frac{2x}{1+x^2} dx = \frac{1}{2} \ln(1+x^2),$$

where we set $f(x) = 1+x^2$, which gives $f'(x) = 2x$, and omit the absolute value, because $f > 0$.

Another application is

$$\int \tan(x) dx = \int \frac{\sin(x)}{\cos(x)} dx = -\ln|\cos(x)|.$$

For this, let $f(x) = \cos(x)$, which implies that $f'(x) = -\sin(x)$.

Remark 6.25. Let us finally recall again, that an antiderivative is not unique. It is just a function, whose derivative satisfies something. However, it can be unique, if we know in advance that it satisfies additional conditions. In particular, one function value is enough. For example, if we are looking for an antiderivative F of e^x , such that $F(0) = 0$, then we obtain $F(x) = e^x - 1$. (Verify yourself!)

This is a special case of so-called *initial value problems* (or *boundary value problems*):

For given $f: I \rightarrow \mathbb{R}$, $x_0 \in I$ and $y_0 \in \mathbb{R}$, we want to find F such that $F' = f$ and $F(x_0) = y_0$.

One may think about the following example: Let $F(t)$ be the position of a train (or car, or particle) at time t moving on the real line. Then, $f(t) = F'(t)$ is the velocity. If we assume that only f is given, i.e., we know only the velocity at every point in time, then we can clearly say something about overall distance traveled, say, between time $t = 0$ and $t = 1$. However, we have no chance to say, where the train actually is at time $t = 1$ (or any other time), if we do not know where the train departed.

6.3 A first definition of the integral

As noted in the beginning of this chapter, antiderivatives are very much connected to the *integration* of a function. That is, for a given function $f: \mathbb{R} \rightarrow [0, \infty)$, and a given interval $[a, b]$, the *integral* $\int_a^b f(x) dx$ is the *area* between the x -axis and the graph of f , i.e., the area of the set $\{(x, y) \in \mathbb{R}^2 : a < x < b, 0 < y < f(x)\}$, see Figure 41.

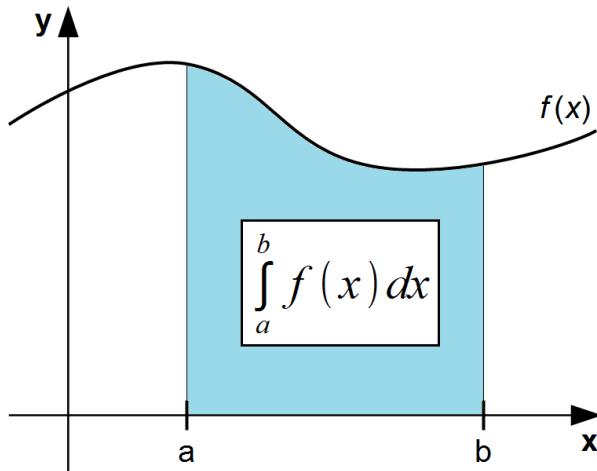


Figure 41: The area of the shaded region is denoted by $\int_a^b f(x) dx$

Clearly (and as we all hopefully already got used to), we have to define precisely what this means. In particular, we have to clarify which functions are *integrable*, and which functions (or sets) do not allow for the definition of a meaningful area.

Remark 6.26. It might come as surprise that there are sets, such that it is not possible to assign them an area (or a volume in higher dimension). This corresponds to the existence of functions such that *we just cannot say how big the area below the graph is*. However, we already were in a similar situation when we discussed in Chapter 5 that not all functions are differentiable (Consider, e.g., $f(x) = |x|$ at $x = 0$). And, similarly to there, the answer to “Is f integrable?” very much depends on the precise definitions. As we discussed in Remark 6.1, there were several attempts to obtain “the most general” or “the most useful” definition of an integral. (Presently, we consider the *Lebesgue integral*, which we will introduce later, the “best”. But who knows...) However, it was a great advance in mathematics to understand that there cannot be a universal definition that assigns an integral to any function. And this insight has a rather fascinating reformulation, if we talk about volumes in three dimensions:

The *Banach-Tarski paradox* states that we can split a ball into five disjoint pieces and, without deforming one of them, we can put them back together in a different way to obtain two *identical copies* of the same ball. In particular, if we would be able to assign a volume to each of the (impossible to draw) pieces, then this construction would show that the volume of the ball equals twice the volume of the same ball. An obvious contradiction. A proof of this statement (which appeared first in 1924), goes far beyond the scope of this lecture.

It is obvious, why this is called a paradox: The same is clearly not possible in our *real (physical) world*, as we cannot just double a ball. This is one of the most prominent examples of a very counterintuitive mathematical result and shows that we have to be careful with our definitions.

We now introduce one possibility of defining an integral. Actually, this is probably the most simple and straightforward definition, which is therefore restricted to “easy” functions. Later, when we need more powerful mathematical tools, we have to give also a more involved definition. However, note that for “easy” functions (that one can draw) it is imperative that all these different definitions should give the same result.

Let us consider a continuous function $f: \Omega \rightarrow \mathbb{R}$, and a closed interval $[a, b] \subset \Omega$. (Note that f is therefore bounded on $[a, b]$.) If we now want to calculate the area between the graph of f and the x -axis, then we could divide the interval $[a, b]$ into equal subintervals, and *approximate* the area in this subinterval just by the area of a *suitable* rectangle. Clearly, there are many reasonable choices. Figure 42, e.g., shows the (bad) approximation of the integral by using the smallest and the largest rectangle in each interval (when we divide it only into four subintervals).

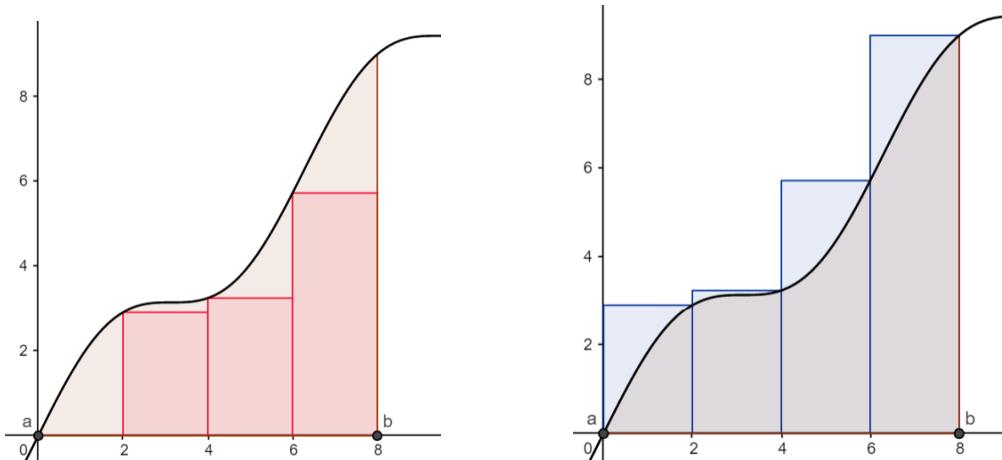


Figure 42: Using the smallest and the largest rectangle.

Although the resulting approximation of the integral might be quite different, this difference gets smaller and smaller if we increase the number of subintervals, see Figure 43.

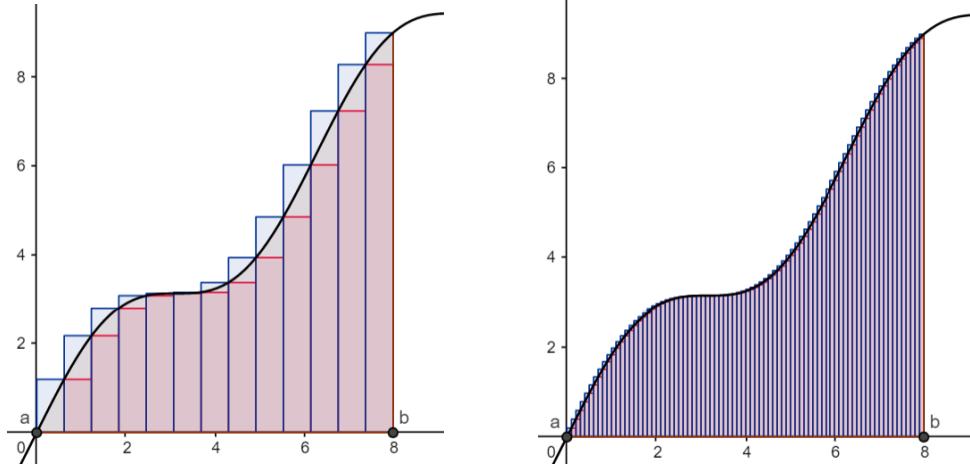


Figure 43: Smallest and the largest rectangles

This suggests that it is actually unimportant which of the rectangles we take, and we will prove that this is in fact the case when we consider **continuous functions on a closed interval**.

Therefore we may just use the rectangles whose height is given by the **left** endpoint of the subinterval. Note that, in the case of monotonically increasing functions, these are the same as the lower rectangles, see Figure 42(left).

Remark 6.27. Note that all the reasoning also makes sense for possibly **negative functions**. In this case, the area between f and the x -axis is counted *negatively*. In particular, if the integral of a function is zero, then this only means that the area above the x -axis equals the area below.

Let us put this into formulas:

Assume we divide the interval $I = [a, b]$, which has length $(b - a)$, into $n \in \mathbb{N}$ subintervals of equal length, which are therefore all of length $\frac{b-a}{n}$. That is, we use the partition

$$\begin{aligned} [a, b] &= \left[a, a + \frac{b-a}{n} \right] \cup \left[a + \frac{b-a}{n}, a + \frac{2(b-a)}{n} \right] \cup \dots \cup \left[a + \frac{(n-1)(b-a)}{n}, b \right] \\ &= \bigcup_{k=0}^{n-1} \left[a + \frac{k(b-a)}{n}, a + \frac{(k+1)(b-a)}{n} \right]. \end{aligned}$$

Note that in the special case $[a, b] = [0, 1]$, this partition has the simple form $\bigcup_{k=0}^{n-1} \left[\frac{k}{n}, \frac{k+1}{n} \right]$.

For illustration purposes, let us stick to the case $[a, b] = [0, 1]$. To approximate the integral of a function $f: [0, 1] \rightarrow \mathbb{R}$, we first consider the first subinterval $[0, \frac{1}{n}]$. In this interval, we approximate the area below the graph by the area of the rectangle $[0, \frac{1}{n}] \times [0, f(0)]$, which is clearly $\frac{1}{n} \cdot (f(0) - 0)$. See again Figure 42, where this area is just zero. We then consider the second subinterval $[\frac{1}{n}, \frac{2}{n}]$, approximate the area by the rectangle $[\frac{1}{n}, \frac{2}{n}] \times [0, f(\frac{1}{n})]$, which is $\frac{1}{n}f(\frac{1}{n})$, and so on. If we add up the areas of all these rectangles, and call the sum $Q_n(f)$, we obtain

$$Q_n(f) := \frac{1}{n} \sum_{k=0}^{n-1} f\left(\frac{k}{n}\right). \quad (6.1)$$

Similarly, we obtain for functions $f: [a, b] \rightarrow \mathbb{R}$ on arbitrary intervals the sum

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

To assign the function f its integral over $[a, b]$, it remains to show that these sums *converge* if we make n larger and larger. This is given in the following lemma. To keep things simple, we only show the case $[a, b] = [0, 1]$. The general case, can be proven along the same lines.

Moreover, as we already discussed above, our choice of the left endpoint to determine the height of the rectangles was somewhat arbitrary, and we have to justify that it is indeed irrelevant. For this, we show that one might also take the smallest or the largest of these rectangles in each subinterval, and the result would still be the same. For this, define the *lower sums*

$$L_n(f) := \frac{1}{n} \sum_{k=0}^{n-1} \min \left\{ f(x) : x \in \left[\frac{k}{n}, \frac{k+1}{n} \right] \right\}$$

and the *upper sums*

$$U_n(f) := \frac{1}{n} \sum_{k=0}^{n-1} \max \left\{ f(x) : x \in \left[\frac{k}{n}, \frac{k+1}{n} \right] \right\}.$$

(Note that the minima and maxima exist, since f is continuous on the closed (sub-)intervals.) Consider again the above pictures where the lower and upper sums are depicted.

In particular, we have

$$L_n(f) \leq Q_n(f) \leq U_n(f)$$

for every continuous function. (If this is not obvious to you, verify it!) So, if $L_n(f)$ and $U_n(f)$ converge to the same value, then (by the sandwich rule) $Q_n(f)$ also converges to that value.

Lemma 6.28. *Let $f: [0, 1] \rightarrow \mathbb{R}$ be continuous. Then,*

$$\lim_{n \rightarrow \infty} L_n(f) = \lim_{n \rightarrow \infty} U_n(f).$$

In particular, both limits exist. Therefore, also the sequence $(Q_n(f))_{n \in \mathbb{N}}$ converges.

Proof. To prove that the limits of $L_n(f)$ and $U_n(f)$ are equal, we will show that the difference $U_n(f) - L_n(f)$ converges to zero, i.e., that for all $\varepsilon > 0$ there is some $n_0 \in \mathbb{N}$ such that $|U_n(f) - L_n(f)| < \varepsilon$ for all $n \geq n_0$. First of all, note that f is continuous on a closed interval, and therefore uniformly continuous, see Theorem 4.65.

Let us now fix some $\varepsilon > 0$. By the uniform continuity we obtain that there is some $\delta > 0$ such that $|x - y| < \delta$ implies $|f(x) - f(y)| < \varepsilon$. If we now take some $n_0 > \frac{1}{\delta}$, we see that $\frac{1}{n} \leq \frac{1}{n_0} < \delta$ for every $n \geq n_0$. Since the interval $[\frac{k}{n}, \frac{k+1}{n}]$ has length $\frac{1}{n}$, we obtain that $|f(x) - f(y)| < \varepsilon$ for all $x, y \in [\frac{k}{n}, \frac{k+1}{n}]$, if $n \geq n_0$. (We use that $|x - y| < \delta$ for all such x, y and n .) In particular,

$$\max_{x \in [\frac{k}{n}, \frac{k+1}{n}]} f(x) - \min_{x \in [\frac{k}{n}, \frac{k+1}{n}]} f(x) < \varepsilon$$

for all $n \geq n_0$, and therefore

$$U_n(f) - L_n(f) = \frac{1}{n} \sum_{k=0}^{n-1} \left(\max_{x \in [\frac{k}{n}, \frac{k+1}{n}]} f(x) - \min_{x \in [\frac{k}{n}, \frac{k+1}{n}]} f(x) \right) < \frac{1}{n} \sum_{k=0}^{n-1} \varepsilon = \varepsilon.$$

As this holds for all $\varepsilon > 0$, this proves the claim. \square

By the above lemma, it does not matter which specific points we choose in the respective intervals. We always obtain the same limit. Therefore, we can *define* the integral of a function as the limit of the given average of the function values.

Definition 6.29. Let $f: \Omega \rightarrow \mathbb{R}$ be continuous, and consider an interval $[a, b] \subset \Omega$. Then, we define by

$$\int_a^b f(x) dx := \lim_{n \rightarrow \infty} \frac{b-a}{n} \sum_{k=0}^{n-1} f\left(a + \frac{k(b-a)}{n}\right)$$

the **(definite) integral of f over $[a, b]$** . We call a and b the **limits of the integral**.

Note that $\int_a^b f(x) dx$ (if it exists) is a **number**, i.e., the area between the graph and the x -axis. Therefore, the “ x ” is just a placeholder, and one may use any other letter. That is, e.g.,

$$\int_a^b f(x) dx = \int_a^b f(t) dt = \int_a^b f(\xi) d\xi = \dots$$

We sometimes even omit the integration variable, and write $\int_a^b f dx = \int_a^b f(x) dx$.

Although the definition of the integral looks rather simple, it is not a very practical one. The involved limit is usually hard to determine. However, we will see in the following section that the integral can be given in terms of the antiderivative of a function. This is also the typical way of calculating integrals, and justifies the similarity of the notations. But always bear in mind that antiderivatives are functions (more precisely: classes of functions) and not just a number.

Example 6.30. Let us consider the function $f: [0, 1] \rightarrow \mathbb{R}$ given by $f(x) = x$.

One does not need higher mathematics to *see* that the integral (i.e. the area below the graph) is $\frac{1}{2}$. In fact, it is just half of the square with side-length 1. Let us see if this fits our definition. Since f is continuous, its integral is given by

$$\int_0^1 f(x) dx = \int_0^1 x dx = \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{k=0}^{n-1} f\left(\frac{k}{n}\right) = \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{k=0}^{n-1} \frac{k}{n} = \lim_{n \rightarrow \infty} \frac{1}{n^2} \sum_{k=0}^{n-1} k.$$

From the formula $\sum_{k=0}^{n-1} k = \frac{(n-1)n}{2}$ (which is called “Gaußsche Summenformel” in German, but doesn’t seem to have a name in English) we then obtain

$$\int_0^1 f(x) dx = \int_0^1 x dx = \lim_{n \rightarrow \infty} \frac{1}{n^2} \frac{(n-1)n}{2} = \lim_{n \rightarrow \infty} \frac{1 - \frac{1}{n}}{2} = \frac{1}{2}.$$

□

From the definition of the integral as a limit, we obtain that it satisfies a list of rules, like linearity. Most of them may even already be clear from the “graphical definition”. We state them without a proof.

Lemma 6.31. *Let $f, g: \Omega \rightarrow \mathbb{R}$ be continuous functions, and let $[a, b] \subset \Omega$. Then,*

- 1) $f = 0$ on $[a, b] \implies \int_a^b f dx = 0$
- 2) $\int_a^a f dx = 0$ (*That's the area over an interval with length 0.*)
- 3) $\int_a^b f + g dx = \int_a^b f dx + \int_a^b g dx$ (*Linearity*)
- 4) $\int_a^b \lambda \cdot f(x) dx = \lambda \cdot \int_a^b f dx$ for $\lambda \in \mathbb{R}$. (*Homogeneity*)
- 5) $c \in [a, b] \implies \int_a^b f dx = \int_a^c f dx + \int_c^b f dx$ (*Splitting the area in two parts.*)
- 6) $f \leq g \implies \int_a^b f dx \leq \int_a^b g dx$ (*Monotonicity w.r.t. the function*)
- 7) $f \geq 0$ and $[c, d] \subset [a, b] \implies \int_c^d f dx \leq \int_a^b f dx$ (*Monotonicity w.r.t. the limits*)

There is one case of the above lemma that is particularly important. In analogy to the very similar inequality for (finite) sums, this is also called **triangle inequality**.

Corollary 6.32. *Let $f: \Omega \rightarrow \mathbb{R}$ be a continuous function, and let $[a, b] \subset \Omega$. Then,*

$$\left| \int_a^b f dx \right| \leq \int_a^b |f| dx.$$

Proof. Clearly, $f \leq |f|$ and $-f \leq |f|$. The inequality therefore follows from $|x| = \max\{x, -x\}$, Lemma 6.31(4) with $\lambda = -1$, and Lemma 6.31(6). \square

Let us finally state some remarks on difficulties with and variants of the above definition.

Remark 6.33 (Non-continuous functions). Continuity seems to be an unnecessary assumption for the considerations above. For example, the definitions from above make perfect sense for indicator functions of intervals: For a given set A , the **indicator function**, which is denoted by χ_A , satisfies $\chi_A(x) = 1$ for $x \in A$, and $\chi_A(x) = 0$ for $x \notin A$. Clearly, χ_A is not continuous if $A \subsetneq \mathbb{R}$. If we consider an interval $[a, b] \subset [0, 1]$, then it is not hard to verify (Do it!) that

$$\int_0^1 \chi_{[a,b]}(x) dx = b - a,$$

which equals the *true* area. (Note that Lemma 6.28 holds also for such functions.)

We will comment on such “piecewise functions” in detail in Section 6.6.

If we consider the *Dirichlet function* $\chi_{\mathbb{Q}}$ instead, i.e., the indicator function of the rational numbers, then, with our definition, we would obtain $\int_0^1 \chi_{\mathbb{Q}} dx = 1$, because all the function values we compute are at rational points. And therefore, $\int_0^1 \chi_{\mathbb{R} \setminus \mathbb{Q}} dx = 1 - \int_0^1 \chi_{\mathbb{Q}} dx = 0$. However, this is unsatisfactory (and in contrast to our intuition), since there are more irrational than rational numbers. One might check that Lemma 6.28 fails for this function. Hence, the outcome of our “integration procedure” depends on the chosen function values, and we have to be careful how we define an integral.

To solve this issue (at least partially) one must be more elaborate and think about a definition of **integrable functions**. One could then define the integral for a much larger class of functions.

(Still not for all functions, see Remark 6.26.) We will talk about a more powerful definition, i.e., the Lebesgue integral, later. But, as stated many times, every generalization of the above concept should lead to the same result when applied to a continuous function.

Remark 6.34 (Improper integrals). In the definitions above it was essential that we talk about continuous (and therefore bounded) functions on a closed (and therefore bounded) interval. This implies that all the rectangles that were used for the approximation of the integral have finite size. Otherwise, the above definition is clearly useless. However, this can be relaxed a bit by considering *improper integrals*. With this, it is also possible to define the integral of some unbounded functions on unbounded intervals, i.e., we may determine the area of unbounded sets. We will shortly come back to this when we know how to compute integrals fast.

Remark 6.35. The “Q” in $Q_n(f)$ in (6.1) stands for “quadrature rule”. This is how such averages over function values are usually called. Quadrature rules are also very much used in practice, but then one sometimes has to think about “more clever” averages. In particular, this is important as we cannot compute the above limits in general, and therefore have to work with *finite* n , which leads to errors that we might want to minimize. However, as we only work with the limits here, Lemma 6.28 shows that such optimizations are not needed.

Remark 6.36. Even if we would work with more *clever* choices of points in the definition of the quadrature rule, and therefore ultimately in the definition of the integral, this would still be not enough for a “satisfactory definition”. (We do not want to comment here on what this means exactly.) One of the first definitions that met these standards is the **Riemann integral**, see Remark 6.1. For this definition we not only have to consider arbitrary points in each subinterval, we also have to consider arbitrary subintervals (i.e., partitions) of the given interval. As this is more technical than needed here, we skip the precise definitions and basic results, which can be found in most beginners books on calculus.

6.4 The fundamental theorem of calculus

We now turn to the *fundamental theorem of calculus*, which provides us with an easier way of determining integrals, and which shows the existence of antiderivatives for continuous functions. Recall that a differentiable function F is an antiderivative of f if and only if $F' = f$. We start with the following additional result, which is of independent interest.

Theorem 6.37 (Mean value theorem for definite integrals). *Let $f, g: [a, b] \rightarrow \mathbb{R}$ be continuous functions, and assume that $g \geq 0$. Then, $\int_a^b f(x)g(x) dx$ exists and there exists some $\xi \in [a, b]$ such that*

$$\int_a^b f(x)g(x) dx = f(\xi) \int_a^b g(x) dx.$$

In particular, we have (for $g = 1$) that

$$\frac{1}{b-a} \int_a^b f(x) dx = f(\xi)$$

for some $\xi \in [a, b]$.

Proof. Since f is continuous, it attains its extrema on $[a, b]$. Let us denote them by $m := \min_{x \in [a, b]} f(x)$ and $M := \max_{x \in [a, b]} f(x)$. It follows that $m g(x) \leq f(x)g(x) \leq M g(x)$ for all $x \in [a, b]$, and therefore

$$m \int_a^b g(x) dx \leq \int_a^b f(x)g(x) dx \leq M \int_a^b g(x) dx,$$

see Lemma 6.31. We define $I = \int_a^b g(x) dx$ and obtain

$$m \cdot I \leq \int_a^b f(x)g(x) dx \leq M \cdot I.$$

If $I = 0$, then $g = 0$ and any $\xi \in [a, b]$ can be used to obtain the result. Otherwise we divide by I and get

$$m \leq \frac{1}{I} \int_a^b f(x)g(x) dx \leq M.$$

Due to the intermediate value theorem f attains every value in the interval $[m, M]$ (i.e., between its extreme values), so in particular the value $\frac{1}{I} \int_a^b f(x)g(x) dx$. □

This is all we need to formulate the main result of this section. For this, we set

$$\int_a^b f dx := - \int_b^a f dx,$$

whenever $b < a$. (Note that we have defined the left hand side only for $a < b$.)

Theorem 6.38 (Fundamental theorem of calculus). *Let f be continuous on some open interval $I \subset \mathbb{R}$, and $a \in I$. Then, the function $F: I \rightarrow \mathbb{R}$ defined by*

$$F(x) = \int_a^x f(y) dy$$

is an antiderivative of f , i.e., $F' = f$.

Moreover, for any $a, b \in I$ and any antiderivative F of f , we have

$$\int_a^b f(x) dx = F(b) - F(a),$$

and we write $[F]_a^b := F(b) - F(a)$.

Proof. First we show that F as given is an antiderivative of f . Therefore we calculate the derivative of F , by considering its difference quotient. For $h \neq 0$, we have

$$\begin{aligned} \frac{F(x+h) - F(x)}{h} &= \frac{1}{h} \left(\int_a^{x+h} f(y) dy - \int_a^x f(y) dy \right) \\ &= \frac{1}{h} \left(\int_a^x f(y) dy + \int_x^{x+h} f(y) dy - \int_a^x f(y) dy \right) \\ &= \frac{1}{h} \int_x^{x+h} f(y) dy, \end{aligned}$$

where we used Lemma 6.31(5) in the second equation.

The mean value theorem implies the existence of some $\xi_h \in [x, x+h]$ such that

$$\frac{1}{h} \int_x^{x+h} f(y) dy = f(\xi_h).$$

Clearly, $\xi_h \rightarrow x$ as $h \rightarrow 0$. The continuity of f therefore yields $f(\xi_h) \rightarrow f(x)$ for $h \rightarrow 0$. This shows the first part of the theorem, i.e.,

$$F'(x) = \lim_{h \rightarrow 0} \frac{F(x+h) - F(x)}{h} = \lim_{h \rightarrow 0} f(\xi_h) = f(x)$$

for all $x \in I$. For the second part, we first plug in a and b into the antiderivative that we just defined and obtain

$$F(b) - F(a) = \int_a^b f(y) dy - \int_a^a f(y) dy = \int_a^b f(y) dy.$$

Moreover, note that different antiderivatives differ only by a constant. Therefore, the right hand side is the same for any antiderivative of f . □

Remark 6.39. The last theorem shows that integration gives us an antiderivative of a continuous function f , and one may ask for an interpretation of this. One somehow sketchy possibility, which originates from a graphical point of view, is the following:

If we *start* at $F(a)$ and then *add all the changes* that F makes between a and b (if this would be possible), then we *arrive* at $F(b)$. Now, these changes (or slopes) are precisely the values of the derivative of F , and summing all of them up is like integration. So, one might guess that $F(b) = F(a) + \int_a^b F'(x) dx$. But, since $F' = f$, this is exactly what the fundamental theorem of calculus is about.

With this very important and powerful theorem we can calculate many integrals easily. (At least if you know many antiderivatives.) Let us consider again the example from above.

Example 6.40. Let us consider the function $f: [0, 1] \rightarrow \mathbb{R}$ given by $f(x) = x$.

We already showed in Example 6.30 that $\int_0^1 x dx = \frac{1}{2}$. This may also be shown by considering $F(x) := \frac{x^2}{2}$, which is an antiderivative of f . We therefore have from the fundamental theorem that

$$\int_0^1 x dx = \left[\frac{x^2}{2} \right]_0^1 = \frac{1^2}{2} - \frac{0^2}{2} = \frac{1}{2}.$$

We now consider some more complicated examples, which may be difficult to compute only with the definition of the integral as a limit.

Example 6.41. We know from Example 6.10 the antiderivative $\int (x^3 + x^2) dx = \frac{x^4}{4} + \frac{x^3}{3}$ on \mathbb{R} . We therefore obtain, for example, the values

$$\int_0^1 (x^3 + x^2) dx = \left[\frac{x^4}{4} + \frac{x^3}{3} \right]_0^1 = \frac{1}{4} + \frac{1}{3} = \frac{7}{12}.$$

or

$$\int_{-1}^1 (x^3 + x^2) dx = \left[\frac{x^4}{4} + \frac{x^3}{3} \right]_{-1}^1 = \frac{2}{3}.$$

(Try to find these values using the definition of the integral, or by any other means.)

Example 6.42. We now want to compute definite integrals of the natural logarithm $\ln(x)$. From Example 6.14 we know that $\int \ln(x) dx = x(\ln(x) - 1)$ for all $x > 0$, i.e., all x such that $\ln(x)$ is defined. From this, we obtain, e.g., the value

$$\int_1^e \ln(x) dx = [x(\ln(x) - 1)]_1^e = e(\ln(e) - 1) - (\ln(1) - 1) = 1.$$

These examples already show that, with the techniques we've just learned, we can calculate integral (or areas) precisely without much effort. (However, it is again essential that you know many (anti)derivatives and how to obtain them.)

We will finally present (again) the rules for integration, that we already discussed in the section about antiderivatives. Although they can be directly deduced from there, we state them again for definite integrals to clarify their meaning.

Let us start with *integration by parts*, which is also called partial integration.

Corollary 6.43 (Integration by parts). *Let f, g be continuously differentiable on $[a, b]$. Then,*

$$\int_a^b f'(x)g(x) dx = [fg]_a^b - \int_a^b f(x)g'(x) dx.$$

Proof. This follows from Lemma 6.13 together with Theorem 6.38. □

Remark 6.44. Note again that we did not define the derivative at a boundary point. Hence, whenever we assume a function to be “continuously differentiable on $[a, b]$ ”, we actually mean that the function is “continuous on $[a, b]$ and continuously differentiable on (a, b) ”. We think this should not lead to any confusion.

Example 6.45. If we want to calculate $\int_0^\pi x \sin x dx$, we set $f'(x) = \sin x$ and $g(x) = x$, which implies that $g'(x) = 1$ and $f(x) = -\cos x$, see Example 6.15. Partial integration yields

$$\begin{aligned} \int_0^\pi x \sin x dx &= [(-\cos x)x]_0^\pi - \int_0^\pi (-\cos x) dx \\ &= [(-\cos x)x]_0^\pi + [\sin x]_0^\pi = -\cos(\pi) \cdot \pi + \sin \pi = \pi. \end{aligned}$$

Example 6.46. Again, we can use this formula also more than once, see Example 6.16. However, the involved formulas for definite integrals can sometimes be much simplified, in contrast to antiderivatives. For example, if we want to calculate $\int_0^1 x^2 e^x dx$, we set $g(x) = x^2$ and $f'(x) = e^x$, which implies that $g'(x) = 2x$ and $f(x) = e^x$. Partial integration yields

$$\int_0^1 x^2 \cdot e^x dx = [x^2 e^x]_0^1 - 2 \int_0^1 x e^x dx = e - 2 \int_0^1 x e^x dx.$$

Applying the rule once more we end up with

$$\int_0^1 x^2 \cdot e^x dx = e - 2[x e^x - e^x]_0^1 = e - 2.$$

(Check this in detail!)

Next we consider again the substitution rule, see Lemma 6.19.

Corollary 6.47 (Substitution rule). *Let $I = [a, b]$, f be continuous and g be continuously differentiable on I . Then,*

$$\int_a^b f(g(x)) g'(x) dx = \int_{g(a)}^{g(b)} f(y) dy$$

Note that, usually, we have a (complicated looking) integral like the one on the left and want to transform it to a more easy one, like the one on the right. We will come to some examples soon.

Proof. Although this follows directly from Lemma 6.19 together with Theorem 6.38, we present a proof here, because this one is slightly different and one can see where the fundamental theorem comes in. First of all, we use the chain rule, see Theorem 5.14, to calculate

$$(F \circ g)'(x) = (F' \circ g)(x)g'(x).$$

If now F is an antiderivative of f , i.e., $F' = f$, then we can apply the fundamental theorem of calculus two times to see

$$\int_{g(a)}^{g(b)} f(x) dx = F(g(b)) - F(g(a)) = \int_a^b (F(g(x)))' dx = \int_a^b f(g(x))g'(x) dx.$$

□

When we use this rule we can use the following (formally not completely correct) strategy:

1. We want to calculate $\int_a^b f(g(x))g'(x) dx$, i.e., we assume that the integral is of this form for some f, g .
2. Set $y = g(x)$.
3. Calculate $\frac{dy}{dx} = \frac{d}{dx}g(x) = g'(x)$. (Now check again if the integral is of the above form!)
4. Regroup to $dx = \frac{1}{g'(x)} dy$ and plug in.
5. Since the new integration variable is $y = g(x)$, we have to replace the limits of the integral by $g(a)$ and $g(b)$.
6. Consider the new integral $\int_{g(a)}^{g(b)} f(y) dy$.

Let us see some examples.

Example 6.48. Consider the integral $\int_0^\pi \sin(2x) dx$.

This is of the form above with $f(x) = \sin(x)$ and $g(x) = 2x$. Following the above procedure, we set $y = 2x$ and calculate $\frac{dy}{dx} = 2$, which implies $dx = \frac{dy}{2}$. Hence,

$$\int_0^\pi \sin(2x) dx = \frac{1}{2} \int_0^{2\pi} \sin(y) dy = \frac{1}{2} [-\cos(y)]_0^{2\pi} = \frac{1}{2} (\cos(0) - \cos(2\pi)) = 0.$$

Example 6.49. Consider the integral $\int_1^2 \frac{1}{2x+42} dx$.

Setting $y = 2x + 42$ we obtain $\frac{dy}{dx} = 2$, which implies $dx = \frac{dy}{2}$.

$$\int_1^2 \frac{1}{2x+42} dx = \frac{1}{2} \int_{44}^{46} \frac{1}{y} dy = \frac{1}{2} (\ln(46) - \ln(44)).$$

The above examples can be summarized in the following general rule.

Example 6.50 (Linear substitution). Let F be an antiderivative of some continuous function f , and consider the integral

$$\int_a^b f(\alpha x + \beta) dx,$$

where $\alpha \neq 0$. We set $y = \alpha x + \beta$. This implies $dx = \frac{1}{\alpha} dy$, and hence

$$\int_a^b f(\alpha x + \beta) dx = \frac{1}{\alpha} \int_{\alpha a + \beta}^{\alpha b + \beta} f(y) dy = \frac{1}{\alpha} (F(\alpha b + \beta) - F(\alpha a + \beta)).$$

Once one gets used to this kind of substitutions, it should be no problem to calculate integrals like

$$\int_0^2 \sin(7x - 11) dx = \frac{-\cos(3) + \cos(-11)}{7}$$

in short time.

Let us finish this section with a **warning** regarding a typical mistake.

Example 6.51. Consider the function $f(x) = \frac{1}{x^2}$, and try to compute $\int_{-1}^1 \frac{1}{x^2} dx$.

When we try to use the strategies we used so far, and are not careful enough, then we might use the antiderivative of f , which we know to be $F(x) = \frac{1}{-x}$, and compute

$$\int_{-1}^1 \frac{1}{x^2} dx = \left[\frac{1}{-x} \right]_{-1}^1 = -1 - 1 = -2.$$

However, this is clearly wrong since the integral of a positive function must be positive. But where is the mistake? The problem is that f is not continuous on $[-1, 1]$, which is necessary for the fundamental theorem. In fact, the function is not even defined at $x = 0$.

Although this looks like a drastical example, it also shows that it is important to verify all assumption before we apply such a theorem. Otherwise, the result might be completely senseless.

6.5 Improper integrals

So far we discussed how to calculate integrals (or areas below graphs), whenever the function and the corresponding interval is bounded. However, one might imagine that we can also calculate integrals of functions over unbounded intervals if the function is “small enough” for large x . By a similar reasoning, we can integrate functions with a *pole* at the boundary, i.e., functions that diverge to infinity at the boundary of the interval, if the divergence is “fast enough”. We discuss both cases.

Let us start with unbounded intervals. In this case, we define the integrals

$$\begin{aligned}\int_a^\infty f dx &:= \lim_{b \rightarrow \infty} \int_a^b f dx, \\ \int_{-\infty}^b f dx &:= \lim_{a \rightarrow -\infty} \int_a^b f dx, \\ \int_{-\infty}^\infty f dx &:= \int_{-\infty}^0 f dx + \int_0^\infty f dx\end{aligned}$$

whenever the integrals and limits on the right hand side exist. We then say that the integrals on the left *converge*.

From the fundamental theorem of calculus, we know how to compute the finite integrals on the right hand side. That is, if F is an antiderivative of f (on the corresponding interval), then $\int_a^b f dx = F(b) - F(a)$. To compute the above limits, it is therefore enough to compute the limits for the antiderivative. That is, if we denote

$$F(-\infty) := \lim_{a \rightarrow -\infty} F(a) \quad \text{and} \quad F(\infty) := \lim_{b \rightarrow \infty} F(b),$$

then we obtain

$$\begin{aligned}\int_a^\infty f dx &= [F]_a^\infty = F(\infty) - F(a), \\ \int_{-\infty}^b f dx &= [F]_{-\infty}^b = F(b) - F(-\infty), \\ \int_{-\infty}^\infty f dx &= [F]_{-\infty}^\infty = F(\infty) - F(-\infty),\end{aligned}$$

whenever the corresponding limits exist.

Let us see some examples.

Example 6.52. Consider $f(x) = \frac{1}{x^\alpha} = x^{-\alpha}$ on $[1, \infty)$. An antiderivative of f is given by $F(x) = \frac{1}{1-\alpha}x^{1-\alpha}$ for $\alpha \neq 1$, and $F(x) = \ln(x)$ for $\alpha = 1$. Noting that $F(x)$ converges to 0 when $x \rightarrow \infty$ if $\alpha > 1$, and diverges otherwise, we obtain

$$\int_1^\infty \frac{dx}{x^\alpha} = F(\infty) - F(1) = \begin{cases} \frac{1}{\alpha-1} & \text{for } \alpha > 1, \\ \infty & \text{for } \alpha \leq 1. \end{cases}$$

From this example we can deduce the following general rule for improper integrals. Note that we had a very similar statement for series, see Lemma 3.91.

Lemma 6.53. Let $a > 0$ and $f: [a, \infty) \rightarrow \mathbb{R}$ be a continuous function. Then,

- $\int_a^\infty f dx$ is convergent if $f(x) \leq \frac{c}{x^\beta}$ for some $c < \infty$, $\beta > 1$.
- $\int_a^\infty f dx$ is divergent if $f(x) \geq \frac{c}{x}$ for some $c > 0$ and all large enough x .

Proof. Exercise. □

As another example, show the following.

Example 6.54.

$$\int_{-\infty}^\infty \frac{dx}{1+x^2} = \pi.$$

By the above arguments, we see that it is not necessary for the antiderivative to be defined at the limits of the integral. (Note that a function on \mathbb{R} is never *defined* at $\pm\infty$. We can only compute limits.) This can also be used if a function is defined, and has an antiderivative, on an open interval (a, b) , but not on the boundary points.

Example 6.55. Consider for example the functions $f(x) = \frac{1}{x}$ or $f(x) = \ln(x)$ on $(0, 1)$. Both functions are continuous on $(0, 1)$, and therefore have an antiderivative on $(0, 1)$. However, for computing the integral $\int_0^1 f dx$ by using the fundamental theorem directly, we would need the values $F(0)$ and $F(1)$. But, since both functions are not even defined at 0, it makes no sense to ask for a function whose derivative equals f at 0, i.e., there cannot be an antiderivative at 0.

Consider a continuous function $f: (a, b) \rightarrow \mathbb{R}$, and let $F: (a, b) \rightarrow \mathbb{R}$ one of its antiderivatives. We then define

$$F(a) := \lim_{x \searrow a} F(x) \quad \text{and} \quad F(b) := \lim_{x \nearrow b} F(x),$$

if these limits exist, and set

$$\int_a^b f dx = F(b) - F(a).$$

One can say that we replace F by its continuous extension to $[a, b]$, when it exists, and then use this extension in the fundamental theorem.

Remark 6.56. One could prove that the above limits of the function F exist if and only F is uniformly continuous on (a, b) . We omit the details here.

Let us again see an example.

Example 6.57. Consider again $f(x) = \ln x$ on $(0, 1)$. From Example 6.42 we know that $F(x) = x(\ln x - 1)$ is an antiderivative of f . First of all, $F(1) = 1(0 - 1) = -1$, but $F(0)$ (“= 0($-\infty - 1$)”) is not defined. However, we can use l’Hospitals rule, see Theorem 5.36, to calculate the corresponding limit. We obtain

$$\lim_{x \searrow 0} F(x) = \lim_{x \searrow 0} \frac{\ln x - 1}{1/x} = \lim_{x \searrow 0} \frac{1/x}{-1/x^2} = \lim_{x \searrow 0} -x = 0$$

and therefore

$$\int_0^1 \ln x dx = F(1) - F(0) = -1.$$

Example 6.58. Consider $f(x) = \frac{1}{x^\alpha} = x^{-\alpha}$ on $(0, 1]$. An antiderivative of f is given by $F(x) = \frac{1}{1-\alpha}x^{1-\alpha}$ for $\alpha \neq 1$, and $F(x) = \ln(x)$ for $\alpha = 1$. Therefore, $F(1) = \frac{1}{1-\alpha}$ for $\alpha \neq 1$. Noting that $F(x)$ converges to 0 when $x \rightarrow 0$ if $\alpha < 1$, and diverges otherwise, we obtain

$$\int_0^1 \frac{dx}{x^\alpha} = F(1) - F(0) = \begin{cases} \frac{1}{1-\alpha} & \text{for } \alpha < 1, \\ \infty & \text{for } \alpha \geq 1. \end{cases}$$

In particular, we obtain $\int_0^1 \frac{1}{\sqrt{x}} dx = 2$ (for $\alpha = \frac{1}{2}$) and $\int_0^1 x dx = \frac{1}{2}$ (for $\alpha = -1$), but $\int_0^1 \frac{1}{x} dx$ does not converge.

Remark 6.59. Note that, by the Examples 6.52 and 6.58, we have that $\int_0^\infty \frac{dx}{x^\alpha} = \infty$ for all $\alpha \in \mathbb{R}$.

Example 6.60. Show that the results of Examples 6.52 and 6.58 are equivalent by using the substitution rule with $f(x) = g(x) = \frac{1}{x}$.

6.6 Piecewise continuous functions

Let us finally comment on functions, which have some desired properties only *piecewise*.

Definition 6.61. Let $I = [a, b]$. We say that a function $f: I \rightarrow \mathbb{R}$ is **piecewise continuous** if and only if there are exist a finite number of points $x_1, \dots, x_m \in I$ such that

1. f is continuous on every subinterval $[a, x_1], (x_m, b]$ and (x_k, x_{k+1}) for $k = 1, \dots, m-1$,
2. the limits $\lim_{x \nearrow x_k} f(x)$ and $\lim_{x \searrow x_k} f(x)$ exist and are finite.

We call x_1, \dots, x_m the **(finite) discontinuities** of f .

A simple example of a function for which such piecewise considerations might be necessary is the indicator function of an interval $[c, d] \subset \mathbb{R}$, i.e.

$$\chi_{[c,d]}(x) := \begin{cases} 1, & \text{if } x \in [c, d], \\ 0, & \text{if } x \notin [c, d]. \end{cases}$$

However, one might also think about other *piecewise defined* functions, like

$$f(x) := \begin{cases} -x^2, & \text{if } x < 0, \\ 2x^2 + 1, & \text{if } x \in [0, 1], \\ x, & \text{if } x > 1. \end{cases}$$

These functions are clearly not continuous on \mathbb{R} . However, when restricting to the individual “pieces” of the functions then they are continuous. (Actually, both function are infinitely often differentiable on each subinterval.) Since also the needed (one-sided) limits are finite, both functions are piecewise continuous.

Now, to compute the integral of such piecewise continuous functions, we can just split the integral into the corresponding parts and then use the respective rules for calculating integrals. That is, if $f: [a, b] \rightarrow \mathbb{R}$ is a piecewise continuous function with exceptions x_1, \dots, x_m , then we use

$$\int_a^b f dx = \int_a^{x_1} f dx + \sum_{k=1}^{m-1} \int_{x_k}^{x_{k+1}} f dx + \int_{x_m}^b f dx.$$

Since, f is now a continuous function in each subinterval. This expression is well-defined. For example, we easily obtain

$$\int_{-\infty}^{\infty} \chi_{[c,d]} dx = \int_{-\infty}^c 0 dx + \int_c^d 1 dx + \int_d^{\infty} 0 dx = d - c.$$

However, note that the subintervals (x_k, x_{k+1}) are (by Definition 6.61) open intervals. Therefore, formally, we need to treat the integrals as *improper integrals*. The assumption about the one-sided limits ensures that these integrals always exist. (One might argue with an continuous extension of f to the closed interval $[x_k, x_{k+1}]$.)

Remark 6.62. The assumption about the one-sided limits could be relaxed a bit. However, since we want to say that “every piecewise continuous function is integrable”, we need an assumption that excludes, e.g., $\frac{1}{x}$ of being piecewise continuous.

7 Fourier series

As discussed several times, it is the major task of natural and applied science to give an approximation of reality (which is actually a complicated function). We have already used derivatives to obtain approximations of functions by using Taylor's theorem, see Theorem 5.53 and Corollary 5.57. Although this is often very useful in theory, it has several drawbacks when it comes to actual computations. First of all, the *quality* of the Taylor polynomial of a function is limited by its smoothness, i.e., how often the function is differentiable, and the size of the domain. This is clearly unsatisfactory, as an actual computational problem may not be "nice" in this respect. For example, *classification problems* are naturally not concerned with smooth functions.

In this section we introduce and discuss *Fourier series*, which is a particular famous way of writing functions as series based on certain integrals. The main idea is that functions can be written as superpositions (or sums) of *wave functions*, represented by certain multiples of cos and sin. That is, for many functions $f: [0, 1] \rightarrow \mathbb{R}$, we can find numbers a_k, b_k such that the *trigonometric polynomials*

$$\sum_{k=0}^n (a_k \cos(2\pi kx) + b_k \sin(2\pi kx)) \approx f(x)$$

are "good" approximations of f for large enough n , where a_k, b_k must depend on f , much like the coefficients in the Taylor polynomial. In particular, one may recover f if we send $n \rightarrow \infty$.

Before we come to precise definitions, let us give some comments on the (partly recent) history of the theory of Fourier series.

Remark 7.1 (History). Fourier series and the corresponding *Fourier analysis* are nowadays of remarkable significance in a lot of applied sciences, especially in physics (acoustics, optics, astrophysics) but also in signal processing, cryptography, oceanography and economics.

The first attempts of using trigonometric series for the approximations of functions dates back to 1740 when the mathematicians *Daniel Bernoulli* (1700–1782) and *Jean-Baptiste le Rond d'Alembert* (1717–1783) discussed this possibility. But only when *Jean Baptiste Joseph Fourier* (1768–1830) presented his famous work "Théorie analytique de la chaleur" ("The analytical theory of heat") in 1822, it became apparent how powerful these techniques are. Fourier managed to solve the *heat equation* (in one dimension) by using the series, which are by now referred to as *Fourier series*.

At this time it was the general thought that every continuous function can be written as such a series. However, one of the first actual convergence results is due to *Peter Gustav Lejeune Dirichlet* (1805–1859). In 1829 he proved that the Fourier series of a Lipschitz continuous function converges. In order to treat Fourier series, *Bernhard Riemann* (1826–1866) actually invented his definition of an integral (the Riemann integral) and discovered the so called localization theorem in 1853. It took until 1876 for *Paul du Bois-Reymond* (1831–1889) to find a continuous function, whose Fourier series did not converge at any point. This was a big surprise at that time. However, in 1904 the Hungarian mathematician *Leopold Fejér* (1880–1959) could show that for any continuous function, its Fourier series converges on arithmetic average. This means, that we can recover any continuous function from its Fourier coefficients, but we need to be careful *how to use them*. This was a major breakthrough, and lead to big advances in theoretical and applied sciences.

The final word on this problem was given only in 1966 when the Swedish mathematician *Lennart Carleson* (1928–now) showed that the Fourier series of any *square-integrable function* converges *almost everywhere*. (We will see later what this means.) This was a question posed in 1915 by *Nikolai Nikolajewitsch Lusin* (1883–1950), and Carleson became world-famous for this. \square

7.1 Periodic functions and trigonometric polynomials

As all the theory so far, and all that will come, we need assumptions about the functions under consideration. Here, we will mostly assume that the functions $f: [0, 1] \rightarrow \mathbb{R}$ are (piecewise) continuous. This is to ensure that the integrals that we use are well-defined. This assumption is not necessary for many claims, but we do not have the theoretical background to treat more general cases, so far.

Moreover, as we want to write functions as sums of cosine and sine functions, which are obviously *periodic* when considered as functions on the real line, it is natural to assume the same for the functions under consideration.

Definition 7.2. A function $f: \mathbb{R} \rightarrow \mathbb{C}$ is called **periodic**, or **1-periodic**, if

$$f(x + 1) = f(x) \quad \text{for all } x \in \mathbb{R}.$$

Note that a periodic function is *completely known* if we know its function values in $[0, 1]$. All other function values on \mathbb{R} follow from periodicity. E.g., $f(0) = f(1)$ and $f(\frac{7}{3}) = f(\frac{1}{3})$ for periodic functions. That's why we also call functions defined on $[0, 1]$ periodic functions, and mean by this its **periodic extension** to \mathbb{R} , i.e., we define $f(x) := f(\{x\})$ for $x \in \mathbb{R} \setminus [0, 1]$, where $\{x\} := x - \lfloor x \rfloor \in [0, 1)$ is the *fractional part* of x . See Figure 44 for the periodic extensions of the functions $f, g: [0, 1] \rightarrow \mathbb{R}$ with $f(x) = x$ and $g(x) = \frac{1}{2} - |\{x\} - \frac{1}{2}|$, which are called *sawtooth wave* and *triangle wave*.

(Note that, by the periodic extension, we have $f(2) = f(0) = 0$ and not $f(2) = 2$.)

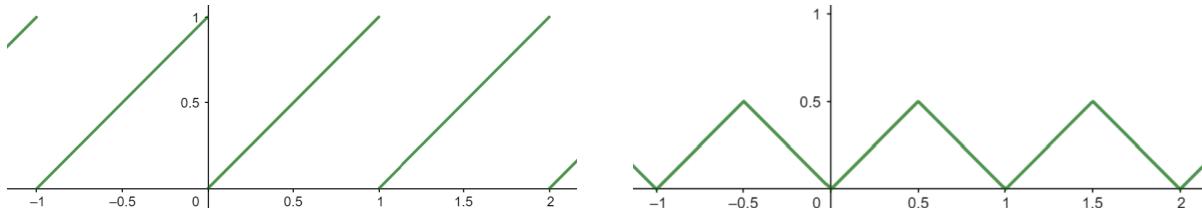


Figure 44: The sawtooth wave and the triangle wave.

This is useful when describing properties of a function, which include the boundary points. For example, the above functions show that, while both f and g are continuous on $[0, 1]$, only the triangle wave g is continuous when considered as a periodic function. We fix this notation in the following definition.

Definition 7.3. We say that a periodic function $f: [0, 1] \rightarrow \mathbb{C}$ has a property if and only if its periodic extension to \mathbb{R} , i.e., the function $f: \mathbb{R} \rightarrow \mathbb{C}$ with $f(x) = f(\{x\})$, has this property.

For example, when we say that a **periodic function f is continuous**, then this also implies that the function values at the boundary coincide, or, more precisely, that $\lim_{x \searrow 0} f(x) = \lim_{x \nearrow 1} f(x)$. The same statements hold for differentiability and so on. In particular, note that the left function in Figure 44, i.e., the sawtooth wave, is also (infinitely often) differentiable if considered as a function on $(0, 1)$. However, as a periodic function it is not even continuous.

Let us come to the most important examples, which are called *trigonometric monomials*, in correspondence to the (*algebraic*) monomials x^k for $k \in \mathbb{N}$.

Example 7.4 (Trigonometric monomials). The functions $\cos(2\pi kx)$ and $\sin(2\pi kx)$ are infinitely often differentiable and 1-periodic functions for any $k \in \mathbb{Z}$. Verify this yourself and make plots of the functions for different k .

Remark 7.5 (Other periods). It is somewhat arbitrary to choose 1 as the length of a period. Another standard choice would be to consider 2π -periodic functions, i.e., functions that satisfy $f(x) = f(x + 2\pi)$, like e.g. $\cos(x)$. We chose this *normalization* to ease the notation.

More general, one can study functions with an arbitrary period $\omega > 0$, i.e., we have $f(x + \omega) = f(x)$ for all $x \in \mathbb{R}$, but note that, in this case, we always have that the function $x \mapsto f(\omega \cdot x)$ is a 1-periodic function.

Although we considered so far mostly real-valued functions, it is very natural to **study Fourier series directly for complex-valued functions**. For this, note that every complex-valued function $f: \mathbb{R} \rightarrow \mathbb{C}$ can be written as $f(x) = u(x) + i \cdot v(x)$, where $u, v: \mathbb{R} \rightarrow \mathbb{R}$ are both real-valued, and $i = \sqrt{-1}$. We call such a function f continuous/differentiable/etc., if the same is true for the *real part* u and the *imaginary part* v .

An especially important class of periodic functions are the *trigonometric polynomials*. These are sums of the cosine and sine functions discussed in Example 7.4. Recall that (*algebraic*) polynomials were functions $p: \mathbb{R} \rightarrow \mathbb{R}$ of the form $p(x) = \sum_{k=0}^n a_k x^n$. We showed that, under certain assumptions, we have that a function can be approximated quite well by using algebraic polynomials. This was Taylor's theorem 5.53.

We now want to have similarly simple “building blocks” also for periodic functions, and it turns out that the functions $\cos(2\pi kx)$ and $\sin(2\pi kx)$ are suitable. However, as it simplifies the notation a lot and is a very useful tool, we use **Euler's formula**

$$e^{it} = \cos(t) + i \sin(t) \quad \text{for } t \in \mathbb{R},$$

and use $e^{2\pi ikx} = \cos(2\pi kx) + i \sin(2\pi kx)$, which are 1-periodic, as building blocks for trigonometric polynomials. But note that we also have to consider “negative frequencies” then.

Definition 7.6. A **trigonometric polynomial** $p: \mathbb{R} \rightarrow \mathbb{C}$ is a periodic function of the form

$$p(x) = \sum_{k=-n}^n c_k e^{2\pi ikx},$$

where $n \in \mathbb{N}$ and $c_{-n}, \dots, c_n \in \mathbb{C}$ are called the **coefficients** of the trigonometric polynomial.

Note that this very short notation for a trigonometric polynomial can clearly be written out by using cosine and sine terms. In particular, by using Euler's formula, we obtain that

$$\begin{aligned} p(x) &= \sum_{k=-n}^n c_k e^{2\pi ikx} = \sum_{k=-n}^n c_k (\cos(2\pi kx) + i \sin(2\pi kx)) \\ &= \sum_{k=-n}^n c_k \cos(2\pi kx) + i \sum_{k=-n}^n c_k \sin(2\pi kx) \end{aligned}$$

However, this representation can be simplified further by using that cosine is **even**, i.e., $\cos(-x) = \cos(x)$, and that sine is an **odd function**, i.e., $\sin(-x) = -\sin(x)$. From this we get

$$p(x) = c_0 + \sum_{k=1}^n (c_k + c_{-k}) \cos(2\pi kx) + i \sum_{k=1}^n (c_k - c_{-k}) \sin(2\pi kx).$$

(Check this in detail!) We may therefore write trigonometric polynomials, as in Definition 7.6, in the form

$$p(x) = a_0 + \sum_{k=1}^n a_k \cos(2\pi kx) + \sum_{k=1}^n b_k \sin(2\pi kx).$$

if we set $a_0 = c_0$ and, for $k \geq 1$,

$$\begin{aligned} a_k &:= c_k + c_{-k}, \\ b_k &:= i(c_k - c_{-k}). \end{aligned} \tag{7.1}$$

Example 7.7. The trig. polynomial $p(x) = e^{6\pi ix} + e^{-6\pi ix}$ can be written as $p(x) = 2 \cos(6\pi x)$. More generally, every trigonometric polynomial with all $c_k \in \mathbb{R}$ such that $c_k = c_{-k}$, gives a sum of cosines with real coefficients.

Example 7.8. Write $p(x) = \sin(2\pi x)$ in the form given in Definition 7.6.

Example 7.9. Why is $e^{\pi ix}$ (or $\sin(x)$, or x^2) not a trigonometric polynomial?

From the above relations, we can deduce quite some information of a trigonometric polynomial, just by looking at its coefficients. In particular, we can say if a trigonometric polynomial is indeed real-valued, i.e., if $p(x) \in \mathbb{R}$ for every $x \in \mathbb{R}$.

Lemma 7.10. Let p be a trigonometric polynomial in the form given in Definition 7.6. Then, p is real-valued, i.e., $p: \mathbb{R} \rightarrow \mathbb{R}$, if and only if

$$\operatorname{Re}(c_k) = \operatorname{Re}(c_{-k}) \quad \text{and} \quad \operatorname{Im}(c_k) = -\operatorname{Im}(c_{-k}).$$

We leave the proof for the reader.

In particular, a real-valued trigonometric polynomial p is uniquely defined by the values of c_0, c_1, \dots , because the c_k for $k < 0$ follow from the above equations.

Remark 7.11 (Function defined on a circle). Another way of looking at periodic functions is to assume they are defined on a *circle*. For this note that, instead of assuming that the function has “the same behavior” at the endpoints of the interval $[0, 1)$, we may also assume that the endpoints are just the same, i.e., we assume $0 = 1$. We can then talk about properties like continuity when “ x goes around the circle”, and there is no distinguished point like a boundary. Mathematically, there are many ways of modeling this. The most prominent is to consider functions defined on the complex unit circle $S_1 := \{z \in \mathbb{C}: |z| = 1\}$. This also gives a justification of the name *trigonometric polynomial* since, with the parametrization $z = e^{2\pi it}$, it can be written as $z \mapsto \sum_{k=-n}^n c_k z^k$, $z \in S_1$, which looks like an algebraic polynomial.

7.2 Fourier coefficients and Fourier series

We now turn to the approximation of periodic functions by trigonometric polynomials. For this, we need the so-called **Fourier coefficients** of a function. These values are then used to build up an approximation of the function, the Fourier series. Note that in a similar way, we used derivative values to obtain approximations by algebraic polynomials, by using Taylor's theorem.

Definition 7.12. Let $f: [0, 1] \rightarrow \mathbb{C}$. Then, for a given integer $k \in \mathbb{Z}$, we call

$$\hat{f}(k) := \int_0^1 f(x) e^{-2\pi i k x} dx$$

the **k -th Fourier coefficient of f** .

For $n \geq 0$, we call the trigonometric polynomial

$$S_n f(x) := \sum_{k=-n}^n \hat{f}(k) e^{2\pi i k x}$$

the **n -th partial sum of the Fourier series of f** .

Moreover, we call the series

$$Sf(x) := \lim_{n \rightarrow \infty} S_n f(x) \quad \left(= \sum_{k=-\infty}^{\infty} \hat{f}(k) e^{2\pi i k x} \right)$$

the **Fourier series** of f . (We use this notation also if the limit does not exist.)

We say that the **Fourier series equals f** (pointwise) if $f(x) = Sf(x)$ for all $x \in [0, 1]$.

Note that the Fourier coefficients, and therefore the partial sums of the Fourier series, are well-defined if the involved integrals are. So, in particular, for any piecewise continuous function f on $[0, 1]$. (We do not even need that f is continuous as a periodic function.) However, the Fourier series does not necessarily make sense (aka. converge), or even then, **it must not be equal to f** everywhere. Before we see this with an easy example, let us state the derivative and the (indefinite) integral of our basic building blocks.

Lemma 7.13. Let $k \in \mathbb{Z} \setminus \{0\}$. Then,

$$\frac{d}{dx} e^{2\pi i k x} = (2\pi i k) e^{2\pi i k x} \quad \text{and} \quad \int e^{2\pi i k x} dx = \frac{e^{2\pi i k x}}{2\pi i k}$$

for all $x \in \mathbb{R}$. In particular, we obtain

$$\int_0^1 e^{2\pi i k x} dx = 0$$

for $k \neq 0$, and $\int_0^1 e^{2\pi i 0 x} dx = 1$.

If we accept that Theorem 5.11 and Lemma 6.9, i.e., linearity of differentiation and integration, also hold for *complex-valued* functions, then a proof is straightforward, and we omit it.

We first discuss the Fourier series of a trigonometric polynomial. Similarly as algebraic polynomials can be represented exactly by some finite Taylor polynomial, it is probably no surprise that some (large enough) partial sum of the Fourier series is exact for trigonometric polynomials. We give a detailed proof here for demonstration.

Example 7.14. For $N \in \mathbb{N}$ and $c_k \in \mathbb{C}$, consider the trigonometric polynomial p given by

$$p(x) = \sum_{k=-N}^N c_k e^{2\pi i kx}.$$

Then, $\hat{p}(k) = c_k$ for all k , and

$$S_n p = p \quad \text{for all } n \geq N.$$

In particular, $Sp = p$.

Proof. We compute the Fourier coefficients of p :

$$\begin{aligned} \hat{p}(k) &= \int_0^1 \left(\sum_{j=-N}^N c_j e^{2\pi i jx} \right) e^{-2\pi i kx} dx \\ &= \sum_{j=-N}^N c_j \int_0^1 e^{2\pi i jx} e^{-2\pi i kx} dx \\ &= \sum_{j=-N}^N c_j \int_0^1 e^{2\pi i (j-k)x} dx. \end{aligned}$$

(Note that we were able to switch integral and sum, only because the sum is finite.) We have

$$\int_0^1 e^{2\pi i (j-k)x} dx = \begin{cases} 1 & \text{if } j = k, \\ \left[\frac{1}{2\pi i (j-k)} e^{2\pi i (j-k)x} \right]_0^1 = 0 & \text{if } j \neq k, \end{cases}$$

such that

$$\hat{p}(k) = \sum_{j=-N}^N c_j \delta_{jk},$$

with δ_{jk} denoting the **Kronecker- δ** ,

$$\delta_{jk} = \begin{cases} 1 & \text{if } j = k, \\ 0 & \text{if } j \neq k. \end{cases}$$

In other words

$$\hat{p}(k) = \begin{cases} c_k & \text{if } -N \leq k \leq N, \\ 0 & \text{otherwise.} \end{cases}$$

This shows that $S_n p = p$ for all $n \geq N$, and therefore $Sp = \lim S_n p = p$. □

We continue by computing the Fourier series of a function that is not a trigonometric polynomial. This example will show, when we finally finish it, that Fourier series are sometimes very helpful in **computing complicated sums**.

Example 7.15. Consider the periodic function $f(x) = x$ on $[0, 1)$ (i.e., we consider the function $f(x) = \{x\}$ on \mathbb{R}). By using the above, and integration by parts, we obtain $\hat{f}(0) = \frac{1}{2}$ and, for $k \neq 0$,

$$\begin{aligned}\hat{f}(k) &= \int_0^1 f(x)e^{-2\pi ikx} dx = \int_0^1 x e^{-2\pi ikx} dx \\ &= \left[\frac{x e^{-2\pi ikx}}{-2\pi ik} \right]_0^1 - \frac{1}{-2\pi ik} \int_0^1 e^{-2\pi ikx} dx \\ &= \frac{1}{-2\pi ik}.\end{aligned}$$

We therefore obtain the Fourier series

$$Sf(x) = \frac{1}{2} + \sum_{k \neq 0} \frac{e^{2\pi ikx}}{-2\pi ik}$$

or, after some computations

$$Sf(x) = \frac{1}{2} - \sum_{k=1}^{\infty} \frac{1}{\pi k} \sin(2\pi kx).$$

Although the last series looks like the (divergent) harmonic series, we will see later that it is actually convergent, and equals $f(x) = x$, for any $x \in (0, 1)$. However, we see already now, that the series is not equal to f at all $x \in [0, 1]$, as $f(x) = Sf(x)$ is false for $x = 0$. To see this, note that all terms in the above series equal 0 if $x = 0$. Therefore, the series converges to $Sf(0) = \frac{1}{2}$. (Indeed, $S_n f(0) = \frac{1}{2}$ for all n , see also Figure 45.) However, we clearly have $f(0) = 0$. This shows that we need to be careful with the points of convergence of a Fourier series. \square

Remark 7.16 (Non-convergence). The above example shows that continuity in $[0, 1)$ is not enough such that all functions are representable as its Fourier series. The reason is, that $f(x) = \{x\}$ is not continuous when considered as a periodic function. (It has jumps at $x \in \mathbb{Z}$.) And if we plot the first partial sums of the Fourier series, see Figure 45, we see that also the approximation close to these jumps is not very good. One may even prove that $S_n f$ is *unbounded*, i.e., that $\lim_{n \rightarrow \infty} \max_{x \in [0, 1)} |S_n f(x)| = \infty$, which shows that an approximation can be arbitrarily bad for finite n . Moreover, it is possible to show that there is also a periodic and continuous function whose Fourier series diverges in a point. Both statements go far beyond this lecture.

Let us consider another Fourier series before we turn to statements about convergence.

Example 7.17. Consider the periodic function $f: [0, 1) \rightarrow \mathbb{C}$ that is given by $f(x) = (x - \frac{1}{2})^2$. Note that f is continuous (as a periodic function). We obtain the Fourier coefficients

$$\begin{aligned}\hat{f}(k) &= \int_0^1 \left(x - \frac{1}{2} \right)^2 e^{-2\pi ikx} dx = \int_{-1/2}^{1/2} t^2 e^{-2\pi ik(t+\frac{1}{2})} dt \\ &= e^{-\pi ik} \int_{-1/2}^{1/2} t^2 e^{-2\pi ikt} dt = (-1)^k \int_{-1/2}^{1/2} t^2 e^{-2\pi ikt} dt,\end{aligned}$$

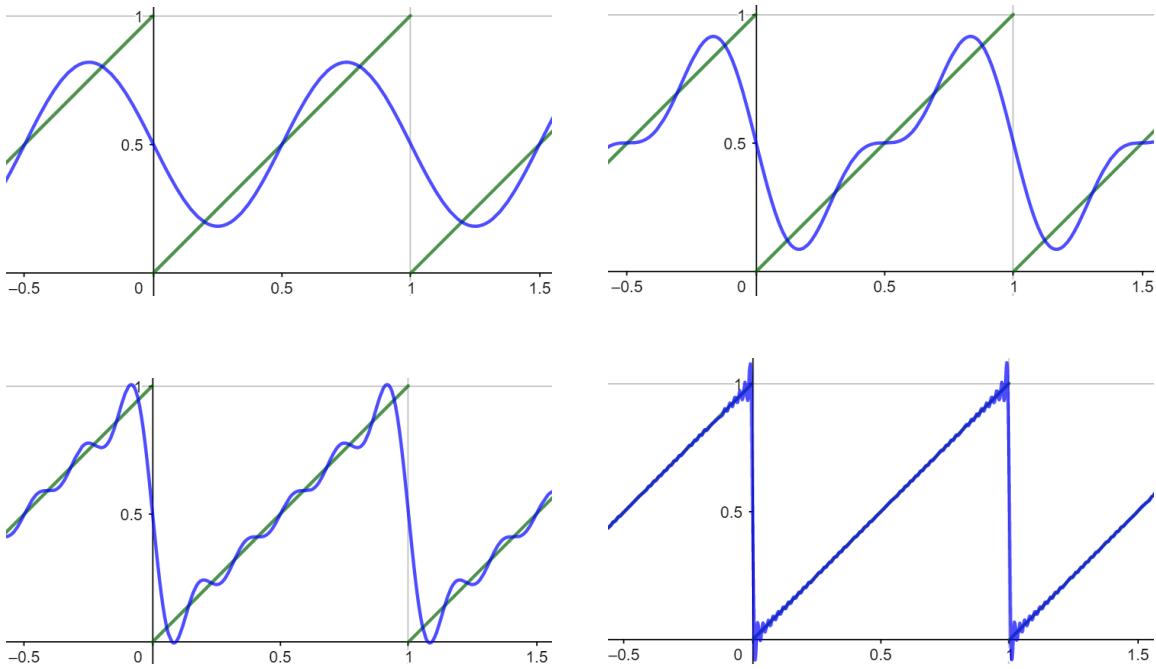


Figure 45: $S_n f$ for $f(x) = \{x\}$ and $n = 1, 2, 5, 50$.

where we used the substitution $t = x - \frac{1}{2}$. For $k = 0$ we obtain that

$$\hat{f}(0) = \int_{-1/2}^{1/2} t^2 dt = \frac{1}{12}.$$

For $k \neq 0$ we get

$$\begin{aligned} \int_{-1/2}^{1/2} t^2 e^{-2\pi i k t} dt &= \left[t^2 \frac{e^{-2\pi i k t}}{-2\pi i k} \right]_{-1/2}^{1/2} - \int_{-1/2}^{1/2} 2t \frac{e^{-2\pi i k t}}{-2\pi i k} dt \\ &= 0 + \frac{1}{\pi i k} \int_{-1/2}^{1/2} t e^{-2\pi i k t} dt \\ &= \frac{1}{\pi i k} \left(\left[t \frac{e^{-2\pi i k t}}{-2\pi i k} \right]_{-1/2}^{1/2} - \int_{-1/2}^{1/2} \frac{e^{-2\pi i k t}}{-2\pi i k} dt \right) \\ &= \frac{1}{\pi i k} \left(\frac{e^{-\pi i k}}{-4\pi i k} - \frac{e^{\pi i k}}{4\pi i k} - \frac{1}{-2\pi i k} \int_{-1/2}^{1/2} e^{-2\pi i k t} dt \right) \\ &= \frac{1}{\pi i k} \left(\frac{(-1)^{k+1}}{2\pi i k} - 0 \right) \\ &= \frac{(-1)^k}{2(\pi k)^2}, \end{aligned}$$

and therefore

$$\hat{f}(k) = (-1)^k \int_{-1/2}^{1/2} t^2 e^{-2\pi i k t} dt = \frac{1}{2(\pi k)^2}.$$

The Fourier series of $f(x) = (\{x\} - \frac{1}{2})^2$ is therefore

$$Sf(x) = \frac{1}{12} + \sum_{k \neq 0} \frac{1}{2(\pi k)^2} e^{2\pi i k x} = \frac{1}{12} + \sum_{k=1}^{\infty} \frac{1}{(\pi k)^2} \cos(2\pi k x).$$

Note that these sums are clearly absolutely convergent (Why?), but, so far, we do not know if Sf equals f at any point.

However, when we have a look at the first partial sums of this Fourier series, then it seems to converge very fast, i.e., already for $n = 20$ we see almost no difference to the original function, see Figure 46. It even looks like the partial sums would converge uniformly to f .

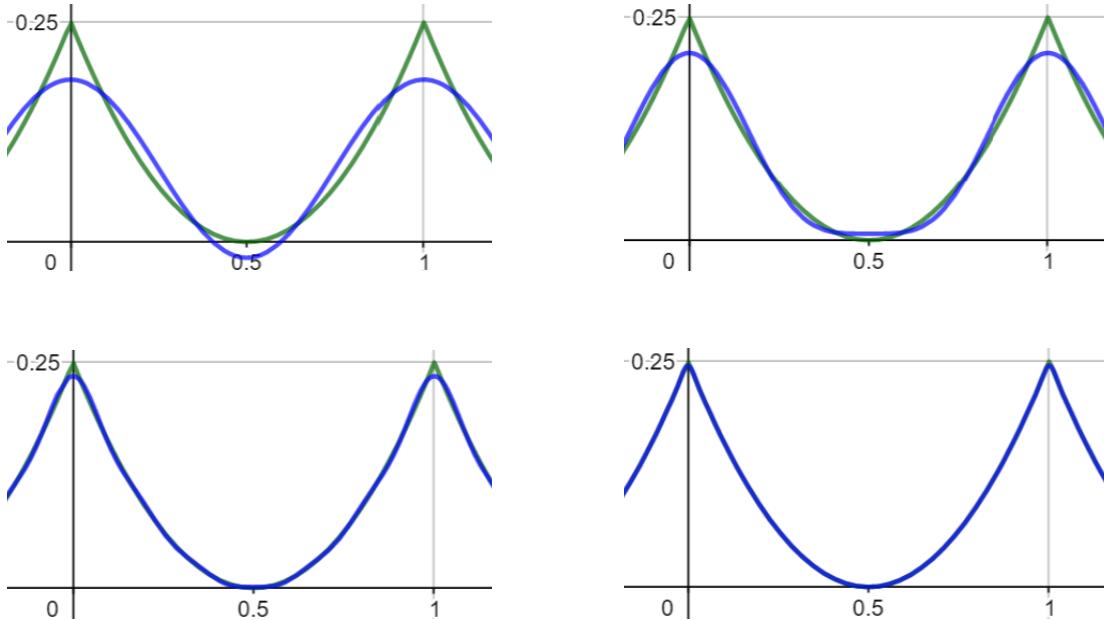


Figure 46: $S_n f$ for $f(x) = (\{x\} - \frac{1}{2})^2$ and $n = 1, 2, 5, 20$.

If we could prove that Sf equals f , i.e., that the partial sums converge pointwise to the function, then, in particular, we would have that $Sf(0) = f(0)$. Noting that $f(0) = \frac{1}{4}$ and that all the cosine terms in the above series equal 1 at $x = 0$, this would imply that

$$\frac{1}{12} + \sum_{k=1}^{\infty} \frac{1}{(\pi k)^2} = Sf(0) \stackrel{?}{=} f(0) = \frac{1}{4},$$

and therefore

$$\sum_{k=1}^{\infty} \frac{1}{k^2} \stackrel{?}{=} \frac{\pi^2}{6}.$$

This formula is correct (and by the way quite beautiful), but **we did not prove it so far**. It still remains to show that the Fourier series converges to f at any point. We will do this by presenting a general rule, based on an assumption on the Fourier coefficient, that a Fourier series converges at all points. This assumption will be fulfilled, e.g., by twice-differentiable and periodic functions.

Remark 7.18. The method above is probably the most powerful technique for computing certain infinite sums. In some cases, there is even no other (manageable) way. One may try, e.g., to compute $\sum_{k=1}^{\infty} \frac{1}{k^2}$ by any other means.

However, we can clearly not evaluate every series by this. One example of this kind is $\sum_{k=1}^{\infty} \frac{1}{k^3}$. (For practice, try to find the value of this series. But do not try too long! There's no explicit form of this number.)

Remark 7.19. Using the computations from page 214 (with $c_k := \hat{f}(k)$), we can write the partial sums of the Fourier series solely by sums of sine and cosine functions, as indicated in the introduction. That is, we can write

$$S_n f(x) = a_0 + \sum_{k=1}^n a_k \cos(2\pi kx) + \sum_{k=1}^n b_k \sin(2\pi kx)$$

with

$$a_k := 2 \int_0^1 f(x) \cos(2\pi kx) dx \quad \text{and} \quad b_k := 2 \int_0^1 f(x) \sin(2\pi kx) dx$$

for $k \in \mathbb{N}$. (Verify this using (7.1).) However, this form has almost no advantages and it is often more work to compute a_k and b_k separately, instead of just the c_k . That's why we usually work with the Fourier coefficients as given above.

Remark 7.20. As stated in the beginning, it was rather arbitrary to choose the period 1. If one chooses the period 2π (which is another prominent choice), then the Fourier coefficients are usually defined by

$$\hat{f}(k) := \frac{1}{2\pi} \int_0^{2\pi} f(x) e^{-ikx} dx.$$

This implies that also the Fourier series of functions may look different. (And they are because of the different domain/period.) Therefore, you should be careful when using other literature.

7.3 First convergence theorems

We now turn to a result about **pointwise convergence** of Fourier series, i.e., we want to know if the partial sums of the Fourier series of a function converge to this function at **all points**. This is clearly a desirable property, and we will show that this holds for functions whose Fourier coefficients are *absolutely summable*, i.e., $\sum_{k \in \mathbb{Z}} |\hat{f}(k)| < \infty$.

Under this assumption, we can prove even more:

We show that the Fourier series **converges uniformly** to the function. That is, the difference between $f(x)$ and $S_n f(x)$ is, for large enough n , “small” for all x simultaneously. This is a very important property when it comes to approximations, where we want to approximate a function on the whole domain, e.g., by the partial Fourier series $S_n f$, and we want that $|f(x) - S_n f(x)| < \varepsilon$, for some given $\varepsilon > 0$, for every $x \in \mathbb{R}$. That is, we want that $\sup_x |f(x) - S_n f(x)| < \varepsilon$.

Before we discuss that with some examples, let us put this into a definition. Since we need this also later, we phrase it a bit more general for arbitrary **sequences of functions** $(f_n)_{n \in \mathbb{N}}$, i.e., every term of this sequence is a function $f_n: D \rightarrow \mathbb{C}$ (Later, we choose $f_n = S_n f$ and $D = \mathbb{R}$) and we may ask if such a sequence converges. But, we need to specify what it means for a sequence of functions to converge.

Definition 7.21. Let $(f_n)_{n=1}^{\infty}$ be a sequence of functions on a set D , and let $f: D \rightarrow \mathbb{C}$.

(i) If

$$\lim_{n \rightarrow \infty} f_n(x) = f(x) \quad \text{for all } x \in D,$$

then we say that (f_n) **converges pointwise** to f . We use the notation $f_n \xrightarrow{\text{pw}} f$ or “ $f_n \rightarrow f$ pointwise”.

(ii) If

$$\lim_{n \rightarrow \infty} \sup_{x \in D} |f_n(x) - f(x)| = 0,$$

then we say that (f_n) **converges uniformly** to f . We write “ $f_n \rightarrow f$ uniformly”.

Uniform convergence implies pointwise convergence. In particular, if a sequence converges uniformly to a function f , then this function is the pointwise limit, i.e., $f(x) = \lim_{n \rightarrow \infty} f_n(x)$. (Just note that we have $|f_n(x) - f(x)| \leq \sup_{y \in D} |f_n(y) - f(y)|$ for every $n \in \mathbb{N}$ and $x \in D$.) However, it is not obvious that these are really different concepts. For this, let us consider the following simple example.

Example 7.22. Consider $f_n(x) = x^n$ on $D = [0, 1]$. Then, since we know that $x^n \rightarrow 0$ for every $x \in [0, 1)$, we have that $f_n \xrightarrow{\text{pw}} 0$. However, we have for every fixed $n \in \mathbb{N}$ that

$$\sup_{x \in [0, 1)} |x^n - 0| = \sup_{x \in [0, 1)} x^n = 1.$$

Since this does not converge to 0, we obtain that f_n is not uniformly convergent (to 0).

Roughly speaking, the sequence x^n converges arbitrarily slow to 0 (depending on x), and therefore not uniformly.

Example 7.23. Another example of a sequence of functions, and its limit, that we discussed already is the difference quotient: For this, let $f: (a, b) \rightarrow \mathbb{R}$ be a continuous function, and define the sequence of functions

$$f_n(x) = \frac{f(x + \frac{1}{n}) - f(x)}{\frac{1}{n}}.$$

We know that $f_n \xrightarrow{\text{pw}} f'$, i.e., f_n converges pointwise to the derivative of f , if f is differentiable. One can also show that f_n converges uniformly to f' if f is continuously differentiable. (We do not need that here.)

We do not want to go too much into detail here. In most cases, we will just talk about pointwise convergence, which should be easy to comprehend, since we actually already work with this type of convergence for some time. However, many results hold directly for the *much stronger* uniform convergence, and that's why we also state it. Moreover, it is sometimes a helpful tool in proving our results. Let us briefly comment on the *power* of this type of convergence: In general, we do not know in advance properties of the pointwise limit of a sequence of functions. However, if $f_n \rightarrow f$ uniformly, then **some properties are preserved**. For example, if every f_n is continuous, then we obtain that also f is continuous. This is a very powerful insight!

(To see that this is false without uniform convergence, consider e.g., $f_n(x) = x^n$ on $[0, 1]$ which converges pointwise to a discontinuous function. Which one?)

In the sequel we will need only two of the properties that are preserved under uniform convergence. We state them in the following lemma.

Lemma 7.24. *Let $(f_n)_{n=1}^{\infty}$ be a sequence of continuous functions on a set D , and let $f: D \rightarrow \mathbb{C}$ be such that $f_n \rightarrow f$ uniformly. Then,*

- (i) *f is continuous, and*
- (ii) *for all $a, b \in D$ with $[a, b] \subset D$, we have*

$$\int_a^b f dx = \lim_{n \rightarrow \infty} \int_a^b f_n dx.$$

Note that the second part of this lemma actually states that we can “interchange” the limit (in $f(x) = \lim_{n \rightarrow \infty} f_n(x)$) and the integral. Since the terms of the sequence are usually much easier functions than the limit, this gives a nice way of computing integrals. Moreover, the first enables us to show that the limit of continuous functions is continuous, even if we do not know the limit precisely. This is particularly useful when working with Fourier series, because in this case ($f_n = S_n f$) the functions f_n are obviously continuous. Hence, uniform convergence of the partial sums of a Fourier series directly implies continuity of the limit.

Proof. To prove part (i), we need to show that $\lim_{m \rightarrow \infty} |f(x_m) - f(x_0)| = 0$ for every $x_0 \in D$ and every sequence $(x_m)_{m \in \mathbb{N}} \subset D$ with $x_m \rightarrow x_0$. Fix some $\varepsilon > 0$. By the uniform convergence of $f_n \rightarrow f$, we obtain that there is some $n_0 \in \mathbb{N}$ such that $|f_n(y) - f(y)| < \frac{\varepsilon}{2}$ for all $n \geq n_0$ and all $y \in D$. Therefore,

$$\begin{aligned} |f(x_m) - f(x_0)| &= |f(x_m) - f_n(x_m) + f_n(x_m) - f_n(x_0) + f_n(x_0) - f(x_0)| \\ &\leq |f(x_m) - f_n(x_m)| + |f_n(x_m) - f_n(x_0)| + |f_n(x_0) - f(x_0)| \\ &< \frac{\varepsilon}{2} + |f_n(x_m) - f_n(x_0)| + \frac{\varepsilon}{2} = |f_n(x_m) - f_n(x_0)| + \varepsilon \end{aligned}$$

for all $n \geq n_0$. Now, since all f_n are continuous, we have that the limit $m \rightarrow \infty$ of the right hand side converges to ε . That is, we have

$$\lim_{m \rightarrow \infty} |f(x_m) - f(x_0)| < \varepsilon.$$

Since this holds for all $\varepsilon > 0$, this proves part (i).

For the second part, we use the triangle inequality to obtain

$$\left| \int_a^b f dx - \int_a^b f_n dx \right| \leq \int_a^b |f(x) - f_n(x)| dx.$$

By uniform convergence, $\sup_x |f(x) - f_n(x)| < \frac{\varepsilon}{b-a}$ for all $n \geq n_0$ and n_0 large enough. This implies

$$\left| \int_a^b f dx - \int_a^b f_n dx \right| < \frac{\varepsilon}{b-a} \int_a^b 1 dx = \varepsilon$$

for $n \geq n_0$. This implies the result. □

We now turn to the convergence of Fourier series.

Theorem 7.25. *Let $f: [0, 1] \rightarrow \mathbb{C}$ be continuous. If $\sum_{k=-\infty}^{\infty} |\hat{f}(k)| < \infty$ holds, then*

$$Sf(x) = \lim_{n \rightarrow \infty} S_n f(x) = f(x)$$

for every $x \in [0, 1]$. Moreover, $S_n f \rightarrow f$ uniformly.

The proof of this theorem makes use of the following lemma, which says that two distinct functions can be distinguished by their Fourier coefficients. This is a natural statement, and one of the fundamental bases of Fourier analysis. However, a formal proof would be rather complicated, and we omit it here.

Lemma 7.26. *Let $f, g: [0, 1] \rightarrow \mathbb{C}$ be continuous functions such that $\hat{f}(k) = \hat{g}(k)$ for all $k \in \mathbb{Z}$. Then,*

$$f(x) = g(x) \quad \text{for all } x \in [0, 1].$$

In particular, $\hat{f}(k) = 0$ for all $k \in \mathbb{Z}$ if and only if $f = 0$.

Remark 7.27. The statement of Lemma 7.26 is equivalent to the statement that for every function $f \neq 0$, there exists some $k \in \mathbb{Z}$ such that $\hat{f}(k) \neq 0$.

Proof of Theorem 7.25. To simplify notation, we set $a_k := \hat{f}(k)$ for $k \in \mathbb{Z}$. According to the assumptions, the sequence of partial sums $(\sum_{k=-n}^n |a_k|)_{n \geq 0}$ converges, because it is a monotone and bounded sequence. In particular, it is a Cauchy sequence. This means that for every $\varepsilon > 0$ there exists $n_0 \in \mathbb{N}$, such that for all $n \geq n_0$ we have

$$\sum_{k: |k| > n} |a_k| \leq \varepsilon.$$

Thus for all $n \geq n_0$, we obtain

$$\sup_{x \in [0, 1]} |Sf(x) - S_n f(x)| = \sup_{x \in [0, 1]} \left| \sum_{|k| > n} a_k e^{2\pi i k x} \right| \leq \sum_{|k| > n} |a_k| < \varepsilon,$$

where we use $|e^{2\pi i k x}| = 1$. Since this holds for all $\varepsilon > 0$, we see that $S_n f$ converges uniformly to Sf . It remains to show that $Sf(x) = f(x)$ for all x . Since we want to use Lemma 7.26 for this, we need to show that both f and Sf are continuous, and that their Fourier coefficients coincide. First, f is continuous by assumption, and by Lemma 7.24, Sf is continuous as the uniform limit of continuous functions. (Trigonometric polynomials are always continuous.)

Let us consider the Fourier coefficients. It is easy to see that, if a sequence of functions (g_n) converges uniformly to g , then, for fixed $\ell \in \mathbb{Z}$, $g_n e^{2\pi i \ell \cdot}$ converges also uniformly to $g e^{2\pi i \ell \cdot}$. (Verify this!) By Lemma 7.24(ii), this implies that

$$\lim_{n \rightarrow \infty} \int_0^1 g_n(x) e^{-2\pi i \ell x} dx = \int_0^1 g(x) e^{-2\pi i \ell x} dx.$$

In other words, $\hat{g}_n(\ell) \rightarrow \hat{g}(\ell)$ for all $\ell \in \mathbb{Z}$. We now apply this to $g_n = S_n f$ and $g = Sf$. Moreover, recall from Example 7.14 that

$$\widehat{S_n f}(\ell) = \int_0^1 S_n f(x) e^{-2\pi i \ell x} dx = \begin{cases} a_\ell, & \text{if } |\ell| \leq n, \\ 0, & \text{otherwise.} \end{cases}$$

(This just means that the “first n ” Fourier coefficients of the partial sum $S_n f$ coincide with the corresponding Fourier coefficients of f .) Since $\widehat{S_n f}(\ell) = a_\ell$ for all $n \geq |\ell|$, we clearly obtain $\lim_{n \rightarrow \infty} \widehat{S_n f}(\ell) = a_\ell$ for all $\ell \in \mathbb{Z}$. Together with the above, and the uniform convergence of $S_n f$ to Sf , we have

$$\widehat{Sf}(k) = \lim_{n \rightarrow \infty} \widehat{S_n f}(k) = a_k = \hat{f}(k)$$

for all $k \in \mathbb{Z}$. This finally shows that all Fourier coefficients of Sf and f coincide. Together with their continuity, we obtain from Lemma 7.26 that $Sf(x) = f(x)$ for all $x \in [0, 1]$. \square

With Theorem 7.25 we can finally prove that the example from the end of Section 7.2 was correct.

Corollary 7.28. *We have that*

$$\sum_{k=1}^{\infty} \frac{1}{k^2} = \frac{\pi^2}{6}.$$

Proof. As discussed above, the Fourier coefficients of the function $f(x) = (\{x\} - \frac{1}{2})^2$ satisfy

$$\hat{f}(k) = (-1)^k \int_{-1/2}^{1/2} t^2 e^{-2\pi i k t} dt = \frac{1}{2(\pi k)^2}.$$

We know from Lemma 3.91 that $\sum_{k=-\infty}^{\infty} |\hat{f}(k)| < \infty$. Therefore, we know from Theorem 7.25 that $Sf(x) = f(x)$ for every $x \in [0, 1]$. In particular, $Sf(0) = f(0)$. This implies that

$$\frac{1}{4} = f(0) = Sf(0) = \frac{1}{12} + \sum_{k \neq 0} \frac{1}{2(\pi k)^2} = \frac{1}{12} + \sum_{k=1}^{\infty} \frac{1}{(\pi k)^2}.$$

Rearranging yields the result. \square

Note that we can deduce a bit more from the proof of Theorem 7.25. In particular, we showed that, given an absolutely summable sequence of complex numbers, then the trigonometric polynomials with these coefficients converge to a continuous function. This is a helpful statement when a function is given by its Fourier coefficients. (Note that the theorem shows that, under the given assumptions, a function might be uniquely defined by its Fourier coefficients.) We state this in the following lemma.

Lemma 7.29. Let $(a_k)_{k \in \mathbb{Z}}$ be a complex-valued sequence such that

$$\sum_{k=-\infty}^{\infty} |a_k| < \infty.$$

Then, the sequence of trigonometric polynomials $(g_n)_{n \geq 0}$ with

$$g_n(x) = \sum_{k=-n}^n a_k e^{2\pi i k x}$$

converges uniformly to a continuous periodic function $g: [0, 1] \rightarrow \mathbb{C}$ with

$$\hat{g}(k) = a_k \quad \text{for all } k \in \mathbb{Z}.$$

Example 7.30. Given the sequence $(a_k)_{k \in \mathbb{Z}}$ with $a_k = 0$ for $k < 0$, and $a_k = e^{-k}$ for $k \geq 0$. From Lemma 7.29 we obtain that

$$g(x) = \sum_{k=0}^{\infty} e^{-k} e^{2\pi i k x}$$

describes a continuous periodic function. To see this, we do not even need to make any computations regarding the infinite series. It is enough that a_k is non-negative and summable. (One could use geometric series to obtain the explicit form $g(x) = \frac{1}{1 - e^{2\pi i x - 1}}$.)

Although Theorem 7.25 is useful and gives a simple criterion for a Fourier series to converge, it is still not satisfactory as it gives a property of the Fourier coefficients as a sufficient condition. In some cases, one would prefer to check only a condition of the function itself, like differentiability. The next result, which follows almost immediately from Theorem 7.25, shows that the Fourier series of twice differentiable functions converges uniformly. However, note that this would not be helpful for proving the result in Corollary 7.28, because this function is not differentiable at 0, see Figure 46. In this respect, Theorem 7.25 is more general.

Theorem 7.31. Let $f: [0, 1] \rightarrow \mathbb{C}$ be periodic and twice continuously differentiable. Then,

$$Sf(x) = \lim_{n \rightarrow \infty} S_n f(x) = f(x)$$

for every $x \in [0, 1]$. Moreover, $S_n f \rightarrow f$ uniformly.

For the proof of this result it is enough to show that the Fourier coefficients of a twice differentiable function are absolutely summable. As this is of independent interest, we state it in the following lemma in a more general form.

Lemma 7.32. Let $s \in \mathbb{N}$ and $f: [0, 1] \rightarrow \mathbb{C}$ be periodic and s -times continuously differentiable. Then, we have

$$|\hat{f}(k)| \leq \frac{M}{|2\pi k|^s} \quad \text{for all } k \neq 0,$$

where

$$M := \max_{x \in [0, 1]} |f^{(s)}(x)|.$$

Proof. Let $k \neq 0$. Using integration by parts, we obtain

$$\begin{aligned}\hat{f}(k) &= \int_0^1 f(x)e^{-2\pi ikx} dx \\ &= \left[f(x) \frac{e^{-2\pi ikx}}{-2\pi ik} \right]_0^1 - \int_0^1 f'(x) \frac{e^{-2\pi ikx}}{-2\pi ik} dx \\ &= \frac{1}{2\pi ik} \int_0^1 f'(x) e^{-2\pi ikx} dx,\end{aligned}$$

where we've used the periodicity of f (and $e^{-2\pi ikt}$) in the last equation. If we repeat the integration by parts additional $(s-1)$ -times, we get

$$\hat{f}(k) = \frac{1}{(2\pi ik)^s} \int_0^1 f^{(s)}(x) e^{-2\pi ikx} dx.$$

The triangle inequality finally implies

$$|\hat{f}(k)| \leq |2\pi k|^{-s} \int_0^1 |f^{(s)}(x)| |e^{-2\pi ikx}| dx \leq \frac{M}{|2\pi k|^s}.$$

□

With this it is easy to prove Theorem 7.31.

Proof of Theorem 7.31. According to the assumptions, f'' is continuous, such that there exists $M < \infty$ with

$$|f''(x)| \leq M, \quad \forall x \in [0, 1].$$

From Lemma 7.32 we have $\hat{f}(k) \leq |2\pi k|^{-2} M \leq |k|^{-2} M$ for all $k \neq 0$, and consequently

$$\sum_{k=-\infty}^{\infty} |\hat{f}(k)| \leq \hat{f}(0) + 2M \sum_{k=1}^{\infty} \frac{1}{k^2} < \infty.$$

Therefore, we obtain from Theorem 7.25 that we have $S_n f \rightarrow f$ uniformly.

□

Remark 7.33. Let us state again that the assumption of the last theorem is that the function f under consideration has to be continuously differentiable *as a periodic function*. This means, that also its derivative f' (or derivatives of higher order) have to be periodic functions and, in particular, have to satisfy $f'(0) = f'(1)$. (It is a typical mistake to forget this property.) For example, the periodic function $f: [0, 1] \rightarrow \mathbb{R}$ with $f(x) = (x - \frac{1}{2})^2$ is continuous, since $\lim_{x \searrow 0} f(x) = \lim_{x \nearrow 1} f(x) = \frac{1}{4}$. However, its derivative satisfies $f'(x) = 2x - 1$ for $x \in (0, 1)$, which implies $\lim_{x \searrow 0} f'(x) = -1 \neq 1 = \lim_{x \nearrow 1} f'(x)$. Therefore, f is not continuously differentiable as a periodic function.

Theorem 7.25 and Theorem 7.31 only state the *qualitative* statements that the Fourier series converge. However, for applications of the theory for actual computations, it is necessary to have also *quantitative* results. That is, we want to know how large we have to choose n such that the *error* is small when we approximate f by $S_n f$. Fortunately, such error bounds can be obtained if we are a bit more careful in bounding the corresponding infinite sums.

For this we state the following lemma, which is a useful tool for bounding (infinite) sums by certain integrals. Note that this can also be used as a *convergence test* for series, similarly to the ones given in Section 3.7, to verify if a sequence of numbers is summable. But this time, this may also lead to reasonable bounds on the sum of the series.

Lemma 7.34. *Let $h: [0, \infty) \rightarrow [0, \infty)$ be a continuous and non-increasing function. Then, for all $m, n \in \mathbb{N}_0$ with $m < n$, we have*

$$\sum_{k=m+1}^n h(k) \leq \int_m^n h(x) dx \leq \sum_{k=m}^{n-1} h(k).$$

In particular, if H is an antiderivative of h , then,

$$\sum_{k=m+1}^{\infty} h(k) \leq H(\infty) - H(m),$$

where $H(\infty) := \lim_{x \rightarrow \infty} H(x)$.

Proof. First, we split the integral into several integrals over intervals of length 1 to obtain

$$\int_m^n h(x) dx = \sum_{k=m}^{n-1} \int_k^{k+1} h(x) dx.$$

For each k , the mean value theorem (Theorem 6.37) yields that there is some $\xi_k \in [k, k+1]$ such that $\int_k^{k+1} h(x) dx = h(\xi_k)$. Since h is non-increasing, we have $h(k+1) \leq h(\xi_k) \leq h(k)$, and therefore

$$\sum_{k=m}^{n-1} h(k+1) \leq \int_m^n h(x) dx \leq \sum_{k=m}^{n-1} h(k).$$

An index shift implies the first statement. The second follows by taking the limit $n \rightarrow \infty$. \square

Example 7.35. The most prominent application of the last lemma is to bound the *power series* $\sum_{k=n+1}^{\infty} k^{-s}$ for $s > 1$. We obtain for every $n \in \mathbb{N}$ that

$$\sum_{k=n+1}^{\infty} k^{-s} \leq \frac{1}{s-1} n^{-s+1}.$$

(Verify this!)

With these bounds at hand, we are able to give explicit quantitative bounds on the error of Fourier series. Recall that we have presented a similar bound already for Taylor polynomials in Corollary 5.57. However, with this bound we were not able to give *good bounds* for functions that are not very smooth. In contrast, the next bound shows that we can give arbitrarily good approximations of a function, as long as it is twice differentiable. (Of course, we need to compute enough Fourier coefficients for this. We do not discuss here, how this could be done.)

Corollary 7.36. Let $f: [0, 1] \rightarrow \mathbb{C}$ be continuous such that for some $s > 1$ and $B < \infty$ we have

$$|\hat{f}(k)| \leq \frac{B}{|k|^s}$$

for all $k \neq 0$. Then,

$$|f(x) - S_n f(x)| \leq \frac{2B}{s-1} \frac{1}{n^{s-1}}$$

for all $x \in [0, 1]$ and $n \in \mathbb{N}$.

Proof. Combining the techniques we have learned so far, we obtain

$$\begin{aligned} |f(x) - S_n f(x)| &= \left| \sum_{|k|>n} \hat{f}(k) e^{2\pi i k x} \right| \leq \sum_{|k|>n} |\hat{f}(k)| \\ &\leq \sum_{|k|>n} \frac{B}{|k|^s} = 2B \sum_{k=n+1}^{\infty} \frac{1}{k^s} \\ &\leq \frac{2B}{s-1} \frac{1}{n^{s-1}}. \end{aligned}$$

□

Let us finally again discuss the example from the beginning, see Example 7.17, and see how good an approximation by a partial sum of the Fourier series would be.

Example 7.37. We consider the periodic function $f: [0, 1] \rightarrow \mathbb{C}$ that is given by $f(x) = (x - \frac{1}{2})^2$. We already showed in Example 7.17 that its Fourier coefficients equal

$$\hat{f}(k) = \frac{1}{2(\pi k)^2}.$$

Therefore, they satisfy the bound of Corollary 7.36 with $s = 2$ and $B = \frac{1}{2\pi^2}$. We therefore obtain the bound

$$|f(x) - S_n f(x)| \leq \frac{1}{\pi^2 n}$$

for all $n \in \mathbb{N}$. One might check, that $n = 11$ suffices to obtain $|f(x) - S_{11} f(x)| < \frac{1}{100}$ (or $n = 102$ for $|f(x) - S_{102} f(x)| < \frac{1}{1000}$). This finally justifies that an approximation of this function can already be good for rather small n , see Figure 46.

Example 7.38. Bounds like those in the last example can also be useful when we want to approximate certain series by finite sums. Consider again the same example, but only at $x = 0$. At this point the Fourier series reads $S f(0) = \frac{1}{12} + \sum_{k=1}^{\infty} \frac{1}{(\pi k)^2}$, and we have $f(0) = \frac{1}{4}$, see Example 7.17. Therefore,

$$|f(0) - S_n f(0)| = \left| \frac{1}{6} - \sum_{k=1}^n \frac{1}{(\pi k)^2} \right| \leq \frac{1}{\pi^2 n},$$

which implies

$$\left| \frac{\pi^2}{6} - \sum_{k=1}^n \frac{1}{k^2} \right| \leq \frac{1}{n}.$$

We can therefore give a rather good approximation of $\frac{\pi^2}{6}$ with an error of at most $\frac{1}{n}$ by just computing the sum of the first n terms of the series. Note that this can actually be done by hand, and it was done like this before calculators were invented. At these times, such error bounds were essential.

7.4 The theorem of Dirichlet

We now shortly comment on some sort of final result regarding the convergence of Fourier series. This is **Dirichlet's theorem**, which states that it is actually enough to consider only *local properties* of a function to obtain pointwise convergence of the Fourier series. Note that, in contrast to this, all the previous results required some *global* knowledge about the function: either through its Fourier coefficients, or because we assumed the function to be differentiable everywhere. The following theorem, which is only a special case of Dirichlet's theorem, states that the latter assumption would be enough at a point.

Theorem 7.39. *Let $f: [0, 1] \rightarrow \mathbb{C}$ be piecewise continuous. Then, if f is differentiable at $x_0 \in (0, 1)$, we have*

$$Sf(x_0) = \lim_{n \rightarrow \infty} S_n f(x_0) = f(x_0).$$

Moreover, if f is differentiable (as a periodic function), then $S_n f \rightarrow f$ uniformly.

We do not prove this statement here. Actually, a proof of this theorem requires quite a lot of prerequisites, and it would fill several lectures to tackle every single detail.

Let us see an example that shows how useful this theorem is.

Example 7.40. Consider again the periodic function $f(x) = x$ on $[0, 1]$, see Example 7.15 and Figure 45. We have shown already that $\hat{f}(0) = \frac{1}{2}$ and $\hat{f}(k) = \frac{1}{-2\pi ik}$ for $k \neq 0$, which yields the Fourier series

$$Sf(x) = \frac{1}{2} + \sum_{k \neq 0} \frac{e^{2\pi i k x}}{-2\pi i k} = \frac{1}{2} - \sum_{k=1}^{\infty} \frac{1}{\pi k} \sin(2\pi k x).$$

This series is not convergent at $x = 0$, because $Sf(0) = \frac{1}{2} \neq 0 = f(0)$. However, since f is differentiable at all $x \in (0, 1)$, we know from Dirichlet's theorem that $Sf(x) = f(x)$ for all $x \in (0, 1)$. This can be used, e.g., to easily compute the series

$$\sum_{k=1}^{\infty} \frac{\sin(kt)}{k} = \frac{\pi - t}{2}$$

for every $t \in (0, 2\pi)$. (Verify this!) □

Remark 7.41 (Advanced Fourier series). In the last sections we discussed partial sums of Fourier series, and how they converge to the original function. Although we presented several results that show this convergence, we have also indicated that this is not always true. Let us finally add that, if we would combine the Fourier coefficients in a more clever way, then the corresponding series converge pointwise for every periodic and continuous function. To be precise, if consider the average of the first n partial sums

$$\sigma_n f(x) = \frac{1}{n} \sum_{m=0}^{n-1} S_m f(x),$$

which are called *Cesaro means of the partial sums*, then we might prove (with a lot of effort) that $\sigma_n f \rightarrow f$ uniformly for every continuous function. This is called **Fejer's theorem**. This result, and its more advanced variants, are heavily used in everyday applications (like JPEG or MP3), in particular, due to their nice convergence guarantee. This shows that one might be careful how to build up an approximation, based on given information.

8 Multivariate Calculus

In this chapter we initiate the study of functions that depend on more than one variable, which are called **multivariate** functions. In analogy to the case of real-valued functions of one variable, we will introduce some concepts, like continuity or differentiability, which will then lead to results on, e.g., extreme values of multivariate functions. Note that the study and computation of minima and maxima of functions that depend on many parameters is one of the main subjects of **optimization**, and therefore particularly important in AI applications.

The general type of multivariate functions are **vector-valued functions**, which have the form

$$V: \mathbb{R}^d \rightarrow \mathbb{R}^m$$

for some $d, m \in \mathbb{N}$. That is, V maps a vector of length d to a vector of length m . We already studied the case $m = d = 1$ in detail, and such functions will be called **univariate**. A special case of vector-valued function are **vector fields**, which have $d = m$, and have the nice interpretation of attaching a vector to each point in space.

The easiest vector-valued functions are given by matrix-vector multiplication with a matrix $A \in \mathbb{R}^{m \times d}$. That is, we define $V: \mathbb{R}^d \rightarrow \mathbb{R}^m$ by $V(x) = Ax \in \mathbb{R}^m$ for $x \in \mathbb{R}^d$. Such functions are called **linear functions**. (Recall that univariate linear functions are just multiplication with a scalar.) However, we also need to discuss non-linear functions, for which the components of the output may result from arbitrary operations with the input, and not only linear combinations.

For this, we start by considering the special case $m = 1$, i.e., we consider **multivariate real-valued functions**

$$f: \mathbb{R}^d \rightarrow \mathbb{R}$$

with $d \in \mathbb{N}$. (We use “ d ” for *dimension*.) Although this might seem to be a very special case, we will see later that general vector-valued functions can be handled rather easily (component-wise) once we have the right tools for real-valued functions.

Let us start with an example, and some comments on the visualization of functions.

Example 8.1. Define $f: \mathbb{R}^2 \rightarrow \mathbb{R}$ by

$$f(x_1, x_2) = x_1^2 + x_2^2 - x_1 \cdot x_2.$$

For example, we have $f(0, 0) = 0$, $f(1, 1) = 1$ and $f(1, \frac{1}{2}) = \frac{3}{4}$.

In contrast to univariate functions, visualization of multivariate might not always be easy or meaningful. However, one might at least want to try in the case $d = 2$, where there are two popular ways of plotting a function, see Figure 47. For the first, we just plot the *graph* $\{(x_1, x_2, x_3) \in \mathbb{R}^3 : x_3 = f(x_1, x_2)\}$ of the function f in three-dimensional space. On the other hand, we can also have a look at the level sets $N_c = \{(x_1, x_2) \in \mathbb{R}^2 : f(x_1, x_2) = c\}$ for $c \in \mathbb{R}$, i.e., N_c contains all points of the function f which are at the ‘height’ c . (For continuous functions, these points form lines.) The level sets can now be used to get a two dimensional plot of f , which is called a *contour plot*.

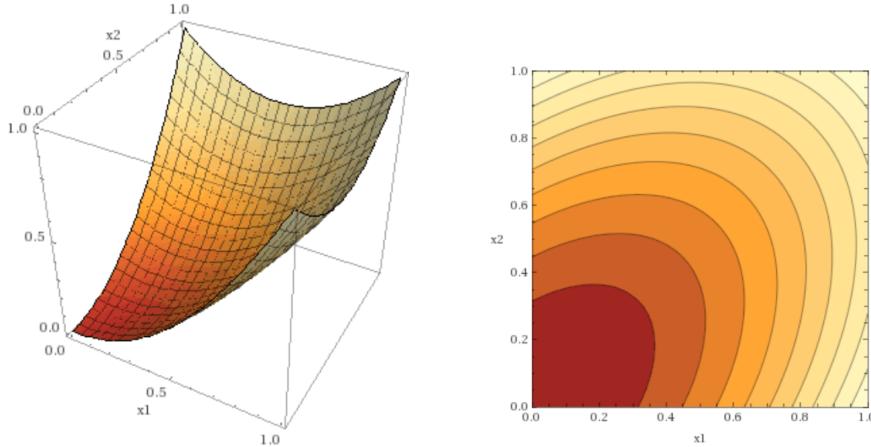


Figure 47: A plot of the graph and the contour plot of $f(x_1, x_2) = x_1^2 + x_2^2 - x_1x_2$ in $[0, 1]^2$.

Remark 8.2. Note that, as usual, vectors $x \in \mathbb{R}^d$ are considered as column vectors when it comes to matrix-vector multiplication. However, when considered as input of some function we use the notation $x = (x_1, \dots, x_d)$ and $f(x) = f(x_1, \dots, x_d)$. This should not lead to confusion.

8.1 Sequences in \mathbb{R}^d

Since we want to investigate continuity and other properties of multivariate functions, we have to mimic the concepts that we introduced in the univariate setting. In particular, we need the concept of a **limit** of a (convergent) sequence. Recall that in the univariate case, we used often that, for a sequence $(a_n)_{n \in \mathbb{N}} \subset \mathbb{R}$, the convergence $\lim a_n = a$ is equivalent to $\lim |a_n - a| = 0$, and we want to do the same here. Therefore, we actually only need to replace the *absolute value* by another quantity that allows for measuring 'how large' a vector is. This can then be used to measure 'how close' two vectors are.

Recall that we already introduced in Section 1.9 the **Euclidean norm** and the corresponding **inner product**

$$\|x\|_2 = \sqrt{\langle x, x \rangle} = \sqrt{\sum_{i=1}^d |x_i|^2}, \quad \text{where} \quad \langle x, y \rangle = \sum_{i=1}^d x_i y_i$$

for $x = (x_1, \dots, x_d) \in \mathbb{R}^d$ and $y = (y_1, \dots, y_d) \in \mathbb{R}^d$.

In analogy with the univariate case, we have the **triangle inequality**

$$\|x + y\|_2 \leq \|x\|_2 + \|y\|_2.$$

for all $x, y \in \mathbb{R}^d$, see Theorem 1.75. Moreover, the **Cauchy-Schwarz inequality** (Lemma 1.76) states that

$$|\langle x, y \rangle| \leq \|x\|_2 \|y\|_2$$

for all $x, y \in \mathbb{R}^d$, and we have the equality $|\langle x, y \rangle| = \|x\|_2 \|y\|_2$ if and only if $y = c \cdot x$ for some $c \in \mathbb{R}$.

We are now ready to treat **limits of sequences of vectors**. But note that we have to be careful with the indices: In the following, we assume that $(x_k)_{k \in \mathbb{N}} \subset \mathbb{R}^d$ is a sequence of vectors, i.e., $x_k \in \mathbb{R}^d$ for every $k \in \mathbb{N}$. Now, these vectors have d entries, which will be denoted by

$$x_k = \begin{pmatrix} x_{k,1} \\ x_{k,2} \\ \vdots \\ x_{k,d} \end{pmatrix}.$$

In analogy to the convergence of univariate sequences, we define the following.

Definition 8.3 (Convergence). Let $(x_k)_{k \in \mathbb{N}} \subset \mathbb{R}^d$ be a sequence of vectors. If there exists some $y = (y_1, \dots, y_d) \in \mathbb{R}^d$ such that

$$x_{k,i} \rightarrow y_i \quad \text{for all } i = 1, \dots, d.$$

then we say (x_k) converges to y , and write $x_k \rightarrow y$ (as $k \rightarrow \infty$), or $\lim_{k \rightarrow \infty} x_k = y$.

If there is no such vector y the sequence is not convergent, or divergent.

That is, a sequence (x_k) converges if and only if every component $(x_{k,i})$, $i = 1, \dots, d$, of the sequence converges.

Let us see an example.

Example 8.4. We consider the sequence given by

$$x_k = \left(\frac{k-2}{k^2+1}, \cos\left(\frac{1}{k} + \pi\right), \frac{e^k + k^2}{e^k} \right).$$

Using our knowledge about real sequences, and the above lemma, we see that $x_k \rightarrow x$ with

$$x = (0, -1, 1),$$

which are the limits of $\frac{k-2}{k^2+1}$, $\cos\left(\frac{1}{k} + \pi\right)$ and $\frac{e^k + k^2}{e^k}$.

With this in mind, we can compute limits of vectors just by computing d usual limits. However, it is sometimes handy to have a characterization of convergence that is based on the Euclidean norm. For illustration, let us see the following easy example.

Example 8.5. We have a look at the sequence which is given by

$$x_k = \frac{1}{k} \begin{pmatrix} 1 \\ 1 \end{pmatrix}.$$

This sequence clearly converges to $0 \in \mathbb{R}^2$. If we now look at the norms of the differences $x_k - 0$ we see that

$$\|x_k - 0\|_2 = \sqrt{\left(\frac{1}{k} - 0\right)^2 + \left(\frac{1}{k} - 0\right)^2} = \sqrt{\frac{1}{k^2} + \frac{1}{k^2}} = \frac{1}{k}\sqrt{2}.$$

Thus $\|x_k - 0\| \rightarrow 0$ as $k \rightarrow \infty$.

Having a closer look to the above example, or the definition of the norm in general, we see that a sequence (x_k) converges to a vector y if and only if the norms $\|x_k - y\|$ converge to 0 (a number). We state this in the following lemma. The proof is an easy exercise.

Lemma 8.6. *Let $(x_k)_{k \in \mathbb{N}} \subset \mathbb{R}^d$ be a sequence of vectors and $y \in \mathbb{R}^d$. Then,*

$$x_k \rightarrow y \iff x_k - y \rightarrow 0 \iff \|x_k - y\|_2 \rightarrow 0.$$

By this equivalence, we can see the similarity to the univariate situation: We had that a sequence of numbers (a_n) converges to a if and only if the absolute values $|a_n - a|$ converge to zero. This similar appearance will be helpful, as many proofs of results on continuity and differentiation that follow will just look very similar to the proofs from the univariate setting.

Let us also comment on **other norms** that one can consider for vectors. We will mostly use the Euclidean norm for the considerations below, but it is sometimes useful (or necessary) to consider another 'distance' between points. However, note that the Cauchy-Schwarz inequality is special to the Euclidean norm.

Let us shortly say what we consider a norm. In particular, to collect all properties that are shared by these quantities. (We discuss that later more detailed.)

Definition 8.7 (Norm). Let $\|\cdot\| : \mathbb{R}^d \rightarrow \mathbb{R}^+$ is called a **norm** if there holds

- 1) $\|x\| = 0 \iff x = 0$
- 2) For any $\lambda \in \mathbb{R}$ and any $x \in \mathbb{R}^d$ we have $\|\lambda x\| = |\lambda| \cdot \|x\|$.
- 3) For all $x, y \in \mathbb{R}^d$ we have $\|x + y\| \leq \|x\| + \|y\|$.

The properties 1) - 3) are called **definiteness**, **homogeneity** and **triangle inequality**.

As shown above, the Euclidean norm $\|\cdot\|_2$, which is also called the **2-norm**, is a norm in this sense.

There are two other important norms. The first is the **1-norm**, which is sometimes called *Manhattan norm*, given by

$$\|x\|_1 = \sum_{i=1}^d |x_i|$$

for $x = (x_1, \dots, x_d) \in \mathbb{R}^d$. The other one is the **maximum norm**, or ∞ -norm, given by

$$\|x\|_\infty = \max\{|x_1|, |x_2|, \dots, |x_d|\}.$$

It is easy to see that both fulfill the first requirement for being a norm, the definiteness. For the homogeneity note that

$$\|\lambda x\|_1 = \sum_{i=1}^d |\lambda x_i| = |\lambda| \sum_{i=1}^d |x_i| = |\lambda| \cdot \|x\|_1.$$

and

$$\|\lambda x\|_\infty = \max\{|\lambda x_1|, |\lambda x_2|, \dots, |\lambda x_d|\} = |\lambda| \max\{|x_1|, |x_2|, \dots, |x_d|\} = |\lambda| \|x\|_\infty.$$

Finally, the triangle inequality follows from

$$\|x + y\|_1 = \sum_{i=1}^d |x_i + y_i| \leq \sum_{i=1}^d (|x_i| + |y_i|) = \sum_{i=1}^d |x_i| + \sum_{i=1}^d |y_i| = \|x\|_1 + \|y\|_1$$

and

$$\begin{aligned} \|x + y\| &= \max\{|x_1 + y_1|, |x_2 + y_2|, \dots, |x_d + y_d|\} \\ &\leq \max\{|x_1| + |y_1|, |x_2| + |y_2|, \dots, |x_d| + |y_d|\} \\ &\leq \max\{|x_1|, |x_2|, \dots, |x_d|\} + \max\{|y_1|, |y_2|, \dots, |y_d|\}, \end{aligned}$$

where we used the univariate triangle inequality several times.

It is important to note that differences between these norms cannot be too large, as the following lemma shows.

Lemma 8.8. *For any $x \in \mathbb{R}^d$ we have*

- 1) $\|x\|_\infty \leq \|x\|_2 \leq \sqrt{d} \|x\|_\infty$,
- 2) $\frac{1}{\sqrt{d}} \|x\|_1 \leq \|x\|_2 \leq \sqrt{d} \|x\|_1$,
- 3) $\|x\|_\infty \leq \|x\|_1 \leq d \|x\|_\infty$.

Note that, in contrast to what may be suggested by the name, the maximum norm is the smallest of these norms. However, as all these norms only differ by a multiplicative constant (that only depends on the dimension), we obtain that for a sequence (x_k) we have

$$x_k \rightarrow x \iff \|x_k - x\|_p \rightarrow 0,$$

where p stands for 1, 2 or ∞ . Therefore, when talking about convergence, it does not matter which norm we use.

Let us finally prove the last lemma.

Proof. Let k be such that $|x_k| = \max\{|x_1|, |x_2|, \dots, |x_d|\}$. Then

$$|x_k| \leq \sqrt{\sum_{i=1}^d |x_i|^2} \leq \sqrt{\sum_{i=1}^d |x_k|^2} = |x_k| \sqrt{\sum_{i=1}^d 1} = \sqrt{d} |x_k|.$$

(Note the indices.) Thus the first point follows. For the third point we calculate

$$|x_k| \leq \sum_{i=1}^d |x_i| \leq \sum_{i=1}^d |x_k| = d |x_k|.$$

The upper bound in the second point follows by combining point 1) and 3) as

$$\|x\|_2 \leq \sqrt{d} \|x\|_\infty \leq \sqrt{d} \|x\|_1.$$

For the lower bound we use the Cauchy-Schwarz inequality, see Lemma 1.76, to obtain

$$\|x\|_1 = \sum_{i=1}^d |x_i| = \sum_{i=1}^d |x_i| \cdot 1 \leq \|x\|_2 \cdot \sqrt{\sum_{i=1}^d 1^2} = \sqrt{d} \|x\|_2.$$

Dividing by \sqrt{d} gives the lower bound on $\|x\|_2$. □

8.2 Continuous functions

The definition of continuity in the multivariate case is the same as in the one dimensional case. That is, we require that we can interchange the limit with the function.

Definition 8.9. Let $\Omega \subset \mathbb{R}^d$, $f: \Omega \rightarrow \mathbb{R}^m$ and $x_0 \in \Omega$.

Then we call f **continuous at x_0** if for any sequence $(x_k)_{k \in \mathbb{N}} \subset \mathbb{R}^d$, such that $x_k \in \Omega$ for all k and $x_k \rightarrow x_0$, we have

$$\lim_{k \rightarrow \infty} f(x_k) = f\left(\lim_{k \rightarrow \infty} x_k\right) = f(x_0).$$

If $U \subset \Omega$ and f is continuous at any $x_0 \in U$, then we say that f is continuous on U .

Again, to have a shorter notation, we use the concept of a limit of a function. Let us recall what we mean by this, see Definition 4.27.

Definition 8.10 (Limit of functions). Let $\Omega \subset \mathbb{R}^d$ and $f: \Omega \rightarrow \mathbb{R}^m$. Moreover, let $y \in \mathbb{R}^m$ and $x_0 \in \mathbb{R}^d$ be an **accumulation point** of Ω , i.e., there is a sequence $(x_k)_{k \in \mathbb{N}} \subset \Omega \setminus \{x_0\}$ with $x_k \rightarrow x_0$. Then, we call y the **limit of f as $x \rightarrow x_0$** , if for all sequences $(x_k) \subset \Omega \setminus \{x_0\}$ with $x_k \rightarrow x_0$, we have

$$f(x_k) \rightarrow y.$$

In this case we use the notation

$$\lim_{x \rightarrow x_0} f(x) = y.$$

Note that, for technical reasons, we have to exclude the limit from the sequences, i.e., we only consider sequences (x_k) with $x_k \rightarrow x_0$ and $x_k \neq x_0$ to define the above limit. If no such sequence exists for x_0 , then we call x_0 an **isolated point** of Ω , and the limit is not defined. However, as discussed in Remark 4.29, a function is always continuous at isolated points, and we therefore see that a function is continuous on $U \subset \Omega$ if and only if

$$\lim_{x \rightarrow x_0} f(x) = f(x_0).$$

for any accumulation point $x_0 \in \Omega$.

Note that in the definitions above we also consider vector-valued functions, i.e. $m \geq 2$, as it really makes no difference here. One just has to consider all limits component-wise, as by definition. However, we focus on real-valued functions for some time now.

Already the similarity of all the definitions to the univariate case may indicate that several of the rules and concepts will also work here, and this is indeed true. The only additional difficulty is that there are now **more indices and variables that have to be considered separately**. But, if one uses a good notation and keeps track of all details, then there are almost no additional concepts needed in this section.

Let us start with some easy examples for $d = 2$.

Let the functions $f, g: \mathbb{R}^2 \rightarrow \mathbb{R}$ and $h: \mathbb{R} \times (\mathbb{R} \setminus \{0\}) \rightarrow \mathbb{R}$ be given by

- $f(x_1, x_2) = x_1 + x_2$,

- $g(x_1, x_2) = x_1 \cdot x_2$, and
- $h(x_1, x_2) = \frac{x_1}{x_2}$.

(Note that $h: \mathbb{R} \times (\mathbb{R} \setminus \{0\}) \rightarrow \mathbb{R}$ means that the second input is not allowed to be zero.)

These functions are particularly simple as they are just a sum/product/quotient of functions depending on one variable. From our knowledge about real sequences, we can easily check that these functions are continuous. For this, let us assume that $(x_k)_{k \in \mathbb{N}}$ is an arbitrary sequence converging to some $x_0 = (x_{0,1}, x_{0,2}) \in \mathbb{R}^2$. We use the notation $x_k = (x_{k,1}, x_{k,2})$. Using that $(x_{k,1}, x_{k,2}) \rightarrow (x_{0,1}, x_{0,2})$ if and only if $x_{k,1} \rightarrow x_{0,1}$ and $x_{k,2} \rightarrow x_{0,2}$, we obtain

$$\lim_{k \rightarrow \infty} f(x_k) = \lim_{k \rightarrow \infty} (x_{k,1} + x_{k,2}) = \lim_{k \rightarrow \infty} (x_{k,1}) + \lim_{k \rightarrow \infty} (x_{k,2}) = x_{0,1} + x_{0,2} = f(x_0).$$

Since this holds for arbitrary sequences converging to arbitrary $x_0 \in \mathbb{R}^2$, we see that f is continuous. The same calculations can be done for g and h .

Similar computations can be used to prove continuity of sum/product/quotient of more general functions. This leads to a statement very similar to the case univariate functions.

Lemma 8.11. *Let $\Omega \subset \mathbb{R}^d$ and $f, g: \Omega \rightarrow \mathbb{R}$ be continuous at $x_0 \in \Omega$.*

Then $f + g$ and $f \cdot g$ are continuous at x_0 .

If additionally $g(x_0) \neq 0$, then $\frac{f}{g}$ is also continuous at x_0 .

Proof. The proof is exactly the same as the proofs of Theorem 4.12 and Theorem 4.17. □

One can consider the above lemma also for vector-valued functions $f, g: \mathbb{R}^d \rightarrow \mathbb{R}^m$ any $m \geq 2$. Clearly, by considering all components separately, we see that the lemma holds also for $f + g$. However, **product and quotient do not make sense for vector-valued functions**.

Additionally, we can again consider the composition of functions. Recall that $g \circ f(x) := g(f(x))$. For vector-valued functions this makes sense only if the dimensions of the corresponding functions agree. That is, if f 'outputs' a vector from \mathbb{R}^p , then g must 'accept' such vectors as 'input'. However, in this case one can show analogously to the univariate case that the **composition of continuous functions** is again a continuous function.

Lemma 8.12. *Let $\Omega \subset \mathbb{R}^d$, $f: \Omega \rightarrow \mathbb{R}^p$ and $g: \mathbb{R}^p \rightarrow \mathbb{R}^m$ for some $d, p, m \in \mathbb{N}$.*

If f is continuous at x_0 and g is continuous at $y_0 = f(x_0)$, then $g \circ f$ is continuous at x_0 .

The most important example for now is with $p = m = 1$ and $h: \mathbb{R} \rightarrow \mathbb{R}$ with $h(y) = |y|$. This implies that the absolute value $|f|$ is continuous for every real-valued continuous function $f: \mathbb{R}^d \rightarrow \mathbb{R}$.

For convenience, we repeat the simple proof.

Proof. If $\lim_{k \rightarrow \infty} x_k = x_0$, then $\lim_{k \rightarrow \infty} f(x_k) = f(x_0) = y_0$, since f is continuous at x_0 . Since g is continuous at y_0 , this implies $\lim_{k \rightarrow \infty} g(f(x_k)) = g(y_0) = g(f(x_0))$, proving the claim.

In a more compact form, one could write

$$\lim_{k \rightarrow \infty} g(f(x_k)) = g\left(\lim_{k \rightarrow \infty} f(x_k)\right) = g\left(f\left(\lim_{k \rightarrow \infty} x_k\right)\right) = g(f(x_0)).$$

□

With these results, one can easily prove that certain functions are continuous.

Example 8.13. First, we consider the norms discussed above, which are clearly also real-valued functions on \mathbb{R}^d . For notational convenience, let us write $f_1(x) := \|x\|_1$, $f_2(x) := \|x\|_2$ and $f_\infty(x) := \|x\|_\infty$ for $x \in \mathbb{R}^d$, or short $f_p := \|\cdot\|_p$ for $p \in \{1, 2, \infty\}$.

All these functions are compositions of the functions $g_i(x_1, \dots, x_d) := |x_i|$, which are clearly continuous (and actually univariate). f_1 and f_∞ are their sum and maximum, respectively, which are therefore continuous. (One might proof by induction that the maximum of d numbers is continuous.) For f_2 , note that also g_i^2 and hence $\sum_{i=1}^d g_i^2$ are continuous on \mathbb{R}^d . Continuity of f_2 then follows by the last part of Lemma 8.11 with $D = \mathbb{R}_+$ and $h(t) := \sqrt{t}$.

Example 8.14. Have a look at the function

$$f(x_1, x_2, x_3) = x_1 \cos(x_2) + e^{x_3} - \frac{\sqrt{x_1}}{1 + (x_2 + x_1)^2}.$$

This function consists only of continuous functions, and the denominator is bounded away from zero, which implies that f is continuous.

One important class of continuous functions are **multivariate (algebraic) polynomials**.

Example 8.15. A multivariate algebraic polynomial has the form

$$p(x_1, \dots, x_d) = \sum_{\|k\|_1 \leq r} c_k \cdot x_1^{k_1} x_2^{k_2} \cdots x_d^{k_d},$$

where the sum is over all $k = (k_1, \dots, k_d) \in \mathbb{N}_0^d$ with $\|k\|_1 = k_1 + \cdots + k_d \leq r \in \mathbb{N}_0$, and the $c_k \in \mathbb{R}$ are the coefficients of the polynomial. We call r the degree of this polynomial (at least if $c_k \neq 0$ for at least one k with $\|k\|_1 = r$). One may even write such polynomials in a shorter way by introducing the so-called **multi-index notation**. That is, for $x \in \mathbb{R}^d$ and $k \in \mathbb{N}_0^d$, we define

$$x^k := \prod_{i=1}^d x_i^{k_i} = x_1^{k_1} x_2^{k_2} \cdots x_d^{k_d}.$$

With this, we can write multivariate polynomials in the familiar form $p(x) = \sum_{\|k\|_1 \leq r} c_k x^k$. A simple example for $d = 2$ is given by

$$p(x_1, x_2) = 2x_1^3 x_2 + x_1^2 x_2^2 + 3x_2 - 42.$$

This polynomial is in the above form with $c_{(3,1)} = 2$, $c_{(2,2)} = 1$, $c_{(0,1)} = 3$, $c_{(0,0)} = -42$ and all other c_k 's equal zero. The degree of this polynomial is therefore 4.

By Lemma 8.11, multivariate polynomials are clearly continuous.

However, in general it is not so easy to prove continuity of multivariate functions. The reason is that there are 'too many' sequences converging to a point. (While there were only left- and right-handed limits in the univariate case, there are infinitely many 'directions' for multivariate functions.) Let us see the following example.

Example 8.16. We consider the function defined by

$$f(x_1, x_2) = \begin{cases} \frac{x_1 \cdot x_2}{x_1^2 + x_2^2} & \text{if } x \neq 0 \\ 0 & \text{if } x = 0. \end{cases}$$

This function is clearly continuous at any $x_0 \neq 0$. If we want to prove continuity at 0, i.e., that $\lim_{x \rightarrow 0} f(x) = 0$, we need to consider all null sequences. In a first try, we may consider the sequences

given by $y_k = (\frac{1}{k}, 0)$ and $z_k = (0, \frac{1}{k})$, which correspond to the limits in *coordinate direction*. Since $f(y_k) = f(z_k) = 0$ for all k (because one of the components is always zero) we see that $f(y_k) \rightarrow 0$ and $f(z_k) \rightarrow 0$, which might indicate that the function is continuous. However, if we consider the sequence $t_k = (1/k, 1/k)$ (i.e., the limit 'from the diagonal'), then we see that

$$f(t_k) = \frac{\frac{1}{k^2}}{\frac{1}{k^2} + \frac{1}{k^2}} = \frac{1}{2}.$$

This implies that the limit $\lim_{x \rightarrow 0} f(x)$ does not exist, i.e., f is not continuous at 0.

This shows that proving continuity of a function is sometimes an issue, and one needs kind of intuition to give a detailed proof. In most cases, one even has to try several approaches, as there is no general rule for this. However, in some cases a function depends on its input only by the norm of the input, and for such functions one might easily obtain continuity.

Example 8.17. We consider the function defined by

$$f(x_1, x_2) = \begin{cases} \frac{1-\cos(|x_1|+|x_2|)}{\sqrt{x_1^2+x_2^2}} & \text{if } x \neq 0 \\ 0 & \text{if } x = 0. \end{cases}$$

Again, the function is continuous at $x_0 \neq 0$. For $x_0 = 0$, note that Lemma 8.8 implies that $\sqrt{x_1^2+x_2^2} = \|(x_1, x_2)\|_2 \geq \frac{1}{\sqrt{2}}\|(x_1, x_2)\|_1$, and therefore, with $x = (x_1, x_2) \neq 0$,

$$|f(x)| = \frac{|1 - \cos(\|x\|_1)|}{\|x\|_2} \leq \sqrt{2} \frac{|1 - \cos(\|x\|_1)|}{\|x\|_1}.$$

If we now take into account that $x \rightarrow 0$ if and only if $\|x\|_1 \rightarrow 0$ (and use the substitution $t := \|x\|_1$), we obtain that

$$\lim_{x \rightarrow 0} |f(x)| \leq \lim_{t \rightarrow 0} \sqrt{2} \frac{|1 - \cos(t)|}{t} = 0.$$

(Here, we used l'Hospital's rule.) This implies $\lim_{x \rightarrow 0} f(x) = 0$, and thus that f is continuous.

We want to finish this subsection with the ε - δ -criterion for multivariate functions, as this will again be needed in some of the upcoming proofs. Note that this is again very similar to the univariate criterion, see Theorem 4.63.

Theorem 8.18. Let $f: \Omega \rightarrow \mathbb{R}$ and $p \in \{1, 2, \infty\}$. Then, f is continuous at $x_0 \in \Omega$ if and only if for any $\varepsilon > 0$ there exists a $\delta > 0$ such that for all $x \in \Omega$, we have

$$\|x - x_0\|_p < \delta \implies |f(x) - f(x_0)| < \varepsilon.$$

In a formula,

$$\forall \varepsilon > 0 \exists \delta > 0 \forall x \in \Omega: \|x - x_0\|_p < \delta \implies |f(x) - f(x_0)| < \varepsilon.$$

The proof is almost the same as for the one-dimensional case, but we still want to give it here. Moreover, note that by Lemma 8.8 continuity is just equivalent for all $p \in \{1, 2, \infty\}$. Therefore, it is enough to prove it for $p = 2$ (or any other of them).

Proof. First we show that the ε - δ -criterion holds if f is continuous. Therefore we assume the opposite, i.e., that there exists $\varepsilon_0 > 0$ such that for any $\delta > 0$ there is some $y \in \Omega$ with $\|y - x_0\| < \delta$ but $|f(y) - f(x_0)| > \varepsilon_0$, and show that this contradicts the continuity of f . In particular, by assumption, we may choose $\delta = \frac{1}{n}$ and find some $y_n \in \Omega$ with the property

$$\|y_n - x_0\| \leq \frac{1}{n} \quad \text{and} \quad |f(y_n) - f(x_0)| > \varepsilon_0.$$

This implies that $y_n \rightarrow x_0$, but $f(y_n) \not\rightarrow f(x)$. So, f is not continuous at x , which is a contradiction.

Next we assume that the ε - δ -criterion holds and show that in this case f is continuous. For this, fix some $\varepsilon > 0$. By definition, if a sequence (x_k) converges to x_0 , then for all k large enough it holds that $\|x_k - x_0\| < \delta$, where $\delta > 0$ is as given by the ε - δ -criterion. Therefore $|f(x_k) - f(x_0)| < \varepsilon$. As this holds for all $\varepsilon > 0$ and all sequences (x_k) , we obtain

$$\lim_{y \rightarrow x} f(y) = f(x),$$

so f is continuous at x .

□

8.3 Differential calculus

Now we are able to start our discussion about differentiating multivariate functions. As there are several variables one could imagine that there are different concepts and approaches to do so. Basically we want to discuss three different kinds of derivatives in the sequel, which are *partial derivatives*, derivatives which are seen as *linear mappings*, i.e., the *total derivative*, and the *directional derivatives*. All of them are somehow related to each other in some way, but as we will see some of them can exist, whilst others don't.

Note that, in the same way as in the univariate case, there are some technical problems related to *boundary points* of the domain of a function. Therefore, we consider at first only *open sets* in \mathbb{R}^d , which are the replacement of open intervals on the real line.

Definition 8.19. Let $x \in \mathbb{R}^d$, $\varepsilon > 0$ and $\|\cdot\|$ be a norm on \mathbb{R}^d . Then we define

$$U_\varepsilon(x) = \{y \in \mathbb{R}^d : \|x - y\| < \varepsilon\}$$

as **open neighborhood** or **open ball** (w.r.t. $\|\cdot\|$) of radius ε around x .

A set $G \subset \mathbb{R}^d$ is called **open** if for every $x \in G$ there exists some $\varepsilon > 0$ such that $U_\varepsilon(x) \subset G$.

A set $C \subset \mathbb{R}^d$ is called **closed** if $\mathbb{R}^d \setminus C$ is open.

In particular, every element $x_0 \in G$ of an open set G is an accumulation point of G .

8.3.1 Partial derivatives

Since this topic includes a lot of indices and so on we use the following notation which will make everything a bit shorter. If not indicated otherwise $\|\cdot\|$ always denotes the Euclidean norm, i.e. $\|x\| = \|x\|_2$, and $G \subset \mathbb{R}^d$ denotes an open subset of \mathbb{R}^d . Moreover, we want to remind you that the i -th unit vector, write e_i , was defined to have a 1 in the i -th coordinate and zeros else.

Finally, let us note that, due to the different indices needed, there might be some confusion between the *coordinates of a vector* and the *elements of a sequence*, as already explained above. In what follows we always write (or at least try to)

$$x = (x_1, \dots, x_d) \in \mathbb{R}^d,$$

such that $x_i \in \mathbb{R}$, $i = 1, \dots, d$, is a number, i.e., the i th coordinate of x , and we use (again) $x_0 \in \mathbb{R}^d$ for a specific point.

Definition 8.20 (Partial derivative). Let $G \subset \mathbb{R}^d$ be an open set, $f: G \rightarrow \mathbb{R}$ and $x \in G$. If the limit

$$\frac{\partial f}{\partial x_i}(x) := \lim_{h \rightarrow 0} \frac{f(x + he_i) - f(x)}{h}$$

exists, then we call the function f **partially differentiable** at x w.r.t. the i -th coordinate.

If $\frac{\partial f}{\partial x_i}(x)$ exists for any $1 \leq i \leq n$ and all $x \in M \subset G$, then we call f **partially differentiable** in M and simply partially differentiable if $M = G$.

If f is partially differentiable (in a neighborhood of x) and all partial derivatives are continuous (at x), then we say f is **continuously partially differentiable** (at x).

Sometimes, also other notations for the partial derivatives $\frac{\partial f}{\partial x_i}$ are used, like $D_i f$ or $D_{e_i} f$ or $\partial_{x_i} f$, or just $\delta_i f$ or f_i . Therefore, one needs to be careful when using other literature.

Example 8.21. Let us have a look at the function

$$f(x_1, x_2) = x_1 \cdot x_2 - e^{-x_1},$$

considered on $G = \mathbb{R}^2$. To calculate $\frac{\partial f}{\partial x_1}(x)$, using the definition of partial derivatives, we have to compute the limit $h \rightarrow 0$ of

$$\begin{aligned} \frac{f(x + he_1) - f(x)}{h} &= \frac{f(x_1 + h, x_2) - f(x_1, x_2)}{h} = \frac{(x_1 + h) \cdot x_2 - e^{-(x_1+h)} - (x_1 \cdot x_2 - e^{-x_1})}{h} \\ &= \frac{(x_1 + h - x_1) \cdot x_2 - (e^{-x_1-h} - e^{-x_1})}{h} \\ &= \frac{h \cdot x_2 - \frac{e^{-x_1}(e^{-h} - 1)}{h}}{h}, \end{aligned}$$

where $x = (x_1, x_2)$. This implies, by recalling known limits, that

$$\frac{\partial f}{\partial x_1}(x) = x_2 + e^{-x_1}.$$

Note that we only send $h \rightarrow 0$ above, and that x_1 and x_2 are fixed.

In the same way, we obtain

$$\frac{\partial f}{\partial x_2}(x) = \lim_{h \rightarrow 0} \frac{f(x_1, x_2 + h) - f(x_1, x_2)}{h} = \lim_{h \rightarrow 0} \frac{x_1 \cdot (x_2 + h - x_2) - (e^{-x_1} - e^{-x_1})}{h} = x_1.$$

□

A detailed inspection of the above example shows that there is a simpler way to compute partial derivatives, than just to plug in the definition. In fact, note that partial differentiation w.r.t. x_1 actually does not 'touch' x_2 . Therefore, we may just consider x_2 as a (fixed) constant and differentiate the univariate function depending only on x_1 . To be precise, consider the expression

$$\lim_{h \rightarrow 0} \frac{f(x + he_i) - f(x)}{h} = \lim_{h \rightarrow 0} \frac{f(x_1, \dots, x_i + h, \dots, x_d) - f(x_1, \dots, x_i, \dots, x_d)}{h}.$$

where $x = (x_1, x_2, \dots, x_d)$ is a fixed point. In this expression, the 'inputs' $x_1, \dots, x_{i-1}, x_{i+1}, \dots, x_d$ are 'untouched', allowing to treat them as fixed. So by defining the univariate function

$$g(x_i) := f(x),$$

we see that

$$g'(x_i) = \frac{\partial f}{\partial x_i}(x).$$

Thus we can compute partial derivatives by calculating one dimensional derivatives, which allows to use all our knowledge from the previous chapter.

Let us see how this procedure helps us if we want to compute partial derivatives.

Example 8.22. We come back to the example 8.21, where we considered

$$f(x_1, x_2) = x_1 \cdot x_2 - e^{-x_1}.$$

Observing x_2 as constant we can immediately see that

$$\frac{\partial f}{\partial x_1}(x) = x_2 + e^{-x_1}.$$

In almost the same way we see that

$$\frac{\partial f}{\partial x_2}(x) = x_1.$$

So indeed this way of calculating the partial derivatives is much easier to handle.

Example 8.23. Sometimes we also have to use some (one dimensional) calculation rules to compute partial derivatives. We have a look at

$$f(x_1, x_2) = \sin(x_1^3 + x_2).$$

Applying the chainrule for one dimensional functions we compute

$$\frac{\partial f}{\partial x_1}(x) = 3x_1^2 \cos(x_1^3 + x_2).$$

where we again considered x_2 as constant. Analogously it follows that

$$\frac{\partial f}{\partial x_2}(x) = \cos(x_1^3 + x_2).$$

Example 8.24. We define $G = \mathbb{R}^d \setminus \{0\}$ (note that this is an open set) and compute the partial derivatives of $f(x) = \|x\|_2$. Since

$$f(x) = (x_1^2 + x_2^2 + \cdots + x_d^2)^{1/2},$$

the chainrule implies that

$$\frac{\partial f}{\partial x_i}(x) = \frac{1}{2} \frac{1}{(x_1^2 + x_2^2 + \cdots + x_d^2)^{1/2}} \cdot 2x_i = \frac{1}{\|x\|_2} x_i.$$

Example 8.25. The function defined by

$$f(x_1, x_2) = \begin{cases} \frac{x_1 \cdot x_2}{x_1^2 + x_2^2} & \text{if } x \neq 0 \\ 0 & \text{if } x = 0. \end{cases}$$

Using the product rule for one dimensional functions we easily see that f is partially differentiable in $\mathbb{R}^2 \setminus \{0\}$. Furthermore, we observe that

$$\frac{f(h e_i) - f(0)}{h} = \frac{0}{h} = 0$$

for any $h \in \mathbb{R} \setminus \{0\}$. Thus $\frac{\partial f}{\partial x_1}(0)$ and $\frac{\partial f}{\partial x_2}(0)$ also exist, making f partially differentiable on \mathbb{R}^2 .

However, f is not continuous at 0 as we will show by using the sequence $x_k = (1/k, 1/k)$, which converges to 0. But

$$f(x_k) = \frac{\frac{1}{k^2}}{2 \frac{1}{k^2}} = \frac{1}{2}.$$

This implies that $f(x_k)$ cannot converge to $0 = f(0)$.

Writing all partial derivatives of a function in a (row) vector, we obtain a compact notation.

Definition 8.26 (Gradient). Let $f: G \rightarrow \mathbb{R}$ be partially differentiable in G . Then, we call

$$(\text{grad } f)(x) = \nabla f(x) = \left(\frac{\partial f}{\partial x_1}(x), \frac{\partial f}{\partial x_2}(x), \dots, \frac{\partial f}{\partial x_d}(x) \right)$$

the **gradient of f** at the point $x \in G$.

Remark 8.27. Some authors prefer to write the gradient as a column vector.

Example 8.28. We consider again Example 8.23, where we considered

$$f(x_1, x_2) = \sin(x_1^3 + x_2).$$

We already computed the partial derivatives and saw that they exist for any $(x_1, x_2) \in \mathbb{R}^2$. Thus

$$\nabla f(x) = (3x_1^2 \cos(x_1^3 + x_2), \cos(x_1^3 + x_2)).$$

There is a result for gradients which can be considered as generalization of the product rule.

Theorem 8.29 (Product rule). *Let $f, g: G \rightarrow \mathbb{R}$ be partially differentiable functions. Then we have*

$$\nabla(fg) = (\nabla f) \cdot g + (\nabla g) \cdot f.$$

Proof. By the definition of the gradient it is sufficient to proof the statement for each coordinate. We use the product rule for one dimensional functions to compute

$$\frac{\partial fg}{\partial x_i}(x) = \frac{\partial f}{\partial x_i}(x) \cdot g(x) + \frac{\partial g}{\partial x_i}(x) \cdot f(x).$$

□

Example 8.30. Consider the easy example

$$f(x_1, x_2) = x_1 \cdot x_2,$$

for which we clearly have $\nabla f(x) = (x_2, x_1)$. However, if we write $f = g \cdot h$ with $g(x_1, x_2) = x_1$ and $h(x_1, x_2) = x_2$, we see that $\nabla g(x) = (1, 0)$ and $\nabla h(x) = (0, 1)$ for all $x = (x_1, x_2) \in \mathbb{R}^2$. With the product rule, we obtain

$$\begin{aligned} \nabla f(x) &= \nabla(gh)(x) = \nabla g(x) \cdot h(x) + \nabla h(x) \cdot g(x) \\ &= (h(x), 0) + (0, g(x)) = (h(x), g(x)) = (x_2, x_1). \end{aligned}$$

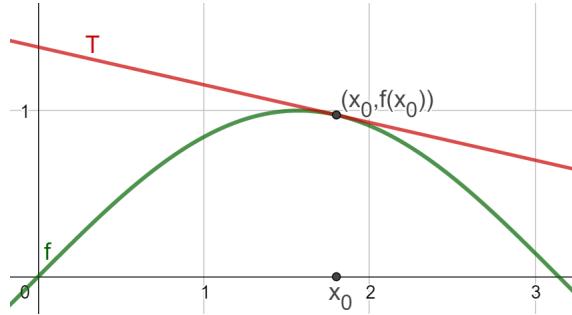
Note that in these computations, it is somehow obvious that everything is done at the (fixed) point x . Therefore, the '(x)' is unnecessary and we write in short

$$\begin{aligned} \nabla f &= \nabla(gh) = \nabla g \cdot h + \nabla h \cdot g \\ &= (h, 0) + (0, g) = (h, g) = (x_2, x_1). \end{aligned}$$

8.3.2 (Total) differentiability

Example 8.25 shows that existence of partial derivatives does not imply continuity of functions, which was a necessary condition for differentiating univariate functions. This, in particular, shows why we want to find a somehow 'better' generalization of the one dimensional differentiability. Therefore, recall that a function $f: \mathbb{R} \rightarrow \mathbb{R}$ is differentiable at $x_0 \in \mathbb{R}$ if and only if

$$\lim_{h \rightarrow 0} \frac{f(x_0 + h) - f(x_0)}{h} = \lim_{x \rightarrow x_0} \frac{f(x) - f(x_0)}{x - x_0}$$

Figure 48: Tangent line T to a function f at x_0

exists and in this case we can write

$$f'(x_0) = \lim_{x \rightarrow x_0} \frac{f(x) - f(x_0)}{x - x_0} \iff \lim_{x \rightarrow x_0} \frac{f(x) - f(x_0) - (x - x_0)f'(x_0)}{x - x_0} = 0.$$

One way to interpret this, is that $T(x) := f(x_0) + (x - x_0)f'(x_0)$ is the tangent line to f at x_0 .

The tangent is by definition an *affine function*, i.e., a linear function plus a constant. Therefore, one can say that the tangent line to a function f at x_0 is given by $T(x) = f(x_0) + D(x - x_0)$, where $D: \mathbb{R} \rightarrow \mathbb{R}$ is a *linear function*. Note that $D(x - x_0) = 0$ for $x = x_0$ (since D is linear), and therefore $T(x_0) = f(x_0)$.

We will now use this to define differentiability in the multidimensional case. Recall that a *linear mapping* $D: \mathbb{R}^d \rightarrow \mathbb{R}$ is characterized by the property that $D(x + y) = D(x) + D(y)$ for any $x, y \in \mathbb{R}^d$, and can always be described by $D(x) = \sum_{i=1}^d a_i x_i = a^T x = \langle a, x \rangle$ for some $a \in \mathbb{R}^d$.

This definition does not appear very handy, but we will see shortly its relation to the gradient.

Definition 8.31. Let $G \subset \mathbb{R}^d$ be an open set, $f: G \rightarrow \mathbb{R}$ and $x \in G$.

We call f (**totally**) **differentiable** at x if there exists a linear mapping $df_x: \mathbb{R}^d \rightarrow \mathbb{R}$ such that

$$\lim_{y \rightarrow 0} \frac{f(x + y) - f(x) - df_x(y)}{\|y\|} = 0.$$

The mapping df_x is called (**total**) **derivative** (or differential) of f at x .

Equivalently, f is (**totally**) **differentiable** at x with derivative df_x if there exists a function $r: \mathbb{R}^d \rightarrow \mathbb{R}$ such that

$$f(x + y) = f(x) + df_x(y) + r(y)$$

(whenever $x + y \in G$) and

$$\lim_{y \rightarrow 0} \frac{r(y)}{\|y\|} = 0.$$

If f is differentiable at every point of G , then we simply say f is differentiable.

Remark 8.32. There is again some other commonly used notation for the derivative df_x . For example, $f'(x)$, $D_x f$, or (even without the point x) Df or df . So, again be careful when using other literature.

In the same way as for univariate functions, i.e., $d = 1$, the derivative df_x can be used to define the **tangent plane** T to a function f at the point $x \in \mathbb{R}^d$ by $T(y) := f(x) + df_x(y - x)$, which is the **best approximation by an affine function** at x . We come back to this and give a handy formula for the tangent plane using partial derivatives.

Let us discuss an example.

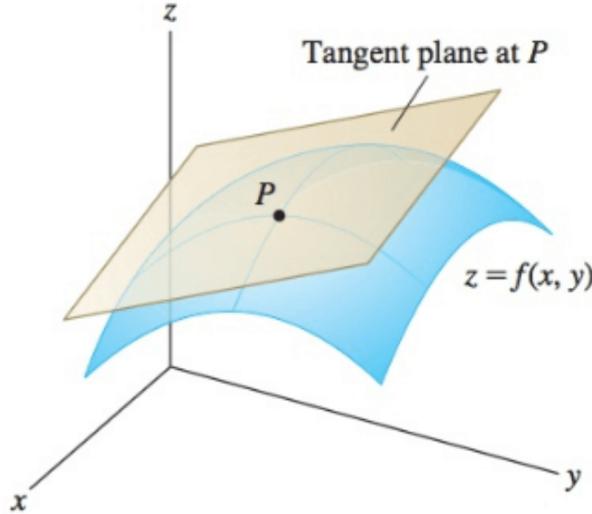


Figure 49: Tangent plane T to f at x intersects f at $P = (x, f(x))$

Example 8.33. Let $f: \mathbb{R}^3 \rightarrow \mathbb{R}$ be given by

$$f(x) = \|x\|^2 = x_1^2 + x_2^2 + x_3^2$$

with $x = (x_1, x_2, x_3)$. We see that

$$\begin{aligned} f(x+y) &= (x_1 + y_1)^2 + (x_2 + y_2)^2 + (x_3 + y_3)^2 \\ &= x_1^2 + 2x_1y_1 + y_1^2 + x_2^2 + 2x_2y_2 + y_2^2 + x_3^2 + 2x_3y_3 + y_3^2 \\ &= x_1^2 + x_2^2 + x_3^2 + 2x_1y_1 + 2x_2y_2 + 2x_3y_3 + y_1^2 + y_2^2 + y_3^2 \\ &= f(x) + 2x_1y_1 + 2x_2y_2 + 2x_3y_3 + f(y). \end{aligned}$$

(We first collected the terms that do not depend on y , then the linear terms, and then the rest.) We see that $r(y) = f(y) = \|y\|^2$ satisfies

$$\lim_{y \rightarrow 0} \frac{r(y)}{\|y\|} = \lim_{y \rightarrow 0} \frac{\|y\|^2}{\|y\|} = \lim_{y \rightarrow 0} \|y\| = 0.$$

Therefore, the linear mapping/function

$$df_x(y) = 2(x_1y_1 + x_2y_2 + x_3y_3) = 2 \sum_{i=1}^3 x_i y_i$$

is the (total) derivative of f at x . The tangent plane T to f at a point $x \in \mathbb{R}^3$ is therefore

$$T(y) = f(x) + df_x(y - x) = \|x\|^2 + 2 \sum_{i=1}^3 x_i(y_i - x_i) = -\|x\|^2 + 2 \sum_{i=1}^3 x_i y_i.$$

This example was particularly simple, because the *error* of the linear approximation, i.e., the function r , did not depend on x . However, this may clearly happen, as the next example shows.

Example 8.34. Let $f: \mathbb{R}^2 \rightarrow \mathbb{R}$ be given by

$$f(x) = x_1^2 \cdot x_2$$

with $x = (x_1, x_2)$. We see that

$$\begin{aligned} f(x+y) &= (x_1 + y_1)^2(x_2 + y_2) \\ &= (x_1^2 + 2x_1y_1 + y_1^2)(x_2 + y_2) \\ &= f(x) + 2x_1x_2y_1 + x_1^2y_2 + x_2y_1^2 + 2x_1y_1y_2 + y_1^2y_2. \end{aligned}$$

(Again, we collected the terms that do not depend on y , then the linear terms, then the rest.) We see that $r(y) = x_2y_1^2 + 2x_1y_1y_2 + y_1^2y_2$ (which depends on x) satisfies

$$\lim_{y \rightarrow 0} \frac{r(y)}{\|y\|} = \lim_{y \rightarrow 0} \left(x_2 \frac{y_1^2}{\|y\|} + 2x_1 \frac{y_1y_2}{\|y\|} + \frac{y_1^2y_2}{\|y\|} \right) = 0.$$

(Verify this yourself using $y_1y_2 = \max\{y_1, y_2\} \min\{y_1, y_2\}$ and $\|y\| \geq \|y\|_\infty$.) Therefore, the linear function

$$df_x(y) = 2x_1x_2y_1 + x_1^2y_2$$

is the (total) derivative of f at x .

Let us now show that differentiability is indeed a stronger condition than the existence of partial derivatives, and implies continuity.

Theorem 8.35. *Let $G \subset \mathbb{R}^d$ be open and $f: G \rightarrow \mathbb{R}$ be (totally) differentiable at $x \in G$. Then,*

- 1) *f is continuous at x ,*
- 2) *all partial derivatives of f exist at x , and*
- 3) *the (total) derivative of f is given by the gradient by*

$$df_x(y) = \nabla f(x) \cdot y = \langle \nabla f(x), y \rangle = \sum_{i=1}^d \frac{\partial f}{\partial x_i}(x) \cdot y_i.$$

Proof. Since f is differentiable at x we know that df_x and r exist such that

$$f(x+y) = f(x) + df_x(y) + r(y),$$

df_x is linear and that $\lim_{y \rightarrow 0} \frac{r(y)}{\|y\|} = 0$. Thus we have that

$$\lim_{y \rightarrow x} f(y) = \lim_{y \rightarrow 0} f(x+y) = f(x) + \lim_{y \rightarrow 0} df_x(y) + \lim_{y \rightarrow 0} r(y).$$

Since df_x is linear, and therefore continuous, we have that $\lim_{y \rightarrow 0} df_x(y) = df_x(0) = 0$. For the second limit we use $\lim_{y \rightarrow 0} |r(y)| \leq \lim_{y \rightarrow 0} \frac{|r(y)|}{\|y\|} = 0$. This shows that all in all we have that

$$\lim_{y \rightarrow x} f(y) = f(x), \quad \text{i.e., } f \text{ is continuous at } x.$$

To show that all partial derivatives exist, and how they are related to the total derivative, first observe that, due to linearity, df_x can be written by $df_x(y) = \sum_{i=1}^d a_i y_i$ for some $a_1, \dots, a_d \in \mathbb{R}$. Using again the representation $f(x+y) = f(x) + \sum_{i=1}^d a_i y_i + r(y)$ for the specific sequences $y = he_i$ for $h \rightarrow 0$ ($h \in \mathbb{R}$), we obtain

$$\frac{f(x+he_i) - f(x)}{h} = \frac{ha_i + r(he_i)}{h} = a_i + \frac{r(he_i)}{h}$$

(e_j denotes the j -th unit vector) Since $\lim_{h \rightarrow 0} \frac{r(he_i)}{h} = \lim_{y \rightarrow 0} \frac{r(y)}{\|y\|} = 0$ we see that $\frac{\partial f}{\partial x_i}(x)$ exists and

$$\frac{\partial f}{\partial x_i}(x) = \lim_{h \rightarrow 0} \frac{f(x+he_i)}{h} = a_i,$$

proving the theorem. □

Example 8.36. If we have a look at Example 8.33 we saw that f , which was given by

$$f(x) = \|x\|^2,$$

was differentiable for any $x \in \mathbb{R}^3$. Moreover, we calculated that $df_x(y) = 2\langle x, y \rangle = 2 \sum_{i=1}^3 x_i y_i$. The components are given by $2(x_1, x_2, x_3)$, which is exactly the gradient of f .

Still we would like to have a simple criterion to decide if f is differentiable.

Theorem 8.37. *Let $f: G \rightarrow \mathbb{R}$ be a continuously partially differentiable function at $x \in G$. Then, f is (totally) differentiable at x and the derivative is given by the gradient.*

Proof. Due to Theorem 8.35 it is sufficient to show that f is differentiable.

Since G is open there exists a neighbourhood of x which is contained in G , so for $\delta > 0$ small enough and $\|y\| < \delta$ the points

$$z^{(k)} = x + \sum_{i=1}^k y_i e_i, \quad k = 0, \dots, d,$$

are contained in G . Again, e_i denotes the i -th unit vector. Furthermore, we see that

$$z^{(k)} - z^{(k-1)} = y_k e_k, \quad z^{(0)} = x \quad \text{and} \quad z^{(d)} = y.$$

Thus the mean value theorem of differential calculus, see Theorem 5.34, implies that there exists some $\xi_k \in [0, 1]$ such that

$$f(z^{(k)}) - f(z^{(k-1)}) = \frac{\partial f}{\partial x_k} \left(z^{(k-1)} + \xi_k y_k e_k \right) \cdot y_k.$$

(Make sure you understand why we can apply the one dimensional mean value theorem here!)

We set $\eta_k = z^{(k-1)} + \xi_k y_k e_k$ and use a telescoping trick to see

$$f(x+y) - f(x) = \sum_{k=1}^{d-1} \left(f(z^{(k+1)}) - f(z^{(k)}) \right) = \sum_{k=1}^d \frac{\partial f}{\partial x_k}(\eta_k) \cdot y_k.$$

Now we define

$$a_k = \frac{\partial f}{\partial x_k}(x) \quad \text{and} \quad r(y) = \sum_{k=1}^d \left(\frac{\partial f}{\partial x_k}(\eta_k) - a_k \right) y_k,$$

such that

$$f(x+y) - f(x) = \sum_{k=1}^d a_k \cdot y_k + r(y) = \langle \nabla f(x), y \rangle + r(y).$$

Due to the continuity of all partial derivatives it follows that $a_k = \lim_{y \rightarrow 0} \frac{\partial f}{\partial x_k}(\eta_k)$. Moreover, an application of the Cauchy-Schwarz inequality yields

$$r(y) \leq \|\nabla f(\eta)\| \cdot \|y\|,$$

where we set $\eta = (\eta_1, \dots, \eta_d)$ and $a = (a_1, \dots, a_d)$. Putting everything together we obtain that

$$\lim_{y \rightarrow 0} \frac{r(y)}{\|y\|} = 0,$$

and therefore that $df_x(y) = \langle \nabla f(x), y \rangle$.

□

Remark 8.38. The opposite of the above theorem does not hold, i.e., there exist differentiable functions which are not continuously partially differentiable. One example that shows this is the function $f(x) = \|x\|^2 \sin\left(\frac{1}{\|x\|}\right)$, continuously extended to $x = 0$ by setting $f(0) = 0$, which is differentiable at 0, but whose partial derivatives are not continuous. We omit the details.

The theorem which we just showed also allows to make notation a bit easier, i.e. shorter. In particular we now know that, if a function is continuously partially differentiable, then it is differentiable and we can interpret the gradient as derivative. In this case the gradient is also continuous, since all components are continuous. So from here on we call a function **continuously differentiable**, if it is continuously partial differentiable. Nevertheless it is important to be very precise if we do not have continuous partial derivatives.

Let us see some more examples.

Example 8.39. Let C be a $d \times d$ matrix and consider

$$f(x) = x^T C x = \sum_{i=1}^d x_i \sum_{j=1}^d c_{ij} x_j = \sum_{i=1}^d \sum_{j=1}^d c_{ij} x_i x_j,$$

for $x = (x_1, \dots, x_d) \in \mathbb{R}^d$. Functions of this form are of particular interest for *quadratic optimization problems*.

Linearity and the product rule implies that the k -th component of ∇f is given by

$$\frac{\partial f}{\partial x_k}(x) = \sum_{i=1}^d \sum_{j=1}^d c_{ij} \frac{\partial(x_i x_j)}{\partial x_k} = \sum_{i=1}^d \sum_{j=1}^d c_{ij} \left(x_i \cdot \frac{\partial x_j}{\partial x_k} + x_j \cdot \frac{\partial x_i}{\partial x_k} \right)$$

Using that $\frac{\partial x_j}{\partial x_k} = 1$ for $j = k$ and $\frac{\partial x_j}{\partial x_k} = 0$ for $j \neq k$, i.e., $\frac{\partial x_j}{\partial x_k} = \delta_{jk}$, we obtain

$$\begin{aligned} \frac{\partial f}{\partial x_k}(x) &= \sum_{i=1}^d \sum_{j=1}^d c_{ij} (x_i \cdot \delta_{jk} + x_j \cdot \delta_{ik}) = \sum_{i=1}^d c_{ik} x_i + \sum_{j=1}^d c_{kj} x_j \\ &= (x^T C)_k + (C x)_k = (x^T C)_k + (x^T C^T)_k = (x^T (C + C^T))_k, \end{aligned}$$

where $(y)_k$ denotes the k -th entry of the (row/column) vector y . Therefore,

$$\nabla f(x) = x^T (C + C^T) = ((C + C^T)x)^T.$$

(Note that this is a row vector, as required.)

In the important special case that C is a symmetric matrix, i.e., $C = C^T$, we obtain

$$\nabla f(x) = 2x^T C = 2(Cx)^T.$$

Clearly, all partial derivatives (i.e., entries of ∇f) are linear and therefore continuous functions. By Theorem 8.37 this implies that f is (totally) differentiable.

Example 8.40. Now we have a look at the function

$$f(x) = e^{-\|x\|^2} = \exp(-(x_1^2 + x_2^2 + \dots + x_d^2)).$$

It follows that

$$\frac{\partial f}{\partial x_k}(x) = -2x_k e^{-\|x\|^2}.$$

Since all partial derivatives, and thus the gradient, are continuous, we see that f is differentiable for any $x \in \mathbb{R}^d$. Moreover,

$$df_x(y) = \langle \nabla f(x), y \rangle = -2e^{-\|x\|^2} \sum_{k=1}^d x_k y_k = -2e^{-\|x\|^2} \langle x, y \rangle$$

is the (total) derivative of f at x .

8.3.3 Directional derivatives

We finally discuss how to use multidimensional derivatives to describe the slope (or increase) of a function in a fixed direction. Note that in the multidimensional setting we need to decide in which direction we want to measure the slope. Just imagine you go for a walk on a mountain. Then there might be a different slope in each direction, and it might be of interest in which direction is the largest increase or decrease. It will turn out, that the gradient actually points to the direction of largest increase/decrease.

Definition 8.41 (Directional derivatives). Let $G \subset \mathbb{R}^d$ be open, $f: G \rightarrow \mathbb{R}$ and $x \in G$.

A vector $v \in \mathbb{R}^d$ with $\|v\|_2 = 1$ is called a **direction**, and we define the set of all directions, i.e., the $(d - 1)$ -dimensional **unit sphere**, by

$$\mathbb{S}^{d-1} := \{v \in \mathbb{R}^d : \|v\|_2 = 1\}.$$

The **directional derivative** of f at $x \in G$ w.r.t. v is given by

$$D_v f(x) = \lim_{h \rightarrow 0} \frac{f(x + hv) - f(x)}{h}$$

if the limit exists.

Remark 8.42. Again, there are many different notations for $D_v f(x)$, like $\frac{\partial f}{\partial v}(x)$ and $\nabla_v f(x)$.

Remark 8.43. If we choose $v = e_i$, we see that $D_{e_i} f = D_i f$ is the i -th partial derivative.

Remark 8.44. The intuition might be, that $D_v f(x)$ is the height change if we make 'one step' of length 1 on the tangent plane starting at x in direction v .

As before, we can give a formula for directional derivatives using the gradient.

Theorem 8.45. Let $G \subset \mathbb{R}^d$ be open, $f: G \rightarrow \mathbb{R}$ be differentiable at $x \in G$ and $v \in \mathbb{S}^{d-1}$. Then, the directional derivative at x w.r.t. v can be computed as

$$D_v f(x) = df_x(v) = \langle \nabla f(x), v \rangle = \sum_{i=1}^d \frac{\partial f}{\partial x_i}(x) \cdot v_i.$$

Proof. As for the proof of Theorem 8.35 we use a special choice of y in the definition of the (total) derivative to obtain the result.

By Definition 8.31 with $y = h \cdot v$ and $h \in \mathbb{R}$ small enough, we see that

$$\frac{f(x + hv) - f(x)}{h} = \frac{df_x(hv) + r(hv)}{h} = df_x(v) + \frac{r(hv)}{h},$$

where we used that $df_x(hv) = h \cdot df_x(v)$ since df_x is linear. Moreover, $\lim_{h \rightarrow 0} \frac{r(hv)}{h} = \lim_{y \rightarrow 0} \frac{r(y)}{\|y\|} = 0$ implies that $D_v f(x) = df_x(v)$. The rest of the statement follows from Theorem 8.35. \square

This result allows us to easily compute directional derivatives of functions.

Example 8.46. Let us consider the function from Example 8.22, i.e.,

$$f(x) = x_1 x_2 - e^{-x_1}.$$

We have that $\nabla f(x) = (x_2 + e^{-x_1}, x_1)$. So the directional derivative w.r.t. $v = (1/\sqrt{2}, 1/\sqrt{2})$ is given by

$$D_v f(x) = \frac{1}{\sqrt{2}}(x_2 + e^{-x_1} + x_1).$$

Example 8.47. As in Example 8.39 consider $f(x) = x^T C x$ where C was a symmetric matrix. The gradient of f is

$$\nabla f(x) = 2Cx,$$

which implies that for any v with $\|v\| = 1$ the directional derivative is given by

$$D_v f(x) = 2x^T Cv.$$

Interestingly, it turns out that the gradient points to the direction of the largest slope. That is, the gradient $\nabla f(x)$ is the direction such that $|D_v f(x)|$ is maximized.

Theorem 8.48. Let $G \subset \mathbb{R}^d$ and $f: G \rightarrow \mathbb{R}$ be differentiable at $x \in G$. Then,

$$|D_v f(x)| \leq \|\nabla f(x)\| \quad \text{for all } v \in \mathbb{S}^{d-1} = \{v \in \mathbb{R}^d : \|v\| = 1\}.$$

Moreover, $D_v f(x) = \pm \|\nabla f(x)\|$ if and only if $v = \pm \frac{\nabla f(x)}{\|\nabla f(x)\|}$.

Proof. Since $D_v f(x) = \langle \nabla f(x), v \rangle$, see Theorem 8.45, we obtain from the Cauchy-Schwarz inequality (Lemma 1.76) that

$$|D_v f(x)| = |\langle \nabla f(x), v \rangle| \leq \|\nabla f(x)\| \|v\| = \|\nabla f(x)\|,$$

since $\|v\| = 1$. Moreover, we have equality if and only if $v = c \cdot \nabla f(x)$, and this c must be $\pm \frac{1}{\|\nabla f(x)\|}$ since, again, we require $\|v\| = 1$. Now, for $c = \frac{1}{\|\nabla f(x)\|}$ we obtain $D_v f(x) = \|\nabla f(x)\|$, and for $c = -\frac{1}{\|\nabla f(x)\|}$ we obtain $D_v f(x) = -\|\nabla f(x)\|$.

□

8.3.4 Higher order partial derivatives

Clearly, and as in the univariate case, we sometimes want to differentiate a function more than once, which leads to the theory of higher order partial derivatives. Again, as in the univariate case, this can be done by iterating the differentiation procedure. However, since there are more coordinates than one, it seems that we need to be careful in which order we calculate the derivative. Luckily, this is not the case if the functions under consideration are 'nice enough', in which case we have

$$\frac{\partial}{\partial x_i} \frac{\partial f}{\partial x_j} = \frac{\partial}{\partial x_j} \frac{\partial f}{\partial x_i}.$$

That is interchanging the order of differentiation does not change the partial derivative. In particular, we will see that this is true if all involved (second-order) partial derivatives are continuous.

Definition 8.49. Let $G \subset \mathbb{R}^d$ be open and $f: G \rightarrow \mathbb{R}$ be partially differentiable at $x \in G$. If for any $i, j \in \{1, 2, \dots, d\}$ we have that the **second-order partial derivatives**

$$\frac{\partial^2 f}{\partial x_j \partial x_i}(x) := \frac{\partial}{\partial x_j} \frac{\partial f}{\partial x_i}(x)$$

exist, then we call f **twice partially differentiable** at x .

If all first-order partial derivatives are totally differentiable (at x), then we call f **twice differentiable** (at x).

If all second-order partial derivatives are continuous (at x), then we call f **twice continuously differentiable** (at x).

(Make sure that you understand the difference between these definitions.)

Example 8.50. We want to compute all second-order partial derivatives of

$$f(x_1, x_2) = x_1^2 x_2^2 + x_2 - x_1.$$

We observe that

$$\frac{\partial f}{\partial x_1} = 2x_1 x_2^2 - 1 \quad \text{and} \quad \frac{\partial f}{\partial x_2} = 2x_1^2 x_2 + 1.$$

Differentiating $\frac{\partial f}{\partial x_1}$ once more w.r.t. x_1, x_2 , we see

$$\frac{\partial^2 f}{\partial x_1^2} = \frac{\partial}{\partial x_1} \frac{\partial f}{\partial x_1} = \frac{\partial}{\partial x_1} (2x_1 x_2^2 - 1) = 2x_2^2.$$

and

$$\frac{\partial^2 f}{\partial x_2 \partial x_1} = \frac{\partial}{\partial x_2} (2x_1 x_2^2 + 1) = 4x_1 x_2.$$

Analogously we compute

$$\frac{\partial^2 f}{\partial x_1 \partial x_2} = 4x_1 x_2 \quad \text{and} \quad \frac{\partial^2 f}{\partial x_2^2} = 2x_1^2.$$

This example shows that we have to systematically compute and write down second-order partial derivatives, especially for large d . However, since we have to derive w.r.t. x_i and x_j for $i, j \in \{1, 2, \dots, n\}$ we can use a matrix to collect all these functions.

Definition 8.51 (Hessian). Let $G \subset \mathbb{R}^d$ and $f: G \rightarrow \mathbb{R}$ be twice partially differentiable at $x \in G$. The matrix

$$H_f(x) = \left(\frac{\partial^2 f}{\partial x_i \partial x_j}(x) \right)_{i,j=1}^d = \begin{pmatrix} \frac{\partial^2 f}{\partial x_1^2}(x) & \frac{\partial^2 f}{\partial x_1 \partial x_2}(x) & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_d}(x) \\ \frac{\partial^2 f}{\partial x_2 \partial x_1}(x) & \frac{\partial^2 f}{\partial x_2^2}(x) & \cdots & \frac{\partial^2 f}{\partial x_2 \partial x_d}(x) \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_d \partial x_1}(x) & \frac{\partial^2 f}{\partial x_d \partial x_2}(x) & \cdots & \frac{\partial^2 f}{\partial x_d^2}(x) \end{pmatrix}$$

is called the **Hessian (matrix)** (german: Hesse-Matrix) of f at the point x .

Example 8.52. The Hessian of the function from Example 8.50 above, which was given by $f(x) = x_1^2 x_2^2 + x_2 - x_1$ with $x = (x_1, x_2)$, is

$$H_f(x) = \begin{pmatrix} 2x_2^2 & 4x_1 x_2 \\ 4x_1 x_2 & 2x_1^2 \end{pmatrix}.$$

In this example, the Hessian is a symmetric matrix, which raises the question for which functions we have $\frac{\partial^2}{\partial x_i \partial x_j} f = \frac{\partial^2}{\partial x_j \partial x_i} f$. We will see soon that this is guaranteed under rather weak assumptions.

However, we first show that the Hessian is related to **second-order directional derivatives**, which is differentiating twice into given directions. That is, for two directions $u, v \in \mathbb{S}^{d-1}$, we compute the directional derivative w.r.t. u of the directional derivative $D_v f$. This is similar to Theorem 8.45, where we showed that the gradient is connected to the (first-order) directional derivatives.

Theorem 8.53. Let $G \subset \mathbb{R}^d$, $f: G \rightarrow \mathbb{R}$ be twice differentiable at $x \in G$ and $u, v \in \mathbb{S}^{d-1}$. Then, the second-order directional derivative at x w.r.t. u and v can be computed as

$$D_u(D_v f)(x) = u^T H_f v = \sum_{i=1}^d \sum_{j=1}^d \frac{\partial^2 f}{\partial x_i \partial x_j}(x) \cdot u_i v_j,$$

where $H_f := H_f(x)$ is the Hessian of f at x .

In particular, $D_v^2 f(x) = v^T H_f v$.

Proof. Since f is twice differentiable, which implies that all first-order partial derivatives are (totally) differentiable, we obtain that also $D_v f(x) = \sum_{i=1}^d \frac{\partial f}{\partial x_i}(x) \cdot v_i$ is differentiable, because it is a sum of differentiable functions. This implies that we can use the gradient of $D_v f$ to compute the directional derivatives $D_v f$, see Theorem 8.45. That is,

$$D_u(D_v f)(x) = \sum_{j=1}^d \frac{\partial(D_v f)}{\partial x_j}(x) \cdot u_j$$

We therefore obtain with the Hessian $H_f := H_f(x)$ that

$$\begin{aligned} u^T H_f v &= v^T \begin{pmatrix} \frac{\partial}{\partial x_1} \frac{\partial f}{\partial x_1}(x) & \frac{\partial}{\partial x_1} \frac{\partial f}{\partial x_2}(x) & \dots & \frac{\partial}{\partial x_1} \frac{\partial f}{\partial x_d}(x) \\ \frac{\partial}{\partial x_2} \frac{\partial f}{\partial x_1}(x) & \frac{\partial}{\partial x_2} \frac{\partial f}{\partial x_2}(x) & \dots & \frac{\partial}{\partial x_2} \frac{\partial f}{\partial x_d}(x) \\ \vdots & & & \vdots \\ \frac{\partial}{\partial x_d} \frac{\partial f}{\partial x_1}(x) & \frac{\partial}{\partial x_d} \frac{\partial f}{\partial x_2}(x) & \dots & \frac{\partial}{\partial x_d} \frac{\partial f}{\partial x_d}(x) \end{pmatrix} v \\ &= u^T \begin{pmatrix} \sum_{i=1}^d \frac{\partial}{\partial x_1} \frac{\partial f}{\partial x_i}(x) v_i \\ \sum_{i=1}^d \frac{\partial}{\partial x_2} \frac{\partial f}{\partial x_i}(x) v_i \\ \vdots \\ \sum_{i=1}^d \frac{\partial}{\partial x_d} \frac{\partial f}{\partial x_i}(x) v_i \end{pmatrix} = u^T \begin{pmatrix} \frac{\partial}{\partial x_1} \sum_{i=1}^d \frac{\partial f}{\partial x_i}(x) v_i \\ \frac{\partial}{\partial x_2} \sum_{i=1}^d \frac{\partial f}{\partial x_i}(x) v_i \\ \vdots \\ \frac{\partial}{\partial x_d} \sum_{i=1}^d \frac{\partial f}{\partial x_i}(x) v_i \end{pmatrix} \\ &= u^T \begin{pmatrix} \frac{\partial}{\partial x_1} D_v f(x) \\ \frac{\partial}{\partial x_2} D_v f(x) \\ \vdots \\ \frac{\partial}{\partial x_d} D_v f(x) \end{pmatrix} = \sum_{j=1}^d \frac{\partial}{\partial x_j} D_v f(x) \cdot u_j \\ &= D_u(D_v f)(x). \end{aligned}$$

□

The next theorem due to *Hermann Schwarz* (1843–1921) from 1873, which has a rather long history with several earlier incomplete proof attempts, shows that the Hessian is symmetric (i.e., one can interchange partial derivatives), whenever f is twice continuously differentiable.

Theorem 8.54 (Schwarz's theorem). Let $G \subset \mathbb{R}^d$ and $f: G \rightarrow \mathbb{R}$ be twice continuously differentiable at $x \in G$. Then, for any $i, j \in \{1, 2, \dots, d\}$, it holds that

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

In particular, this shows that the Hessian of f is a symmetric matrix, i.e. $H_f(x) = (H_f(x))^T$.

Proof. First note that we can prove the statement individually for every pair $(i, j) \in \{1, \dots, d\}^2$ and we can treat the other components as constants. Therefore, it is sufficient to prove the statement for the case that $d = 2$, $i = 1$, and $j = 2$. This means we consider a function $f: \mathbb{R}^2 \rightarrow \mathbb{R}$. Additionally we assume that w.l.o.g. $x = 0 \in G$, which will save a lot of notation. (If $x \neq 0$, then we consider the function $g(\cdot) = f(\cdot + x)$.) We will use the notation f_1 and f_2 to denote the partial derivatives w.r.t. x_1 or x_2 , respectively.

Observe that by definition of partial derivatives, we have

$$\begin{aligned}\frac{\partial^2 f}{\partial x_1 \partial x_2}(0, 0) &= \frac{\partial f_2}{\partial x_1}(0, 0) = \lim_{k \rightarrow 0} \frac{f_2(k, 0) - f_2(0, 0)}{k} \\ &= \lim_{k \rightarrow 0} \lim_{h \rightarrow 0} \frac{f(k, h) - f(k, 0) - f(0, h) + f(0, 0)}{kh}.\end{aligned}$$

Now we fix small enough $k, h \neq 0$, such that $(k, h) \in G$, and have a look at the univariate function $g(t) = f(t, h) - f(t, 0)$. An application of the mean value theorem of differential calculus, see Theorem 5.34, yields that there exists some $\xi \in [0, k]$ such that

$$g(k) - g(0) = kg'(\xi) = k(f_1(\xi, h) - f_1(\xi, 0)).$$

(Here we use that g is continuous and differentiable on the interval $(0, k)$.)

Another application of the mean value theorem shows that there exists some $\eta \in [0, h]$ such that

$$f_1(\xi, h) - f_1(\xi, 0) = h \frac{\partial f_1}{\partial x_2}(\xi, \eta) = h \frac{\partial^2 f}{\partial x_2 \partial x_1}(\xi, \eta).$$

Plugging in for $g(k) - g(0)$ we see that

$$f(k, h) - f(0, k) - f(h, 0) + f(0, 0) = kh \left(\frac{\partial^2 f}{\partial x_2 \partial x_1}(\xi, \eta) \right).$$

If $k, h \rightarrow 0$ then $\xi, \eta \rightarrow 0$ and due to continuity of the second-order derivatives it follows that

$$\begin{aligned}\frac{\partial^2 f}{\partial x_1 \partial x_2}(0, 0) &= \lim_{k \rightarrow 0} \lim_{h \rightarrow 0} \frac{f(k, h) - f(0, k) - f(h, 0) + f(0, 0)}{kh} \\ &= \lim_{\xi \rightarrow 0} \lim_{\eta \rightarrow 0} \frac{\partial^2 f}{\partial x_2 \partial x_1}(\xi, \eta) \\ &= \frac{\partial^2 f}{\partial x_2 \partial x_1}(0, 0).\end{aligned}$$

□

Remark 8.55. The continuity of the second-order partial derivatives is a necessary condition for the above result to hold for all such functions. If all second-order partial derivatives exist, but are not necessarily continuous, then there are examples where the Hessian is symmetric or not, depending on the point x .

Example 8.56. Let us have a look at the function

$$f(x_1, x_2) = \sin(x_1) \cos(x_2).$$

We see that the Hessian is given by

$$H(x) = (-1) \begin{pmatrix} \sin(x_1) \cos(x_2) & \cos(x_1) \sin(x_2) \\ \cos(x_1) \sin(x_2) & \sin(x_1) \cos(x_2) \end{pmatrix}.$$

Note that, since all second-order partial derivatives are continuous (which should be known also before their computation), we don't need to compute $\frac{\partial^2 f}{\partial x_1 \partial x_2}$ and $\frac{\partial^2 f}{\partial x_2 \partial x_1}$ separately. They are just equal, so computation of one of them is enough to write down the Hessian.

Analogously, one can also define partial derivatives of arbitrary order.

Definition 8.57. Let $G \subset \mathbb{R}^d$ and $f: G \rightarrow \mathbb{R}$ be k -times partially differentiable (at $x \in G$). If all partial derivatives of each k -th order partial derivative exist (at $x \in G$), then we say that f is **$(k+1)$ -times partially differentiable** (at $x \in G$). We use the notation

$$D_{i_{k+1}} \dots D_{i_2} D_{i_1} f(x) := \frac{\partial}{\partial x_{i_{k+1}}} \dots \frac{\partial}{\partial x_{i_2}} \frac{\partial f}{\partial x_{i_1}}(x)$$

for $i_j \in 1, \dots, d$ with $j = 1, \dots, k+1$.

If all k -th order partial derivatives are totally differentiable (at x), then we call f **$(k+1)$ -times differentiable** (at x).

If all $(k+1)$ -st order partial derivatives are continuous (at x), then we call f **$(k+1)$ -times continuously differentiable** (at x).

The Schwarz theorem also holds in this case.

Theorem 8.58. Let $G \subset \mathbb{R}^d$ and $f: G \rightarrow \mathbb{R}$ be k -times continuously differentiable. Then for any $i_1, i_2, \dots, i_k \in \{1, 2, \dots, d\}$ and any permutation σ of $\{1, \dots, k\}$ we have

$$D_{i_{\sigma(k)}} \dots D_{i_{\sigma(2)}} D_{i_{\sigma(1)}} f(x) = D_{i_k} \dots D_{i_2} D_{i_1} f(x).$$

Proof. This follows by inductively applying the Schwarz theorem, see Theorem 8.54. □

8.4 Extrema

As in the one dimensional case, one can use differential calculus to study *optimization problems*, i.e., to find extrema. Although some of the techniques used here are (or at least look) more complicated than the corresponding parts of Section 5, the overall strategy is the same:

1. We use the derivative to find *candidates* for (local) extrema, i.e., stationary points,
2. if possible, we use the second derivatives to check if (local) maximum or minimum.
3. Finally, we consider the *boundary* of the considered domain separately.

As we learned in Section 5 in the univariate case, stationary points are the candidates for local extrema, if they are in the domain of the function. But also functions without stationary points may have extreme points. (Consider the easy example of a linear function on a closed interval.) For this we always needed to compute the function values at the boundary points and/or consider the limit to $\pm\infty$ to verify if a function has global/local extrema.

Unfortunately, all these objects are more difficult to handle than in the univariate case. First of all, there are several (partial) derivatives in a given point, in contrast to the univariate case, and so we need to generalize the previous concepts. We will see that, here, the gradient plays the role of the derivative and the Hessian matrix will substitute the second derivative.

An additional difficulty comes from multivariate domains. While the boundary of a bounded interval, which was the typical domain for univariate function, consists only of two points, the boundary in the multivariate setting is more complex and needs more care. We will see an example in a second.

However, let us first recall the definition of global extrema from Definition 4.54.

Definition 8.59. Let D be any set and $f: D \rightarrow \mathbb{R}$. Then, f has a **(global) minimum at** $x_0 \in D$ if

$$f(x) \geq f(x_0) \quad \text{for all } x \in D,$$

and f has a **(global) maximum at** $x_0 \in D$ if

$$f(x) \leq f(x_0) \quad \text{for all } x \in D.$$

The point x_0 is called **(global) minimum/maximum point**, or **global extreme point**.

The value $f(x_0)$ is called **minimum/maximum**, or, collectively, **extreme value** of f .

Note that this is word by word the same definition; it works for arbitrary domains.

Also the definition of local extrema comes only with minimal modifications, see Definition 5.24.

Definition 8.60. Let $\Omega \subset \mathbb{R}^d$ and $f: \Omega \rightarrow \mathbb{R}$.

Then, f has a **local minimum at** $x_0 \in \Omega$ if there exists $\varepsilon > 0$ such that

$$f(x) \geq f(x_0) \quad \text{for all } x \in U_\varepsilon(x_0) \cap \Omega,$$

and a **strict local minimum** if $f(x) > f(x_0)$ for all $x \in U_\varepsilon(x_0) \cap \Omega \setminus \{x_0\}$.

Analogously, we say f has a **local maximum at** $x_0 \in \Omega$ if there exists $\varepsilon > 0$ such that

$$f(x) \leq f(x_0) \quad \text{for all } x \in U_\varepsilon(x_0) \cap \Omega,$$

and a **strict local maximum** if $f(x) < f(x_0)$ for all $x \in U_\varepsilon(x_0) \cap \Omega \setminus \{x_0\}$.

The point $x_0 \in \Omega$ is called **local maximum/minimum point**, or **local extreme point**.

We use Ω as a domain here, to make clear that this set does not need to be an open set. (Note that we assumed that G is always open.)

Before we turn to the actual computation of extrema and extreme values, let us first consider the question **whether a function has an extremum or not**. For this, recall from extreme value theorem (Theorem 4.56) that every univariate function $f: [a, b] \rightarrow \mathbb{R}$, defined on a closed interval, attains its minimum and maximum.

Theorem 8.61 (Extreme value theorem). *Let $C \subset \mathbb{R}^d$ be a closed and bounded set and $f: C \rightarrow \mathbb{R}$ be a continuous function. Then there exist $x_{\min}, x_{\max} \in C$ such that*

$$\begin{aligned} f(x_{\min}) &= \inf_{x \in C} f(x) := \inf \{f(x): x \in C\}, \\ f(x_{\max}) &= \sup_{x \in C} f(x) := \sup \{f(x): x \in C\}. \end{aligned}$$

In other words, continuous functions attain their extreme values on closed and bounded sets.

Proof. Coming soon... □

We start with an “easy” example.

Example 8.62. Consider

$$f(x) := e^{-(x_1^2 + 2x_2^2)}$$

with $x = (x_1, x_2)$ on the set $\Omega := \{x: \|x\| \leq 1\} = \{(x_1, x_2): x_1^2 + x_2^2 \leq 1\}$.

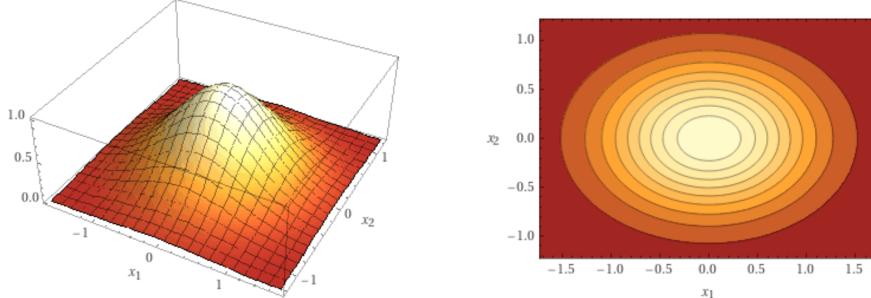


Figure 50: 3d and contour plot of $f(x_1, x_2) = \exp(-(x_1^2 + 2x_2^2))$.

From what we know about exponentials, we see that $f(x) \leq 1$ for all $x \in \mathbb{R}^2$, and that $f(x) = 1$ if and only if $x = (0, 0)$. Therefore, $f(0) > f(x)$ for every $x \neq 0$, which implies that $x_0 = (0, 0)$ is a strict local maximum as well as the global maximum of f .

When looking for a minimum we first realize that $f(x_1, x_2)$ converges to zero, when x_1 and/or x_2 tend to infinity, but $f > 0$. Therefore, if we would consider f on \mathbb{R}^2 , then f would not have a minimum. (One would say its *infimum* is 0.) However, we consider extreme values on Ω and, since we have $|x_1| \leq 1$ and $|x_2| \leq 1$ for all $x = (x_1, x_2) \in \Omega$, it is obvious that

$$f(x) = e^{-(x_1^2 + 2x_2^2)} \geq e^{-(1+2)} = e^{-3} \geq 0.049 \quad \forall x \in \Omega.$$

Moreover, due to monotonicity, we see that the minimum is on the *boundary* $\{x: \|x\| = 1\}$.

It might already be seen (from the contour plot above) that f decreases fastest in x_2 -direction, and therefore that the global/local minima are at $x_0 = (0, \pm 1)$ such that $f(x_0) = e^{-2}$. However, it remains to find a way to verify this in a systematic way. \square

This example shows that even in simple examples, it might be non-trivial to determine all extrema when we are working on bounded sets. Therefore, **we first consider here functions that are defined on the whole \mathbb{R}^d** , i.e., we consider extrema of functions $f: \mathbb{R}^d \rightarrow \mathbb{R}$. In the context of optimization, this is sometimes called *free* (or *unconstrained*) *optimization*. We will come back to extrema on subsets $\Omega \subset \mathbb{R}^d$, which corresponds to *constrained optimization*, afterwards.

Remark 8.63. It might be interesting to note here that the set \mathbb{R}^d is open and closed at the same time, see Definition 8.19. This follows from the fact that the empty set \emptyset is, by definition, open. (As there is no $x \in \emptyset$, the requirement to be open set is trivially true.) Therefore, all the results from the previous subsections, which were stated for open sets, are valid for $G = \mathbb{R}^d$.

Before we turn to the generalization of the concepts from Section 5, let us discuss another simple example that shows that there might be infinitely many (local) extrema, but none of them is a *strict* local extremum.

Example 8.64. Consider $f: \mathbb{R}^2 \rightarrow \mathbb{R}$ with $f(x_1, x_2) := \sin(x_1 + x_2)$.

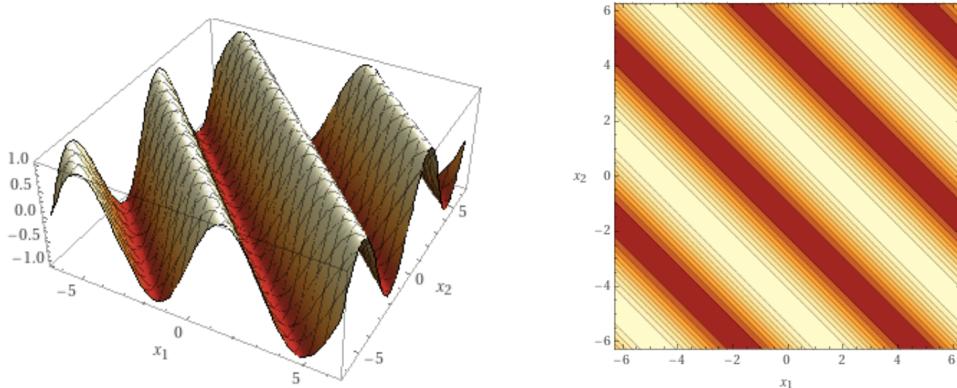


Figure 51: 3d and contour plot of $f(x_1, x_2) = \sin(x_1 + x_2)$.

We already know that $\sin(t)$, $t \in \mathbb{R}$, is maximal whenever $t = \frac{\pi}{2} + 2k\pi$ and minimal when $t = \frac{3\pi}{2} + 2k\pi$, where $k \in \mathbb{Z}$. So all possible maxima $x_0 = (x_1, x_2)$ have to satisfy $x_1 + x_2 = \frac{\pi}{2} + 2k\pi$ for some $k \in \mathbb{Z}$, and for such a point we have $f(x_0) = 1$. This implies that $f(x) \leq f(x_0)$ for all $x \in \mathbb{R}^2$, i.e., such x_0 are (local) maxima. (Recall that $\sin(t) \in [-1, 1]$ for all $t \in \mathbb{R}$.)

However, f does not have any strict local maxima. To see this, let $x_0 = (x_1, x_2)$ be a local maximum and $y = (x_1 + \delta, x_2 - \delta)$ with $\delta > 0$ be another point. We obtain

$$f(y) = \sin(x_1 + \delta + x_2 - \delta) = \sin(x_1 + x_2) = f(x_0) = 1.$$

Hence, y is also a maximum and, since y can be arbitrary close to x_0 for δ small enough, we obtain that x_0 cannot be a strict local maximum. The same arguments work for minima of f . □

As for $d = 1$, we need criteria to check if some $x \in \mathbb{R}^d$ is a minimum/maximum of a function f . Recall that in the univariate case we saw that if the slope of a function was different from 0 at some point then it was not possible for this point to be a minimum/maximum, see Theorem 5.25. We will now show a very similar result, which gives a **necessary condition for being a (local) maximum/minimum**.

Theorem 8.65 (Necessary condition for an extreme point). *Let $G \subset \mathbb{R}^d$ be open and $f: G \rightarrow \mathbb{R}$ be partially differentiable. If $x_0 \in G$ is a local extremum, then*

$$\nabla f(x_0) = 0.$$

*A point $x_0 \in G$ such that $\nabla f(x_0) = 0$ is called **stationary point** of f .*

This means that, if $\nabla f(x) \neq 0$ for some point x , then this point cannot be an extremum.

Proof. We show that the statement is true if x_0 is a local maximum. The same arguments can be used if x_0 is a local minimum. For this, we consider each entry of the gradient separately and use the known results from the univariate setting. Fix $i \in \{1, \dots, d\}$ and let $U = U_\varepsilon(x_0)$ be a neighborhood of x_0 such that $f(x_0) \geq f(x)$ for all $x \in U$.

Since $x_0 + te_i \in U$ for $t \in (-\varepsilon, \varepsilon)$, we have by assumption that the function $g: (-\varepsilon, \varepsilon) \rightarrow \mathbb{R}$,

$$g(t) := f(x + te_i)$$

exist and is differentiable. (We use that the i -th partial derivative of f exists.) Moreover, again by $x_0 + te_i \in U$, we see that

$$g(0) = f(x_0) \geq f(x_0 + te_i) = g(t)$$

for any $t \in (-\varepsilon, \varepsilon)$. Thus, g has a local maximum at 0 implying that

$$g'(0) = 0,$$

see Theorem 5.25. Since $g'(0) = \frac{\partial f}{\partial x_i}(x_0)$ and the above holds for all $1 \leq i \leq d$ the result follows. \square

Let us see how we can use this result to determine local extrema.

Example 8.66. We consider the function $f(x) = e^{-x_1^2 - 2x_2^2}$, $x = (x_1, x_2)$, from Example 8.62. Computing the gradient we see that

$$\nabla f(x) = \left(-2x_1 e^{-x_1^2 - 2x_2^2}, -4x_2 e^{-x_1^2 - 2x_2^2} \right) = -2f(x) \cdot (x_1, 2x_2).$$

Since $f(x) = e^{-x_1^2 - 2x_2^2} \neq 0$ for any $x \in \mathbb{R}^2$, we see that

$$\nabla f(x) = 0 \iff x = 0.$$

Hence, the only possible local extremum of f is at $x_0 = 0$. Although we already know for this function that $x_0 = 0$ is a maximum, we still need a systematic way for general functions to verify if a stationary point is indeed a maximum or a minimum. \square

We already mentioned that Theorem 8.65 is only a necessary condition for x_0 to be an extremum. Unfortunately, this is not a sufficient condition as the next example shows.

Example 8.67. We want to determine all extrema of $f: \mathbb{R}^2 \rightarrow \mathbb{R}$,

$$f(x_1, x_2) = x_1^2 \cdot x_2.$$

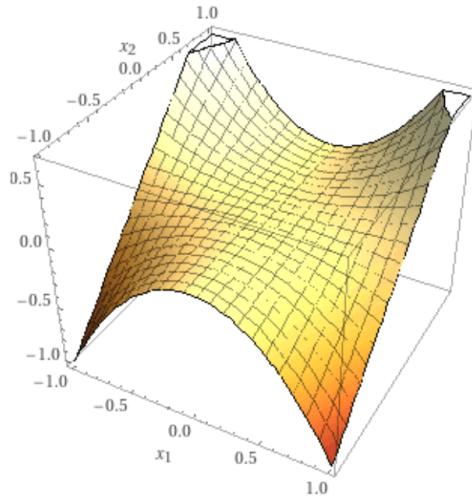


Figure 52: Plot of $f(x_1, x_2) = x_1^2 x_2$.

To do so we compute

$$\nabla f(x) = (2x_1 x_2, x_1^2),$$

implying that $\nabla f(x) = 0 \iff x_1 = 0$. Note that x_2 can be chosen arbitrary. Therefore, all points in the set $\{(0, x_2) : x_2 \in \mathbb{R}\}$ are stationary points and they are the only candidates for local extrema. However, we have $f(0, x_2) = 0$ for every x_2 and therefore, as indicated by the plot, none of these points is a global extremum as the function attains smaller and larger values than 0. Regarding local extrema, note that $f(x_1, x_2) \geq 0$ whenever $x_2 \geq 0$ and $f(x_1, x_2) \leq 0$ whenever $x_2 \leq 0$. Therefore, if $x_2 > 0$, then there is a neighborhood $U_\varepsilon(x_0)$ around $x_0 = (0, x_2)$ such that $f(x) \geq 0 = f(x_0)$ in $U_\varepsilon(x_0)$. (One can choose

$\varepsilon = x_2$.) This shows that x_0 is a local minimum. In the same way, we obtain that every $x_0 = (0, x_2)$ with $x_2 < 0$ is a local maximum.

It remains to check the point $x_0 = (0, 0)$. For this not that, for every $\varepsilon > 0$, we have $f(\varepsilon, \varepsilon) > 0$ and $f(\varepsilon, -\varepsilon) < 0$. Therefore, every neighborhood of $(0, 0)$ contains points with smaller and larger function values, which shows that $(0, 0)$ is not a (local) extremum, although $\nabla f(0, 0) = 0$.

□

We now turn to a method, based on derivatives, to verify if a function has a maximum or a minimum, or no extremum at all. This method is, similarly to the univariate case, called the **second partial derivative test**, see Theorem 5.29. Recall that in the univariate case we used positivity/negativity of the second derivative to decide if a stationary point is a minimum/maximum. However, since the second derivative of a multivariate function is represented by a matrix, i.e., the Hessian matrix, we first need a notion of *positivity* of a matrix.

Definition 8.68. Let $A \in \mathbb{R}^{d \times d}$ be a symmetric matrix. We call A ...

- **positive definite** if

$$v^T A v > 0 \quad \text{for all } v \in \mathbb{R}^d \setminus \{0\}.$$

- **positive semi-definite** if

$$v^T A v \geq 0 \quad \text{for all } v \in \mathbb{R}^d.$$

- **negative definite** if

$$v^T A v < 0 \quad \text{for all } v \in \mathbb{R}^d \setminus \{0\}.$$

- **negative semi-definite** if

$$v^T A v \leq 0 \quad \text{for all } v \in \mathbb{R}^d.$$

If A is neither positive semi-definite nor negative semi-definite, then A is called **indefinite**.

Remark 8.69. Clearly, A is positive definite if and only if $-A$ is negative definite.

Remark 8.70. One time-saving argument to verify that a matrix is indefinite is that it has entries on the **diagonal with different signs**. To see this, one just needs to consider the unit vectors e_i . Using that $e_i^T A e_i = A_{ii}$, i.e., the i -th diagonal entry of A , we see that different signs on the diagonal of A imply that there are i, j such that $e_i^T A e_i < 0$ and $e_j^T A e_j > 0$, which makes the matrix indefinite. However, the converse is not true: A matrix with all entries of the same sign is not always definite.

Example 8.71. Consider the matrix $A = \begin{pmatrix} 1 & 3 \\ 3 & 4 \end{pmatrix}$. We see that

$$v^T A v = (v_1, v_2) \begin{pmatrix} v_1 + 3v_2 \\ 3v_1 + 4v_2 \end{pmatrix} = v_1^2 + 4v_2^2 + 6v_1v_2.$$

This expression is positive, e.g., for $v = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$, but negative, e.g., for $v = \begin{pmatrix} -2 \\ 1 \end{pmatrix}$. Hence, A is indefinite.

Example 8.72. Consider the matrix $A = \begin{pmatrix} 2 & 1 \\ 1 & 1 \end{pmatrix}$. We see that

$$v^T A v = (v_1, v_2) \begin{pmatrix} 2v_1 + v_2 \\ v_1 + v_2 \end{pmatrix} = 2v_1^2 + v_2^2 + 2v_1 v_2.$$

This expression is positive for all $v \in \mathbb{R}^d \setminus \{0\}$, and hence the matrix is positive definite. However, it is not clear how to verify this precisely. We now introduce a direct method. (Still, try to verify it directly.)

Determining if a matrix is pos./neg. definite can be quite time-consuming, and is in most cases not straight-forward. We therefore present an easy method based on the determinant of a matrix, see Section 2.4. This is also the fastest method available, at least for small matrices. Note that this method does not allow for determining semi- or indefiniteness.

Lemma 8.73 (Sylvester's criterion). *Let $A = (a_{ij})_{i,j=1}^d \in \mathbb{R}^{d \times d}$ be a symmetric matrix, and let $A_k = (a_{ij})_{i,j=1}^k \in \mathbb{R}^{k \times k}$ be the (upper left) submatrices of A . Then, A is ...*

- positive definite if and only if $\det(A_k) > 0$ for all $k = 1, \dots, d$.
- negative definite if and only if $\det(A_k) > 0$ for even k and $\det(A_k) < 0$ for odd k .

A proof of this result is out of reach with our present knowledge and can be found in the literature.

Example 8.74. The matrix

$$A = \begin{pmatrix} 2 & 1 \\ 1 & 1 \end{pmatrix}$$

from Example 8.72 is positive definite. Using Sylvester's method, with $A_1 = 2$ and $A_2 = A$, we see that $\det(A_1) = 2 > 0$ and $\det(A_2) = 1 > 0$.

Example 8.75. Consider the matrices

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

Since, $\det(A_1) = 1$, $\det(A_2) = \det(A) = -5$ and $\det(B_1) = 0$, we observe from Sylvester's criterion that A and B both cannot be positive or negative definite.

Indeed, we know from Example 8.71 that A is indefinite. For B , we see that $v^T B v = (v_1, v_2) \begin{pmatrix} v_2 \\ v_1 \end{pmatrix} = 2v_1 v_2$ is positive for $v = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$, and negative for $v = \begin{pmatrix} -1 \\ 1 \end{pmatrix}$, proving that B is indefinite.

Example 8.76. Consider the (symmetric) matrices

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

First, note that A has entries on the diagonal with different signs, and is therefore indefinite, see Remark 8.70. (As $\det(A_1) = 1$ and $\det(A_2) = -2$, also Sylvester's criterion is inconclusive.)

For B , we observe that $\det(B_1) = -1 < 0$, $\det(B_2) = 1 > 0$ and $\det(B_3) = \det(B) = -1 < 0$. By Sylvester's criterion, that A and B both cannot be positive- or negative-definite.

(A proof without Sylvester's criterion would be a mess!)

Based on the notion of definiteness, we can finally give a **sufficient condition for being an extremum**.

Theorem 8.77 (Second (partial) derivative test). *Let $f: G \rightarrow \mathbb{R}$ be twice continuously differentiable and let $x_0 \in G$ such that $\nabla f(x_0) = 0$. Then, we have*

- 1) $H_f(x_0)$ is positive-definite $\implies x_0$ is a strict local minimum
- 2) $H_f(x_0)$ is negative-definite $\implies x_0$ is a strict local maximum
- 3) $H_f(x_0)$ is indefinite $\implies x_0$ is not an extremum of f

In all other cases (i.e., semi-definite but not definite), we do not gain information from the second derivative test.

Remark 8.78. We actually show below that points $x_0 \in G$ such that $\nabla f(x_0) = 0$ and $H_f(x_0)$ indefinite are actually points such that for every $\varepsilon > 0$ there exist $x, y \in U_\varepsilon(x_0)$ with $f(x) > f(x_0)$ and $f(y) < f(x_0)$. That is, in every neighborhood there exist strictly smaller and larger function values. Such points, which are clearly no extrema, are usually called **saddle points**.

Proof of Theorem 8.77. We consider the first case, i.e., that $H_f(x_0)$ is positive-definite. First, note that there is some $\alpha > 0$ such that $v^T H_f(x_0)v \geq \alpha$ for all $v \in \mathbb{S}^{d-1}$, i.e., all v with $v^T v = 1$. (This can be proven by using that $v^T H_f(x_0)v$ must attain its minimum on \mathbb{S}^{d-1} .) Now note that, by continuity of all second-order partial derivatives, there is some $\varepsilon > 0$, such that

$$\left| \frac{\partial^2 f}{\partial x_i \partial x_j}(x) - \frac{\partial^2 f}{\partial x_i \partial x_j}(x_0) \right| < \frac{\alpha}{2d}$$

for all $x \in U_\varepsilon(x_0)$ and all $i, j = 1 \dots, d$. That is, $\frac{\partial^2 f}{\partial x_i \partial x_j}(x)$ is 'close' to $\frac{\partial^2 f}{\partial x_i \partial x_j}(x_0)$ in a small neighborhood around x_0 . From this, we obtain that

$$\begin{aligned} \left| v^T (H_f(x) - H_f(x_0))v \right| &= \left| \sum_{i=1}^d \sum_{j=1}^d \left(\frac{\partial^2 f}{\partial x_i \partial x_j}(x) - \frac{\partial^2 f}{\partial x_i \partial x_j}(x_0) \right) \cdot v_i v_j \right| \\ &\leq \sum_{i=1}^d \sum_{j=1}^d \left| \frac{\partial^2 f}{\partial x_i \partial x_j}(x) - \frac{\partial^2 f}{\partial x_i \partial x_j}(x_0) \right| \cdot |v_i| |v_j| \\ &< \frac{\alpha}{2d} \left(\sum_{i=1}^d |v_i| \right) \left(\sum_{j=1}^d |v_j| \right) \\ &\leq \frac{\alpha}{2} \|v\|^2 = \frac{\alpha}{2}, \end{aligned}$$

see Lemma 8.8. This implies

$$\begin{aligned} v^T H_f(x)v &= v^T H_f(x_0)v + v^T (H_f(x) - H_f(x_0))v \\ &\geq v^T H_f(x_0)v - \left| v^T (H_f(x) - H_f(x_0))v \right| \\ &> \alpha - \frac{\alpha}{2} = \frac{\alpha}{2} > 0 \end{aligned}$$

for all $x \in U_\varepsilon(x_0)$ and $v \in \mathbb{S}^{d-1}$. That is, $H_f(x)$ is also positive-definite in a neighborhood of x .

We now fix some $v \in \mathbb{S}^{d-1}$ and consider the univariate function $g(t) := f(x_0 + tv) - f(x_0)$. We see that $g(0) = 0$ and $g'(0) = D_v f(x_0) = \langle \nabla f(x_0), v \rangle = 0$, by assumption. Taylor's theorem (Theorem 5.53) now shows that

$$g(t) = g(0) + g'(0) \cdot t + \frac{g''(\xi)}{2} \cdot t^2 = \frac{g''(\xi)}{2} \cdot t^2$$

for some $\xi \in (0, t)$. From Theorem 8.53, we obtain that $g''(\xi) = D_v^2 f(x_0 + \xi v) = v^T H_f(x_0 + \xi v)v$. In particular, positive-definiteness implies that $g''(\xi) > 0$ for all $\xi \in [0, \varepsilon]$, independent of v .

Since $t \in (0, \varepsilon)$ implies $\xi \in (0, \varepsilon)$, we have $g(t) > 0$, i.e., $f(x_0 + tv) > f(x_0)$, for all $t \in (0, \varepsilon)$, independent of v . In other words, $f(x) > f(x_0)$ for all $x \in U_\varepsilon(x_0) \setminus \{x_0\}$, which proves the claim.

The case of $H_f(x_0)$ negative-definite follows from considering $-f$ instead. (Note that $H_{-f}(x_0)$ is positive-definite then.)

Finally, if $H(x_0)$ is indefinite, then there exist some $u, v \in \mathbb{S}^{d-1}$ such that

$$v^T H_f v > 0 \quad \text{and} \quad u^T H_f u < 0,$$

again in a neighborhood of x_0 . Thus, the (univariate) function $g(t) = f(x_0 + tv) - f(x_0)$ is positive, and the function $h(t) = f(x_0 + tu) - f(x_0)$ is negative, for every small enough $t > 0$. In other words, every neighborhood of x_0 contains points with smaller and larger function value, respectively, which shows that f cannot have an extremum at x_0 . \square

Now we want to use the above result to calculate some extrema.

Example 8.79. We have a look at the function

$$f(x_1, x_2) = 2x_1^2 + 3x_2^3 = (x_1, x_2) \begin{pmatrix} 2 & 0 \\ 0 & 3 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}.$$

Thus

$$\nabla f(x_1, x_2) = \begin{pmatrix} 4x_1 \\ 6x_2 \end{pmatrix} \quad \text{and} \quad H_f(x_1, x_2) = \begin{pmatrix} 4 & 0 \\ 0 & 6 \end{pmatrix} = 2 \begin{pmatrix} 2 & 0 \\ 0 & 3 \end{pmatrix}.$$

The only possible candidate for a local extremum is $x_0 = (0, 0)$. Now, note that $H := H_f(0, 0)$ satisfies $v^T H v = (v_1, v_2) \begin{pmatrix} 4v_1 \\ 6v_2 \end{pmatrix} = 4v_1^2 + 6v_2^2$, which is clearly positive whenever $v = \begin{pmatrix} v_1 \\ v_2 \end{pmatrix} \neq 0$. This can also be seen by Sylvester's criterion (Lemma 8.73). Hence, $H_f(0, 0)$ is positive definite, and x_0 is a strict local minimum.

Example 8.80. The previous example can be generalized by considering functions of the form

$$f(x) = x^T C x,$$

where C is a symmetric matrix, see also Example 8.39. Moreover, we already showed there that $\nabla f(x) = 2Cx$, which implies that all possible extrema have to satisfy $Cx = 0$. Hence, we always have that $x_0 = 0$ is a stationary point and that $f(0) = 0$. Moreover, we obtain that $H_f(x) = 2C$ for any x , i.e., $H_f(x)$ is independent of x .

If C is positive-definite, then C is an invertible matrix (We prove that later.), which implies that 0 is the only stationary point. Moreover, by positive-definiteness of the Hessian $H_f = 2C$, we obtain from Theorem 8.77 that $x_0 = 0$ is a strict local minimum. (This could also be seen from $x^T C x > 0$ for every $x \neq 0$.) If C is negative-definite it follows analogously that 0 is a strict local maximum and that there are no other extrema. If C is indefinite and invertible there are no extrema at all.

Finally, if C is 'only' positive/negative semi-definite, or, more general, not invertible, then we cannot say anything just by using the second derivative test.

In the special case $d = 2$, we can employ Sylvester's criterion to obtain a very useful formulation of the second derivative test.

Corollary 8.81 (Second derivative test for $d = 2$). *Let $G \subset \mathbb{R}^2$, $f: G \rightarrow \mathbb{R}$ be twice continuously differentiable and let $x_0 \in G$ such that $\nabla f(x_0) = 0$.*

Moreover, let $H := H_f(x_0)$ be the Hessian of f at x_0 with upper left entry H_{11} . Then, we have

- 1) $\det(H) > 0$ and $H_{11} > 0 \implies x_0$ is a strict local minimum
- 2) $\det(H) > 0$ and $H_{11} < 0 \implies x_0$ is a strict local maximum
- 3) $\det(H) < 0 \implies x_0$ is not an extremum of f

If $\det(H) = 0$, we do not gain information from the second derivative test.

Again, we omit the proof. (The third point cannot be proven here.)

Example 8.82. With this we can now easily verify that

$$f(x) := e^{-(x_1^2 + 2x_2^2)}$$

with $x = (x_1, x_2)$ from Example 8.62 has a (global) maximum at $x_0 = (0, 0)$. We already know from Example 8.66 that $\nabla f(x) = -2f(x)(x_1, 2x_2)$, and that this implies that $x_0 = (0, 0)$ is the only stationary point of f . Computing the Hessian matrix of f we obtain

$$H_f(x) = f(x) \cdot \begin{pmatrix} 4x_1^2 - 2 & 8x_1x_2 \\ 8x_1x_2 & 16x_2^2 - 4 \end{pmatrix}$$

and therefore $H_f(0) = \begin{pmatrix} -2 & 0 \\ 0 & -4 \end{pmatrix}$. Since $\det(H_f(0)) = 8 > 0$ and $H_{11} = -2 < 0$ (where we use the notation from Corollary 8.81), we see that $x_0 = (0, 0)$ is a strict local maximum.

Example 8.83. Consider again the function $f: \mathbb{R}^2 \rightarrow \mathbb{R}$,

$$f(x_1, x_2) = x_1^2 \cdot x_2,$$

from Example 8.67. Since $\nabla f(x) = (2x_1x_2, x_1^2)$, we saw that all points in the set $\{(0, x_2): x_2 \in \mathbb{R}\}$ are stationary points, and we also verified which of them are extrema.

Moreover, we easily obtain the Hessian

$$H_f(x_1, x_2) = \begin{pmatrix} 2x_2 & 2x_1 \\ 2x_1 & 0 \end{pmatrix},$$

and therefore $H_f(0, x_2) = \begin{pmatrix} 2x_2 & 0 \\ 0 & 0 \end{pmatrix}$. This matrix has $\det(H_f(0, x_2)) = 0$ for every $x_2 \in \mathbb{R}$. Therefore, the second derivative test does not lead to an answer whether some of the stationary points are extrema or not.

Example 8.84. We want to compute the extrema of $f: \mathbb{R}^2 \rightarrow \mathbb{R}$ with

$$f(x_1, x_2) = \sin(x_1) \cos(x_2).$$

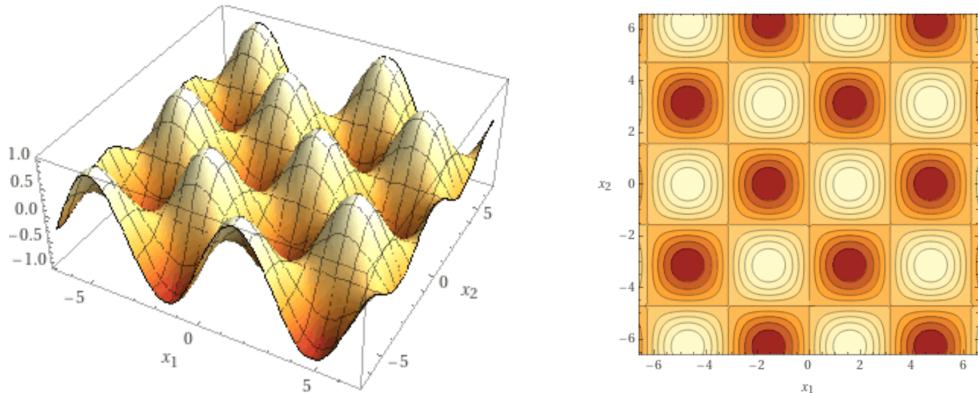


Figure 53: 3d and contour plot of $f(x_1, x_2) = \sin(x_1) \cos(x_2)$.

The gradient of f is

$$\nabla f(x) = \begin{pmatrix} \cos(x_1) \cos(x_2) \\ -\sin(x_1) \sin(x_2) \end{pmatrix}$$

and the Hessian, which we already computed in Example 8.56, is

$$H_f(x) = (-1) \begin{pmatrix} \sin(x_1) \cos(x_2) & \cos(x_1) \sin(x_2) \\ \cos(x_1) \sin(x_2) & \sin(x_1) \cos(x_2) \end{pmatrix}.$$

Let us compute all stationary points of f , i.e. all $x = (x_1, x_2)$ such that $\nabla f(x) = 0$. Since $\sin(t)$ and $\cos(t)$ cannot be zero at the same time (i.e., for the same t), we obtain that $\nabla f(x) = 0$ if either

$$\cos x_1 = 0 \quad \text{and} \quad \sin x_2 = 0$$

or

$$\sin x_1 = 0 \quad \text{and} \quad \cos x_2 = 0.$$

We start with the first case, i.e. $\cos x_1 = 0$ and $\sin x_2 = 0$, which implies that

$$x_1 = k_1\pi + \frac{\pi}{2} = \frac{(2k_1+1)\pi}{2} \quad \text{and} \quad x_2 = k_2\pi,$$

for some $k_1, k_2 \in \mathbb{Z}$. (For example, $x_1 = \frac{\pi}{2}$ and $x_2 = 0$.)

Plugging this into the Hessian, we see that for such an $x = (x_1, x_2)$ we have

$$H_f(x) = (-1) \begin{pmatrix} (-1)^{k_1+k_2} & 0 \\ 0 & (-1)^{k_1+k_2} \end{pmatrix}.$$

(Make sure that you understand the basic properties of cos/sin that lead to this.)

This shows $H_f(x) = \begin{pmatrix} -1 & 0 \\ 0 & -1 \end{pmatrix}$ if $k_1 + k_2$ is even, and that $H_f(x) = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$ if $k_1 + k_2$ is odd. Therefore,

$$x_0 = \left(k_1\pi + \frac{\pi}{2}, k_2\pi \right) \text{ is a strict local } \begin{cases} \text{maximum,} & \text{if } k_1 + k_2 \text{ is even,} \\ \text{minimum,} & \text{if } k_1 + k_2 \text{ is odd.} \end{cases}$$

(Verify that the corresponding matrices are pos./neg. definite!)

For example, the point $(\frac{25\pi}{2}, 42\pi)$ is a strict local maximum ($k_1 = 12, k_2 = 42$).

If we consider the second case, i.e., $\sin x_1 = 0$ and $\cos x_2 = 0$, we need that

$$x_1 = k_1\pi \quad \text{and} \quad x_2 = k_2\pi + \frac{\pi}{2} = \frac{(2k_2+1)\pi}{2},$$

for some $k_1, k_2 \in \mathbb{Z}$. (For example, $x_1 = 0$ and $x_2 = \frac{\pi}{2}$.) Note that the contour plot above already shows, that there cannot be an extremum at such points, because every point on these 'lines' has points with smaller and larger function value around it.

However, to prove this, plug these (x_1, x_2) into the Hessian, to obtain

$$H_f(x) = (-1) \begin{pmatrix} 0 & (-1)^{k_1+k_2} \\ (-1)^{k_1+k_2} & 0 \end{pmatrix}.$$

This shows that either $H_f(x) = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$ or $H_f(x) = \begin{pmatrix} 0 & -1 \\ -1 & 0 \end{pmatrix}$, but both matrices have an eigenvalue 1 and an eigenvalue -1 , making $H_f(x)$ indefinite. Therefore, all points $x_0 = (k_1\pi, k_2\pi + \frac{\pi}{2})$ with $k_1, k_2 \in \mathbb{Z}$ are no extrema. (They are actually *saddle points*, see Remark 8.78.)

□

8.4.1 Extrema subject to constraints

The results above allow to find extrema of functions which are defined on open sets, like $G = \mathbb{R}^d$, and to verify if these extreme are minima or maxima. However, in many applications we are interested in extrema which are contained in some given (closed) set $\Omega \subset G$. For this, we have to consider the boundary of the set separately. Recall from the univariate case that a function defined on a closed interval can have a minimum/maximum at the boundary points. For example, the function $f(t) = 2t$ on $[1, 2]$ has a minimum at 1 and a maximum at 2, but the derivative of f is nowhere zero. These boundary points are easy to check, but for multivariate functions the boundary is more complex.

In what follows, we consider only functions defined on $G = \mathbb{R}^d$ and we want to find its extrema in sets that are given by

$$\Omega = \{x \in \mathbb{R}^d : g(x) \leq c\}.$$

for some function $g: \mathbb{R}^d \rightarrow \mathbb{R}$ and $c \in \mathbb{R}$. For the sake of simplicity, we assume here that the function g is a continuously differentiable function. From this we obtain, in particular, that the **interior**

$$\Omega^o := \{x \in \mathbb{R}^d : g(x) < c\}$$

of the set Ω is an open set. We can therefore find all extrema of a function inside Ω^o by the techniques from above. It remains to consider the **boundary** of Ω , i.e.,

$$\partial\Omega := \{x \in \mathbb{R}^d : g(x) = c\}$$

Extrema of a function on the boundary $\partial\Omega$, which are defined in the same way as in Definition 8.60 with Ω replaced by $\partial\Omega$, are called **extrema subject to the constraint** $g(x) = c$.

Before we come to a more systematic way of finding the extrema of a function, let us note that the equation $g(x) = g(x_1, \dots, x_d) = c$ does sometimes lead to an 'easy' restriction of just one coordinate, which we may just plug into our original problem to find the extrema on $\partial\Omega$. That is, $g(x_1, \dots, x_d) = c$ can sometimes be written as $x_d = h(x_1, \dots, x_{d-1})$ for some function $h: \mathbb{R}^{d-1} \rightarrow \mathbb{R}$. In this case, finding an extrema of f of $\partial\Omega$ is just the same as finding an extrema of

$$F(x_1, \dots, x_{d-1}) := f\left(x_1, \dots, x_{d-1}, h(x_1, \dots, x_{d-1})\right)$$

on \mathbb{R}^{d-1} . That is, the restriction just reduces the dimension by one. Let us see an example.

Example 8.85. We want to compute the extrema of the function

$$f(x_1, x_2) = x_1^2 + x_2^2$$

in the set $\Omega = \{(x_1, x_2) \in \mathbb{R}^2 : x_1 + x_2 \geq 2\}$. In the notation from above, we have

$$\Omega = \{x \in \mathbb{R}^2 : g(x) \leq c\} \quad \text{with} \quad g(x_1, x_2) = -x_1 - x_2 \quad \text{and} \quad c = -2.$$

To find the extrema of f in Ω we first compute the stationary points of f on \mathbb{R}^2 , i.e., all points such that ∇f vanishes. We already saw that $x = (0, 0)$ is the only stationary point of f . However, $(0, 0)$ is not contained in Ω , because it satisfies $g(0, 0) = 0 \not\leq -2$, and therefore cannot be an extremum of f in Ω . Hence, there are no (local) extrema in Ω^o .

To treat the boundary, i.e., $\partial\Omega = \{x \in \mathbb{R}^2 : g(x) = c\}$, note that

$$g(x_1, x_2) = -2 \iff x_2 = 2 - x_1.$$

Therefore, every point (x_1, x_2) with $g(x_1, x_2) = -2$ is of the form $(x_1, 2 - x_1)$. To find an extremum of f is therefore the same as finding an extremum of the (univariate) function

$$F(x_1) := f(x_1, 2 - x_1) = x_1^2 + (2 - x_1)^2 = 2x_1^2 - 4x_1 + 4.$$

We see that this function has a minimum at $x_1 = 1$. Since $x_2 = 2 - x_1$, we obtain that f has a minimum subject to the constraint $x_1 + x_2 = 2$ at the point $(1, 1)$. By geometric reasoning, we see that $(1, 1)$ is a global minimum of f in Ω , and that f has no maximum. (Make a plot!) □

The last example shows that it is sometimes easy to incorporate the constraint and then find the extrema on the boundary. However, this is clearly not always the case, and we need a way to compute extrema subject to a constraint $g(x) = c$ systematically. This is done by the **method of Lagrange multipliers**. For this, we define the **Lagrange function**

$$\mathcal{L}(x, \lambda) := f(x) + \lambda \cdot (g(x) - c),$$

where the number $\lambda \in \mathbb{R}$ is called the *Lagrange multiplier*. Note that the function $\mathcal{L}: \mathbb{R}^{d+1} \rightarrow \mathbb{R}$ now depends on $d + 1$ variables, namely the d 'original' variables and λ .

It turns out that finding extrema subject to constraints is just the same as finding the extrema of the Lagrange function \mathcal{L} . For this we need to compute the gradient, i.e., all partial derivatives, of \mathcal{L} and find points where it is zero. Note that the gradient of $\mathcal{L}(x, \lambda)$ is given by

$$\nabla \mathcal{L}(x, \lambda) = \left(\frac{\partial \mathcal{L}}{\partial x_1}, \dots, \frac{\partial \mathcal{L}}{\partial x_d}, \frac{\partial \mathcal{L}}{\partial \lambda} \right) (x, \lambda)$$

and that the partial derivative w.r.t. λ is just

$$\frac{\partial \mathcal{L}(x, \lambda)}{\partial \lambda} = g(x) - c.$$

Therefore, setting this partial derivative to zero is equivalent to $g(x) = c$, which is precisely the constraint. This partial derivative is therefore not of much interest and we mostly need only the first entries of the gradient, which we denote by

$$\nabla_x \mathcal{L} := \left(\frac{\partial \mathcal{L}}{\partial x_1}, \dots, \frac{\partial \mathcal{L}}{\partial x_d} \right)$$

and, by the definition of the Lagrange function, we see that

$$\nabla_x \mathcal{L}(x, \lambda) = \nabla f(x) + \lambda \cdot \nabla g(x).$$

To compute $\nabla_x \mathcal{L}$ it is therefore enough to compute the gradients of f and g .

It remains to show that this is useful to find extrema of f under a constraint $g(x) = c$. This is based on the following theorem, which is again only a necessary condition.

Theorem 8.86 (Necessary condition). *Let $f, g: \mathbb{R}^d \rightarrow \mathbb{R}$ be continuously differentiable functions, and x_0 be a local extremum of f subject to the constraint $g(x) = c$ for some $c \in \mathbb{R}$. Then, if $\nabla g(x_0) \neq 0$, there exists a constant λ such that*

$$\nabla f(x_0) = -\lambda \nabla g(x_0).$$

The equation from this theorem means that the gradients of f and g at the point x_0 are parallel, i.e., they show in the same or the opposite directions. The additional constant λ is necessary, because the gradients might be of different length.

A formal proof of Theorem 8.86 is out of reach at the moment, as it is based on the (rather involved) *implicit function theorem* that we don't discuss here. However, we can give a geometric explanation of this necessary condition, see Figure 54.

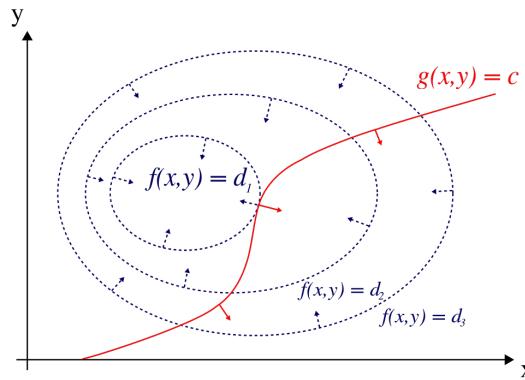


Figure 54: Contour plot of a function and a constraint with their gradients.

Note that the gradient of a function f at x_0 is always perpendicular to the 'surface' $\{x: f(x) = f(x_0)\}$, i.e., the *level set* at height $f(x_0)$. Therefore, if f has an extremum, say a maximum, subject to $g(x) = c$

at x_0 , then the level set $\{x: f(x) = f(x_0)\}$ 'touches' the set $\{x: g(x) = c\}$ in a single point, implying that the gradients of f and g at x_0 are parallel. If this would not be fulfilled, then one could 'wander' along $\{x: g(x) = c\}$ and increase the function value, which contradicts that x_0 is a maximum.

Theorem 8.86 implies that for each extrema at x_0 subject to a constraint $g(x) = c$, with $\nabla g(x_0) \neq 0$, there is some λ such that $\nabla f(x_0) + \lambda \nabla g(x_0) = 0$. That is, if x_0 is a (local) extremum subject to $g(x) = c$, then there is some λ such that

$$\nabla \mathcal{L}(x_0, \lambda) = 0.$$

It is therefore necessary to find all *stationary points* of \mathcal{L} . Moreover, the above theorem makes no statement about points with $\nabla g(x) = 0$, and they have to be considered separately.

In summary, to find all extrema of f subject to a constraint $g(x) = c$ we need to consider all **critical points of \mathcal{L}** , which are

1. all points x_0 with $\nabla g(x_0) = 0$ and $g(x_0) = c$, and
2. all points x_0 with $\nabla g(x_0) \neq 0$ and $\nabla \mathcal{L}(x_0, \lambda) = 0$ for some λ .

Note that, as in the unconstrained optimization, not each of these points is an extremum, and one needs a different reasoning to verify if they are (local) minima or maxima.

Remark 8.87. There is also a variant of the second derivative test for constraint optimization. This method is based on the so-called *bordered Hessian matrix*, which is the Hessian of \mathcal{L} . Unfortunately, it is not as simple as before (by verifying positive/negative-definiteness) to determine the type of an extremum subject to constraints. We do not discuss the details here.

Let us consider some examples.

Example 8.88. We consider again the function from Example 8.85. That is, we want to find the extrema of $f(x_1, x_2) = x_1^2 + x_2^2$ subject to the constraint $g(x_1, x_2) := x_1 + x_2 = 2$, see Figure 55. (For sake of notation, we use a different g here than in Example 8.85.)

First, we check if the constraint has a vanishing gradient by computing

$$\nabla g(x) = (1, 1).$$

Clearly, this is never zero, and therefore all critical points are points $x = (x_1, x_2)$ where the gradient of the Lagrange function

$$\mathcal{L}(x, \lambda) = f(x) + \lambda(g(x) - c) = x_1^2 + x_2^2 + \lambda(x_1 + x_2 - 2)$$

is zero for some λ . Differentiation yields

$$\nabla_x \mathcal{L}(x, \lambda) = (2x_1 + \lambda, 2x_2 + \lambda)$$

and

$$\frac{\partial \mathcal{L}(x, \lambda)}{\partial \lambda} = x_1 + x_2 - 2.$$

Setting $\nabla_x \mathcal{L}(x, \lambda) = 0$ we see that the equation is solved by $x_1 = x_2 = -\frac{\lambda}{2}$.

We finally need to find λ such that this point satisfies the constraint $g(x) = x_1 + x_2 = 2$. Clearly, this leads to $\lambda = -2$, and therefore to the unique critical point $(1, 1)$, as already shown in Example 8.85. This point is clearly a minimum, see Figure 55.

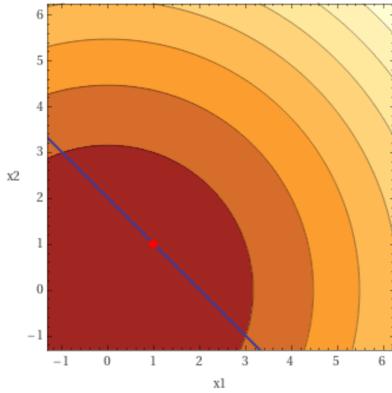


Figure 55: Extrema of $f(x_1, x_2) = x_1^2 + x_2^2$ subject to $x_1 + x_2 = 2$.

Example 8.89. We want to determine all extrema of $f: \mathbb{R}^2 \rightarrow \mathbb{R}$,

$$f(x_1, x_2) = x_1^2 \cdot x_2 \quad \text{subject to} \quad x_1^2 + x_2^2 \leq 3.$$

Recall from Example 8.67 that all points from $\{(0, x_2) : x_2 \in \mathbb{R}\}$ are stationary points of f and that all of them except $(0, 0)$ are local extrema. However, all of these points satisfy $f(x) = 0$ and are therefore clearly no global extrema. As f is unbounded it actually does not have global extrema (without constraints). We now consider the bounded and closed domain $\Omega := \{x : x_1^2 + x_2^2 \leq 3\}$, see Figure 56. The function clearly has global minima and maxima in Ω , and they are not in the interior Ω° , since all stationary points have function value 0. (However, they are still local extrema in Ω .) Therefore, the global extrema are on the boundary $\partial\Omega = \{x : x_1^2 + x_2^2 = 3\}$, and to find them, we need to find the extrema of f subject to $g(x) = x_1^2 + x_2^2 = 3$.

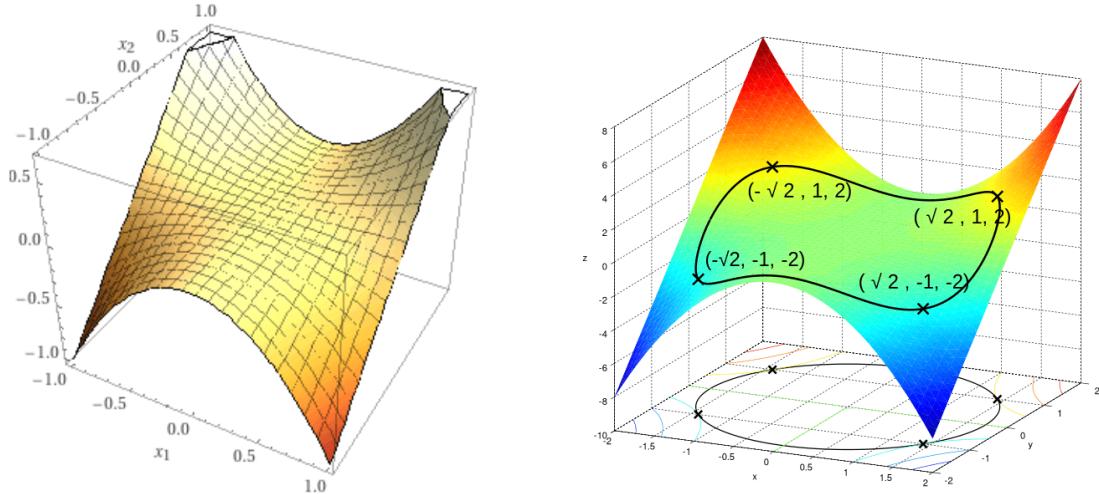


Figure 56: Plot of $f(x_1, x_2) = x_1^2 x_2$ subject to $x_1^2 + x_2^2 = 3$.

To do so we compute

$$\nabla f(x) = (2x_1 x_2, x_1^2)$$

and, for $g(x) = x_1^2 + x_2^2$,

$$\nabla g(x) = (2x_1, 2x_2).$$

First of all, $\nabla g(x) = 0$ if and only if $x = (0, 0)$, but $x = (0, 0)$ does not satisfy $g(x) = 3$, i.e., $x \notin \partial\Omega$, and can therefore be ignored. Now we consider the Lagrange function

$$\mathcal{L}(x, \lambda) = f(x) + \lambda(g(x) - 3) = x_1^2 x_2 + \lambda(x_1^2 + x_2^2 - 3),$$

which has the gradient

$$\nabla_x \mathcal{L}(x, \lambda) = (2x_1x_2 + 2\lambda x_1, x_1^2 + 2\lambda x_2).$$

First note that $\lambda = 0$ corresponds to the (local) extrema of f without constraints, because $\nabla_x \mathcal{L}(x, 0) = \nabla f(x)$. Therefore, solving the equation $\nabla_x \mathcal{L}(x, 0) = 0$ gives the set of solutions $\{(0, x_2) : x_2 \in \mathbb{R}\}$, and the only points of this set on $\partial\Omega$ are $P_1 = (0, \sqrt{3})$ and $P_2 = (0, -\sqrt{3})$. As in Example 8.67, we see that P_1 is a local maximum and P_2 is a local minimum with $f(P_1) = f(P_2) = 0$, see also Figure 56.

Let us consider the general equation $\nabla_x \mathcal{L}(x, \lambda) = 0$, i.e., the system of equations

$$\begin{aligned} 2x_1x_2 + 2\lambda x_1 &= 0, \\ x_1^2 + 2\lambda x_2 &= 0. \end{aligned}$$

We see that the first equation reads $2x_1(x_2 + \lambda) = 0$, which is satisfied if either $x_1 = 0$ or $x_2 + \lambda = 0$. Since $x_1 = 0$ only leads to local extrema (see above), we consider the second case, which implies $x_2 = -\lambda$. Putting this into the second equation gives $x_1^2 - 2\lambda^2 = 0$, i.e., $x_1^2 = 2\lambda^2$ and therefore $x_1 = \pm\sqrt{2} \cdot |\lambda|$. It remains to find suitable λ by putting into the constraint $x_1^2 + x_2^2 = 2\lambda^2 + (-\lambda)^2 = 3$. This gives the solutions $\lambda_1 = 1$ and $\lambda_2 = -1$. For $\lambda = 1$ we obtain the points $P_3 = (\sqrt{2}, -1)$ and $P_4 = (-\sqrt{2}, -1)$, and for $\lambda = -1$ we obtain $P_5 = (\sqrt{2}, 1)$ and $P_6 = (-\sqrt{2}, 1)$.

Finally, by computing the function values $f(P_3) = f(P_4) = -2$ and $f(P_5) = f(P_6) = 2$, we see that f has global minima in $\Omega = \{x : x_1^2 + x_2^2 \leq 3\}$ at P_3 and P_4 , and global maxima in Ω at P_5 and P_6 . \square

Example 8.90. Let us finally consider again our initial example from Example 8.62, i.e.,

$$f(x) := e^{-(x_1^2 + 2x_2^2)}$$

with $x = (x_1, x_2)$ on the set $\Omega := \{x : \|x\| \leq 1\} = \{(x_1, x_2) : x_1^2 + x_2^2 \leq 1\}$.

We already know that f has a global maximum at $(0, 0)$ with $f(0) = 1$, see Example 8.82, and no other stationary point. It therefore has a global minimum at the boundary $\partial\Omega = \{x : \|x\| = 1\}$. To find the minima, first note that ∇g only vanishes at $(0, 0)$, which is not in $\partial\Omega$, and can be ignored. Now consider the Lagrange function

$$\mathcal{L}(x, \lambda) = e^{-(x_1^2 + 2x_2^2)} + \lambda(x_1^2 + x_2^2 - 1)$$

such that

$$\begin{aligned} \nabla_x \mathcal{L}(x, \lambda) &= (-2x_1 f(x) + 2\lambda x_1, -4x_2 f(x) + 2\lambda x_2) \\ &= (-2x_1(f(x) + \lambda), -2x_2(2f(x) + \lambda)). \end{aligned}$$

So, $\nabla \mathcal{L}(x, \lambda) = 0$ holds if either $x_1 = 0$ and $x_2 = \pm 1$, or $x_2 = 0$ and $x_1 = \pm 1$. (Verify why there are no other possibilities!) Since $f(0, \pm 1) = e^{-2}$ and $f(\pm 1, 0) = e^{-1}$, we see that f has a global minimum in Ω at $x_0 = (0, \pm 1)$.

8.5 Differential calculus for vector-valued functions

So far we mostly considered functions of the form $f: G \rightarrow \mathbb{R}$. However, as mentioned in the very beginning of this chapter, there are also the so called vector valued functions, e.g. vector fields, which often appear in physics and data analysis. These functions map vectors of \mathbb{R}^d to vectors in \mathbb{R}^m , where we assume $m > 1$. From here on we will always use the notation

$$f(x) = \begin{pmatrix} f_1(x) \\ f_2(x) \\ \vdots \\ f_m(x) \end{pmatrix},$$

where for all $1 \leq i \leq m$ we have $f_i: \mathbb{R}^d \rightarrow \mathbb{R}$. This implies that for each component of f , i.e. for each f_i , we are able to use all results we proved so far. We will see that, it is often enough to study all components, i.e. we can reduce our questions to problems which only contain real functions. Therefore it makes sense to use the following definition.

Definition 8.91. Let $G \subset \mathbb{R}^d$ and $f: G \rightarrow \mathbb{R}^m$.

We call f **continuous, differentiable, etc.** if and only if for every $i = 1, \dots, m$ we have that f_i is continuous, differentiable, etc., respectively.

Example 8.92. The function given by

$$f(x_1, x_2) = (x_1 \cos x_2, x_1 \sin x_2)$$

consists of the components

$$f_1(x_1, x_2) = x_1 \cos x_2 \quad \text{and} \quad f_2(x_1, x_2) = x_1 \sin x_2.$$

Both of these components are continuous making f a continuous function. This function is an example of an so called **vector field**, as it maps \mathbb{R}^2 into itself. Such a function can be visualized by drawing a vector at each point, see Figure 57.

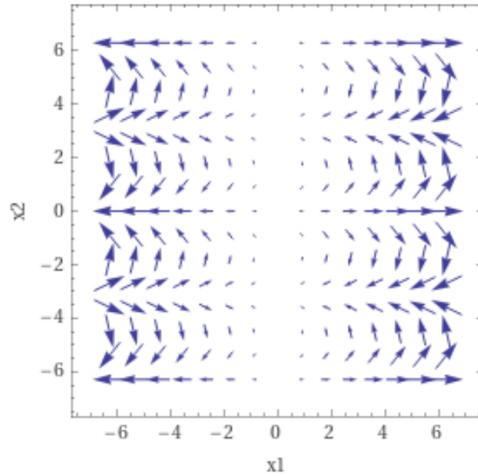


Figure 57: Vector plot of $f(x_1, x_2) = (x_1 \cos x_2, x_1 \sin x_2)$

Example 8.93. Another type of interesting vector-valued functions are representations of **curves**, which are mappings $h: \mathbb{R} \rightarrow \mathbb{R}^m$, i.e., the map a number to a point in \mathbb{R}^m . One prominent example is the **helix**

$$h(t) = (r \cos(2\pi t), r \sin(2\pi t), t)$$

with radius $r > 0$, see Figure 58. This function is clearly continuous, since all components are continuous.

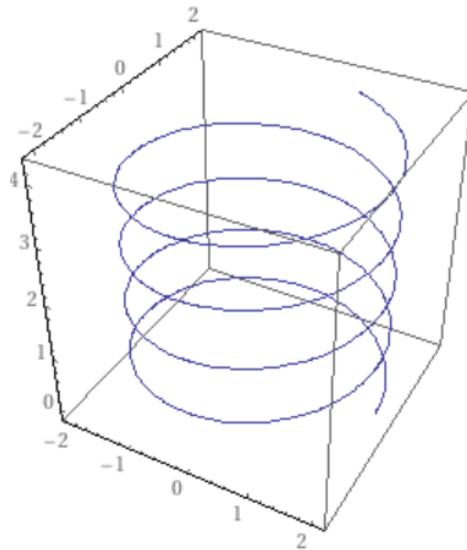


Figure 58: Plot of the curve $h(t) = (2 \cos(2\pi t), 2 \sin(2\pi t), t)$

The analysis of individual vector-valued function goes along the same lines than for real-valued (multivariate) functions. However, as there are now m different components that we need to keep track of, we need some more definitions.

First, we discuss the partial derivatives for vector-valued functions. The only difference between real- and vector-valued functions is that we need to consider all partial derivatives of every component f_i of f , and we use again a matrix to collect them. (Recall that the partial derivatives of a real-valued function were collected in the gradient, i.e., a $1 \times d$ matrix aka. vector.) Nevertheless the computation of partial derivatives is not harder in this setting as we will immediately see from the next definition.

Definition 8.94. Let $G \subset \mathbb{R}^d$, $f: G \rightarrow \mathbb{R}^m$ and $x \in G$.

If for any $1 \leq j \leq d$ and any $1 \leq i \leq m$ the partial derivative

$$\frac{\partial f_i}{\partial x_j}(x) = \lim_{h \rightarrow 0} \frac{f_i(x + he_j) - f_i(x)}{h}$$

exists we call f **partially differentiable** at x .

If this is the case for any $x \in G$ we simply say that f is partial differentiable.

Moreover, if f is partial differentiable at $x \in G$, we call

$$J_f(x) := \begin{pmatrix} \frac{\partial f_1}{\partial x_1}(x) & \frac{\partial f_1}{\partial x_2}(x) & \dots & \frac{\partial f_1}{\partial x_d}(x) \\ \frac{\partial f_2}{\partial x_1}(x) & \frac{\partial f_2}{\partial x_2}(x) & \dots & \frac{\partial f_2}{\partial x_d}(x) \\ \vdots & & \ddots & \vdots \\ \frac{\partial f_m}{\partial x_1}(x) & \frac{\partial f_m}{\partial x_2}(x) & \dots & \frac{\partial f_m}{\partial x_d}(x) \end{pmatrix}$$

the **Jacobian matrix** or functional matrix of f at x .

We see that the i -th row of the Jacobian is given by the gradient of f_i . We will illustrate this by the following example.

Example 8.95. Let us compute the Jacobian of

$$f(x_1, x_2, x_3) = (x_1 x_2, e^{x_3}),$$

which is a function from \mathbb{R}^3 to \mathbb{R}^2 with $f_1(x) = x_1 x_2$ and $f_2(x) = e^{x_3}$, where $x = (x_1, x_2, x_3)$. We see that

$$\nabla f_1(x) = (x_2, x_1, 0) \quad \text{and} \quad \nabla f_2(x) = (0, 0, e^{x_3}).$$

Therefore we see that the Jacobian is given by

$$J(x) = \begin{pmatrix} x_2 & x_1 & 0 \\ 0 & 0 & e^{x_3} \end{pmatrix}.$$

In fact, we already saw some vector fields and their Jacobian, although a bit hidden.

Example 8.96. Let $f: G \rightarrow \mathbb{R}$ be a twice-partially differentiable function. (Note that f is not vector-valued!) The gradient is then a mapping from G to \mathbb{R}^d , as for every $x \in G \subset \mathbb{R}^d$ we have $\nabla f(x) \in \mathbb{R}^d$, i.e. gradients are always vector fields if they exist.

Now, since every component of the gradient is partially differentiable by assumption, we can compute the Jacobian of $\nabla f(x)$, and we see that it is actually given by the Hessian of f , i.e.

$$J_{\nabla f}(x) = H_f(x).$$

Note that under the additional assumption that f is twice-continuously differentiable, we even have that the Jacobian of ∇f is symmetric. (Why is this the case?)

Similar to multivariate real functions it is not sufficient to only consider partial derivatives. Let us adapt the definitions to this vector-valued case, which is quite straightforward.

Definition 8.97. Let $G \subset \mathbb{R}^d$, $f: G \rightarrow \mathbb{R}^m$ be a continuous function.

If there exists a linear mapping $D: \mathbb{R}^d \rightarrow \mathbb{R}^m$ and a mapping $r: \mathbb{R}^d \rightarrow \mathbb{R}^m$ such that

$$f(x + y) = f(x) + D(y) + r(y),$$

where r satisfies

$$\lim_{y \rightarrow 0} \frac{r(y)}{\|y\|} = 0,$$

then we call f **differentiable** at x . We call $D = df_x$ the **(total) derivative** of f at x .

Remark 8.98. Recall that a linear mapping from $\mathbb{R}^d \rightarrow \mathbb{R}^m$ is always described by a matrix. Moreover, $r(y) \in \mathbb{R}^m$ is a vector and $\lim_{y \rightarrow 0} \frac{r(y)}{\|y\|} = 0$ if and only if $\lim_{y \rightarrow 0} \frac{\|r(y)\|}{\|y\|} = 0$

We see that f is differentiable in the above sense if and only if all components $f_i: \mathbb{R}^d \rightarrow \mathbb{R}$ are differentiable, where $1 \leq i \leq m$. This allows to use all results for real functions when considering vector-valued ones. In particular, we are able to generalize the results which were used to connect partial derivatives and differentiability, see Section 8.3.2. All the proof are basically the same, with the additional observation that the Jacobian of a vector-valued function can be written as

$$J_f(x) = \begin{pmatrix} \nabla f_1(x) \\ \nabla f_2(x) \\ \vdots \\ \nabla f_m(x) \end{pmatrix}.$$

Let us summarize this in the following theorem, which we state without proof.

Theorem 8.99. Let $G \subset \mathbb{R}^d$ and $f: G \rightarrow \mathbb{R}^m$ be differentiable at $x \in G$ with derivative df_x . Then,

- 1) f is continuous at x ,
- 2) all partial derivatives of all components f_i exist at x , and
- 3) the (total) derivative of f is given by matrix multiplication with the Jacobian by

$$df_x(y) = J_f(x) \cdot y = \begin{pmatrix} \langle \nabla f_1(x), y \rangle \\ \langle \nabla f_2(x), y \rangle \\ \vdots \\ \langle \nabla f_m(x), y \rangle \end{pmatrix}.$$

Moreover, if $f: G \rightarrow \mathbb{R}^m$ is a mapping such that all partial derivatives of all components are continuous at $x \in G$, then f is also differentiable at x .

Example 8.100. We consider the function

$$f(x_1, x_2) = \begin{pmatrix} x_1 x_2 \\ e^{-x_1-x_2} \end{pmatrix}.$$

The Jacobian is given by

$$J(x) = \begin{pmatrix} x_2 & x_1 \\ -e^{-x_1-x_2} & -e^{-x_1-x_2} \end{pmatrix}.$$

Clearly, all entries of the Jacobian are continuous functions, which shows that f is (totally) differentiable.

By the above theorem, we see that many computations related to the derivative of a vector-valued function can be performed by computations of the (real-valued) component-functions f_i . For example, we easily see that $d(f+g)_x = df_x + dg_x$ for all $f, g: \mathbb{R}^d \rightarrow \mathbb{R}^m$ that are differentiable at x , by using the statements for each component individually.

In addition, using that the derivative is given by the Jacobi matrix, we obtain a particularly useful formula for the derivative of the **composition of functions**. In fact, the derivative (i.e. Jacobi matrix) of the composition $g \circ f$ is given by the matrix-product of the Jacobi matrices f and g (at appropriate points).

As in the univariate case, it is called the **chain rule**.

Theorem 8.101 (Chain rule). *Let $f: \mathbb{R}^d \rightarrow \mathbb{R}^p$ be differentiable at $x \in \mathbb{R}^d$ and $g: \mathbb{R}^p \rightarrow \mathbb{R}^m$ be differentiable at $y = f(x) \in \mathbb{R}^p$. Then, their composition is also differentiable at x and*

$$J_{g \circ f}(x) = J_g(f(x)) \cdot J_f(x).$$

In short, this can be written as $J_{g \circ f} = (J_g \circ f) \cdot J_f$.

Proof. To keep track of all the variables we use the following notation throughout this proof, $A = J_f(x) \in \mathbb{R}^{p \times d}$, $B = J_g(y) \in \mathbb{R}^{m \times p}$, and

$$\begin{aligned} f(x + \xi) &= f(x) + A\xi + r(\xi) \\ g(y + \eta) &= g(y) + B\eta + s(\eta) \end{aligned}$$

with $\xi \in \mathbb{R}^d$ and $\eta \in \mathbb{R}^p$, where $\frac{r(\xi)}{\|\xi\|} \rightarrow 0$ as $\xi \rightarrow 0$ and $\frac{s(\eta)}{\|\eta\|} \rightarrow 0$ as $\eta \rightarrow 0$. We now choose $\eta = f(x + \xi) - f(x) = A\xi + r(\xi)$ and compute

$$\begin{aligned} (g \circ f)(x + \xi) &= g(f(x + \xi)) \\ &= g(f(x) + \eta) \\ &= g(f(x)) + BA\xi + Br(\xi) + s(A\xi + r(\xi)). \end{aligned}$$

Now we define $\varphi(\xi) = Br(\xi) + s(A\xi + r(\xi))$ and if we show that

$$\frac{\varphi(\xi)}{\|\xi\|} \rightarrow 0, \quad \text{as } \xi \rightarrow 0,$$

it follows that BA is the derivative of $g \circ f$ as required.

We observe that the property $\frac{r(\xi)}{\|\xi\|} \rightarrow 0$ as $\xi \rightarrow 0$ (and that B is continuous as linear mapping) implies that also

$$\frac{Br(\xi)}{\|\xi\|} = B \left(\frac{r(\xi)}{\|\xi\|} \right) \rightarrow 0 \quad \text{as } \xi \rightarrow 0.$$

Now, note that

$$\frac{\|s(A\xi + r(\xi))\|}{\|\xi\|} = \frac{\|s(A\xi + r(\xi))\|}{\|A\xi + r(\xi)\|} \cdot \frac{\|A\xi + r(\xi)\|}{\|\xi\|}$$

and since $A\xi + r(\xi) \rightarrow 0$ for $\xi \rightarrow 0$, we now that the first term on the right hand side goes to zero as $\xi \rightarrow 0$. It remains to prove that the second term is bounded. For this note that $\|r(\xi)\| \leq \|\xi\|$ by assumption and that $\|A\xi\| \leq C_A \|\xi\|$ for some constant $C_A < \infty$. The latter follows from $\|A\xi\| \leq \sqrt{d} \|A\xi\|_\infty \leq \sqrt{dd} \max(|a_{i,j}|) \|\xi\|_\infty \leq \sqrt{dd} \max(|a_{i,j}|) \|\xi\|_2$.

We finally obtain

$$\frac{\|s(A\xi + r(\xi))\|}{\|\xi\|} \leq \frac{\|s(A\xi + r(\xi))\|}{\|A\xi + r(\xi)\|} \cdot \frac{\|A\xi + r(\xi)\|}{\|\xi\|} \leq (C_A + 1) \cdot \frac{\|s(A\xi + r(\xi))\|}{\|A\xi + r(\xi)\|} \rightarrow 0,$$

as $\xi \rightarrow 0$. All in all this shows that $\frac{\varphi(\xi)}{\|\xi\|} \rightarrow 0$ as $\xi \rightarrow 0$, which finishes the proof. \square

Example 8.102. We consider the functions $f: \mathbb{R}^2 \rightarrow \mathbb{R}^3$ and $g: \mathbb{R}^3 \rightarrow \mathbb{R}^3$ given by

$$f(x) = \begin{pmatrix} x_1^2 \\ x_2^2 \\ x_1 x_2 \end{pmatrix} \quad \text{and} \quad g(y) = \begin{pmatrix} y_1^2 \\ y_2^2 \\ y_3^2 \end{pmatrix},$$

where $x = (x_1, x_2)$ and $y = (y_1, y_2, y_3)$.

The Jacobians are given by

$$J_f(x) = \begin{pmatrix} 2x_1 & 0 \\ 0 & 2x_2 \\ x_2 & x_1 \end{pmatrix} \quad \text{and} \quad J_g(y) = \begin{pmatrix} 2y_1 & 0 & 0 \\ 0 & 2y_2 & 0 \\ 0 & 0 & 2y_3 \end{pmatrix}.$$

Note that in order to compute $g \circ f(x)$ and anything that is related to this function we have to set $y = f(x)$. Thus we obtain

$$J_g(f(x)) = \begin{pmatrix} 2x_1^2 & 0 & 0 \\ 0 & 2x_2^2 & 0 \\ 0 & 0 & 2x_1 x_2 \end{pmatrix}$$

and finally by the chain rule

$$J_{g \circ f}(x) = J_g(f(x)) \cdot J_f(x) = \begin{pmatrix} 2x_1^2 & 0 & 0 \\ 0 & 2x_2^2 & 0 \\ 0 & 0 & 2x_1 x_2 \end{pmatrix} \cdot \begin{pmatrix} 2x_1 & 0 \\ 0 & 2x_2 \\ x_2 & x_1 \end{pmatrix} = \begin{pmatrix} 4x_1^3 & 0 \\ 0 & 4x_2^3 \\ 2x_1 x_2^2 & 2x_1^2 x_2 \end{pmatrix}.$$

Having a look at

$$(g \circ f)(x) = \begin{pmatrix} x_1^4 \\ x_2^4 \\ x_1^2 x_2^2 \end{pmatrix}$$

and computing the Jacobian of this function directly, we obtain exactly the same matrix.

Example 8.103. Now we have a look at

$$h(x) = \begin{pmatrix} e^{x_1 - x_2} \\ \cos(\sin(x_1)) \\ x_3^2 + \sin(x_1) \end{pmatrix}$$

This function can be written as $g \circ f$, where

$$f(x) = \begin{pmatrix} x_1 - x_2 \\ x_3^2 \\ \sin x_1 \end{pmatrix} \quad \text{and} \quad g(y) = \begin{pmatrix} e^{y_1} \\ \cos y_3 \\ y_2 + y_3 \end{pmatrix}$$

with $x = (x_1, x_2, x_3)$ and $y = (y_1, y_2, y_3)$. (There are many different choices for f and g . Try to find some!)

If we compute the corresponding Jacobian matrices

$$J_g(y) = \begin{pmatrix} e^{y_1} & 0 & 0 \\ 0 & 0 & -\sin y_3 \\ 0 & 1 & 1 \end{pmatrix} \quad \text{and} \quad J_f(x) = \begin{pmatrix} 1 & -1 & 0 \\ 0 & 0 & 2x_3 \\ \cos x_1 & 0 & 0 \end{pmatrix},$$

we obtain

$$J_{g \circ f}(x) = J_g(f(x)) \cdot J_f(x) = \begin{pmatrix} e^{x_1 - x_2} & -e^{x_1 - x_2} & 0 \\ -\sin(\sin(x_1)) \cos(x_1) & 0 & 0 \\ \cos x_1 & 0 & 2x_3 \end{pmatrix}.$$

Of course we would also be able to compute this Jacobian directly, but this example suggests a quite useful method one can use. We want to compute the Jacobian of a complicated function, in this case h . To do so we rewrite it as the composition of two easier functions f, g and use the chainrule to compute $J_{g \circ f}$. This may lead to more computations and a matrix multiplication, but we only have to compute very easy derivatives if we make a clever choice of f, g . Thus the problem may become easier in some cases as many derivatives can be written down immediately.

8.6 Taylor series

One of the main applications where one needs higher-order partial derivatives is the generalization of Taylor's theorem to the multidimensional case. This is a very commonly used technique in classical and modern physics and chemistry, but also a large class of technical problems requires a method to approximate functions. Here, we again only discuss real-valued functions $f: \mathbb{R} \rightarrow \mathbb{R}$.

Before we start we want to introduce the commonly used **multi-index notation**, which will be needed later on. From here on we assume that $\alpha \in \mathbb{N}_0^d$ is a d -dimensional vector of natural numbers (or zeros). For vectors of this kind we introduce the following quantities

$$|\alpha| = \alpha_1 + \alpha_2 + \cdots + \alpha_d \\ \alpha! = \alpha_1! \cdot \alpha_2! \cdots \alpha_d!.$$

If f is $|\alpha|$ -times continuously differentiable then we define

$$D^\alpha f(x) = D_1^{\alpha_1} D_2^{\alpha_2} \cdots D_d^{\alpha_d} f(x) = \frac{\partial^{\alpha_1}}{\partial x_1^{\alpha_1}} \frac{\partial^{\alpha_2}}{\partial x_2^{\alpha_2}} \cdots \frac{\partial^{\alpha_d}}{\partial x_d^{\alpha_d}} f(x),$$

i.e. we differentiate α_1 times w.r.t. x_1 , α_2 times w.r.t. x_2 and so on. Since f is $|\alpha|$ -times continuously differentiable it follows from Theorem 8.58 that we could change the order of differentiation. Moreover, we will use the following useful notation

$$x^\alpha = x_1^{\alpha_1} x_2^{\alpha_2} \cdots x_d^{\alpha_d}.$$

To show Taylor's theorem we need the following auxiliary result.

Lemma 8.104. *Let $G \subset \mathbb{R}^d$ be open, $f: G \rightarrow \mathbb{R}$ be k -times continuously differentiable and let $x, y \in G$ such that for any $t \in [0, 1]$ we have that $x + ty \in G$. Then, $g: [0, 1] \rightarrow \mathbb{R}$ with $g(t) := f(x + ty)$ is k -times continuously differentiable and we have*

$$g^{(k)}(t) = \frac{d^k g}{dt^k}(t) = \sum_{|\alpha|=k} \frac{k!}{\alpha!} D^\alpha f(x + ty) \cdot y^\alpha.$$

Remark 8.105. This formula for the derivatives of g makes sense also at the endpoints of the interval $[0, 1]$, although we usually avoided to consider derivatives of the endpoints of the domain of the function. For this note that, since G is open, and $x, y \in G$, we know that also small neighborhoods around x and y are contained in G . In particular, there is some $\varepsilon > 0$ such that $x + ty \in G$ for all $t \in (-\varepsilon, 1 + \varepsilon)$. If we now define the function $g: (-\varepsilon, 1 + \varepsilon) \rightarrow \mathbb{R}$, we can consider its derivatives also at $t = 0$ and $t = 1$.

Proof. We split the proof in two parts:

First part: We show by induction that

$$\frac{d^k g}{dt^k}(t) = \sum_{i_1, \dots, i_k=1}^d D_{i_k} \cdots D_{i_1} f(x + ty) \cdot y_{i_1} \cdots y_{i_k}.$$

If $k = 1$ this is just an application of the definition of the total derivative. For this, write $\bar{x} := x + ty$ and note that we can write $f(x + (t + h)y) = f(\bar{x} + hy) = f(\bar{x}) + d_{\bar{x}}f(hy) + r(hy)$, and so

$$g'(t) = \lim_{h \rightarrow 0} \frac{g(t + h) - g(t)}{h} = \lim_{h \rightarrow 0} \frac{f(x + (t + h)y) - f(x + ty)}{h} = d_{(x+ty)}f(y) \\ = \sum_{i=1}^d D_i f(x + ty) \cdot y_i,$$

see Definition 8.31 and Theorem 8.35 (with x replaced by $x + ty$). Now we assume that the statement is true for $k - 1$, i.e.

$$\frac{d^{k-1}g}{dt^{k-1}}(t) = \sum_{i_1, \dots, i_{k-1}=1}^d D_{i_{k-1}} \dots D_{i_1} f(x + ty) \cdot y_{i_1} \cdots y_{i_{k-1}}.$$

This function is still differentiable, since $D_{i_{k-1}} \dots D_{i_1} f$ is so for every choice of i_1, \dots, i_{k-1} . By the same computation than above, with f replaced by $D_{i_{k-1}} \dots D_{i_1} f$, we see

$$\begin{aligned} g^{(k)}(t) &= \frac{d}{dt} \frac{d^{k-1}g}{dt^{k-1}}(t) = \frac{d}{dt} \sum_{i_1, \dots, i_{k-1}=1}^d D_{i_{k-1}} \dots D_{i_1} f(x + ty) \cdot y_{i_1} \cdots y_{i_{k-1}} \\ &= \sum_{i_1, \dots, i_{k-1}=1}^d y_{i_1} \cdots y_{i_{k-1}} \cdot \frac{d}{dt} (D_{i_{k-1}} \dots D_{i_1} f(x + ty)) \\ &= \sum_{i_1, \dots, i_{k-1}=1}^d y_{i_1} \cdots y_{i_{k-1}} \cdot \sum_{j=1}^d D_j (D_{i_{k-1}} \dots D_{i_1} f(x + ty)) \cdot y_j \\ &= \sum_{i_1, \dots, i_{k-1}=1}^d \sum_{j=1}^d D_j D_{i_{k-1}} \dots D_{i_1} f(x + ty) \cdot y_{i_1} \cdots y_{i_{k-1}} \cdot y_j \end{aligned}$$

Using the index i_k instead of j the first part of the proof follows.

Second part: Now we have to show that

$$\sum_{i_1, \dots, i_k=1}^d D_{i_k} \dots D_{i_1} f(x + ty) y_{i_1} \cdots y_{i_k} = \sum_{|\alpha|=k} \frac{k!}{\alpha!} (D^\alpha f)(x + ty) y^\alpha.$$

For this, we need to count the number of different partial derivatives. Let $\alpha \in \mathbb{N}_0^d$ with $|\alpha| = k$. By Schwarz' theorem (Theorem 8.58) we can interchange all these derivatives, i.e., if i_j appears α_j times ($1 \leq j \leq k$)

$$D_{i_1} D_{i_2} \dots D_{i_k} f(x + ty) y_{i_1} \cdots y_{i_k} = D_1^{\alpha_1} D_2^{\alpha_2} \dots D_k^{\alpha_k} f(x + ty) y_1^{\alpha_1} y_2^{\alpha_2} \dots y_d^{\alpha_d},$$

using the multi-index notation from above. Moreover, the number of tuples (i_1, \dots, i_k) such that i_j appears exactly α_j times for $1 \leq j \leq d$ is $\frac{k!}{\alpha_1! \alpha_2! \dots \alpha_d!}$. This proves the desired formula. \square

With the help of this result we are able to show the multivariate version of Taylor's theorem.

Theorem 8.106 (Taylor's theorem). *Let $G \subset \mathbb{R}^d$ be open, $f: G \rightarrow \mathbb{R}$ be $(n+1)$ -times continuously differentiable and $x, y \in G$ such that for any $t \in [0, 1]$ the point $x + ty$ is also contained in G . Then there exists some $\theta \in (0, 1)$ such that*

$$f(x + y) = \sum_{|\alpha| \leq n} \frac{D^\alpha f(x)}{\alpha!} \cdot y^\alpha + \sum_{|\alpha|=n+1} \frac{D^\alpha f(x + \theta y)}{\alpha!} \cdot y^\alpha.$$

Proof. Let $g(t) = f(x + ty)$, $t \in [0, 1]$, which is $(n+1)$ -times continuously differentiable, see Lemma 8.104. Hence we are able to apply the univariate Taylor's theorem, see Theorem 5.53, implying

$$g(1) = \sum_{k=1}^n \frac{g^{(k)}(0)}{k!} + \frac{g^{(n+1)}(\theta)}{(n+1)!},$$

for some $\theta \in (0, 1)$. (Here, we use the Taylor polynomial of order n at $x_0 = 0$.) Lemma 8.104 implies that for $1 \leq k \leq n$

$$\frac{g^{(k)}(0)}{k!} = \sum_{|\alpha|=k} \frac{1}{\alpha!} D^\alpha f(x) \cdot y^\alpha$$

and

$$\frac{g^{(n+1)}(\theta)}{(n+1)!} = \sum_{|\alpha|=n+1} \frac{1}{\alpha!} D^\alpha f(x + \theta y) \cdot y^\alpha.$$

Observing that $g(1) = f(x+y)$ the result follows. \square

A simple reformulation of the above result, using $y = x - x_0$ for some fixed x_0 , gives a statement similar to the univariate Taylor's theorem 5.53.

Corollary 8.107. *Let $G \subset \mathbb{R}^d$, $f: G \rightarrow \mathbb{R}$ be $(n+1)$ -times continuously differentiable, $x_0 \in G$ and let $U = U(x_0)$ be a neighborhood of x_0 which is completely contained in G . Then, for any $x \in U$ we have the representation*

$$f(x) = \sum_{|\alpha| \leq n} \frac{D^\alpha f(x_0)}{\alpha!} (x - x_0)^\alpha + \sum_{|\alpha|=n+1} \frac{D^\alpha f(\xi)}{\alpha!} (x - x_0)^\alpha$$

for some ξ between x_0 and x , i.e., $\xi = x_0 + \theta(x - x_0)$ for some $\theta \in (0, 1)$. We call

$$T_n(x) := \sum_{|\alpha| \leq n} \frac{D^\alpha f(x_0)}{\alpha!} (x - x_0)^\alpha,$$

the **Taylor polynomial of f of order n (at x_0)**.

The term

$$R_n(x) := f(x) - T_n(x) = \sum_{|\alpha|=n+1} \frac{D^\alpha f(\xi)}{\alpha!} (x - x_0)^\alpha$$

is called the **remainder** of the Taylor polynomial.

Note that ξ is not explicitly known here, and depends (somehow) on f .

Proof. Note that since $x \in U$ there exists some $y \in \mathbb{R}^d$ with $x = x_0 + y$ and such that for any $t \in [0, 1]$ we have that $x_0 + ty \in U$. An application of Theorem 8.106, with x replaced by x_0 and $y = x - x_0$, yields the result. \square

We now turn to error bounds in this approximation. Note that, in contrast to the explicit formula for the remainder of T_n above, where we needed f to be $(n+1)$ -times continuously differentiable, we only need that f is n -times continuously differentiable to obtain *error bounds*.

Corollary 8.108. *Let $G \subset \mathbb{R}^d$, $f: G \rightarrow \mathbb{R}$ be n -times continuously differentiable, $x_0 \in G$ and let $U = U(x_0)$ be a neighborhood of x_0 which is completely contained in G .*

Moreover, assume additionally that $|D^\alpha f(x)| \leq M$ for some $M < \infty$, all $\alpha \in \mathbb{N}_0^d$ with $|\alpha| = n$, and all $x \in U(x_0)$. Then for any $x \in U(x_0)$ we have

$$|f(x) - T_n(x)| \leq \frac{2M \cdot d^n}{n!} \cdot \|x - x_0\|^n.$$

Note that this upper bound might be very large (i.e., bad) for large d and small n , in which case we only have reasonable bounds for x very close to x_0 .

Proof. First of all, writing f as its order $n - 1$ Taylor polynomial, and regrouping, we obtain

$$\begin{aligned} f(x) &= \sum_{|\alpha| \leq n-1} \frac{D^\alpha f(x_0)}{\alpha!} (x - x_0)^\alpha + \sum_{|\alpha|=n} \frac{D^\alpha f(\xi)}{\alpha!} (x - x_0)^\alpha \\ &= \sum_{|\alpha| \leq n} \frac{D^\alpha f(x_0)}{\alpha!} (x - x_0)^\alpha + \sum_{|\alpha|=n} \frac{D^\alpha f(\xi) - D^\alpha f(x_0)}{\alpha!} (x - x_0)^\alpha \\ &= T_n(x) + \sum_{|\alpha|=n} \frac{D^\alpha f(\xi) - D^\alpha f(x_0)}{\alpha!} (x - x_0)^\alpha. \end{aligned}$$

That is, f can also be written by its Taylor polynomial of order n , but with a different error/remainder. To bound all the terms separately, we need that

$$(x - x_0)^\alpha = \prod_{i=1}^d (x_i - x_{0,i})^{\alpha_i} \leq \prod_{i=1}^d \|x - x_0\|^{\alpha_i} = \|x - x_0\|^n$$

for all $\alpha \in \mathbb{N}_0^d$ with $|\alpha| = n$. In addition, we need a special case of the *multinomial theorem*, i.e.,

$$\sum_{\alpha \in \mathbb{N}_0^d : |\alpha|=n} \frac{n!}{\alpha!} = d^n,$$

which follows from combinatorial arguments. (We omit details here.) We finally obtain

$$\begin{aligned} |f(x) - T_n(x)| &= \left| \sum_{|\alpha|=n} \frac{D^\alpha f(\xi) - D^\alpha f(x_0)}{\alpha!} (x - x_0)^\alpha \right| \\ &\leq 2M \|x - x_0\|^n \sum_{|\alpha|=n} \frac{1}{\alpha!} \leq 2M \|x - x_0\|^n \frac{d^n}{n!}. \end{aligned}$$

□

Example 8.109 (Second order approximation). We have a look at a twice continuously differentiable function $f: G \rightarrow \mathbb{R}$, $G \subset \mathbb{R}^d$, and assume that $x_0 = 0$ and that all second-order partial derivatives of f are bounded in G . We want to compute the Taylor representation of f for $k = 2$, which is formally given by

$$f(x) = \sum_{|\alpha| \leq 2} \frac{D^\alpha f(0)}{\alpha!} \cdot x^\alpha + R_2(x).$$

First of all, the only term with $|\alpha| = 0$ is $f(0)$. Considering the terms with $|\alpha| = 1$, where any α contains exactly one non-zero entry with a 1. Thus, all α we have to consider are given by the unit vectors and $\alpha! = 1$. Therefore we see that

$$T_2(x) = f(0) + \sum_{i=1}^d D_i f(0) \cdot x_i + \sum_{|\alpha|=2} \frac{D^\alpha f(0)}{\alpha!} x^\alpha.$$

For $|\alpha| = 2$ we see that any such vector can be obtained as $e_i + e_j$, where $1 \leq i \leq j \leq d$. If $i = j$ then $\alpha! = (2e_i)! = 2$, and if $i < j$ then $\alpha! = (e_i + e_j)! = 1!$. This shows that

$$\begin{aligned} \sum_{|\alpha|=2} \frac{D^\alpha f(0)}{\alpha!} x^\alpha &= \frac{1}{2} \sum_{i=1}^d D_i^2 f(0) \cdot x_i^2 + \sum_{i < j} D_i D_j f(0) \cdot x_i x_j \\ &= \frac{1}{2} \sum_{i=1}^d D_i^2 f(0) \cdot x_i^2 + \frac{1}{2} \sum_{i \neq j} D_i D_j f(0) \cdot x_i x_j, \end{aligned}$$

since all second-order partial derivatives are continuous, and can therefore be interchanged. It follows that

$$T_2(x) = f(0) + \nabla f(0) \cdot x + \frac{1}{2}x^T H_f(0)x,$$

where we write $\nabla f(0) \cdot x$ for $\langle \nabla f(0), x \rangle$. (This makes sense since the gradient is a row vector.) In general, i.e., with $x_0 \neq 0$, we have

$$T_2(x) = f(x_0) + \nabla f(x_0) \cdot (x - x_0) + \frac{1}{2}(x - x_0)^T H_f(x_0)(x - x_0).$$

Note that the error (for $x_0 = 0$) can be estimated by

$$|f(x) - T_2(x)| \leq M \cdot d^2 \cdot \|x\|^2,$$

if all second-order partial derivatives are bounded by M , see Corollary 8.108.

Example 8.110. We now want to use this general formula to calculate the second-order Taylor approximation for

$$f(x_1, x_2, x_3) = x_1 x_2 + e^{x_3},$$

at $x_0 = (2, 1, 0)$. It is easy to compute that

$$\nabla f(x) = (x_2, x_1, e^{x_3}) \quad \text{and} \quad H_f(x) = \begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & e^{x_3} \end{pmatrix}.$$

Thus, $f(x_0) = 3$,

$$\nabla f(x_0) = (1, 2, 1) \quad \text{and} \quad H_f(x_0) = \begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

The second-order Taylor approximation of f at $x_0 = (2, 1, 0)$ is therefore given by

$$T_2(x) = 3 + 1 \cdot (x_1 - 2) + 2 \cdot (x_2 - 1) + 1 \cdot (x_3 - 0) + (x_1 - 2)(x_2 - 1) + \frac{1}{2} \cdot x_3^2$$

Moreover, note that all second-order partial derivatives are bounded by $\max\{1, e^{x_3}\}$. In particular, for the special choice

$$G = \{x \in \mathbb{R}^3 : \|x\| < r\},$$

where $r > 0$, we have $M := \sup_{x \in G} |D^\alpha f(x)| \leq e^r$ for all $\alpha \in \mathbb{N}_0^3$ with $|\alpha| = 2$. Therefore, we have the error bound

$$|f(x) - T_2(x)| \leq 9e^r \cdot \|x - x_0\|^2,$$

according to Corollary 8.108. We see that the error of the approximation is smaller than $\varepsilon > 0$, if $\|x - x_0\| < \sqrt{\frac{\varepsilon}{9}}e^{-r}$.

We already saw that being able to use higher order derivatives leads to a better approximation. This suggests, like in the one dimensional case, to try to write certain functions as the limit of Taylor polynomials. A necessary condition is then that one has to be able to compute arbitrary derivatives of the function.

Definition 8.111. Let $f: G \rightarrow \mathbb{R}$ be infinitely-often differentiable and let $x_0 \in G$.

The formal series given by

$$Tf(x) = \sum_{\alpha \in \mathbb{N}_0^d} \frac{D^\alpha f(x_0)}{\alpha!} (x - x_0)^\alpha$$

is called **Taylor series** of f at x_0 .

Note that, a priori, we do not know if $Tf(x)$ converges, and even then, we do not know if $Tf(x) = f(x)$ for some $x \in G$. In the same way as in the univariate case, we introduce some criteria that imply the convergence, at least in a neighborhood of x_0 .

Theorem 8.112. Let $G \subset \mathbb{R}^d$ be open, $f: G \rightarrow \mathbb{R}$ be infinitely-often differentiable and $x_0 \in G$. If $r > 0$ is such that

$$\lim_{n \rightarrow \infty} \frac{r^n}{n!} \cdot \max_{\alpha \in \mathbb{N}_0^d : |\alpha|=n} \sup_{\xi \in U_r(x_0)} |D^\alpha f(\xi)| = 0,$$

and $U_r(x_0) = \{x \in \mathbb{R}^d : \|x - x_0\| < r\}$ is completely contained in G .

Then, f can be written by its Taylor series

$$f(x) = \sum_{\alpha \in \mathbb{N}_0^d} \frac{D^\alpha f(x_0)}{\alpha!} (x - x_0)^\alpha \quad \text{for all } x \in U_r(x_0).$$

Proof. We have a look at the Taylor polynomials, which were given by

$$T_n(x) = \sum_{|\alpha| \leq n} \frac{D^\alpha f(x_0)}{\alpha!} (x - x_0)^\alpha.$$

We know from Corollary 8.108 that we have the error bound

$$|f(x) - T_n(x)| \leq \frac{2M_n \cdot n^d}{n!} \cdot \|x - x_0\|^n,$$

where $M_n := \max_{\alpha \in \mathbb{N}_0^d : |\alpha|=n} \sup_{\xi \in U_r(x_0)} |D^\alpha f(\xi)|$. Regrouping leads to

$$|f(x) - T_n(x)| \leq \left(\frac{r^n M_n}{n!} \right) \cdot 2 \left(\frac{\|x - x_0\|}{r} \right)^n n^d,$$

The first term tends to zero by assumption. For the remaining terms note that $x \in U(x_0)$ implies $\frac{\|x - x_0\|}{r} < 1$, and that $\lim_{n \rightarrow \infty} q^n n^d = 0$ for any $d \in \mathbb{N}$ and $0 < q < 1$. Therefore, $|f(x) - T_n(x)| \rightarrow 0$ as $n \rightarrow \infty$, and we obtain

$$Tf(x) = \lim_{n \rightarrow \infty} T_n(x) = f(x).$$

□

Example 8.113. Let $f(x_1, x_2) = e^{x_1+x_2}$. We want to compute the Taylor series of this function for $x_0 = 0$. The partial derivatives of f are

$$\frac{\partial f(x)}{\partial x_1} = e^{x_1+x_2} = \frac{\partial f(x)}{\partial x_2}.$$

Thus, all partial derivatives are given by f itself, which implies $D^\alpha f(x_0) = f(x_0) = 1$ for all $\alpha \in \mathbb{N}_0^d$. We obtain the Taylor series

$$Tf(x) = \sum_{|\alpha| \in \mathbb{N}_0^2} \frac{D^\alpha f(0)}{\alpha!} x^\alpha = \sum_{|\alpha| \in \mathbb{N}_0^2} \frac{x^\alpha}{\alpha!}.$$

To see that this series converges, let $U_r = \{x \in \mathbb{R}^2 : \|x\| < r\}$, for some $r > 0$. We see that $|D^\alpha f(x)| \leq e^{\sqrt{2}r}$ for all $x \in U_r$. (Why?)

If we observe that

$$\lim_{n \rightarrow \infty} \frac{r^n}{n!} \cdot \max_{\alpha \in \mathbb{N}_0^d : |\alpha|=n} \sup_{\xi \in U_r} |D^\alpha f(\xi)| \leq \lim_{n \rightarrow \infty} \frac{r^n e^{\sqrt{2}r}}{n!} = 0,$$

since $n!$ grows faster than any exponential, we obtain from Theorem 8.112 that $Tf(x) = f(x)$ for any $x \in U_r$. Since r was arbitrary the Taylor series of f converges for every point $x \in \mathbb{R}^2$.

8.7 Multiple integrals

We now turn to the integral of real-valued functions $f: \mathbb{R}^d \rightarrow \mathbb{R}$. As for one-dimensional integration, we only discuss a valid definition here for **continuous functions**. Moreover, we first discuss **integration over boxes** (or hyperrectangles) of the form $R = [a_1, b_1] \times \cdots \times [a_d, b_d]$. Later we will consider also more general domains, and in the next chapter we even discuss a more general approach to integration, which gives us many (theoretical) tools to work with integrals (aka. averages). Here we focus on the actual computation.

To keep presentation simple, we will mostly illustrate the two-dimensional case here and consider corresponding examples. In this case, we denote the underlying rectangle by $R = [a, b] \times [c, d]$. Note that, similar to the univariate case, one can think of the integral as the volume that is enclosed between the graph of the function, which represents a **surface** in \mathbb{R}^3 , and the x - y -plane, see Figure 59.



Figure 59: Volume under a surface

For such continuous functions defined on rectangles we can follow exactly the same lines as in Section 6.3 and define the integral as the limit of an average of function values.

However, since the possible domains are much more complicated in the multivariate case, we have to be a bit more precise here, and actually need the precise definition of a *Riemann integral*. Let us only illustrate the two-dimensional case. As for univariate integrals, we split the rectangle $R = [a, b] \times [c, d]$ into smaller parts. These are obtained by partitioning each of the intervals $[a, b]$ and $[c, d]$ into smaller intervals, and consider their Cartesian products. For this, assume we have a partition $a = s_0 < s_1 < \cdots < s_n = b$ of $[a, b]$ and a partition $c = t_0 < t_1 < \cdots < t_n = d$ of $[c, d]$. The n^2 Cartesian products of these univariate intervals, say R_1, \dots, R_{n^2} , are all of the form $R_k = [s_i, s_{i+1}] \times [t_j, t_{j+1}]$ for some $i, j = 0, \dots, n$, with area $|R_k| = (s_{i+1} - s_i)(t_{j+1} - t_j)$. If we now bound the values of a function in each of these rectangles by the smallest or largest one, we obtain the lower and upper sums as in Section 6.3, which are lower and upper bounds for the value of the integral, if it exists, independent from the chosen partition. Here, we take in addition those partitions that lead to the largest and smallest value, respectively. That is, we consider the *lower* and *upper sums*

$$L_n(f) := \sup \sum_{k=1}^{n^2} |R_k| \left(\min_{x \in R_k} f(x) \right) \quad \text{and} \quad U_n(f) := \inf \sum_{k=1}^{n^2} |R_k| \left(\max_{x \in R_k} f(x) \right),$$

where the sup / inf are over all partitions as described above. It is not hard to see that $L_n(f) \leq U_n(f)$ for arbitrary functions f . Moreover, $L_n(f)$ is monotonically increasing with n , and $U_n(f)$ is decreasing, which implies that both sequences converge. So, if their limits are the same, i.e., $\lim_{n \rightarrow \infty} L_n(f) = \lim_{n \rightarrow \infty} U_n(f)$, then we define the integral of f by this common limit.

(The generalization to higher dimensions is straightforward.)

Definition 8.114 (Riemann-integrable functions). Let $R = \prod_{i=1}^d [a_i, b_i]$ be a box and $f: R \rightarrow \mathbb{R}$ be a bounded function. Then, if

$$\lim_{n \rightarrow \infty} L_n(f) = \lim_{n \rightarrow \infty} U_n(f),$$

we call f a **(Riemann-)integrable function** and define the **integral of f over R** by this common limit, i.e.,

$$\int_R f(x) dx := \int_R f(x_1, \dots, x_d) d(x_1, \dots, x_d) := \lim_{n \rightarrow \infty} L_n(f).$$

This definition is quite impractical as the involved limits, suprema and infima are hard to determine. We will discuss shortly how to evaluate integrals easier. However, for continuous functions defined on a box (or rectangle), the definition can be much simplified:

We can, again, define the value of the integral by the limit of *cubature rules* applied to f , see Remark 6.35. Let us assume $R = [a, b] \times [c, d]$ and define the sums

$$Q_n(f) := \frac{(b-a)(d-c)}{n^2} \sum_{i,j=1}^n f\left(a + \frac{i}{n}(b-a), c + \frac{j}{n}(d-c)\right), \quad (8.1)$$

see also (6.1). In the special case $R = [0, 1]^2$, we have $Q_n(f) = \frac{1}{n^2} \sum_{i,j=1}^n f\left(\frac{i}{n}, \frac{j}{n}\right)$.

The following lemma shows that these sums (aka. averages) converge to the integral for continuous f . We state it directly for higher dimensions. The modifications of the above definitions (and the proof sketch below) are straightforward.

Lemma 8.115. Let $R = \prod_{i=1}^d [a_i, b_i]$ and $f: R \rightarrow \mathbb{R}$ be continuous. Then, f is integrable and

$$\int_R f(x) dx = \lim_{n \rightarrow \infty} Q_n(f).$$

Sketch of proof. We use the same ideas as in the univariate case. Moreover, we only prove the case $d = 2$ with $R = [0, 1]^2$ here. If we use an *equidistant partition* R_1, \dots, R_{n^2} of R , i.e., all R_k are of the form $[\frac{i}{n}, \frac{i+1}{n}] \times [\frac{j}{n}, \frac{j+1}{n}]$ for some $i, j = 1, \dots, n$, and denote the corresponding lower and upper sum by $L_n^*(f)$ and $U_n^*(f)$, we obtain

$$L_n^*(f) \leq L_n(f) \leq U_n(f) \leq U_n^*(f).$$

(Recall that L_n involves the supremum over all partitions and is therefore larger, similar for U_n .) Hence, f is integrable if $\lim_{n \rightarrow \infty} L_n^*(f) = \lim_{n \rightarrow \infty} U_n^*(f)$. For this, let $\ell_i = \min\{f(x): x \in R_i\}$ and $u_i = \max\{f(x): x \in R_i\}$. We obtain, for fixed $\varepsilon > 0$, that

$$|u_i - \ell_i| < \varepsilon.$$

for all $i = 1, \dots, n^2$, if n is large enough. (As in Lemma 6.28, we use here that f on a bounded set, here the rectangle, is not only continuous but even uniformly continuous.) This yields that

$$|L_n^*(f) - U_n^*(f)| < \frac{(b-a)(d-c)}{n^2} \sum_{i=1}^{n^2} \varepsilon = (b-a)(d-c) \varepsilon$$

for large enough n . Since this holds for all $\varepsilon > 0$, we obtain that the limits are equal.

To obtain this limit equals $\lim_{n \rightarrow \infty} Q_n(f)$ note that, by the choice of the partition, we have $L_n^*(f) \leq Q_n(f) \leq U_n^*(f)$. The sandwich rule implies the result. \square

Although the above gives us a valid definition, it is not very handy when we want to compute integrals. Unfortunately, there is **no equivalent for the antiderivative** for multivariate functions, which was,

together with the *fundamental theorem of calculus* (Theorem 6.38), the most powerful technique to evaluate integrals. Luckily, we are again in a situation that allows to reduce everything to the **evaluation of one-dimensional integrals**. This is (a special case of) *Fubini's theorem* which is probably the most important result related to multiple integrals.

Theorem 8.116 (Fubini). *Let $R = \prod_{i=1}^d [a_i, b_i]$ be a box and $f: R \rightarrow \mathbb{R}$ be continuous. Then,*

$$\int_R f(x) dx = \int_{a_1}^{b_1} \left(\int_{a_2}^{b_2} \left(\cdots \left(\int_{a_d}^{b_d} f(x_1, \dots, x_d) dx_d \right) \cdots \right) dx_2 \right) dx_1.$$

The order of the iterated integrals can be chosen arbitrary.

In the special case $d = 2$, with a function $f: [a, b] \times [c, d] \rightarrow \mathbb{R}$, this reads

$$\int_R f(x) dx = \int_a^b \left(\int_c^d f(x_1, x_2) dx_2 \right) dx_1 = \int_c^d \left(\int_a^b f(x_1, x_2) dx_1 \right) dx_2.$$

We usually omit the brackets as the order of integral signs and the dx_i 's should leave no room for confusion. However, it is always beneficial to use brackets when things are not clear.

We do not discuss a proof here, because we come back to this important result in the next chapter, where we prove that this holds even more generally.

Fubini's Theorem now allows us to use all the calculation rules for univariate integrals to obtain corresponding results also for multiple integrals. In particular, in view of Lemma 6.31, we obtain the linearity of the integral, i.e., for any continuous functions $f, g: \mathbb{R}^d \rightarrow \mathbb{R}$ and $\lambda, \mu \in \mathbb{R}$ we have that

$$\int_R \lambda f(x) + \mu g(x) dx = \lambda \int_R f(x) dx + \mu \int_R g(x) dx,$$

where R is some box. (To exercise formalism, verify this calculations on your own by using Fubini's theorem and linearity for one dimensional integrals.) Moreover, we obtain that a non-negative continuous function is the zero function if and only if its integral is zero, and, as in Corollary 6.32, that we have the **triangle inequality**

$$\int_R f(x) dx \leq \int_R |f(x)| dx$$

also for multiple integrals. Let us see some examples.

Example 8.117. We start with a continuous function of the form $f(x_1, x_2) = g(x_1)h(x_2)$ on a rectangle $R = [a, b] \times [c, d]$, which are called **product functions**. Such functions are nice to handle with Fubini's theorem since we have

$$\begin{aligned} \int_R f(x) dx &= \int_a^b \int_c^d f(x_1, x_2) dx_2 dx_1 = \int_a^b \int_c^d g(x_1)h(x_2) dx_2 dx_1 \\ &= \int_a^b g(x_1) \left(\int_c^d h(x_2) dx_2 \right) dx_1 = \left(\int_a^b g(x_1) dx_1 \right) \left(\int_c^d h(x_2) dx_2 \right). \end{aligned}$$

That is, **the integral of a product function is the product of the integrals**.

To see a specific example, let $f(x_1, x_2) = \cos(x_1) \sin(x_2)$ and $R = [0, \pi/2] \times [0, \pi]$. We want to calculate

$$\int_R f(x) dx = \int_0^{\pi/2} \int_0^\pi \cos(x_1) \sin(x_2) dx_2 dx_1.$$

Since $\cos(x_1)$ does not depend on x_2 we are allowed to move it outside the inner integral. So,

$$\begin{aligned} \int_0^{\pi/2} \int_0^\pi \cos(x_1) \sin(x_2) dx_2 dx_1 &= \int_0^{\pi/2} \cos(x_1) \int_0^\pi \sin(x_2) dx_2 dx_1 \\ &= \left(\int_0^{\pi/2} \cos(x_1) dx_1 \right) \left(\int_0^\pi \sin(x_2) dx_2 \right). \end{aligned}$$

However, these univariate integrals are easily evaluated to be $\int_0^{\pi/2} \cos t dt = 1$ and $\int_0^\pi \sin t dt = 2$, implying

$$\int_R f(x) dx = 2.$$

Example 8.118. Let us consider the multivariate polynomial $p(x_1, x_2) = x_1^2 x_2^3 + x_1 + x_2 + 3$ on $R = [0, 1]^2 = [0, 1] \times [0, 1]$. We compute

$$\begin{aligned} \int_{[0,1]^2} p(x) dx &= \int_0^1 \int_0^1 (x_1^2 x_2^3 + x_1 + x_2 + 3) dx_1 dx_2 \\ &= \int_0^1 \left(\int_0^1 x_1^2 x_2^3 dx_1 + \int_0^1 x_1 dx_1 + \int_0^1 x_2 dx_1 + \int_0^1 3 dx_1 \right) dx_2 \\ &= \int_0^1 \left(x_2^3 \int_0^1 x_1^2 dx_1 + \int_0^1 x_1 dx_1 + x_2 \int_0^1 1 dx_1 + \int_0^1 3 dx_1 \right) dx_2 \\ &= \int_0^1 \left(x_2^3 \frac{1}{3} + \frac{1}{2} + x_2 + 3 \right) dx_2 \\ &= \int_0^1 \frac{1}{3} x_2^3 + x_2 + \frac{7}{2} dx_2 \\ &= \frac{49}{12}. \end{aligned}$$

Again, every calculation in this example can be reduced to one-dimensional integration theory.

Example 8.119. Clearly, Fubini's theorem is also very helpful for more complicated multivariate functions. Again, we can deduce everything to univariate integrals, but one should be careful with the different variables appearing, and sometimes one should find the correct order of integration to get a simple solution.

Consider e.g. the function $f(x_1, x_2) = x_1 \cos(x_1 x_2)$ on $R = [0, \pi]^2$. We obtain

$$\begin{aligned} \int_{[0,\pi]^2} f(x) dx &= \int_0^\pi \int_0^\pi x_1 \cos(x_1 x_2) dx_1 dx_2 \\ &= \int_0^\pi x_1 \left(\int_0^\pi \cos(x_1 x_2) dx_2 \right) dx_1 \\ &= \int_0^\pi x_1 \left[\frac{\sin(x_1 x_2)}{x_1} \right]_{x_2=0}^\pi dx_1 \\ &= \int_0^\pi x_1 \left(\frac{\sin(\pi x_1)}{x_1} - 0 \right) dx_1 = \int_0^\pi \sin(\pi x_1) dx_1 \\ &= \frac{1 - \cos(\pi^2)}{\pi}. \end{aligned}$$

Here, we computed first the integral w.r.t. x_2 because the corresponding integrand 'looked simpler'. In the same way we could have calculated the integral by starting with x_1 . Then, we would need to work with the antiderivative (w.r.t. t) of $t \cos(tx_2)$, which would result in a more complicated calculation. However, by Fubini's theorem, the result would clearly be the same. \square

Clearly, it is not always the case that one has to integrate w.r.t. a box. However, general domains in higher dimensions are even harder to tackle than already in the univariate case, and it is just not possible to define integrals of arbitrary functions over arbitrary domains. We therefore introduce the following rather general class of domains, which are those sets that can be assigned a volume (or area for $d = 2$) by our definition of an integral (which is by now only defined over boxes). Recall that the **indicator function** $\chi_A: \mathbb{R}^d \rightarrow \mathbb{R}$ for a set $A \subset \mathbb{R}^d$ is defined by

$$\chi_A(x) := \begin{cases} 1, & x \in A, \\ 0, & x \notin A. \end{cases}$$

Definition 8.120 (Jordan-measurable set). Let $A \subset \mathbb{R}^d$ be a bounded domain, i.e., there is a box $R \subset \mathbb{R}^d$ with $A \subset R$. We call A **Jordan-measurable**, if χ_A is Riemann-integrable.

In this case, we define the **volume of A** by

$$\text{vol}_d(A) := \int_R \chi_A(x) dx.$$

Moreover, we call a bounded function $f: A \rightarrow \mathbb{R}$ **integrable over A** , if $f \cdot \chi_A$ is integrable, and we set

$$\int_A f(x) dx := \int_R f(x) \cdot \chi_A(x) dx.$$

(Here we set $(f \cdot \chi_A)(x) = 0$ for $x \notin A$.)

That is, we now also defined integrals over more general domains. However, as this kind of generalization is the topic of the next chapter, we do not go into detail here and only discuss a specific class of domains, which is anyhow quite usual in applications.

Definition 8.121 (Normal domains). A bounded set $A \subset \mathbb{R}^2$ of the form

$$A = \{(x_1, x_2) \in \mathbb{R}^2 : x_1 \in [a, b], \varphi_1(x_1) \leq x_2 \leq \psi_1(x_1)\}$$

for some $a, b \in \mathbb{R}$ and continuous functions $\varphi_1, \psi_1: [a, b] \rightarrow \mathbb{R}$, is called a **normal domain**. (Strict inequalities are also allowed.)

In higher dimensions, we define inductively that $A \subset \mathbb{R}^d$ is called a normal domain if

$$A = \{(x', x_d) \in \mathbb{R}^d : x' \in A', \varphi_d(x') \leq x_d \leq \psi_d(x')\}$$

for some normal domain $A' \in \mathbb{R}^{d-1}$ and continuous functions $\varphi_d, \psi_d: A' \rightarrow \mathbb{R}$. (We used for a simpler notation $x' := (x_1, \dots, x_{d-1})$.)

Before we discuss easier examples, let us see how this definition has to be understood in higher dimensions. It just means that the variables of the domain are restricted in a sequential way, meaning that the k -th variable is contained in an interval that is bounded by some continuous functions of the first $k - 1$ variables. For example, for $d = 3$, a normal domain has the form

$$\{(x, y, z) \in \mathbb{R}^3 : x \in [a, b], y \in [\varphi_1(x), \psi_1(x)], z \in [\varphi_2(x, y), \psi_2(x, y)]\}.$$

We now want to compute integrals over normal domains. Let us start with the case $d = 2$. If we combine Fubini's theorem (on a box that contains A) with the fact that we can restrict the boundaries of the integral, if the function is zero 'outside', we obtain that **normal domains are Jordan-measurable** and that the integral can be written in an easier form.

Lemma 8.122. Let $A \subset \mathbb{R}^2$ be a normal domain of the form

$$A = \{x \in \mathbb{R}^2 : x_1 \in [a, b], \varphi(x_1) \leq x_2 \leq \psi(x_1)\}$$

where $\varphi \leq \psi$ and both are continuous functions. Then, the integral of an integrable function $f: A \rightarrow \mathbb{R}$ equals

$$\int_A f(x) dx = \int_a^b \int_{\varphi(x_1)}^{\psi(x_1)} f(x_1, x_2) dx_2 dx_1.$$

In particular, the area of A equals

$$\text{vol}_2(A) = \int_A 1 dx = \int_a^b (\psi(t) - \varphi(t)) dt.$$

We omit a formal proof.

For a normal domain $A \subset \mathbb{R}^3$ in three dimensions as above, we obtain that the integral of an integrable function $f: A \rightarrow \mathbb{R}$ can be computed by

$$\int_A f(x) dx = \int_a^b \int_{\varphi_1(x)}^{\psi_1(x)} \int_{\varphi_2(x,y)}^{\psi_2(x,y)} f(x, y, z) dz dy dx$$

and corresponding formulas hold for the volume and in higher dimension.

Let us turn to some examples.

Example 8.123. For some given $a \in [-1, 1]$, let us consider

$$A := \{x \in \mathbb{R}^2 : x_1 \in [a, 1], x_1^2 + x_2^2 \leq 1\},$$

which describes a **circular segment** of the circle with radius 1 (centred in the origin), see Figure 60. The second condition can be rewritten to a condition on x_2 depending only on x_1 , which leads to

$$A = \left\{x \in \mathbb{R}^2 : x_1 \in [a, 1], x_2 \in \left[-\sqrt{1-x_1^2}, \sqrt{1-x_1^2}\right]\right\}.$$

This shows that A is a normal domain.

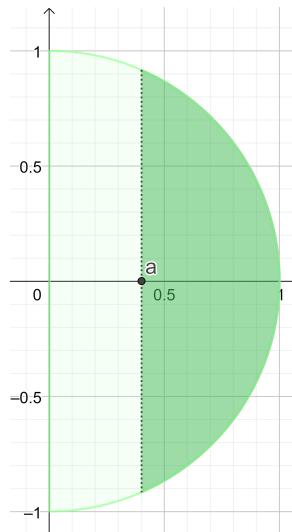


Figure 60: A circular segment.

To compute the area of A , i.e., $\text{vol}_2(A)$, we calculate

$$\int_A 1 dx = \int_a^1 \left(\int_{-\sqrt{1-x_1^2}}^{\sqrt{1-x_1^2}} 1 dx_2 \right) dx_1 = 2 \int_a^1 \sqrt{1-x_1^2} dx_1.$$

We know from Example 6.22 that an antiderivative of $\sqrt{1-t^2}$ is given by $\frac{1}{2}(t\sqrt{1-t^2} + \arcsin(t))$. Therefore,

$$\int_A 1 dx = \frac{\pi}{2} - a\sqrt{1-a^2} + \arcsin(a).$$

In particular, for $a = 0$, we obtain that the area of a half-circle is

$$\text{vol}_2(\{x \in \mathbb{R}^2 : x_1 \in [0, 1], x_1^2 + x_2^2 \leq 1\}) = 2 \int_0^1 \sqrt{1-t^2} dt = \frac{\pi}{2},$$

and the area of the full circle, which we obtain for $a = -1$, is

$$\text{vol}_2(\{x \in \mathbb{R}^2 : x_1^2 + x_2^2 \leq 1\}) = 2 \int_{-1}^1 \sqrt{1-t^2} dt = \pi.$$

Example 8.124. In a similar way, we can compute the volume of the 3-dimensional unit ball

$$A := \{(x, y, z) \in \mathbb{R}^3 : x^2 + y^2 + z^2 \leq 1\}.$$

The ball can again be written as a normal domain by

$$A = \{(x, y, z) : x \in [-1, 1], y \in [-\sqrt{1-x^2}, \sqrt{1-x^2}], z \in [-\sqrt{1-x^2-y^2}, \sqrt{1-x^2-y^2}]\},$$

and we obtain that

$$\begin{aligned} \text{vol}_3(A) &= \int_A 1 dx = \int_{-1}^1 \left(\int_{-\sqrt{1-x^2}}^{\sqrt{1-x^2}} \left(\int_{-\sqrt{1-x^2-y^2}}^{\sqrt{1-x^2-y^2}} 1 dz \right) dy \right) dx \\ &= 2 \int_{-1}^1 \left(\int_{-\sqrt{1-x^2}}^{\sqrt{1-x^2}} \sqrt{1-x^2-y^2} dy \right) dx. \end{aligned}$$

For the inner integral, we use the substitution $t = \frac{y}{\sqrt{1-x^2}}$ to obtain

$$\begin{aligned} \int_A 1 dx &= 2 \int_{-1}^1 (1-x^2) \left(\int_{-1}^1 \sqrt{1-t^2} dt \right) dx = \pi \int_{-1}^1 (1-x^2) dx \\ &= \frac{4\pi}{3}, \end{aligned}$$

where we used again, as above, that $\int_{-1}^1 \sqrt{1-t^2} dt = \frac{\pi}{2}$.

Calculating not only the volume of a set but the integral of a given function, works exactly along the same lines.

Example 8.125. We want to integrate the function $f(x_1 x_2) = x_1^2 x_2^2$ over the triangle with corners $(0,0), (1,0)$ and $(1,1)$, see Figure 61. This set can be modeled as the normal domain $A = \{x \in \mathbb{R}^2 : x_1 \in [0, 1], x_2 \in [0, x_1]\}$.

Using Fubini, we compute

$$\begin{aligned} \int_A f(x) dx &= \int_0^1 \int_0^{x_1} x_1^2 x_2^2 dx_2 dx_1 \\ &= \int_0^1 x_1^2 \left(\int_0^{x_1} x_2^2 dx_2 \right) dx_1 \\ &= \int_0^1 x_1^2 \cdot \left(\frac{1}{3} x_1^3 \right) dx_1 \\ &= \frac{1}{3} \int_0^1 x_1^5 dx_1 = \frac{1}{18}. \end{aligned}$$

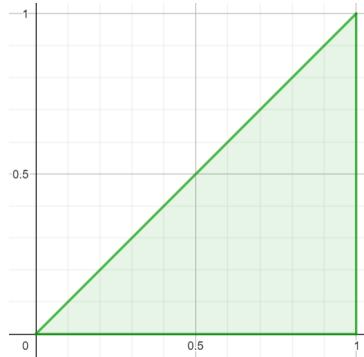


Figure 61: The triangular normal domain

This approach allows us to compute many integrals. However, for this we need to bring the domain in the correct form as a normal domain, and then we need to compute successively all the univariate integrals, which can be very time-consuming.

Such heavy computations can sometimes be avoided by bringing the integral under consideration in a more familiar form. That is, we use some **change of variables** to transform an integral to another integral we already know. As in the univariate case, this is done by **integration by substitution**, which is one of the most powerful ways to compute (difficult) integrals.

Theorem 8.126 (Substitution rule). *Let $G \subset \mathbb{R}^d$ be open and $A \subset G$ be a bounded and Jordan-measurable set. Moreover, let $\Phi: G \rightarrow \mathbb{R}^d$ be a continuously differentiable and injective function such that either $\det J_\Phi(u) > 0$ or $\det J_\Phi(u) < 0$ for any $u \in G$, where $J_\Phi(u)$ is the Jacobi matrix of Φ at $u \in G$.*

Then, $\Phi(A)$ is also Jordan-measurable and, for any bounded and continuous function $f: \Phi(A) \rightarrow \mathbb{R}$, we have

$$\int_{\Phi(A)} f(x) dx = \int_A f(\Phi(u)) \cdot |\det J_\Phi(u)| du.$$

We say that we use the substitution $x = \Phi(u)$.

The proof is quite involved and we omit it here. Note that we need the additional assumption that f is bounded, because we do not know in general that continuous functions on Jordan-measurable domains are integrable. (One might think on the univariate example $f(t) = 1/t$ on $(0, 1]$.)

Let us see how this result can be applied.

Example 8.127 (Polar coordinates). One of the most classical application are **polar coordinates**, which can be used to describe any point in \mathbb{R}^2 . Given some $x \in \mathbb{R}^2$ we determine the distance to the origin, denoted by $r := \|x\| > 0$ and the angle w.r.t. the x_1 -axis, denoted by $\theta \in [0, 2\pi)$. In other words, $x = (x_1, x_2)$ can be written as $x_1 = r \cos \theta$ and $x_2 = r \sin \theta$.

The mapping $\Phi(r, \theta) = (r \cos \theta, r \sin \theta)$, which maps each radius r and angle θ to a point in \mathbb{R}^2 , is continuously differentiable (w.r.t. r and θ) at any $(r, \theta) \in \mathbb{R}^2 =: G$ and

$$\det J_\Phi(r, \theta) = \det \begin{pmatrix} \cos \theta & -r \sin \theta \\ \sin \theta & r \cos \theta \end{pmatrix} = r(\cos^2 \theta + \sin^2 \theta) = r > 0.$$

If we now want to find areas of or integrals over domains that can be easier described in polar coordinates, it is beneficial to use the substitution $x = \Phi(r, \theta)$.

For example, to compute the area of the *annulus* $B := \{x \in \mathbb{R}^2 : R_1 \leq \sqrt{x_1^2 + x_2^2} \leq R_2\}$ for some $0 < R_1 < R_2 < \infty$, see Figure 62, one should realize that $B = \Phi(A)$ with

$$A := \{(r, \theta) : R_1 \leq r \leq R_2, \theta \in [0, 2\pi)\},$$

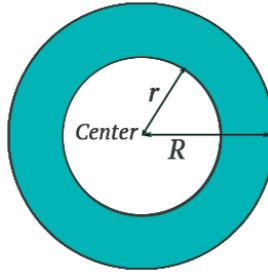


Figure 62: The annulus $\{x \in \mathbb{R}^2 : r \leq \sqrt{x_1^2 + x_2^2} \leq R\}$

which is clearly a normal domain.

We can therefore use the substitution rule with the function $f \equiv 1$ to compute the area

$$\text{vol}_2(B) = \int_B 1 dx = \int_0^{2\pi} \int_{R_1}^{R_2} |\det J_\Phi(r, \theta)| dr d\theta = \int_0^{2\pi} \int_{R_1}^{R_2} r dr d\theta = (R_2^2 - R_1^2)\pi.$$

In the same way, we can use this formula to calculate the integral of functions $f: \mathbb{R}^2 \rightarrow \mathbb{R}$ over such annuli. For example, if we consider $f(x) := \frac{1}{\|x\|} = (x_1^2 + x_2^2)^{-1/2}$ on the set $A = \{x \in \mathbb{R}^2 : R_1 \leq \sqrt{x_1^2 + x_2^2} \leq R_2\}$ with $0 < R_1 < R_2 < \infty$, where it is clearly continuous, we obtain

$$\int_B f(x) dx = \int_0^{2\pi} \int_{R_1}^{R_2} \frac{1}{\sqrt{r^2 \cos^2 \theta + r^2 \sin^2 \theta}} r dr d\theta = \int_0^{2\pi} \int_{R_1}^{R_2} 1 dr d\theta = 2\pi(R_2 - R_1).$$

Clearly, one can also consider circular sections of such annuli, i.e., sets of the form $\Phi(A)$ with

$$A = \{(r, \theta) : R_1 \leq r \leq R_2, \theta_1 \leq \theta \leq \theta_2\}$$

for some $0 < R_1 < R_2$ and $0 \leq \theta_1 < \theta_2 < 2\pi$. Then, Theorem 8.126 shows that

$$\int_{\Phi(A)} f(x) dx = \int_{\theta_1}^{\theta_2} \int_{R_1}^{R_2} f(r \cos \theta, r \sin \theta) \cdot r dr d\theta.$$

Example 8.128 (Linear mappings). Another important class of transformations $\Phi: \mathbb{R}^d \rightarrow \mathbb{R}^d$ are *linear mappings*. Recall that they are given by $d \times d$ -matrices, i.e., $\Phi(x) = Tx$ for some $T \in \mathbb{R}^{d \times d}$. For such a mapping, it is easy to compute that

$$J_\Phi(x) = T \quad (\text{independent of } x).$$

Therefore, if $\det(T) \neq 0$, then we have that Φ is injective and that $\det(J_\Phi(x))$ is smaller or larger than zero for any $x \in \mathbb{R}^d$. We can therefore use Theorem 8.126 to obtain,

$$\int_{T(A)} f(x) dx = |\det(T)| \int_A f(Tu) du$$

for arbitrary continuous functions $f: \mathbb{R}^d \rightarrow \mathbb{R}$ and Jordan-measurable domains $A \subset \mathbb{R}^d$. In particular, for the function $f \equiv 1$, we obtain

$$\text{vol}_d(T(A)) = |\det(T)| \text{vol}_d(A).$$

For example, if one is interested in the area of the set

$$A = \left\{ (x_1, x_2) \in \mathbb{R}^2 : 0 \leq 2x_1 - x_2 \leq 1, 2 \leq x_1 + x_2 \leq 4 \right\},$$

then we can use the substitution $u_1 := 2x_1 - x_2$ and $u_2 := x_1 + x_2$, i.e., $u = Tx$ with

$$T = \begin{pmatrix} 2 & -1 \\ 1 & 1 \end{pmatrix}.$$

We see that

$$T(A) = \left\{ (u_1, u_2) \in \mathbb{R}^2 : 0 \leq u_1 \leq 1, 2 \leq u_2 \leq 4 \right\} = [0, 1] \times [2, 4]$$

and $\det(T) = 3$, and so

$$\text{vol}_2(A) = \frac{\text{vol}_2(T(A))}{|\det(T)|} = \frac{1 \cdot (4-2)}{3} = \frac{2}{3}.$$

Let us discuss one more involved example.

Example 8.129. Let us compute the integral of $f(x, y) = \frac{y}{x}$ over the set

$$B := \left\{ (x, y) \in \mathbb{R}_+^2 : 1 \leq \frac{y}{x} \leq 2, 1 \leq xy \leq 2 \right\}.$$

Note that this function is well-defined and continuous due to the second condition.

To compute this integral we consider the substitution $u := \frac{y}{x}$ and $v := xy$, because the domain is just $A := \{(u, v) : 1 \leq u, v \leq 2\} = [1, 2]^2$ under this substitution. To apply Theorem 8.126 we need to find the mapping Φ with $\Phi(u, v) = (x, y)$ such that with $\Phi(A) = B$.

For this note that $u = \frac{y}{x}$ is equivalent to $y = xu$. Plugging this into $v = xy$ we obtain $v = x^2u$, i.e., $x = \sqrt{\frac{v}{u}}$, and therefore $y = \sqrt{uv}$. Therefore, the desired mapping is given by

$$\Phi(u, v) = \begin{pmatrix} \sqrt{\frac{v}{u}} \\ \sqrt{uv} \end{pmatrix} \quad \text{on } A = [1, 2]^2.$$

(This mapping is continuously differentiable in an open set around A .) By Theorem 8.126 we obtain

$$\int_B f(x, y) d(x, y) = \int_A f\left(\sqrt{\frac{v}{u}}, \sqrt{uv}\right) \cdot |\det J_\Phi(u, v)| d(u, v) = \int_1^2 \int_1^2 u \cdot |\det J_\Phi(u, v)| du dv.$$

It remains to compute the Jacobi matrix and its determinant, which are

$$J_\Phi(u, v) = \begin{pmatrix} -\frac{1}{2}u^{-3/2}v^{1/2} & \frac{1}{2}u^{-1/2}v^{-1/2} \\ \frac{1}{2}u^{-1/2}v^{1/2} & \frac{1}{2}u^{1/2}v^{-1/2} \end{pmatrix} \quad \text{and} \quad \det J_\Phi(u, v) = \frac{-1}{2u}.$$

(Verify this!) Therefore,

$$\int_B f(x, y) d(x, y) = \int_1^2 \int_1^2 u \cdot \frac{1}{2u} du dv = \frac{1}{2} \int_1^2 \int_1^2 1 du dv = \frac{1}{2}.$$

9 Matrices II

Let us turn back to matrices, see Section 2.

We now discuss the concept of *eigenvalues* and *eigenvectors* of matrices, which allow for some powerful techniques to work with (large) matrices. In particular, we will see that also matrices can be written as sums of easier matrices, which could be seen as a kind of analogue to the Taylor or Fourier series of functions. This can then be used to 'easily' compute the inverse or large powers of a given matrix.

9.1 Eigenvalues and eigenvectors

Let us start with the definition.

Definition 9.1. Let $A \in \mathbb{C}^{n \times n}$. A vector $v \in \mathbb{C}^n$ with $v \neq 0_n$ is called **eigenvector** of A to the **eigenvalue** $\lambda \in \mathbb{C}$ if

$$Av = \lambda v.$$

We call the pair (λ, v) an **eigenpair** of A .

Moreover, the set of all eigenvalues of A , i.e.,

$$\sigma(A) := \left\{ \lambda \in \mathbb{C}: \exists v \in \mathbb{C}^n \setminus \{0\} \text{ with } Av = \lambda v \right\},$$

is called the **spectrum** of A .

At first it seems like this is just a special case of the general linear system $Ax = b$, and therefore there must be always a solution x , at least if A is regular. However, this is not true! Since the solution x appears on both sides, such an equation is completely different in nature.

An eigenvector is very special for a given matrix, and can only exist in combination with an unique eigenvalue. To see this, consider A as a mapping that maps a vector $x \in \mathbb{R}^n$ to (another) vector $Ax \in \mathbb{R}^n$. The equation $Ax = \lambda x$ then means that x is mapped to a multiple of itself, and in general, such vectors are rare. Moreover, we will see that we actually "know" the whole matrix (or mapping) A once we know all its eigenvalues and eigenvectors. This is of huge significance when it comes, in particular, to the *approximation* of very big matrices, and therefore essential in applications. But it is also a very powerful theoretical tool.

Remark 9.2. Note that we assumed that eigenvectors and eigenvalues may be **complex**, and this is essential. In general, matrices do not need to have real eigenvalues/-vectors, but we will see that they have, if we allow for complex numbers. We will see an easy example soon. This is similar to the existence of solutions to arbitrary polynomial equations, like $x^2 + 1 = 0$, which is possible over \mathbb{C} , but there might be no solution in \mathbb{R} . However, we will see that in many cases, all eigenvalues/-vectors are real, and that this is always the case for *symmetric matrices*.

Remark 9.3. Sometimes authors require that eigenvalues and eigenvectors have to be considered over the same field as the matrix, i.e., for real-valued matrices we only consider real-valued eigenvalues/-vectors. In this case $\lambda \in \mathbb{C} \setminus \mathbb{R}$ would not be called an eigenvalue of a real matrix. Although this might also lead to interesting insights, and we would save to work with complex numbers, it would also result in several technical difficulties. We will therefore focus on complex matrices, and also allow for complex eigenvectors/-values.

Let us see an example. If we consider the matrix

$$A = \begin{pmatrix} 2 & 1 \\ 0 & 1 \end{pmatrix},$$

then a vector $x = (x_1, x_2)^T \in \mathbb{R}^2$ is mapped to the vector

$$Ax = \begin{pmatrix} 2x_1 + x_2 \\ x_2 \end{pmatrix}.$$

For example, $A\left(\begin{pmatrix} 1 \\ 1 \end{pmatrix}\right) = \begin{pmatrix} 3 \\ 1 \end{pmatrix}$ or $A\left(\begin{pmatrix} 0 \\ 1 \end{pmatrix}\right) = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$, both 'outputs' are clearly not a multiple of the 'input'. However, one may find $A\left(\begin{pmatrix} -1 \\ 1 \end{pmatrix}\right) = \begin{pmatrix} -1 \\ 1 \end{pmatrix}$ and $A\left(\begin{pmatrix} 1 \\ 0 \end{pmatrix}\right) = \begin{pmatrix} 2 \\ 0 \end{pmatrix} = 2 \cdot \begin{pmatrix} 1 \\ 0 \end{pmatrix}$, which shows that $(-1, 1)^T$ is an eigenvector with eigenvalue 1 and that $(1, 0)^T$ is an eigenvector with eigenvalue 2.

The following figure shows how the matrix A maps points x from the unit circle, and indicate which points are mapped to a multiple of itself.

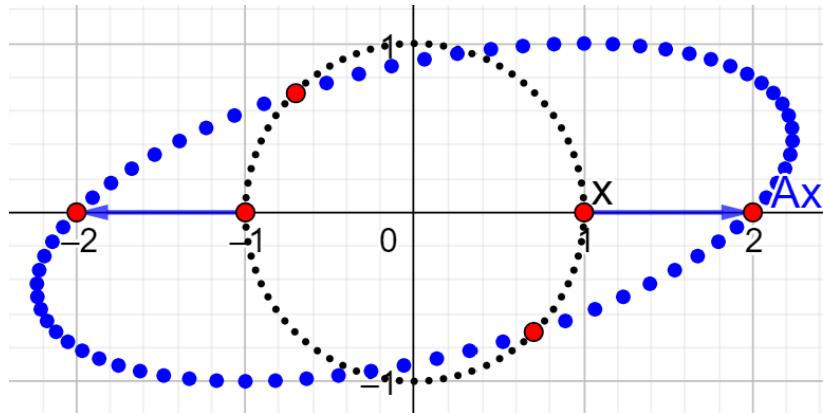


Figure 63: Points x on the unit circle (black), corresponding 'mapped' points Ax (blue), and points that are mapped on (multiples of) itself (red).

Note that eigenvectors to a given eigenvalue are **not unique**, if they exist. For example, we see that in the above example every vector of the form $(-c, c)^T$ with $c \in \mathbb{R}$ is an eigenvector of A to the eigenvalue 1, i.e., $A\left(\begin{pmatrix} -c \\ c \end{pmatrix}\right) = \begin{pmatrix} -c \\ c \end{pmatrix}$. In the same way, every vector of the form $(c, 0)^T$ is an eigenvector of A to the eigenvalue 2.

This is true in general and follows directly from the linearity of the matrix multiplication. For this, let us assume that $v \in \mathbb{R}^n$ is an eigenvector of A to the eigenvalue λ , i.e., (λ, v) is an eigenpair of A . Now, for any $\alpha \in \mathbb{C}$, we see that

$$A(\alpha v) = \alpha Av = \alpha \lambda v = \lambda(\alpha v).$$

This shows that αv is also an eigenvector to the eigenvalue λ , i.e., $(\lambda, \alpha v)$ is an eigenpair for any $\alpha \in \mathbb{R}$. In particular, if v is an eigenvector, then $-v$ is an eigenvector to the same eigenvalue. For example, in the above example we see that $A\left(\begin{pmatrix} -1 \\ 0 \end{pmatrix}\right) = 2 \cdot \begin{pmatrix} -1 \\ 0 \end{pmatrix}$ and therefore, that $(-1, 0)^T$ is also an eigenvector to the eigenvalue 2 (and not -2), see Figure 63.

By this reasoning, we see that, if there exists an eigenvector to an eigenvalue, then there are actually infinitely many eigenvectors to the same eigenvalue, which are at least all multiples of this eigenvector. However, there might also be several different (i.e., not multiples of each other) eigenvectors to a given eigenvalue. These vectors are collected in the so-called *eigenspace*.

Definition 9.4 (Eigenspace). Let $A \in \mathbb{C}^{n \times n}$ and $\lambda \in \mathbb{C}$. Then, we call the set

$$E(A, \lambda) := \left\{ v \in \mathbb{C}^n : Av = \lambda v \right\}$$

the **eigenspace** of A associated with λ , or just eigenspace of λ if the matrix is clear.

Note that $E(A, \lambda) \supset \{0\}$, i.e., $x = 0$ is an element of $E(A, \lambda)$, for any $A \in \mathbb{C}^{n \times n}$ and $\lambda \in \mathbb{C}$. For this just note that $A \cdot 0 = 0 = \lambda \cdot 0$ is fulfilled for any $\lambda \in \mathbb{C}$. Therefore, $v = 0$ is not very interesting, and we are not interested in 0 as an eigenvector (which is also the reason for excluding it in the definition). We see that λ is an eigenvalue if and only if $E(A, \lambda) \neq \{0\}$, i.e., if there exist some $v \in E(A, \lambda)$ with $v \neq 0$.

However, we do clearly not exclude the eigenvalue 0. We have already learned that the corresponding (homogeneous) equation $Ax = 0$ is of huge importance, at least from a theoretical perspective. We therefore give the eigenspace to the eigenvalue 0 a special name.

Definition 9.5 (Kernel). Let $A \in \mathbb{C}^{n \times n}$. Then, we call the set

$$N(A) := \left\{ v \in \mathbb{C}^n : Av = 0 \right\}$$

the **kernel** (or **nullspace**) of A .

As discussed above, the nullspace is clearly the set of solutions of $Ax = 0$, and hence

$$N(A) = E(A, 0) = L(A, 0),$$

where we used again the notations from Definition 2.17 and Definition 9.4.

(Note that the “0” in $E(A, 0)$ is a number, while the “0” in $L(A, 0)$ is a vector.)

Remark 9.6. Don’t be confused that we use different notations for the same set. The kernel is just an important object and appears as a special case of several other definitions. However, and although they agree in this special case, $E(A, \lambda)$ and $L(A, b)$ are in general completely different. In particular, if $v \in E(A, \lambda)$, i.e., (λ, v) is an eigenpair of A then also every multiple of v is in $E(A, \lambda)$. E.g., $-v \in E(A, \lambda)$. The same is not true for $L(A, b)$ with $b \neq 0$. If $Av = b$ for some $v \in \mathbb{R}^n$, i.e., $v \in L(A, b)$, then, e.g., $A(-v) = -Av = -b \neq b$, i.e., $-v \notin L(A, b)$.

Let us see some examples.

Example 9.7. We consider again the matrix

$$A = \begin{pmatrix} 2 & 1 \\ 0 & 1 \end{pmatrix}.$$

We have seen above that $v_1 := \begin{pmatrix} -1 \\ 1 \end{pmatrix}$ is an eigenvector of A to the eigenvalue 1, and that $v_2 := \begin{pmatrix} 1 \\ 0 \end{pmatrix}$ is an eigenvector of A to the eigenvalue 2. We will see soon that a $n \times n$ -matrix can have at most n different eigenvalues. Therefore, we found all eigenvalues in this example, i.e., $\sigma(A) = \{1, 2\}$. For the eigenspaces, we need to find all solutions to $Ax = x$ and $Ax = 2x$, respectively.

Writing $x = (x_1, x_2)$, we see that $Ax = x$ is equivalent to the linear system

$$\begin{aligned} 2x_1 + x_2 &= x_1, \\ x_2 &= x_2. \end{aligned}$$

We obtain that x_2 is arbitrary, but $x_1 + x_2 = 0$ from the first equation. Therefore,

$$E(A, 1) = \{(x_1, x_2) \in \mathbb{C}^2 : x_1 + x_2 = 0\} = \{\alpha \cdot v_1 : \alpha \in \mathbb{C}\},$$

with $v_1 := \begin{pmatrix} -1 \\ 1 \end{pmatrix}$. In the same way, we see that $Ax = 2x$ is equivalent to the linear system

$$\begin{aligned} 2x_1 + x_2 &= 2x_1, \\ x_2 &= 2x_2, \end{aligned}$$

which implies $x_2 = 0$. Moreover, x_1 is arbitrary and therefore

$$E(A, 2) = \{(x_1, x_2) \in \mathbb{C}^2 : x_2 = 0\} = \{\alpha \cdot e_1 : \alpha \in \mathbb{C}\},$$

where $v_2 = e_1 = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$.

We now discuss some more theoretic examples, before we turn to a systematic way for the actual computation of eigenvalues and the corresponding eigenspaces.

Example 9.8. The easiest matrix/mapping one might consider is the identity matrix I_n in \mathbb{R}^n . Just by definition, we know that $I_n x = x = 1 \cdot x$. This shows that every vector is an eigenvector to the eigenvalue 1. In other words, $E(I_n, 1) = \mathbb{R}^n$. Moreover, if we are looking for other eigenvalues $a \neq 1$, i.e., we ask for solutions of $I_n x = x = a \cdot x$, then we see that this equation is equivalent to $(1 - a) \cdot x = 0$, which has, for $a \neq 1$, only the solution $x = 0_n$. Therefore, $E(I_n, a) = \{0_n\}$ for all $a \neq 1$, and $\sigma(I_n) = \{1\}$.

Example 9.9. Next, we consider a diagonal matrix $D := \text{diag}(\lambda_1, \lambda_2, \dots, \lambda_n) \in \mathbb{R}^{n \times n}$ with all the diagonal entries different, i.e., $\lambda_i \neq \lambda_j$ for $i \neq j$, which also very easy to treat. Just by applying the matrix D to the unit vectors $e_1, \dots, e_n \in \mathbb{R}^n$ we see that $D e_k = \lambda_k \cdot e_k$ for all $k = 1, \dots, n$. That is (λ_k, e_k) is an eigenpair of D . Since all the diagonal entries are different, it is not hard to see that every eigenvector to the eigenvalue λ_k is a multiple of e_k . (In other words, an eigenvector to λ_k must have all but the k -th entry equal to zero.) Thus,

$$E(D, \lambda_k) = \{\alpha \cdot e_k : \alpha \in \mathbb{C}\}$$

for all $k = 1, \dots, n$, if all λ_k 's are different, and we obtain $\sigma(D) = \{\lambda_1, \dots, \lambda_n\}$.

The next example gives a combination of the two examples above. For simplicity, we only consider the case $n = 3$ with two different eigenvalues.

Example 9.10. Let us consider the matrix $D = \begin{pmatrix} 2 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 2 \end{pmatrix}$. We see again that $e_2 \in \mathbb{R}^3$ is the only eigenvector of D to $\lambda_2 = 1$, i.e., $E(D, 1) = \{\alpha \cdot e_2 : \alpha \in \mathbb{C}\}$. However, there are more eigenvectors to the eigenvalue $\lambda_1 = 2$, since $D e_1 = 2e_1$ and $D e_3 = 2e_3$. By solving the corresponding linear system $Dx = 2x$, we see that the eigenspace satisfies

$$E(D, 2) = \{(x_1, x_2, x_3) \in \mathbb{R}^3 : x_2 = 0\} = \{\alpha \cdot e_1 + \beta \cdot e_3 : \alpha, \beta \in \mathbb{R}\}.$$

(The last equality might be seen by noting that vectors in $E(D, 2)$ are just arbitrary in the first and last coordinate.) One can check that these are the only eigenvalues, and thus $\sigma(D) = \{1, 2\}$. So, again, the spectrum consists just of the diagonal elements.

The same is true for arbitrary diagonal matrices $D := \text{diag}(\lambda_1, \lambda_2, \dots, \lambda_n) \in \mathbb{C}^{n \times n}$, as we will prove soon. That is, the spectrum consists of the diagonal elements, i.e., $\sigma(D) = \{\lambda_1, \dots, \lambda_n\}$. But note that this set might contain less than n elements, if some of the λ_k 's are equal.

The last example shows that in some cases, the eigenspace to a given eigenvalue is **not just the multiple of one eigenvector**. We see that we actually have more eigenvectors to the same eigenvalue, which are not multiples of each other. (This property will later be called *linear independence*.) Although we already learned how to calculate (and state) also such an eigenspace, it is sometimes desirable to have a set of eigenvectors that satisfy certain assumptions, or at least that the choice of eigenvectors is unique (up to multiplication with a scalar). This is not the case in the above example, as we have some freedom in choosing a *basis* (aka. two linearly independent vectors that *span* the eigenspace). We will come back to this later, when we talk about vector spaces and their bases. In the meantime, we sometimes assume that all the eigenvalues of a matrix are different, which makes things much easier.

However, it is worth noting that an eigenspace $E(A, \lambda)$, for $\lambda \in \mathbb{C}$, is a *linear subspace* of \mathbb{C}^n . That is, for two vectors $v_1, v_2 \in E(A, \lambda)$ and $\alpha, \beta \in \mathbb{C}$, we have that

$$A(\alpha v_1 + \beta v_2) = \alpha A v_1 + \beta A v_2 = \lambda \alpha v_1 + \lambda \beta v_2 = \lambda \cdot (\alpha v_1 + \beta v_2).$$

That is, not only multiples of an eigenvector, but also the sum of two eigenvectors to the same eigenvalue λ , are eigenvectors of A associated with λ . Note that the vector $\alpha v_1 + \beta v_2$ is called a **linear combination** of the vectors v_1 and v_2 .

We now turn to the computation of the eigenvalues. Note that this is the new part, as the computation of the corresponding eigenvectors/-spaces may be done, e.g., by Gaussian elimination, as discussed in Section 2.3. For the calculation of the eigenvalues, we use that the equation $Ax = \lambda x$, whose solutions define the eigenspace to the eigenvalue λ , can be rewritten into the form

$$(A - \lambda I_n)x = 0.$$

(Just subtract λx from both sides and note that $Ax - \lambda x = (A - \lambda I_n)x$.)

Note that $A - \lambda I_n$ is just the matrix $A = (a_{ij})$ where we subtract λ in each diagonal entry, i.e.,

$$A - \lambda I_n = \begin{pmatrix} a_{11} - \lambda & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} - \lambda & \dots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \dots & a_{nn} - \lambda \end{pmatrix}.$$

Now note that this 'new' linear system is a homogeneous linear system, and we already know that this system has the *trivial solution* $x = 0$ for every $\lambda \in \mathbb{C}$. However, this is not an eigenvector. To obtain an eigenvector of a matrix, i.e., a solution $x \neq 0$ to $(A - \lambda I_n)x = 0$, we need that this homogeneous linear system has a *non-trivial* solution.

However, by Lemma 2.43, we know that this is possible if and only if the corresponding matrix $A - \lambda I_n \in \mathbb{R}^{n \times n}$ has rank strictly smaller than n . And by Theorem 2.50, this holds if and only if $A - \lambda I_n$ is not invertible, which is equivalent to $\det(A - \lambda I_n) = 0$. This reasoning leads to the following lemma.

Lemma 9.11. *Let $A \in \mathbb{C}^{n \times n}$. Then,*

$$\lambda \in \sigma(A) \iff \det(A - \lambda I_n) = 0.$$

Recall that $\lambda \in \sigma(A)$ just means that λ is an eigenvalue of A .

This shows that finding the eigenvalues of a matrix is the same as finding the roots of the function $\det(A - \lambda I_n)$ as a function of λ . This leads to the following definition.

Definition 9.12. For a given matrix $A \in \mathbb{C}^{n \times n}$ and $\lambda \in \mathbb{C}$ we define

$$p(\lambda) = p_A(\lambda) := \det(A - \lambda I_n),$$

which is called the **characteristic polynomial** of A .

(We omit the subscript A in p_A if the matrix is clear.)

From the definition of the determinant, see Definition 2.46, we see that p_A is indeed a polynomial. Moreover, we can see that it is of degree n . For this note that λ appears only linearly in every entry of A (if it appears at all). Since the determinant is the sum of products of exactly n entries of A , we see that the highest power of λ in $\det(A - \lambda I_n)$ is λ^n .

(Note that the product over the diagonal, i.e., $\prod_{i=1}^n (a_{ii} - \lambda)$, appears in any case as a summand.)

This will be helpful soon for some theoretical considerations, but let us first see in some examples how this can be used to calculate the eigenvalues.

Example 9.13. We consider again the matrix $A = \begin{pmatrix} 2 & 1 \\ 0 & 1 \end{pmatrix}$. For computing the eigenvalues, we need to find the roots of

$$p_A(\lambda) = \det(A - \lambda I_2) = \det \begin{pmatrix} 2 - \lambda & 1 \\ 0 & 1 - \lambda \end{pmatrix} = (2 - \lambda)(1 - \lambda).$$

It is clear that this equals zero if and only if $\lambda = 1$ or $\lambda = 2$. As these are the only zeros of p_A , we see that these are all eigenvalues of A , i.e., $\sigma(A) = \{1, 2\}$.

Example 9.14. If we consider again a diagonal matrix $D = \text{diag}(\lambda_1, \lambda_2, \dots, \lambda_n)$, we see that

$$p(\lambda) = \det(D - \lambda I_n) = (\lambda_1 - \lambda) \cdot (\lambda_2 - \lambda) \cdots (\lambda_n - \lambda) = \prod_{k=1}^n (\lambda_k - \lambda).$$

Again, $p(\lambda)$ is zero if and only if one of the terms in the product is zero, which holds if and only if $\lambda = \lambda_k$ for some $k = 1, \dots, n$. We obtain (again) that $\sigma(D) = \{\lambda_1, \dots, \lambda_n\}$.

Example 9.15. In the same way as in the last example, we can also compute that the eigenvalues of an upper triangular matrix A are just the diagonal elements, i.e., a_{kk} . (Verify this yourself!) However, note that the eigenvectors are, in general, not just the unit vectors e_k , as they would be for a diagonal matrix. To get the eigenvectors we still need to solve the linear systems

$$(A - a_{kk} I_n)x = 0$$

for $k = 1, \dots, n$. But since these matrices are also upper triangular, these linear systems are easy to solve by Gaussian elimination.

Let us now turn to our favorite running-example.

Example 9.16. Consider

$$A = \begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix}.$$

To find the eigenvalues, we compute

$$p(\lambda) = \det \begin{pmatrix} 1-\lambda & 2 \\ 3 & 4-\lambda \end{pmatrix} = (1-\lambda)(4-\lambda) - 6 = \lambda^2 - 5\lambda - 2.$$

The roots of p can be obtained, e.g., by “completing the square”, i.e., we write

$$p(\lambda) = \left(\lambda - \frac{5}{2}\right)^2 - \left(\frac{5}{2}\right)^2 - 2 = \left(\lambda - \frac{5}{2}\right)^2 - \frac{33}{4}.$$

By this, we see that the roots of p are those λ such that $(\lambda - \frac{5}{2})^2 = \frac{33}{4}$, and these are given by

$$\lambda_1 := \frac{1}{2}(5 + \sqrt{33}) \quad \text{and} \quad \lambda_2 := \frac{1}{2}(5 - \sqrt{33}).$$

(Alternatively, one can also use the quadratic formula: $\lambda^2 + p\lambda + q = 0$ for $\lambda_{1/2} = -\frac{p}{2} \pm \sqrt{\frac{p^2}{4} - q}$.) To compute the corresponding eigenvectors we first have to solve the linear systems $Ax = \lambda_1 x$ and $Ax = \lambda_2 x$. The first system of equations is

$$\begin{array}{rcl} (1 - \lambda_1)x_1 & + & 2x_2 = 0, \\ 3x_1 & + & (4 - \lambda_1)x_2 = 0. \end{array}$$

This system can clearly be solved by Gaussian elimination. We obtain

$$E(A, \lambda_1) = \left\{ (x_1, x_2) \in \mathbb{C}^2 : x_1 = \frac{-3 + \sqrt{33}}{6} \cdot x_2 \right\}.$$

In the same way, we obtain

$$E(A, \lambda_2) = \left\{ (x_1, x_2) \in \mathbb{C}^2 : x_1 = \frac{-3 - \sqrt{33}}{6} \cdot x_2 \right\}.$$

So, one may choose, e.g.,

$$v_1 = \begin{pmatrix} -3 + \sqrt{33} \\ 6 \end{pmatrix}$$

and

$$v_2 = \begin{pmatrix} -3 - \sqrt{33} \\ 6 \end{pmatrix}$$

as eigenvectors, such that we have the eigenpairs (λ_1, v_1) and (λ_2, v_2) . Note again that every multiple of v_k is again an eigenvector to λ_k .

The following example shows that some matrices do not have any real eigenvalues, although they look very simple.

Example 9.17. Consider the matrix

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

The characteristic polynomial is given by

$$p(\lambda) = \det \begin{pmatrix} -\lambda & -1 \\ 1 & -\lambda \end{pmatrix} = \lambda^2 + 1.$$

We see that the zeros of p are those λ with $\lambda^2 = -1$, and there is clearly no real solution, but the complex solutions $\lambda_1 = i$ and $\lambda_2 = -i$, where $i := \sqrt{-1}$. (This equation was actually the motivation for introducing complex numbers.)

For finding the eigenspace of A to $\lambda_1 = i$, we have to solve the linear system

$$\begin{aligned} -i \cdot x_1 - x_2 &= 0, \\ x_1 - i \cdot x_2 &= 0. \end{aligned}$$

The second equation is just the $(-i)$ -th multiple of the first, and we obtain

$$E(A, i) = \{(x_1, x_2) \in \mathbb{C}^2 : x_1 = i \cdot x_2\}.$$

In the same way,

$$E(A, -i) = \{(x_1, x_2) \in \mathbb{C}^2 : x_2 = i \cdot x_1\} = \{(x_1, x_2) \in \mathbb{C}^2 : x_1 = -i \cdot x_2\}.$$

(The last equality follows from the calculation rules for complex numbers.)

Possible eigenvectors are

$$v_1 = \begin{pmatrix} i \\ 1 \end{pmatrix} \quad \text{and} \quad v_2 = \begin{pmatrix} 1 \\ i \end{pmatrix},$$

such that we have the eigenpairs (i, v_1) and $(-i, v_2)$.

(Verify that v_1, v_2 satisfy $Av_1 = iv_1, Av_2 = -iv_2$ and that they are not multiples of each other.)

Let us finish with an “ 3×3 -example” with shorter solution.

Example 9.18. Consider the matrix

$$A = \begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 0 & 0 & 1 \end{pmatrix} \quad \text{such that} \quad A - \lambda I_3 = \begin{pmatrix} 1 - \lambda & 2 & 3 \\ 4 & 5 - \lambda & 6 \\ 0 & 0 & 1 - \lambda \end{pmatrix}$$

To calculate the characteristic polynomial, i.e., the determinant of $A - \lambda I_3$, we can use Laplace expansion along the last row (see Theorem 2.61) to obtain

$$\begin{aligned} p(\lambda) &= \det(A - \lambda I_3) = (-1)^6 \cdot (1 - \lambda) \cdot \det \begin{pmatrix} 1 - \lambda & 2 \\ 4 & 5 - \lambda \end{pmatrix} = (1 - \lambda)((1 - \lambda)(5 - \lambda) - 8) \\ &= (1 - \lambda)(\lambda^2 - 6\lambda - 3). \end{aligned}$$

Since this is zero if and only if one of the factors is zero, we already see that $\lambda_1 = 1$. The solutions of $\lambda^2 - 6\lambda - 3 = 0$ lead to the other two eigenvalues. We obtain

$$\lambda_1 = 1, \quad \lambda_2 = 3 + 2\sqrt{3} \quad \text{and} \quad \lambda_3 = 3 - 2\sqrt{3}.$$

Solving the linear systems $(A - \lambda_k I_3)x = 0$ to obtain some eigenvectors v_k , we obtain, e.g., that

$$v_1 = \begin{pmatrix} 0 \\ -3 \\ 2 \end{pmatrix}, \quad v_2 = \begin{pmatrix} -1 + \sqrt{3} \\ 2 \\ 0 \end{pmatrix} \quad \text{and} \quad v_3 = \begin{pmatrix} 1 + \sqrt{3} \\ -2 \\ 0 \end{pmatrix}.$$

(Again, every multiple of them would also be eigenvectors.)

In most of the above examples, the eigenspaces were just multiples of a single vector. This is not always the case. However, it is an important property that makes many considerations simpler. We now state a property of an eigenvalue that guarantees that the corresponding eigenspace is of this form.

For this, let us now recall the **Fundamental Theorem of Algebra** from Theorem 1.70. This result implies that the characteristic polynomial p , which is a polynomial of degree n , can be written (as long as we allow complex numbers) in the form

$$p(\lambda) = (\lambda_1 - \lambda)(\lambda_2 - \lambda) \cdots (\lambda_n - \lambda),$$

where the λ_k 's are the zeros of p , i.e., the eigenvalues of the corresponding matrix. This implies, in particular, that **there are exactly n eigenvalues**. However, note that some eigenvalues may appear more often. (In the same way as the polynomial x^2 has both of its roots at zero.)

That is, if there are only $k \leq n$ different eigenvalues $\lambda_1, \dots, \lambda_k$, then the characteristic polynomial p can be written as

$$p(\lambda) = (\lambda_1 - \lambda)^{\mu_1} (\lambda_2 - \lambda)^{\mu_2} \cdots (\lambda_k - \lambda)^{\mu_k},$$

where $\mu_1, \dots, \mu_k \in \mathbb{N}$ are such that $\mu_1 + \cdots + \mu_k = n$.

This can be used to define the (algebraic) multiplicity of an eigenvalue.

Definition 9.19. Let $A \in \mathbb{C}^{n \times n}$. Then, we can write the characteristic polynomial of A in the form

$$p(\lambda) = (\lambda_1 - \lambda)^{\mu_1} (\lambda_2 - \lambda)^{\mu_2} \cdots (\lambda_k - \lambda)^{\mu_k}$$

where all $\lambda_i \in \mathbb{C}$ are different.

For $i = 1, \dots, k$, the integer μ_i is called **(algebraic) multiplicity** of the eigenvalue λ_i .

If $\mu_i = 1$ for some $1 \leq i \leq k$, then we say that λ_i is a **simple eigenvalue**.

Example 9.20. Consider the matrix

$$A = \begin{pmatrix} 4 & 0 & 0 \\ 0 & 4 & 0 \\ 0 & 0 & 3 \end{pmatrix},$$

which has clearly the eigenvalues 4 and 3. The characteristic polynomial is $p(\lambda) = (4 - \lambda)(4 - \lambda)(3 - \lambda) = (4 - \lambda)^2(3 - \lambda)$. Therefore, we see that 4 has algebraic multiplicity 2, whereas 3 is a simple eigenvalue.

The multiplicity of an eigenvalue λ_k gives information about the *dimension* of the corresponding eigenspace. We do not want to elaborate on this too much, as we will discuss this in a more general context later. Let us just state here (without proof), that the eigenspace of a simple eigenvalue is particularly simple.

Lemma 9.21. Let $A \in \mathbb{C}^{n \times n}$ and $\lambda \in \mathbb{C}$ be a simple eigenvalue of A . Then, there is some $v \in \mathbb{C}^n$ with $v \neq 0$, such that

$$E(A, \lambda) = \{\alpha v : \alpha \in \mathbb{C}\},$$

i.e., all eigenvectors to a simple eigenvalue are multiples of one vector.

The eigenspace to eigenvalues that are not simple is a bit more involved. Although the multiplicity still gives the number of *linear independent* vectors that are needed to characterize the eigenspace, there is quite some freedom in choosing them.

Let us finally state some useful properties of eigenvalues and eigenvectors.

First, note that there is a useful formula for the determinant of a matrix.

Theorem 9.22. *Let $A \in \mathbb{C}^{n \times n}$ with eigenvalues $\lambda_1, \lambda_2, \dots, \lambda_n$. Then,*

$$\det A = \prod_{k=1}^n \lambda_k.$$

Proof. The characteristic polynomial of A is given by

$$p(\lambda) = (\lambda_1 - \lambda)(\lambda_2 - \lambda) \cdots (\lambda_n - \lambda).$$

Thus,

$$\det A = \det(A - 0 \cdot I_n) = p(0) = \prod_{k=1}^n \lambda_k.$$

□

A very nice consequence of this is the following.

Corollary 9.23. *Let $A \in \mathbb{C}^{n \times n}$. Then, A is invertible (aka. regular) if and only if $0 \notin \sigma(A)$, i.e., 0 is not an eigenvalue of A .*

Proof. We know that $\det(A) = \prod_{k=1}^n \lambda_k$. Now, the product is zero if and only if one of its factors is zero.

□

Moreover, we can calculate all the eigenvalues of the inverse.

Lemma 9.24. *Let $A \in \mathbb{C}^{n \times n}$ be an invertible matrix and λ be an eigenvalue of A , i.e., $\lambda \in \sigma(A)$. Then, λ^{-1} is an eigenvalue of A^{-1} , i.e.,*

$$\frac{1}{\lambda} \in \sigma(A^{-1}).$$

Moreover, if (λ, v) is an eigenpair of A , then $(\frac{1}{\lambda}, v)$ is an eigenpair of A^{-1} .

Proof. Since λ is an eigenvalue, we have that

$$Av = \lambda v,$$

for some vector v . We apply A^{-1} on both sides of the equation to obtain

$$v = \lambda A^{-1} v.$$

Dividing by λ , which cannot be zero due to Corollary 9.23, implies $A^{-1}v = \frac{1}{\lambda} \cdot v$. Therefore, v is an eigenvector of A^{-1} to the eigenvalue λ^{-1} .

□

We will see soon that we can actually calculate the inverse matrix by knowing all the eigenvalues and eigenvectors, at least in some special cases.

9.2 Diagonalization

We will now see how eigenvalues and eigenvectors can be used to write matrices in a particular useful way. That is, we will show that, under some assumptions, a matrix A can be written as $A = V^T D V$, where D is a diagonal matrix, which is usually called *diagonalization* of the matrix. This will be essential to introduce (a special case of) the *singular value decomposition* in the next section.

For now, we restrict ourselves to **real and symmetric matrices**, i.e., we consider $A \in \mathbb{R}^{n \times n}$ with $A^T = A$.

Recall that the *inner product* of two vectors $x, y \in \mathbb{R}^n$ and the *Euclidean norm* of $x \in \mathbb{R}^n$ are defined by

$$\langle x, y \rangle := x^T y = \sum_{i=1}^n x_i y_i$$

and

$$\|x\| := \sqrt{\langle x, x \rangle} = \sqrt{\sum_{i=1}^n x_i^2},$$

respectively. If we consider a matrix $A \in \mathbb{R}^{n \times n}$, then we obtain $(Ax)^T y = x^T A^T y = x^T (A^T y)$, i.e.,

$$\langle Ax, y \rangle = \langle x, A^T y \rangle.$$

From this we easily obtain that two eigenvectors v_1, v_2 to different eigenvalues are *orthogonal*, meaning that $\langle v_1, v_2 \rangle = 0$.

Lemma 9.25. *Let $A \in \mathbb{R}^{n \times n}$ be symmetric, i.e., $A^T = A$, and let (λ_1, v_1) and (λ_2, v_2) be two eigenpairs of A with $\lambda_1 \neq \lambda_2$. Then,*

$$\langle v_1, v_2 \rangle = v_1^T v_2 = 0.$$

That is, eigenvectors to different eigenvalues are orthogonal.

Proof. From the above equation we obtain for symmetric matrices that

$$\langle Ax, y \rangle = \langle x, Ay \rangle$$

for all $x, y \in \mathbb{R}^n$. Moreover, since $Av_1 = \lambda_1 v_1$ and $Av_2 = \lambda_2 v_2$, we have

$$\lambda_1 \langle v_1, v_2 \rangle = \langle \lambda_1 v_1, v_2 \rangle = \langle Av_1, v_2 \rangle = \langle v_1, Av_2 \rangle = \langle v_1, \lambda_2 v_2 \rangle = \lambda_2 \langle v_1, v_2 \rangle.$$

As $\lambda_1 \neq \lambda_2$ by assumption, this is only possible if $\langle v_1, v_2 \rangle = 0$. □

Remark 9.26. By considering the *complex inner product* $\langle x, y \rangle = x^* y$, where x^* denotes the adjoint of $x \in \mathbb{C}^n$, see Remark 2.14, one can prove that **symmetric matrices have only real eigenvalues** and that one can choose corresponding eigenvectors that have only real entries. Since this is not necessary in the following, we omit a detailed proof.

The diagonalization will be based on the orthogonality of the eigenvectors. In fact, we will build a new matrix whose columns are the eigenvectors, and this matrix then has a special property that motivates the next definition.

Definition 9.27. Let $O \in \mathbb{R}^{n \times n}$ be such that

$$O^T O = I_n,$$

then we call O an **orthogonal matrix**.

(Equivalently, O is an orthogonal matrix if and only if $O^{-1} = O^T$.)

Note that orthogonal matrices are clearly invertible. This definition might appear very special, but it is an important type of matrices that appears in many applications. Easy 2×2 -examples of orthogonal matrices are the identity $I_2 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$, or the matrix $\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$. However, if a matrix is orthogonal or not is usually verified by the next equivalent characterization.

Lemma 9.28. *A matrix $O = (c_1, \dots, c_n) \in \mathbb{R}^{n \times n}$ is orthogonal if and only if all columns of O are pairwise orthogonal, i.e.,*

$$\langle c_i, c_j \rangle = 0 \quad \text{for all } i \neq j,$$

and all columns are normalized, i.e.,

$$\langle c_i, c_i \rangle = 1 \quad \text{for all } i = 1, \dots, n.$$

That is, O is orthogonal if and only if $\langle c_i, c_j \rangle = \delta_{ij}$.

Proof. Since the columns of O are c_1, \dots, c_n , we obtain directly from the definition of the matrix product that $(O^T O)_{ij} = c_i^T c_j = \langle c_i, c_j \rangle$. The condition $\langle c_i, c_j \rangle = \delta_{ij}$ is therefore equivalent to O being orthogonal. \square

Remark 9.29. Orthogonal matrices have several important properties that will be discussed later in more detail. Let us just note that an orthogonal matrix $O \in \mathbb{R}^{n \times n}$ satisfies $\langle Ox, Oy \rangle = \langle x, y \rangle$ and therefore $\|Ox\| = \|x\|$ for all $x, y \in \mathbb{R}^n$. (Verify this!) This means that O , considered as a mapping, does not change the length of the input, and does not change the 'angle' between two vectors. Therefore, at least for $n = 2, 3$, orthogonal matrices can be considered just as **rotations and reflections**.

Example 9.30. Verify that $\det(O) \in \{-1, 1\}$ for any orthogonal matrix $O \in \mathbb{R}^{n \times n}$.

From the above, we know that eigenvectors to different eigenvalues are orthogonal. To build up an orthogonal matrix from these vectors, we also need them to be normalized. That is, we need $\langle v, v \rangle = 1$ ($\Leftrightarrow \|v\| = 1$) for every eigenvector v . Such a vector, i.e., $v \in \mathbb{R}^n$ with $Av = \lambda v$ and $\|v\| = 1$, is called **normalized eigenvector** to the eigenvalue λ .

Luckily, we know that every multiple of an eigenvector is also an eigenvector to the same eigenvalue. We can therefore just divide the vector by its norm. That is, if (λ, v) is an eigenpair of A , then

$$\bar{v} := \frac{1}{\|v\|} \cdot v$$

is a normalized eigenvector to λ .

With this, we can now give the following important result.

Theorem 9.31. *Let $A \in \mathbb{R}^{n \times n}$ be symmetric with eigenvalues $\lambda_1, \dots, \lambda_n \in \mathbb{R}$ and corresponding pairwise orthogonal normalized eigenvectors $v_1, \dots, v_n \in \mathbb{R}^n$. Then,*

$$A = VDV^T,$$

where

$$V := (v_1, \dots, v_n)$$

is an orthogonal matrix and

$$D = \begin{pmatrix} \lambda_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \lambda_n \end{pmatrix}.$$

Note that, by Lemma 9.25, the eigenvectors are guaranteed to be pairwise orthogonal, if all eigenvalues are different.

Proof. By the assumptions, as well as Lemma 9.25 and Lemma 9.28, we know that V is an orthogonal matrix. Therefore, $V^T = V^{-1}$ and, by multiplying the equation with V from the right, leads to

$$A = VDV^T \iff AV = VDV^TV = VD.$$

The result finally follows from

$$AV = A(v_1, \dots, v_n) = (Av_1, \dots, Av_n) = (\lambda_1 v_1, \dots, \lambda_n v_n) = VD.$$

□

Let us consider some examples.

Example 9.32. Consider the matrix

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

Clearly, A is symmetric, and one can compute that it has the eigenvalues $\lambda_1 = 1$ and $\lambda_2 = -1$ with corresponding eigenvectors

$$v_1 = \begin{pmatrix} 1 \\ 1 \end{pmatrix} \quad \text{and} \quad v_2 = \begin{pmatrix} -1 \\ 1 \end{pmatrix}.$$

We see that $\|v_1\| = \|v_2\| = \sqrt{2}$, and compute the normalized eigenvectors

$$\bar{v}_1 = \frac{1}{\sqrt{2}} \cdot \begin{pmatrix} 1 \\ 1 \end{pmatrix} \quad \text{and} \quad \bar{v}_2 = \frac{1}{\sqrt{2}} \cdot \begin{pmatrix} -1 \\ 1 \end{pmatrix}.$$

If we define the matrix

$$V = (\bar{v}_1, \bar{v}_2) = \begin{pmatrix} \frac{1}{\sqrt{2}} & \frac{-1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{pmatrix},$$

we obtain that

$$A = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} = \begin{pmatrix} \frac{1}{\sqrt{2}} & \frac{-1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} \begin{pmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ \frac{-1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{pmatrix} = VDV^T$$

with $D = \text{diag}(1, -1)$.

Example 9.33. Consider the matrix

$$A = \begin{pmatrix} 1 & 2 \\ 2 & 1 \end{pmatrix}.$$

Again, A is symmetric, and one can compute that it has the eigenvalues $\lambda_1 = 3$ and $\lambda_2 = -1$. We obtain the normalized eigenvectors

$$v_1 = \begin{pmatrix} \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{pmatrix} \quad \text{and} \quad v_2 = \begin{pmatrix} -\frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{pmatrix},$$

which are the same as in the last example. Therefore, we obtain

$$A = \begin{pmatrix} 1 & 2 \\ 2 & 1 \end{pmatrix} = \begin{pmatrix} \frac{1}{\sqrt{2}} & \frac{-1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{pmatrix} \begin{pmatrix} 3 & 0 \\ 0 & -1 \end{pmatrix} \begin{pmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ \frac{-1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{pmatrix} = VDV^T$$

with $V = (v_1, v_2)$ and $D = \text{diag}(3, -1)$.

The last examples show that it may happen that two matrices can be '*diagonalized*' by using the same matrix V . Such matrices are somehow "acting in the same directions".

Besides the fact that the diagonalization of a matrix can be used to extract certain information of a matrix A , it is also very useful to compute **powers of matrices**, i.e., A^ℓ for some $\ell \in \mathbb{Z}$.

Corollary 9.34. *Let $\ell \in \mathbb{Z}$ and $A = VDV^T \in \mathbb{R}^{n \times n}$ be as in Theorem 9.31, then*

$$A^\ell = VD^\ell V^T,$$

where $D^\ell = \text{diag}(\lambda_1^\ell, \dots, \lambda_n^\ell)$. (For $\ell < 0$ we need here that A is invertible, i.e., $\lambda_k \neq 0$.)

In particular, if (λ, v) is an eigenpair of A , then (λ^ℓ, v) is an eigenpair of A^ℓ .

Proof. Since $V^T V = VV^T = I_n$, we see that

$$A^2 = (VDV^T)(VDV^T) = VD(V^T V)DV^T = VD^2V^T.$$

By induction, this can be extended to all $\ell \in \mathbb{N}$.

Moreover, the statement is clear for $\ell = 0$, i.e., $A^0 = I_n = VV^T$. For $\ell < 0$ we use Lemma 9.24, which shows that A^{-1} has the same eigenvectors as A with corresponding eigenvalues λ replaced by λ^{-1} . This shows $A^{-1} = VD^{-1}V^T$. By the same arguments as above, we obtain the result for all $\ell \in \mathbb{Z}$. \square

We will see below how to actually compute the power of a matrix in a fast way.

Remark 9.35. The last result was only stated for powers of matrices. We will see later that also other functions of matrices can be handled in this way. That is, that $f(A) = Vf(D)V^T$, where $f(D) = \text{diag}(f(\lambda_1), \dots, f(\lambda_n))$, under certain assumptions on the function f .

Remark 9.36. Note that the assumptions posed in Theorem 9.31 are only necessary to state the result using only real numbers. A corresponding result also holds in the complex case, but the statement would be a bit more involved, and would require to work with the complex inner product. We omit the details here.

9.3 Singular value decomposition

We now use the result from the last section to show that matrices can be written as the sum of very easy matrices, namely **rank-one matrices**, at least under some assumption. Recall that a rank-one matrix is a matrix that can be written as $xy^T \in \mathbb{R}^{n \times n}$ for some $x, y \in \mathbb{R}^n$. Such matrices, and operations with them, are very easy to handle, also when it comes to very high-dimensional applications. It is therefore clearly desirable to write large matrices by these 'easy building blocks'. This may remind one to the study of *Fourier series*, where complicated functions are to be written by easier ones.

One of the most fundamental results of this chapter is the following. As we consider only the case of **square, real and symmetric** matrices here, this result is only a special case of the *singular value decomposition*, which is also called **eigendecomposition** of a matrix.

Theorem 9.37. *Let $A \in \mathbb{R}^{n \times n}$ be symmetric with eigenvalues $\lambda_1, \dots, \lambda_n \in \mathbb{R}$ and corresponding pairwise orthogonal normalized eigenvectors $v_1, \dots, v_n \in \mathbb{R}^n$. Then,*

$$A = \sum_{k=1}^n \lambda_k \cdot v_k v_k^T.$$

If the eigenvectors are not normalized, we can write

$$A = \sum_{k=1}^n \frac{\lambda_k}{\langle v_k, v_k \rangle} \cdot v_k v_k^T.$$

Recall that $v_k v_k^T \in \mathbb{R}^{n \times n}$, and that $\langle v_k, v_k \rangle = \|v_k\|^2 = v_k^T v_k$.

Proof. From Theorem 9.31 we know, under the assumption of the theorem, that $A = VDV^T$ with $V = (v_1, \dots, v_n)$ and $D = \text{diag}(\lambda_1, \dots, \lambda_n)$. We now define the matrices E_k as the matrices that contain only a single one on the diagonal in the kk -th entry, and are zero otherwise, i.e., $E_k = (0, \dots, 0, e_k, 0, \dots, 0)$. We therefore have

$$D = \sum_{k=1}^n \lambda_k \cdot E_k,$$

which implies that

$$A = VDV^T = \sum_{k=1}^n \lambda_k \cdot VE_k V^T.$$

If we finally note that $VE_k = (0, \dots, 0, v_k, 0, \dots, 0)$, we obtain that $VE_k V^T = v_k v_k^T$.

The second statement of the theorem follows just by normalization of the (unnormalized) eigenvectors. \square

Note that all terms in the above sum that corresponds to the eigenvalue 0 just disappear. It is therefore not necessary to compute eigenvectors to the eigenvalue 0.

As stated earlier, a similar result would hold for general (even non-square) matrices. We omit the details, but note that the basic idea is the same as stated above.

It turns out that such decompositions are the foundation of the majority of numerical methods for high-dimensional applications. A very basic method to *approximate matrices* is to consider only some of the terms in the above sum (similarly as we have done it with the partial sums of Fourier series). Namely, for a large matrix (i.e., large n) we take only those terms corresponding to large eigenvalues, as they are 'more important'. This is also called *low-rank (matrix) approximation*.

Let us see some examples.

Example 9.38. Consider the matrix

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

We already computed the eigenvalues $\lambda_1 = 1$ and $\lambda_2 = -1$ with corresponding eigenvectors

$$v_1 = \begin{pmatrix} 1 \\ 1 \end{pmatrix} \quad \text{and} \quad v_2 = \begin{pmatrix} -1 \\ 1 \end{pmatrix}.$$

We see that $\|v_1\|^2 = \|v_2\|^2 = 2$, and that

$$v_1 v_1^T = \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix} \quad \text{and} \quad v_2 v_2^T = \begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix}.$$

We obtain that

$$A = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} = \frac{1}{2} \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix} + \frac{-1}{2} \begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix} = \frac{\lambda_1}{\langle v_1, v_1 \rangle} v_1 v_1^T + \frac{\lambda_2}{\langle v_2, v_2 \rangle} v_2 v_2^T.$$

Example 9.39. We consider again the matrix

$$A = \begin{pmatrix} 1 & 2 \\ 2 & 1 \end{pmatrix}$$

with the eigenvalues $\lambda_1 = 3$ and $\lambda_2 = -1$, see Example 9.33. As above, we obtain

$$A = \begin{pmatrix} 1 & 2 \\ 2 & 1 \end{pmatrix} = \frac{3}{2} \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix} + \frac{-1}{2} \begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix}.$$

If we combine this representation with the findings of Corollary 9.34, i.e., that $A^\ell = VD^\ell V^T$ for $A = VDV^T$, we can compute powers, and the inverse, of symmetric matrices in a fast way.

Corollary 9.40. Let $\ell \in \mathbb{Z}$ and $A \in \mathbb{R}^{n \times n}$ be as in Theorem 9.37, then

$$A^\ell = \sum_{k=1}^n \frac{\lambda_k^\ell}{\langle v_k, v_k \rangle} \cdot v_k v_k^T.$$

(For $\ell < 0$ we need here that A is invertible, i.e., $0 \notin \sigma(A)$.)

Example 9.41. Let us compute A^4 for the matrix

$$A = \begin{pmatrix} 1 & 2 \\ 2 & 1 \end{pmatrix},$$

see Example 9.33. Since we know from Corollary 9.34 that the eigenpairs of A^4 are given by $(\lambda_1^4, v_1) = (3^4, v_1) = (81, v_1)$ and $(\lambda_2^4, v_2) = ((-1)^4, v_2) = (1, v_2)$, with v_1 and v_2 as above. We obtain that

$$\begin{aligned} A^4 &= \begin{pmatrix} 1 & 2 \\ 2 & 1 \end{pmatrix}^4 = \frac{\lambda_1^4}{\langle v_1, v_1 \rangle} v_1 v_1^T + \frac{\lambda_2^4}{\langle v_2, v_2 \rangle} v_2 v_2^T = \frac{81}{2} \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix} + \frac{1}{2} \begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix} \\ &= \begin{pmatrix} 41 & 40 \\ 40 & 41 \end{pmatrix}. \end{aligned}$$

In the same way, we can easily calculate the inverse of A by

$$\begin{aligned} A^{-1} &= \begin{pmatrix} 1 & 2 \\ 2 & 1 \end{pmatrix}^{-1} = \frac{\lambda_1^{-1}}{\langle v_1, v_1 \rangle} v_1 v_1^T + \frac{\lambda_2^{-1}}{\langle v_2, v_2 \rangle} v_2 v_2^T = \frac{1}{6} \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix} + \frac{-1}{2} \begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix} \\ &= \frac{1}{6} \begin{pmatrix} -2 & 4 \\ 4 & -2 \end{pmatrix}. \end{aligned}$$

Let us finally comment on the case of **non-symmetric**, or even **non-square** matrices $A \in \mathbb{R}^{m \times n}$ for some $m, n \in \mathbb{N}$. If $m \neq n$, then we cannot write down an eigenvalue problem for this kind of matrices. Since $Ax \in \mathbb{R}^n$ for $x \in \mathbb{R}^m$, it is clear that multiplication by a number (aka. eigenvalue) cannot achieve a change of the dimension. However, what we can do is to have a look at the matrices

$$A^T A \in \mathbb{R}^{n \times n} \quad \text{and} \quad AA^T \in \mathbb{R}^{m \times m}.$$

Since $(AA^T)^T = (A^T)^T A^T = AA^T$ and vice versa, both matrices are symmetric. We can therefore apply Theorem 9.37 to these matrices, and, with this, we can write also non-square and non-symmetric matrices as a sum of rank-one matrices similar to Theorem 9.37. However, formulas and assumptions would become a bit more involved. We omit the details.

10 Basic measure theory and the Lebesgue integral

We now want to introduce a different approach for the definition of an integral which is called the *Lebesgue integral*. This concept is currently considered as *the* appropriate integral for many problems and applications in modern science. It allows to apply the so called *standard machinery*, which is a very general proof technique that allows to give simple proofs, even for strong results. The main advantage of this whole concept is that it provides a **general framework** for many disciplines of mathematics, including integration theory, computation of series, and probability theory. For this, we need to introduce the concept of a *measure*, which generalizes the idea of a volume or length. However, as everything comes at a price, we will always need to specify which classes of sets we are considering and need to consider functions that are *measurable* with respect to these sets. This approach, basically due to *Henri Lebesgue* (1875–1941) and *Emile Borel* (1871–1956), is the foundation of *measure theory* and is one of the cornerstones of theoretical science. In this chapter, we will learn what it means for a function to be *measurable* and *integrable* (for a given measure), and we will show general results for such functions. We will see at the end how this implies powerful tools for working with integrals.

To get some intuition what issues could appear when talking about the measure/area/volume of a set we do a short discussion. We all know how to compute the length (measure) of an interval $[a, b]$: it is just $b - a$. Also if we consider the union or intersection of several intervals it is not so hard to calculate the area of the corresponding set. However, a more complicated question is how to measure the set $\mathbb{Q} \cap [0, 1]$ of rational points in $[0, 1]$. Intuitively, it has to be 'somewhere between 0 and 1'. Moreover, the measures of rational and irrational points should add up to 1, which is the length of $[0, 1]$. Unfortunately, the definition of the area as the common limit of the lower and upper sums (see Section 8.7) applied to the indicator function does not work, because the lower sums would all equal zero and the upper sums equal one, i.e., $\mathbb{Q} \cap [0, 1]$ is not Jordan-measurable. This shows that we need a different concept of a measure.

For this, let us assume that we have some set Ω , which we call **ground set**. (One may think of $\Omega = \mathbb{R}$ or $\Omega = \mathbb{N}$ or $\Omega = \{1, \dots, n\}$.) Ideally, we would like to assign a measure to arbitrary subsets of Ω , i.e., to all $A \in \mathcal{P}(\Omega)$ where $\mathcal{P}(\Omega)$ is the **power set** $\mathcal{P}(\Omega) := \{A : A \subset \Omega\}$.

However, this is in general not possible, and we want to find the 'largest' set of subsets of Ω that can be treated. That is, we need to define which **subsets we want to assign a measure**. For this, let us start with some formalism that is needed to describe the theory.

Definition 10.1. Let Ω be some set and $\mathcal{A} \subset \mathcal{P}(\Omega)$ be a family of subsets of Ω .

We call \mathcal{A} a **σ -algebra** (over Ω) if the following properties hold

- 1) $\Omega \in \mathcal{A}$,
- 2) $A \in \mathcal{A} \implies A^c = \Omega \setminus A \in \mathcal{A}$,
- 3) $A_1, A_2, \dots \in \mathcal{A} \implies \bigcup_{i \in \mathbb{N}} A_i \in \mathcal{A}$.

So, a σ -algebra is a set of sets which is closed under complement and countable unions.

For a set Ω and a σ -algebra $\mathcal{A} \subset \mathcal{P}(\Omega)$, we call the tuple (Ω, \mathcal{A}) a **measurable space**, and the sets $A \in \mathcal{A}$ **measurable sets** (sometimes \mathcal{A} -measurable sets).

Other common ways to denote a σ -algebra are \mathcal{F} or Σ .

The definition of a σ -algebra only considers the complement and the countable union. However, it is important to note that these three properties also imply that

- 4) $\emptyset \in \mathcal{A}$,
- 5) $A_1, A_2, \dots \in \mathcal{A} \implies \bigcap_{i \in \mathbb{N}} A_i \in \mathcal{A}$, and
- 6) $A, B \in \mathcal{A} \implies A \setminus B \in \mathcal{A}$.

That is, the definition of an σ -algebra \mathcal{A} implies that the empty set is also in \mathcal{A} , that countable intersections of sets from \mathcal{A} also belong to \mathcal{A} , and that differences of two sets are in \mathcal{A} . To see this, we only need

to employ the fact that $A^c \in \mathcal{A}$ for all $A \in \mathcal{A}$ (Property 2), *De Morgan's law* $\bigcap_{i \in \mathbb{N}} A_i = (\bigcup_{i \in \mathbb{N}} A_i^c)^c$ and that $A \setminus B = A \cap B^c$.

Remark 10.2. Clearly, by the definition of a σ -algebra \mathcal{A} , we also have that finite unions and intersections are again in \mathcal{A} . That is, $A_1, \dots, A_n \in \mathcal{A}$ implies that $A_1 \cup \dots \cup A_n \in \mathcal{A}$ and $A_1 \cap \dots \cap A_n \in \mathcal{A}$. (Just set $A_{n+1} = A_{n+2} = \dots = \emptyset$ and use the result for countable unions.)

Let us see some examples.

Example 10.3. For any set Ω we have that $\{\emptyset, \Omega\}$ is a σ -algebra, called the *trivial σ -algebra*. (Verify this!)

Example 10.4. For any set Ω we have that $\mathcal{P}(\Omega)$ is a σ -algebra.

Proof. Clearly, $\emptyset, \Omega \in \mathcal{P}(\Omega)$. Furthermore, for any $A \subset \Omega$ we have $\Omega \setminus A \subset \Omega$, hence $A^c \in \mathcal{P}(\Omega)$. Last but not least for $A_1, A_2, \dots \subset \Omega$ we have $\bigcup_{n \in \mathbb{N}} A_n \subset \Omega$. This makes $\mathcal{P}(\Omega)$ a σ -algebra. \square

Example 10.5. To see a specific simple σ -algebra, consider $\Omega = \{1, 2, 3\}$, i.e., a set that only contains 3 elements. We clearly obtain $\mathcal{P}(\Omega) = \{\emptyset, \{1\}, \{2\}, \{3\}, \{1, 2\}, \{1, 3\}, \{2, 3\}, \{1, 2, 3\}\}$, which makes $(\Omega, \mathcal{P}(\Omega))$ a measurable space.

Another σ -algebra over Ω is $\mathcal{A} = \{\emptyset, \{1\}, \{2, 3\}, \{1, 2, 3\}\}$. (Verify that this is a σ -algebra!)

We now define what we mean by a measure. Note that this is actually a very easy and natural definition, as this is really what we would assume a measure/area/volume to satisfy. It is remarkable that such a basic definition (together with the structure of a corresponding σ -algebra) leads to rather powerful results in the sequel.

Let us recall that two sets $A_1, A_2 \subset \Omega$ are called disjoint if $A_1 \cap A_2 = \emptyset$, i.e., if their intersection is empty. If we now have more than two sets, say $A_1, A_2, \dots \subset \Omega$, then we still call them disjoint if $\bigcap_{i=1}^{\infty} A_i = \emptyset$, i.e., there is no $x \in \Omega$ contained in all A_i . However, this is not enough for most purposes, and we call $A_1, A_2, \dots \subset \Omega$ **pairwise disjoint** if $A_i \cap A_j = \emptyset$ for every $i \neq j$. Note that, e.g., the sets $A_1 = \{1\}$, $A_2 = \{2\}$ and $A_3 = \{1, 2\}$ are disjoint, but not pairwise disjoint.

Definition 10.6. Let (Ω, \mathcal{A}) be a measurable space.

A function $\mu: \mathcal{A} \rightarrow [0, \infty]$ is called **measure** if

- 1) $\mu(\emptyset) = 0$
- 2) For pairwise disjoint $A_1, A_2, \dots \in \mathcal{A}$ we have $\mu(\bigcup_{i \in \mathbb{N}} A_i) = \sum_{i \in \mathbb{N}} \mu(A_i)$.

The second property is called σ -additivity.

Such a triple $(\Omega, \mathcal{A}, \mu)$ is called a **measure space**.

A measure is therefore a non-negative function on a σ -algebra \mathcal{A} , i.e., it assigns a non-negative number to each set $A \subset \mathcal{A}$, such that it 'behaves well' if we consider the union of disjoint sets. Note that the measure of a set is allowed to be infinity.

There is one class of sets that will appear rather often, and therefore get their own definition. Those are the sets of measure zero.

Definition 10.7 (Null set). Let $(\Omega, \mathcal{A}, \mu)$ be a measure space.

A set $N \in \mathcal{A}$ such that $\mu(N) = 0$ is called **null set**.

When we consider measures on the real line, then typical examples of null sets will be just single points $\{p\}$, $p \in \mathbb{R}$, which can be seen as closed intervals of 'length' zero. However, by σ -additivity of a measure, we see that also countable unions of disjoint null sets are null sets, showing that also \mathbb{N} or \mathbb{Q} are examples of null sets, and there are even more complicated ones. We come back to this soon.

Before we turn to some examples, let us note that some properties of a measure follow easily from the definition. The proof is left to the reader.

Lemma 10.8. *Let $(\Omega, \mathcal{A}, \mu)$ be a measure space.*

Then, for all $A, B, A_1, A_2, \dots \in \mathcal{A}$, we have

- 1) $\mu(A) \leq \mu(B)$ if $A \subset B$, and (Monotonicity)
- 2) $\mu(B \setminus A) = \mu(B) - \mu(A)$ if $A \subset B$ and $\mu(A) < \infty$, (Differences)
- 3) $\mu(\bigcup_{i \in \mathbb{N}} A_i) \leq \sum_{i \in \mathbb{N}} \mu(A_i)$ for all $A_1, A_2, \dots \in \mathcal{A}$. (Subadditivity)

Let us start with some (discrete) toy examples, where we can state the σ -algebra (and therefore the measure) explicitly for every element of the σ -algebra.

Example 10.9. Let us consider the finite example from above, i.e., $\Omega = \{1, 2, 3\}$ with the σ -algebra $\mathcal{A} = \{\emptyset, \{1\}, \{2, 3\}, \{1, 2, 3\}\}$. A possible measure is just the number of elements $\mu(A) := \#A$ for all $A \in \mathcal{A}$, i.e., $\mu(\emptyset) = 0$, $\mu(\{1\}) = 1$, $\mu(\{2, 3\}) = 2$ and $\mu(\{1, 2, 3\}) = 3$, which clearly satisfies the properties of a measure. However, the function μ_1 with $\mu_1(\emptyset) = 0$, $\mu_1(\{1\}) = 1$, $\mu_1(\{2, 3\}) = 0$ and $\mu_1(\{1, 2, 3\}) = 1$ is also a measure with the additional null set $\{2, 3\}$. In contrast, the function μ_2 with $\mu_2(\emptyset) = 0$, $\mu_2(\{1\}) = 1$, $\mu_2(\{2, 3\}) = 2$ and $\mu_2(\{1, 2, 3\}) = 1$ is not a measure. (Why?)

The next example will later be important to see the close connection of integrals and series.

Example 10.10 (Counting measure). Let us consider the measure that assigns a set the number of its elements more detailed. For this let $\Omega \subset \mathbb{N}$ be some set together with the σ -algebra $\mathcal{P}(\Omega)$. Then, $\mu: \mathcal{P}(\Omega) \rightarrow [0, \infty]$ with

$$\mu(A) := \#A := \sum_{k \in \Omega} \chi_A(k)$$

is a measure, called the **counting measure** (on Ω).

To see this, first note that μ is non-negative and that $\mu(\emptyset) = 0$. So, we only have to check the σ -additivity. We will do a case distinction for $\#\Omega < \infty$ and $\#\Omega = \infty$. Note that $\#A = \infty$ means that A is a (countable) infinite set.

Now assume we have a sequence of pairwise disjoint sets $A_1, A_2, \dots \in \mathcal{P}(\Omega)$.

Case 1: Assume $\#\Omega = n < \infty$. Thus only finitely many A_i of the above sequence can be non-empty, say $m \leq n$ of the sets are non-empty. W.l.o.g. the first m are non-empty. Moreover, the disjointness of these sets yields that if an element x is contained in $\bigcup_{i=1}^m A_i$, then it is contained in exactly one of the A_i .

Hence

$$\mu\left(\bigcup_{i=1}^{\infty} A_i\right) = \mu\left(\bigcup_{i=1}^m A_i\right) = \#\bigcup_{i=1}^m A_i = \sum_{i=1}^m \#A_i = \sum_{i=1}^m \mu(A_i) = \sum_{i=1}^{\infty} \mu(A_i).$$

As this holds for arbitrary $A_1, A_2, \dots \in \mathcal{P}(\Omega)$, we see that μ is a measure in this case.

Case 2: Assume that $\#\Omega = \infty$. If $\bigcup_{i=1}^{\infty} A_i$ contains only finitely many elements, then the same arguments as in the first part of the example can be used to see the σ -additivity.

If on the other hand $\bigcup_{i=1}^{\infty} A_i$ contains infinitely many elements, i.e., $\mu(\bigcup_{i=1}^{\infty} A_i) = \infty$, there are two

possibilities, namely (at least) one of the A_i contains infinitely many elements or all of them contain only a finite number of elements. In the first case, i.e., if there is some A_k with $\#A_k = \infty$, we obtain

$$\infty = \#A_k \leq \sum_{i=1}^{\infty} \#A_i \leq \infty,$$

which implies that

$$\mu\left(\bigcup_{i=1}^{\infty} A_i\right) = \infty = \sum_{i=1}^{\infty} \#A_i = \sum_{i=1}^{\infty} \mu(A_i).$$

If none of the A_i is infinite, i.e., $\#A_i < \infty$, then we can w.l.o.g. assume that all of them contain at least one element. This is, because we can just omit all empty sets, and there cannot be only finitely many sets left, because this would contradict, that the union has infinitely many elements. So, for each A_i we have that $\#A_i \geq 1$, hence

$$\infty = \sum_{i=1}^{\infty} 1 \leq \sum_{i=1}^{\infty} \#A_i \leq \infty.$$

This shows that also in the case $\#\Omega = \infty$ the counting measure satisfies all requirements to be a measure (as expected).

The next example is very much related to the regularly appearing *Dirac delta*.

Example 10.11. Given some measurable space (Ω, \mathcal{A}) , e.g. $(\mathbb{N}, \mathcal{P}(\mathbb{N}))$, and an element $x \in \Omega$, then we define the **Dirac measure** at x by

$$\delta_x(A) = \begin{cases} 1 & \text{if } x \in A \\ 0 & \text{else.} \end{cases}$$

That is, we give measure one to every set that contains the element x , and measure zero to all other sets. Let us check that this is a measure:

The property $\delta_x(\emptyset) = 0$ is clear. Given a sequence of disjoint sets $A_1, A_2, \dots \in \mathcal{A}$ we see that

$$x \in \bigcup_{i=1}^{\infty} A_i \iff x \in A_i \text{ for exactly one index } i \in \mathbb{N}.$$

(Note that this is only true because all the A_i are disjoint!) This immediately implies that

$$\sum_{i=1}^{\infty} \delta_x(A_i) = \begin{cases} 1, & \text{if } x \in \bigcup_{i=1}^{\infty} A_i, \\ 0, & \text{otherwise.} \end{cases}$$

Thus

$$\sum_{i=1}^{\infty} \delta_x(A_i) = \delta_x\left(\bigcup_{i=1}^{\infty} A_i\right),$$

proving that δ_x is a measure on (Ω, \mathcal{A}) .

The above examples are rather easy because the measure of every element of the corresponding σ -algebra can be written down explicitly. But in many cases this is not possible. One may think on our example $\mathbb{Q} \cap [0, 1]$, where we actually don't have a precise definition yet. However, we know what the measure on \mathbb{R} we are looking for should satisfy: It should assign every interval its length. We may now hope that this requirement already fixes the measure for sufficiently general sets. In other words, following a more practical approach, we want to just specify the measure for 'nice' sets of a σ -algebra \mathcal{A} and want to say that there is a unique measure on \mathcal{A} we can work with, although it is not a priori clear how one can calculate the measure of a specific set. Luckily, this is possible and will ultimately lead to the definition (or construction) of the *Lebesgue measure*.

For this, let us assume we have a family of subsets $\mathcal{E} \subset \mathcal{P}(\Omega)$ of some ground set Ω . We now denote by $\sigma(\mathcal{E})$ the **smallest σ -algebra that contains \mathcal{E}** . Roughly speaking, we take \mathcal{E} and check the conditions

of a σ -algebra. If now a complement or union (or intersection) of sets in \mathcal{E} is not in \mathcal{E} , then we include it, and continue this procedure. A formal definition is given by the *intersection of all σ -algebras* that contain \mathcal{E} , i.e.,

$$\sigma(\mathcal{E}) := \bigcap \{\mathcal{A} \subset \mathcal{P}(\Omega): \mathcal{A} \text{ is a } \sigma\text{-algebra with } \mathcal{E} \subset \mathcal{A}\}.$$

Note that $\mathcal{P}(\Omega)$ is always a σ -algebra that contains \mathcal{E} . Moreover, the intersection of σ -algebras is also a σ -algebra (Verify this!), which makes this intersection well-defined. Moreover, if $\mathcal{A} = \sigma(\mathcal{E})$ for some $\mathcal{E} \subset \mathcal{P}(\Omega)$, then \mathcal{E} is called a **generator** of the σ -algebra \mathcal{A} . Although this is a rather informal definition, it is enough for our purposes here. We just need to mind that $\sigma(\mathcal{E})$ is a σ -algebra that contains \mathcal{E} . Let us see some examples.

Example 10.12. For $\Omega = \{1, 2, 3\}$, we obtain that $\sigma(\{1\}) = \{\emptyset, \{1\}, \{2, 3\}, \{1, 2, 3\}\}$, because every σ -algebra contains \emptyset and Ω , and we need to include $\{2, 3\}$ as the complement of $\{1\}$.

Example 10.13. For $\Omega = [0, 1]$, i.e., the closed unit interval, we consider $\mathcal{E} = \{[0, \frac{1}{4}), [\frac{1}{4}, \frac{1}{2}]\}$. To obtain the smallest σ -algebra, we need to include \emptyset , Ω and all unions and complements. We obtain $\mathcal{E}' = \{\emptyset, [0, \frac{1}{4}), [\frac{1}{4}, \frac{1}{2}], [0, \frac{1}{2}), [\frac{1}{4}, 1], [0, \frac{1}{4}) \cup [\frac{1}{2}, 1], [0, 1]\}$. However, as this is still not a σ -algebra, we need to repeat this procedure to obtain

$$\sigma(\mathcal{E}) = \left\{ \emptyset, \left[0, \frac{1}{4}\right), \left[\frac{1}{4}, \frac{1}{2}\right), \left[0, \frac{1}{2}\right), \left[\frac{1}{4}, 1\right), \left[\frac{1}{2}, 1\right], \left[0, \frac{1}{4}\right) \cup \left[\frac{1}{2}, 1\right], [0, 1] \right\},$$

which is a σ -algebra.

If we now have a generator \mathcal{E} that is 'rich enough', then we obtain that the measure is uniquely specified on $\sigma(\mathcal{E})$ by its values on \mathcal{E} . This is due to the powerful *extension theorem* of *Constantin Caratheodory* (1873–1950).

Theorem 10.14 (Caratheodory's extension theorem).

Let $\mathcal{E} \subset \mathcal{P}(\Omega)$ and $\mu': \mathcal{E} \rightarrow [0, \infty]$ be a function with $\mu'(\emptyset) = 0$ that satisfies

1. $A \cup B \in \mathcal{E}$ for all $A, B \in \mathcal{E}$,
2. $A \setminus B \in \mathcal{E}$ for all $A, B \in \mathcal{E}$,
3. there exist $E_1, E_2, \dots \in \mathcal{E}$ with $\Omega = \bigcup_{i=1}^{\infty} E_i$ and $\mu'(E_i) < \infty$ for all i , and
4. $\mu'(\bigcup_{i \in \mathbb{N}} A_i) = \sum_{i \in \mathbb{N}} \mu'(A_i)$ for all pairwise disjoint $A_1, A_2, \dots \in \mathcal{E}$.

Then there exists a unique measure $\mu: \sigma(\mathcal{E}) \rightarrow [0, \infty]$ with $\mu(E) = \mu'(E)$ for all $E \in \mathcal{E}$.

A function μ' as above is also called a **pre-measure**, and μ the **extension** of μ' .

A proof of this theorem goes far beyond the scope of this lecture. Note that, however, this theorem allows to 'define' a (pre-)measure only on a set \mathcal{E} and we immediately obtain a measure μ on the σ -algebra $\sigma(\mathcal{E})$. The advantage is that, since μ is a measure, it satisfies all the properties introduced above for measures like monotonicity, σ -additivity etc. for all measurable sets, which makes it easier to work with, and, moreover, the uniqueness justifies that there is only one measure with this property.

Let us see how this works for the most important example: **The Lebesgue measure**.

To define the Lebesgue measure on sets $\Omega \subset \mathbb{R}^d$, let us define the set

$$\mathcal{E} := \left\{ \Omega \cap \bigcup_{i=1}^n A_i: n \in \mathbb{N} \text{ and } A_i \text{ is of the form } (a_1, b_1) \times (a_2, b_2) \times \cdots \times (a_d, b_d) \right\},$$

which are arbitrary finite unions of **open boxes** intersected with Ω . In one dimension these sets are finite unions of open intervals. (One might also use closed or half-open boxes, but this choice is the most classical one.)

The σ -algebra that is generated by this set, is the most prominent σ -algebra at all.

Definition 10.15. Let $\Omega \subset \mathbb{R}^d$. The σ -algebra $\mathcal{B}(\Omega) := \sigma(\mathcal{E})$, i.e., the σ -algebra generated by all open boxes in Ω , is called the **Borel- σ -algebra**. Sets of the Borel- σ -algebra are called **Borel-measurable** or **Borel-sets**.

($\Omega = \mathbb{R}^d$ is possible.)

Note that we will generally not verify if a given complicated set is Borel-measurable. We will assume instead that this is our 'universe' in which we are working. However, it contains anyhow everything that we would consider as 'nice' sets and even more. Besides unions/intersections of arbitrary many (open/half-open/closed) boxes it contains, for example, the unit ball of \mathbb{R}^d , which can be written as a union of infinitely many boxes. It is also not hard to show that single points, $p \in \mathbb{R}^d$, are also Borel-sets. More precisely, we have $\{p\} \in \mathcal{B}(\Omega)$ for $p \in \Omega$. (Verify this for $d = 1$!)

If we now apply Caratheory's theorem for the set \mathcal{E} above, and with the pre-measure μ' that assigns each box its volume, i.e.,

$$\mu'([a_1, b_1] \times \cdots \times [a_d, b_d]) := \prod_{i=1}^d (b_i - a_i),$$

we obtain the desired measure on Ω . (μ' is just the length of an interval for $d = 1$.)

Corollary 10.16 (Borel-Lebesgue measure). *Let $\mathcal{B}(\Omega)$ be the Borel- σ -algebra over $\Omega \subset \mathbb{R}^d$. Then there exists a unique measure $\mu: \mathcal{B}(\Omega) \rightarrow [0, \infty]$ which assigns each rectangle its volume. This measure is called the (d-dimensional) **Borel-Lebesgue measure**.*

Although $\mathcal{B}(\Omega)$ contains already very complicated sets, **it is necessary to extend the Borel-Lebesgue measure even further**. (We will see later why.) The idea is that we want to define the measure on as general sets as possible. Hence, it seems natural to define

$$\mathcal{L}(\Omega) := \left\{ A \in \mathcal{P}(\Omega) : B_1 \subset A \subset B_2 \text{ for some } B_1, B_2 \in \mathcal{B}(\Omega) \text{ with } \mu(B_1) = \mu(B_2) \right\}$$

as the set of all **Lebesgue-measurable** sets, i.e., the **Lebesgue- σ -algebra**, and assign A the same measure as B_1 and B_2 .

(Prove that $\mathcal{L}(\Omega)$ is indeed a σ -algebra!)

Another interpretation is that $A \in \mathcal{L}$, i.e., A is Lebesgue-measurable, if it differs from a Borel-set only by a null set. That is, if we consider the **Lebesgue-null sets**

$$\mathcal{N} := \{N \in \mathcal{P}(\Omega) : N \subset B \text{ for some } B \in \mathcal{B}(\Omega) \text{ with } \mu(B) = 0\},$$

we have that $A \in \mathcal{L}$ if $A = B \cup N$ for some disjoint $B \in \mathcal{B}(\Omega)$ and $N \in \mathcal{N}$. Extending the Borel-Lebesgue measure also to such sets, by setting the measure of sets from \mathcal{N} to zero, we finally obtain the Lebesgue measure. Roughly speaking, we thereby enable ourselves to measure '*everything that happens on null sets*', but we assign it measure zero.

Corollary 10.17 (Lebesgue measure). *Let $\mathcal{L}(\Omega)$ be the Lebesgue- σ -algebra over $\Omega \subset \mathbb{R}^d$. Then there is a unique measure $\lambda_d: \mathcal{L}(\Omega) \rightarrow [0, \infty]$ which assigns each rectangle its volume, and zero to every set in \mathcal{N} . This measure is called the (d-dimensional) **Lebesgue measure**.*

Remark 10.18. If the dimension is clear, we may omit the d in the notation $\lambda = \lambda_d$.

Remark 10.19. There are still sets that are not Lebesgue-measurable. We do not discuss that here, but one may look for *Vitali sets* or the *Banach-Tarski paradox*. Moreover, note that the Lebesgue- σ -algebra already depends on the measure, while the Borel- σ -algebra does not.

It is important to note here that the Lebesgue measure for an arbitrary measurable set can be given by the formula

$$\lambda_d(A) = \inf \left\{ \sum_{i=1}^{\infty} \mu'(A_i) : A_1, A_2, \dots \text{ are boxes with } A \subset \bigcup_{i=1}^{\infty} A_i \right\},$$

where μ' is the area of a box as above. (Note that for such boxes $\mu'(A_i) = \lambda_d(A_i)$.)

That is, we *cover* a set A by a union of simpler sets, which can easily be assigned a volume, and then make this covering finer and finer.

This definition is does not look handy, but it provides us with a useful property for being a null set, which play a special role when working with the Lebesgue measure.

Lemma 10.20. *A set $N \subset \mathbb{R}^d$ is a null set w.r.t. the Lebesgue measure λ_d , i.e., $N \in \mathcal{N}$, if and only if*

$$\forall \varepsilon > 0 \exists A_1, A_2, \dots \in \mathcal{B}(\Omega) : N \subset \bigcup_{i=1}^{\infty} A_i \quad \text{and} \quad \sum_{i=1}^{\infty} \lambda_d(A_i) \leq \varepsilon.$$

The sets A_i can chosen to be boxes.

Again, we omit a proof.

This finally allows us to come back to our motivating example.

Example 10.21. Let λ be the Lebesgue measure on \mathbb{R} . Then,

$$\lambda(\mathbb{Q} \cap [0, 1]) = \lambda(\mathbb{Q}) = 0,$$

i.e., the set of rational numbers has (Lebesgue) measure zero, and therefore

$$\lambda([0, 1] \setminus \mathbb{Q}) = 1.$$

This is quite reasonable as there are uncountably many irrational numbers, but 'only' countably many rationals. For a proof first note that the measure of a single point $\{p\}$ with $p \in \mathbb{R}$ is a null set, i.e., $\{p\} \in \mathcal{N}$. For this, note that, for all $\varepsilon > 0$, the interval $[p, p + \varepsilon)$ contains p and has length ε . Setting $A_1 = [p, p + \varepsilon)$ and $A_2 = A_3 = \dots = \emptyset$ in the above lemma implies that $\{p\} \in \mathcal{N}$. (One might also argue that $\{p\}$ is a closed interval of length zero, but we wanted to apply the lemma.)

Now, since λ is a measure and \mathbb{Q} is countable, σ -additivity shows that

$$\lambda(\mathbb{Q}) = \lambda\left(\bigcup_{i=1}^{\infty} \{q_i\}\right) = \sum_{i=1}^{\infty} \lambda(\{q_i\}) = 0,$$

i.e., the set of rational numbers is a null set, where q_1, q_2, \dots is an enumeration of the rationals. We also obtain $\lambda(\mathbb{Q} \cap [0, 1]) \leq \lambda(\mathbb{Q}) = 0$ and $\lambda([0, 1] \setminus \mathbb{Q}) = \lambda([0, 1]) - \lambda(\mathbb{Q} \cap [0, 1]) = 1$.

We now have the framework to work with measures of rather arbitrary sets of \mathbb{R}^d . This will be used in the sequel to define corresponding classes of functions, show that everything is fine for continuous functions, and how to define their integral.

10.1 Measurable functions

The next stop on the road to our new integral is to define 'what a function shall look like'. Here, we only consider real-valued functions $f: \Omega \rightarrow \mathbb{R}$, where Ω can be arbitrary. Instead of requiring that f is continuous/differentiable etc., we now assume that the function 'fits' to our given measurable space (Ω, \mathcal{A}) . In detail, we assume that pre-images of measurable sets (from \mathbb{R}) are also measurable (in \mathcal{A}).

Definition 10.22. Let (Ω, \mathcal{A}) be a measurable space and let $f: \Omega \rightarrow \mathbb{R}$ be a real-valued function. If

$$\forall B \in \mathcal{B}(\mathbb{R}): f^{-1}(B) \in \mathcal{A},$$

then we call f a **measurable function**.

We sometimes write $f: (\Omega, \mathcal{A}) \rightarrow \mathbb{R}$ or call f \mathcal{A} -measurable to indicate the σ -algebra.

A measurable function $f: (\Omega, \mathcal{L}(\Omega)) \rightarrow \mathbb{R}$ for $\Omega \in \mathbb{R}^d$ is called **(Lebesgue-)measurable**.

Remark 10.23. The above definition could be generalized to functions between arbitrary measurable spaces $(\Omega_1, \mathcal{A}_1)$ and $(\Omega_2, \mathcal{A}_2)$. Then f is called \mathcal{A}_1 - \mathcal{A}_2 -measurable if

$$\forall A_2 \in \mathcal{A}_2: f^{-1}(A_2) \in \mathcal{A}_1.$$

The special case $\mathcal{A}_2 = \mathcal{B}(\mathbb{R})$ is somehow the 'default parameter' and most authors do not write explicitly that they use the Borel- σ -algebra. It is just the natural choice when working with real-valued functions.

This is not a very useful condition, as verifying it might be difficult. However, this property can be much simplified. Again, we omit the most technical proofs.

Lemma 10.24. Let (Ω, \mathcal{A}) be a measurable space and let $f: \Omega \rightarrow \mathbb{R}$ be a real-valued function. Then, f is measurable if and only if

$$f^{-1}((-\infty, a)) = \{x \in \Omega: f(x) < a\} \in \mathcal{A},$$

for all $a \in \mathbb{R}$.

Even more, it is enough to check the condition for measurability in Definition 10.22 only for all $B \in \mathcal{E}$, where \mathcal{E} is an arbitrary generator of $\mathcal{B}(\mathbb{R})$.

Let us see some examples.

Example 10.25. Consider the measurable space (Ω, \mathcal{A}) with $\Omega = \{1, 2, 3\}$ and $\mathcal{A} = \{\emptyset, \{1\}, \{2, 3\}, \Omega\}$. To verify that a function on (Ω, \mathcal{A}) is measurable, we need that the pre-images of arbitrary Borel-sets are in \mathcal{A} .

For example, the function f given by $f(1) = 0$, $f(2) = 4$ and $f(3) = 4$ satisfies $f^{-1}((-\infty, a)) = \{1\}$ for $a \leq 4$, and $f^{-1}((-\infty, a)) = \{1, 2, 3\}$ for $a > 4$, all of these sets are in \mathcal{A} . In contrast, the function g given by $g(1) = 0$, $g(2) = 4$ and $g(3) = 5$ satisfies $g^{-1}((-\infty, 5)) = \{1, 2\} \notin \mathcal{A}$. Therefore, g is not measurable.

It is in general the case that, a \mathcal{A} -measurable function **must be constant** on sets $A \in \mathcal{A}$ that do not contain any non-empty subset of \mathcal{A} . (Verify this!)

Moreover, we have that sets, where a measurable function is constant, are measurable.

Lemma 10.26. Let (Ω, \mathcal{A}) be a measurable space and let $f: \Omega \rightarrow \mathbb{R}$ be a measurable function. Then, the **level sets**

$$f^{-1}(\{c\}) = \{x \in \Omega: f(x) = c\}$$

are measurable for any $c \in \mathbb{R}$, i.e., $f^{-1}(\{c\}) \in \mathcal{A}$.

Proof. Since f is measurable, we know the pre-images of Borel-sets are measurable. Since $(-\infty, c], [c, \infty) \in \mathcal{B}(\mathbb{R})$, we have that $f^{-1}((-\infty, c]), f^{-1}([c, \infty)) \in \mathcal{A}$. Therefore,

$$\begin{aligned} f^{-1}((-\infty, c]) \cap f^{-1}([c, \infty)) &= \{x \in \Omega: f(x) \leq c\} \cap \{x \in \Omega: f(x) \geq c\} \\ &= \{x \in \Omega: f(x) = c\} \in \mathcal{A}, \end{aligned}$$

where we use that intersections of measurable sets are measurable. \square

Although measurability might be difficult to verify, we see again that all requirements are fulfilled, when we consider our most important application, i.e., continuous real-valued functions on a domain in \mathbb{R}^d .

Lemma 10.27. *Let $\Omega \subset \mathbb{R}^d$ and let $f: \Omega \rightarrow \mathbb{R}$ be continuous. Then f is Lebesgue-measurable.*

Sketch of proof. Since we did not discuss the important steps needed here in detail, we only give an idea of the proof. We would use that:

1. It is enough to verify measurability for all sets from a generator of $\mathcal{B}(\Omega)$.
2. The set of all open sets generates $\mathcal{B}(\Omega)$.
3. The pre-image of open sets under continuous functions is open.

Since all open sets are in $\mathcal{L}(\Omega)$, we obtain that any continuous f is measurable. \square

But measurability is more general than continuity, because indicator functions are clearly not continuous.

Example 10.28. For be a measurable space (Ω, \mathcal{A}) and a set $A \subset \Omega$, we consider the **indicator function of A** , i.e.,

$$\chi_A(x) = \begin{cases} 1 & \text{if } x \in A \\ 0 & \text{else.} \end{cases}$$

This function is measurable if and only if A is a measurable set, i.e., $A \in \mathcal{A}$. (Why?)

We collect some properties of measurable functions.

Lemma 10.29. *Let (Ω, \mathcal{A}) be a measurable space and let f, g be measurable functions. Then $f + g$, $f \cdot g$, $\max\{f, g\}$, $\min\{f, g\}$, $f^+ = \max\{0, f\}$, $f^- = \min\{0, f\}$, $|f|$ and αf for $\alpha \in \mathbb{R}$ are also measurable functions.*

Moreover, if $(f_n)_n$ is a sequence of measurable functions, then $\sup_n f_n$, $\inf_n f_n$, $\limsup_n f_n$ and $\liminf_n f_n$ (considered pointwise) are also measurable functions.

A proof of this lemma is based on Lemma 10.24 together with the properties of σ -algebras. We omit the details.

Note that this, together with the examples above, implies that the function $f \cdot \chi_E: (\Omega, \mathcal{A}) \rightarrow \mathbb{R}$ is measurable if $f: (\Omega, \mathcal{A}) \rightarrow \mathbb{R}$ is measurable and $E \in \mathcal{A}$.

We now turn to some particularly simple measurable function and show that they are actually the building blocks for arbitrary measurable functions.

Definition 10.30 (Simple functions). Let (Ω, \mathcal{A}) be a measurable space, $A_1, \dots, A_n \in \mathcal{A}$ and $\alpha_1, \dots, \alpha_n \in \mathbb{R}$. We call

$$f(x) = \sum_{k=1}^n \alpha_k \chi_{A_k}(x)$$

a **simple function**. If the A_k are pairwise disjoint and $\alpha_i \neq \alpha_j$ for $i \neq j$, then we say that the simple function is in **canonical form**.

By the above Lemma 10.29 and the measurability of the functions χ_{A_k} , we obtain that simple functions are measurable. From this point on we assume that any simple function is given in its canonical form if not stated otherwise. This saves a lot of notation.

Lemma 10.31. *Let (Ω, \mathcal{A}) be a measurable space and $f: \Omega \rightarrow \mathbb{R}$ be a non-negative measurable function. Then, there is a sequence of simple functions $(f_n)_{n \in \mathbb{N}}$ such that $f_1 \leq f_2 \leq f_3 \leq \dots$ and*

$$f(x) = \lim_{n \rightarrow \infty} f_n(x).$$

That is, we can write f pointwise as a monotone limit of simple functions.

Proof. For $n \in \mathbb{N}$ and $0 \leq k \leq 4^n - 1$ we have a look at the measurable sets $A_k = \{k2^{-n} \leq f \leq (k+1)2^{-n}\}$ (those sets are the pre-images of $[k2^{-n}, (k+1)2^{-n}]$). We use them to define the measurable simple functions

$$f_n(x) = \sum_{k=0}^{4^n-1} \frac{k}{2^n} \chi_{A_k}(x) + 2^n \chi_{\{f \geq 2^n\}}.$$

These functions are increasing and for sufficiently large n we have $|f(x) - f_n(x)| \leq 2^{-n}$ if f is finite, i.e. $f(x) < \infty$. In the case that $f(x) = \infty$ we have $f_n(x) \geq 2^n$. In both cases, $\lim_{n \rightarrow \infty} f_n(x) = f(x)$. \square

Let us now consider an arbitrary measurable function $f: \Omega \rightarrow \mathbb{R}$. We can always use the representation

$$f(x) = f^+(x) - f^-(x),$$

where $f^+(x) := \max\{f(x), 0\}$ and $f^-(x) := \max\{-f(x), 0\} = -\min\{f(x), 0\}$ for all $x \in \Omega$. The measurable and non-negative functions f^+ and f^- are called the **positive and negative part** of f , respectively.

By Lemma 10.31 we can approximate f^+ and f^- by increasing sequences of simple functions. Thus we can also approximate f by simple functions. In conclusion we get the following.

Corollary 10.32. *Let (Ω, \mathcal{A}) be a measurable space and let f be a measurable function. Then, f is the pointwise limit of simple functions.*

10.2 The Lebesgue integral

We are now ready to give a precise definition of the Lebesgue integral and to prove some important results for it. Thereto, recall that the idea behind the Riemann integral (for $d = 1$) was to 'measure' the area below the graph of a function by approximating it with small rectangles. The new idea of H. Lebesgue was to split not the domain of a function into smaller pieces, but the codomain. Meaning that we split \mathbb{R} (again) into smaller subintervals, but we now consider the 'set of x ' such that $f(x)$ lies in such a small subinterval. Summing all these measures up and making the partition finer and finer, leads to the new concept for an integral, see Figure 64.

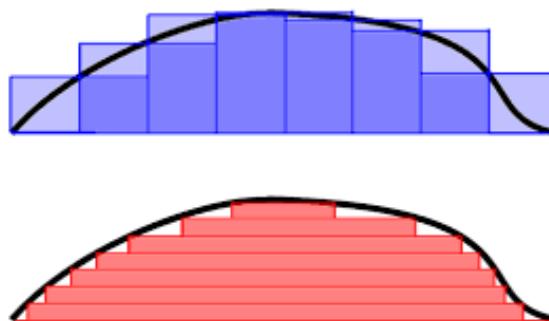


Figure 64: Riemann vs. Lebesgue integral

This also allows to 'integrate' functions that are defined on quite general domains. Let us start with the 'very natural' definition of the integral for simple functions.

Definition 10.33 (Integral for simple functions). Let $(\Omega, \mathcal{A}, \mu)$ be a measure space and let f be a non-negative simple function, i.e.,

$$0 \leq f(x) := \sum_{k=1}^n \alpha_k \chi_{A_k}(x), \quad \alpha_k > 0, A_k \in \mathcal{A}.$$

Then, we define its **integral** over a measurable set $E \in \mathcal{A}$ by

$$\int_E f d\mu := \sum_{k=1}^n \alpha_k \mu(A_k \cap E).$$

The special choice of $\Omega \subset \mathbb{R}^d$, $\mathcal{A} = \mathcal{L}(\Omega)$ and $\mu = \lambda_d$ leads to the **Lebesgue integral** of f . In this case we use the notation

$$\int_E f(x) dx = \int_E f d\lambda_d.$$

Let us mention that we will use the (usual) convention that " $0 \cdot \infty = 0$ ".

Remark 10.34. Recall that we generally assume that all simple functions are given in canonical form. However, the above definition of the integral would also work without this assumption.

Let us see some examples.

Example 10.35 (Counting measure). We have a look at the measure space $(\mathbb{N}, \mathcal{P}(\mathbb{N}), \mu)$, where μ denotes the counting measure. A non-negative simple function f has the general form

$$f(n) = \sum_{i=1}^m \alpha_i \chi_{A_i}(n) \quad \alpha_i > 0, A_i \in \mathcal{P}(\mathbb{N}).$$

Observe that the sets $A_i \in \mathcal{P}(\mathbb{N})$ are not finite in general, since they only have to be measurable. For this function we compute

$$\int_{\mathbb{N}} f d\mu = \sum_{i=1}^m \alpha_i \mu(A_i \cap \mathbb{N}) = \sum_{i=1}^m \alpha_i \cdot \#A_i.$$

This shows, however, that the integral can only be finite if all the A_i are finite.

Example 10.36 (Dirac measure). We have a look at some measurable space (Ω, \mathcal{A}) and the Dirac measure at $y \in \Omega$, denoted by δ_y . For any given simple and non-negative function $f: \Omega \rightarrow \mathbb{R}$ we compute

$$\int_{\Omega} f d\delta_y = \sum_{i=1}^m \alpha_i \delta_y(A_i \cap \Omega) = \begin{cases} \alpha_j & \text{if } y \in A_j \text{ for some } j \\ 0 & \text{else.} \end{cases}$$

On the other hand we observe that

$$f(y) = \sum_{i=1}^m \alpha_i \chi_{A_i}(y) = \begin{cases} \alpha_j & \text{if } y \in A_j \text{ for some } j \\ 0 & \text{else.} \end{cases}$$

(We used here that f is in canonical form.) This shows that

$$\int_{\Omega} f d\delta_y = f(y).$$

Example 10.37. Last but not least we want to study the Lebesgue integral on the real line for simple functions of the form

$$f(x) = \sum_{i=1}^m \alpha_i \chi_{[a_i, b_i)}(x),$$

again we assume that f is in its canonical form, i.e., the intervals $[a_i, b_i)$ are disjoint. We use the definition to see

$$\int_{\mathbb{R}} f(x) dx = \sum_{i=1}^m \alpha_i \lambda([a_i, b_i)) = \sum_{i=1}^m \alpha_i (b_i - a_i).$$

We have the following useful properties for this simple kind of integrals.

Lemma 10.38. Let $(\Omega, \mathcal{A}, \mu)$ be a measure space, $E \in \mathcal{A}$ and f, g be non-negative simple functions. Then, we have

- 1) $\nu: \mathcal{A} \rightarrow [0, \infty]$ with $\nu(E) := \int_E f d\mu$ defines a measure on \mathcal{A} .
- 2) $\int_E f + g d\mu = \int_E f d\mu + \int_E g d\mu$.
- 3) For $\alpha \in \mathbb{R}$ we have $\int_E \alpha \cdot f d\mu = \alpha \cdot \int_E f d\mu$.
- 4) If $f \leq g$ then also $\int_E f d\mu \leq \int_E g d\mu$.

The properties 2) and 3) are called *linearity* and 4) is called *monotonicity* of the integral.

Proof. Let the simple functions f and g be given by

$$f = \sum_{k=1}^n \alpha_k \chi_{A_k} \quad \text{and} \quad g = \sum_{k=1}^m \beta_k \chi_{B_k}.$$

For the first part of the lemma we first observe that by definition $\nu(\emptyset) = 0$ and $\nu(E) \geq 0$ for any measurable set E . Moreover, for $E = \bigcup_{i \in \mathbb{N}} E_i$, where $E_i \in \mathcal{A}$ and all of these sets are pairwise disjoint, we use the properties of μ to see

$$\begin{aligned} \nu(E) &= \sum_{k=1}^n \alpha_k \mu(E \cap A_k) \\ &= \sum_{k=1}^n \alpha_k \sum_{i \in \mathbb{N}} \mu(E_i \cap A_k) \\ &= \sum_{i \in \mathbb{N}} \sum_{k=1}^n \alpha_k \mu(E_i \cap A_k) \\ &= \sum_{i \in \mathbb{N}} \nu(E_i). \end{aligned}$$

Hence ν is a measure. For the second point we define $E_{i,j} = A_i \cap B_j \cap E$ for $i = 0, \dots, n$ and $j = 0, \dots, m$, where we set for completeness $A_0 = E \setminus \bigcup_{i=1}^n A_i$ and $B_0 = E \setminus \bigcup_{j=1}^m B_j$. Setting $\alpha_0 = 0$ and $\beta_0 = 0$, we obtain by the definition of integrals that

$$\begin{aligned} \int_{E_{i,j}} f d\mu &= \alpha_i \mu(E_{i,j}), \\ \int_{E_{i,j}} g d\mu &= \beta_j \mu(E_{i,j}). \end{aligned}$$

We clearly have $E = \bigcup_{i,j} E_{i,j}$, $A_i \cap E = \bigcup_{j=0}^m E_{i,j}$ and $B_j \cap E = \bigcup_{i=0}^n E_{i,j}$, and therefore

$$\int_E f + g \, d\mu = \sum_{i,j} \int_{E_{i,j}} (f + g) \, d\mu = \sum_{i,j} \int_{E_{i,j}} f \, d\mu + \int_{E_{i,j}} g \, d\mu = \int_E f \, d\mu + \int_E g \, d\mu.$$

Note that the first equality is allowed due to the first part of the proof.

The third point follows from the definition. For the last point we observe that $g - f$ is a non-negative simple function and therefore we have that its integral is non-negative. Using linearity (part 2 of the proof) the result follows. \square

So it turns out our new definition of an integral is rather easy to handle for simple functions. But what about complicated, i.e., 'non-simple', functions? We will now define and discuss step by step more general functions.

Definition 10.39 (Integral for non-negative functions). Let $(\Omega, \mathcal{A}, \mu)$ be a measure space and let f be a non-negative measurable function. For a measurable set $E \in \mathcal{A}$ we define by

$$\int_E f \, d\mu := \sup \left\{ \int_E g \, d\mu : g \text{ is a simple function with } 0 \leq g \leq f \text{ on } E \right\}$$

the **integral of f over E w.r.t. μ** .

Note that it is allowed here that $\int_E f \, d\mu = \infty$.

Clearly, if f is a simple function the above definition coincides with the first definition of integrals for simple functions. Moreover, we saw that each measurable function $f \geq 0$ is the pointwise limit of an increasing sequence of simple functions. This is the reason why the above definition makes perfect sense.

Again we want to collect some properties for these integrals. But let us start with the fact, that the integral over a set $E \in \mathcal{A}$ can always be written as the integral of another function over the whole domain Ω

Lemma 10.40. Let $(\Omega, \mathcal{A}, \mu)$ be a measure space, $E \subset \mathcal{A}$ and f be a non-negative measurable function. Then,

$$\int_E f \, d\mu = \int_\Omega \chi_E \cdot f \, d\mu.$$

Proof. The equality clearly holds for the case that f is a indicator function (by definition). Using that the integral is linear we obtain that the result is true for simple functions. Moreover, note that $\chi_E \cdot f$ is a simple function whenever g is simple. Hence,

$$\begin{aligned} \int_E f \, d\mu &= \sup \left\{ \int_E g \, d\mu : g \text{ is simple with } 0 \leq g \leq f \text{ on } E \right\} \\ &= \sup \left\{ \int_\Omega \chi_E \cdot g \, d\mu : g \text{ is simple with } 0 \leq \chi_E \cdot g \leq \chi_E \cdot f \text{ on } \Omega \right\} \\ &= \sup \left\{ \int_\Omega h \, d\mu : h \text{ is simple with } 0 \leq h \leq \chi_E \cdot f \text{ on } \Omega \right\} \\ &= \int_\Omega \chi_E \cdot f \, d\mu. \end{aligned}$$

The first equality is just definition. The second uses the statement for simple functions and the easy fact that $g \leq f$ on E (i.e., $g(x) \leq f(x)$ for all $x \in E$) is equivalent to $\chi_E \cdot g \leq \chi_E \cdot f$ on Ω . (Why?) Finally, we just renamed $\chi_E \cdot g$ to h , and used again the definition. This proves the result. \square

All of the following facts follow directly from the above definition, possibly by using the fact that they are true for simple functions.

Lemma 10.41. *Let $(\Omega, \mathcal{A}, \mu)$ be a measure space, f, g be non-negative measurable functions and $E, F \in \mathcal{A}$ be measurable sets. Then,*

- 1) If $f \leq g$, then $\int_E f d\mu \leq \int_E g d\mu$
- 2) If $E \subset F$, then $\int_E f d\mu \leq \int_F f d\mu$.
- 3) If $\mu(E) = 0$, then $\int_E f d\mu = 0$.

Proof. By definition,

$$\begin{aligned} \int_E f d\mu &= \sup \left\{ \int_E h d\mu : h \text{ is simple with } 0 \leq h \leq f \text{ on } E \right\} \\ &\leq \sup \left\{ \int_E h d\mu : h \text{ is simple with } 0 \leq h \leq g \text{ on } E \right\} \\ &= \int_E g d\mu, \end{aligned}$$

where the inequality come from the fact that we take the supremum over a larger set, making it larger. (That is, every simple function h with $0 \leq h \leq f$ also satisfies $0 \leq h \leq g$. Just because $f \leq g$.)

The second point now follows from the first point, together with Lemma 10.40 and $\chi_E \cdot f \leq \chi_F \cdot f$.

Check the third one yourself by proving it first for indicator and simple functions and then apply the definition. \square

Note that the above monotonicity is already enough to show that our definition of an integral coincides with the Riemann integral considered earlier.

Example 10.42 (Lebesgue vs Riemann integral for continuous functions). We consider a non-negative continuous function $f: [0, 1] \rightarrow \mathbb{R}$. We already know that this is a measurable function and that f is integrable w.r.t. our old definition of the integral, see Section 6.3. There we had a look at the *lower sums*, given by

$$L_n(f) := \frac{1}{n} \sum_{k=0}^{n-1} \min \left\{ f(x) : x \in \left[\frac{k}{n}, \frac{k+1}{n} \right] \right\},$$

and the *upper sums* given by

$$U_n(f) := \frac{1}{n} \sum_{k=0}^{n-1} \max \left\{ f(x) : x \in \left[\frac{k}{n}, \frac{k+1}{n} \right] \right\}.$$

These sums are the Lebesgue integrals of the simple functions

$$g_n(x) := \sum_{k=0}^{n-1} \min \left\{ f(x) : x \in \left[\frac{k}{n}, \frac{k+1}{n} \right] \right\} \chi_{[\frac{k}{n}, \frac{k+1}{n}]}(x)$$

and

$$h_n(x) := \sum_{k=0}^{n-1} \max \left\{ f(x) : x \in \left[\frac{k}{n}, \frac{k+1}{n} \right] \right\} \chi_{[\frac{k}{n}, \frac{k+1}{n}]}(x).$$

We note that these functions are not in canonical form, however their integral would not change if we exclude either the left or the right point of the intervals $[\frac{k}{n}, \frac{k+1}{n}]$ and change the max / min to sup / inf. So we will just use them in the above form.

Moreover, for any n we have that $g_n(x) \leq f(x) \leq h_n(x)$ so by the monotonicity of the integral it follows that

$$L_n(f) = \int_{[0,1]} g_n(x) dx \leq \int_{[0,1]} f(x) dx \leq \int_{[0,1]} h_n(x) dx = U_n(f).$$

(Here we used the Lebesgue integral!)

Since f is continuous we now obtain that

$$\lim_{n \rightarrow \infty} |L_n(f) - U_n(f)| = 0.$$

Thus the Lebesgue integral coincides with the Riemann integral i.e., we have $\int_{[0,1]} f(x) dx = \lim_{n \rightarrow \infty} L_n(f)$. The same holds for arbitrary $[a, b]$ or even boxes in \mathbb{R}^d . We can therefore use all the calculation rules we have already learned, whenever the functions are Riemann-integrable.

Before we turn to actual examples, let us state one of the most important results in measure theory. That is, the *monotone convergence theorem* due to *Beppe Levi* (1875–1961). It shows that we can actually use an **arbitrary monotone sequence instead of a supremum** to compute the integral which is usually much easier to handle. We will see later that this is the key ingredient to develop an easy and general proof technique in the context of integration.

Theorem 10.43 (Monotone convergence theorem). *Let $(\Omega, \mathcal{A}, \mu)$ be a measure space and let $(f_n)_{n \in \mathbb{N}}$ be a sequence of non-negative measurable functions such that for all $x \in \Omega$,*

$$f_1(x) \leq f_2(x) \leq f_3(x) \leq \dots \quad \text{and} \quad \lim_{n \rightarrow \infty} f_n(x) = f(x).$$

Then, for any measurable E , we have

$$\int_E f d\mu = \int_E \lim_{n \rightarrow \infty} f_n d\mu = \lim_{n \rightarrow \infty} \int_E f_n d\mu.$$

Note that $\int_E f d\mu = \infty$ is still allowed.

Proof. Due to the fact that pointwise limits of measurable functions are again measurable functions, we get that f is measurable. The previous lemma, i.e., Lemma 10.41, implies that the sequence

$$I(n) := \int_E f_n d\mu$$

is a non-decreasing sequence since (f_n) is monotone. Thus $I := \sup_n I(n) = \lim_{n \rightarrow \infty} I(n)$ exists in $[0, \infty]$. By monotonicity we have that $\int_E f d\mu \geq I$.

Now we show $\int_E f d\mu \leq I$. In particular we show that for any simple function $0 \leq g \leq f$ we have $\int_E g d\mu \leq I$. To do so we consider $0 \leq c < 1$ and define the sets $E_n = \{x \in E : f_n(x) \geq cg(x)\}$. This is an increasing family of sets, i.e. $E_n \subset E_{n+1}$ and we additionally have that all E_n are contained in \mathcal{A} and $E = \bigcup_{n \in \mathbb{N}} E_n$. Lemma 10.41 implies

$$I \geq \int_E f_n d\mu \geq \int_{E_n} f_n d\mu \geq \int_{E_n} cg d\mu = c \int_{E_n} g d\mu$$

We showed already that $A \mapsto \int_A g d\mu$ is a measure, hence it is σ -additive. Moreover, for any σ -additive function ν (so in particular for measures) and any increasing sequence of sets $A_n \subset A_{n+1}$ with $A = \bigcup_{n \in \mathbb{N}} A_n$ we can define

$$\begin{aligned} B_1 &= A_1 \\ B_2 &= A_2 \setminus A_1 \\ B_3 &= A_3 \setminus (A_1 \cup A_2) \\ &\vdots \end{aligned}$$

The sequence of sets $(B_i)_{i \in \mathbb{N}}$ consists of pairwise disjoint sets such that $A = \bigcup_{i=1}^{\infty} B_i$ and $A_n = \bigcup_{i=1}^n B_i$. Thus

$$\nu(A) = \nu\left(\bigcup_{i=1}^{\infty} B_i\right) = \sum_{i=1}^{\infty} \nu(B_i) = \lim_{n \rightarrow \infty} \sum_{i=1}^n \nu(B_i) = \lim_{n \rightarrow \infty} \nu(A_n).$$

In our case this shows that $\lim_{n \rightarrow \infty} \int_{E_n} g d\mu = \int_E g d\mu$, and therefore $I \geq c \int_E g d\mu$. As $c < 1$ was arbitrary we get that $I \geq \int_E g d\mu$, which finishes the proof. \square

Let us see how it can be applied to compute an integral, or here a sum.

Example 10.44 (Counting measure). We want to integrate a non-negative measurable function $f: \mathbb{N} \rightarrow \mathbb{R}$ w.r.t. the counting measure μ . Thereto we observe that the sequence of functions given by

$$f_n(k) := \sum_{i=1}^n f(i) \chi_{\{i\}}(k) = \begin{cases} f(k), & \text{if } k \leq n, \\ 0, & \text{otherwise,} \end{cases}$$

converges pointwise to f and all f_n are simple functions with $f_n \leq f_{n+1}$. We compute the integral of f_n by

$$\int_{\mathbb{N}} f_n d\mu = \sum_{i=1}^n f(i).$$

Therefore, from the monotone convergence theorem, we obtain that

$$\int_{\mathbb{N}} f d\mu = \lim_{n \rightarrow \infty} \int_{\mathbb{N}} f_n d\mu = \lim_{n \rightarrow \infty} \sum_{i=1}^n f(i) = \sum_{i=1}^{\infty} f(i).$$

This implies that the function f has finite integral w.r.t. the counting measure if the sequence $(f(k))_{k \in \mathbb{N}}$ leads to a convergent series. E.g., if $f(k) = \frac{1}{2^k}$, then

$$\int_{\mathbb{N}} f d\mu = 1.$$

The monotone convergence theorem is the key ingredient in many proofs related to integrals. Let us see an example of this proof technique.

Lemma 10.45. *Let $(\Omega, \mathcal{A}, \mu)$ be a measure space, f, g be non-negative measurable functions and $\alpha \in \mathbb{R}$. Then,*

$$\int_E (f + g) d\mu = \int_E f d\mu + \int_E g d\mu$$

and

$$\int_E \alpha f d\mu = \alpha \int_E f d\mu,$$

for any measurable E .

Proof. We know from Lemma 10.31 that for each non-negative measurable function, say f and g , there are monotone sequences (f_n) and (g_n) of simple functions that converge pointwise to f and g , respectively. Now, we know from Lemma 10.38 that the respective rules are true for simple functions. We therefore obtain from the monotone convergence theorem that

$$\int_E (f + g) d\mu = \lim_{n \rightarrow \infty} \int_E (f_n + g_n) d\mu = \lim_{n \rightarrow \infty} \left(\int_E f_n d\mu + \int_E g_n d\mu \right) = \int_E f d\mu + \int_E g d\mu,$$

where we use that the corresponding limits exist. \square

We can even treat infinite sums aka. series.

Lemma 10.46. *Let $(\Omega, \mathcal{A}, \mu)$ be a measure space and let g_1, g_2, \dots be non-negative measurable functions. Then $g = \sum_{k=1}^{\infty} g_k$ is measurable and non-negative, and for measurable E we have*

$$\int_E g d\mu = \sum_{k=1}^{\infty} \int_E g_k d\mu.$$

Proof. The function g is the pointwise limit of $f_n = \sum_{k=1}^n g_k$. Since all g_k are measurable we get that every f_n is measurable and therefore g is measurable. Clearly, we have

$$\int_E f_n d\mu = \int_E \sum_{k=1}^n g_k d\mu = \sum_{k=1}^n \int_E g_k d\mu.$$

Since all g_k are non-negative the sequence $(f_n)_{n \in \mathbb{N}}$ is increasing with g as pointwise limit. We obtain from the monotone convergence theorem that

$$\begin{aligned} \int_E g d\mu &= \int_E \lim_{n \rightarrow \infty} f_n d\mu = \lim_{n \rightarrow \infty} \int_E f_n d\mu \\ &= \lim_{n \rightarrow \infty} \sum_{k=1}^n \int_E g_k d\mu = \sum_{k=1}^{\infty} \int_E g_k d\mu. \end{aligned}$$

□

Example 10.47. We have a look at the continuous function $x \mapsto e^x$ on $[0, 1]$, which is clearly measurable. We already know that for any $x \in \mathbb{R}$

$$e^x = \sum_{k=0}^{\infty} \frac{x^k}{k!}.$$

So by defining $g_k(x) = \frac{x^k}{k!}$ we obtain that

$$e^x = \sum_{k=0}^{\infty} g_k(x).$$

Moreover, all g_k and e^x are non-negative measurable functions on $[0, 1]$. Thus,

$$\int_{[0,1]} e^x dx = \sum_{k=0}^{\infty} \int_{[0,1]} \frac{x^k}{k!} dx = \sum_{k=0}^{\infty} \frac{1}{(k+1)!} = e - 1.$$

Now we generalize the integral to the case that f is measurable, i.e., we do not assume that f is non-negative any more. Thereto we remind that any measurable f can be decomposed into

$$f = f^+ - f^-, \quad \text{with} \quad f^+ := \max\{f, 0\} \quad \text{and} \quad f^- := \max\{-f, 0\}.$$

Recall that these operations should be understood pointwise, i.e., $f = g$ means $f(x) = g(x)$ for all $x \in \Omega$. Moreover, the functions f^+, f^- are non-negative and measurable if f is measurable.

Definition 10.48. Let $(\Omega, \mathcal{A}, \mu)$ be a measure space, $E \in \mathcal{A}$ and $f = f^+ - f^-$ be a measurable function. We call f **integrable over E** if

$$\int_E f^+ d\mu < \infty \quad \text{and} \quad \int_E f^- d\mu < \infty,$$

and we define

$$\int_E f d\mu := \int_E f^+ d\mu - \int_E f^- d\mu.$$

For $E = \Omega$ we call f just integrable.

Using the representation $|f| = f^+ + f^-$ we see that f is integrable over E if and only if

$$\int_E |f| d\mu < \infty.$$

Example 10.49 (Counting measure continued). For the measurable space $(\mathbb{N}, \mathcal{P}(\mathbb{N}), \mu)$, where μ denotes the counting measure, we see that a function $f: \mathbb{N} \rightarrow \mathbb{R}$ (aka. a sequence) is integrable if and only if

$$\int_{\mathbb{N}} |f(x)| d\mu(x) = \sum_{k=1}^{\infty} |f(k)| < \infty.$$

This is equivalent to saying that the sequence $(f(k))_{k \in \mathbb{N}}$ is absolutely summable. For the integral of f we have in this case

$$\int_{\mathbb{N}} f(x) d\mu(x) = \sum_{k=1}^{\infty} f^+(k) - \sum_{k=1}^{\infty} f^-(k).$$

Both series converge as bounded series with non-negative terms, see Theorem 3.68.

Let us finally collect some properties of this general integral. This will be proved by using earlier results for the positive and negative part, i.e., f^+ and f^- , separately.

Lemma 10.50. *Let $(\Omega, \mathcal{A}, \mu)$ be a measure space, $E \in \mathcal{A}$, f, g be integrable functions and $\alpha \in \mathbb{R}$. Then, we have*

- 1) $f + g$ is integrable with $\int_E f + g d\mu = \int_E f d\mu + \int_E g d\mu$
- 2) $\int_E \alpha f d\mu = \alpha \int_E f d\mu$
- 3) $f \geq 0 \implies \int_E f d\mu \geq 0$
- 4) $|\int_E f d\mu| \leq \int_E |f| d\mu$

Note that, however, **the product of two integrable functions is in general not integrable**. (Find an example, or many.)

Proof. Since f, g are integrable we obtain $\int |f + g| d\mu \leq \int |f| + |g| d\mu = \int |f| d\mu + \int |g| d\mu < \infty$. Hence, $f + g$ is integrable. We now use the decompositions $f = f^+ - f^-$, $g = g^+ - g^-$ and $f + g = h^+ - h^-$ (where we set $h := f + g$), and obtain $h^+ - h^- = f + g = f^+ + g^+ - f^- - g^-$. Regrouping shows $h^+ + f^- + g^- = h^- + f^+ + g^+$. But now, all involved functions are non-negative and measurable, and we obtain from Lemma 10.45 that

$$\begin{aligned} \int_E h^+ d\mu + \int_E f^- d\mu + \int_E g^- d\mu &= \int_E h^+ + f^- + g^- d\mu \\ &= \int_E h^- + f^+ + g^+ d\mu \\ &= \int_E h^- d\mu + \int_E f^+ d\mu + \int_E g^+ d\mu. \end{aligned}$$

Regrouping again, which is allowed since all involved integrals are finite, and using the definition of the integral shows

$$\begin{aligned} \int_E f + g d\mu &= \int_E h^+ d\mu - \int_E h^- d\mu \\ &= \int_E f^+ d\mu + \int_E g^+ d\mu - \int_E f^- d\mu - \int_E g^- d\mu \\ &= \int_E f d\mu + \int_E g d\mu. \end{aligned}$$

For the third point note that $f^- = 0$ if $f \geq 0$. (Recall the definition!) From Lemma 10.41.3) we get that $\int_E f^- d\mu = 0$, and therefore $\int_E f d\mu = \int_E f^+ d\mu \geq 0$.

For the last point, we use the triangle inequality to see

$$\begin{aligned} \left| \int_E f d\mu \right| &= \left| \int_E f^+ d\mu - \int_E f^- d\mu \right| \leq \left| \int_E f^+ d\mu \right| + \left| \int_E f^- d\mu \right| \\ &= \int_E f^+ d\mu + \int_E f^- d\mu = \int_E f^+ + f^- d\mu = \int_E |f| d\mu. \end{aligned}$$

□

Note that most of the proofs above were done by a very similar procedure. This is the so called **standard machinery**, which is a very common (and easy) way to prove results for integrals. It works in the following way:

- 1) Show that the property holds for indicator functions.
- 2) Show that this implies that it holds for non-negative simple functions.
- 3) Show it for non-negative measurable functions by using the monotone convergence theorem (and the fact that measurable functions can be approximated by simple functions).
- 4) Show it for integrable functions by using the representation $f = f^+ - f^-$.

Remark 10.51. The reason why we need that functions are integrable in order to use the standard machinery is that expressions of the form " $\infty - \infty$ " are simply not defined, and there is no satisfying definition for such expressions. This is also the reason why we say f is integrable if and only if $|f|$ has a finite integral. However, one could generalize some of the definitions/theorems to the case where only one of the functions f^+ or f^- has a finite integral.

10.3 Lebesgue's theorem

In this section we establish a result that allows to interchange limit and integration also for functions that are not necessarily non-negative. The so-called *dominated convergence theorem* is one of the most powerful mathematical tools existing, and also leads to some results that help computing complicated integrals.

The main ingredient is the following generalization of the monotone convergence theorem.

Lemma 10.52 (Fatou). *Let $(\Omega, \mathcal{A}, \mu)$ be a measure space, $E \in \mathcal{A}$ and $(f_n)_{n \in \mathbb{N}}$ be a sequence of non-negative measurable functions. Then*

$$\int_E \liminf_n f_n d\mu \leq \liminf_n \int_E f_n d\mu.$$

If there exists a non-negative integrable function g such that $f_n \leq g$ for all $n \in \mathbb{N}$, then

$$\limsup_{n \rightarrow \infty} \int_E f_n d\mu \leq \int_E \limsup_{n \rightarrow \infty} f_n d\mu.$$

Proof. The function $\liminf_n f_n$ is measurable and we calculate

$$\begin{aligned}\int_E \liminf_n f_n d\mu &= \int_E \liminf_k \inf_{n \geq k} f_n d\mu \\ &= \lim_k \int_E \inf_{n \geq k} f_n d\mu \\ &\leq \liminf_k \int_E f_n d\mu = \liminf_n \int_E f_n d\mu.\end{aligned}$$

In the first line we simply plugged in the definition of \liminf , in the second we used the monotone convergence theorem, see Theorem 10.43, for the sequence of functions $g_k = \inf_{n \geq k} f_n$. By taking out the \inf , the integral gets at most bigger because we do not consider the infimum pointwise anymore. In the last step we used again the definition of \liminf .

For the second part of the lemma, first note that $0 \leq f_n \leq g$ implies that all f_n are integrable. Now consider the functions $h_n := g - f_n$ which are non-negative and measurable. (h_n is indeed integrable, but this is not important here.) Note that $\liminf_{n \rightarrow \infty} h_n = g - \limsup_{n \rightarrow \infty} f_n$ (Verify this precisely!), and that also $\limsup_{n \rightarrow \infty} f_n \leq g$, which implies that $\limsup_{n \rightarrow \infty} f_n$ is integrable. Hence, we get from the first part that

$$\begin{aligned}\int_E g d\mu - \int_E \limsup_{n \rightarrow \infty} f_n d\mu &= \int_E g - \limsup_{n \rightarrow \infty} f_n d\mu = \int_E \liminf_{n \rightarrow \infty} h_n d\mu \\ &\leq \liminf_{n \rightarrow \infty} \int_E h_n d\mu = \liminf_{n \rightarrow \infty} \left(\int_E g d\mu - \int_E f_n d\mu \right) \\ &= \int_E g d\mu - \limsup_{n \rightarrow \infty} \int_E f_n d\mu,\end{aligned}$$

where the first and the next to last equality use integrability of the involved functions and Lemma 10.50. Rearranging yields the result. \square

Before we now come to the main theorem of this chapter, we need one more definition.

Definition 10.53. A measure space $(\Omega, \mathcal{A}, \mu)$ is called **complete** if

$$A \subset N \text{ for some } N \in \mathcal{A} \text{ with } \mu(N) = 0 \implies A \in \mathcal{A}.$$

That is, arbitrary subsets of null sets are measurable in a complete measure space.

Note that every measure space $(\Omega, \mathcal{P}(\Omega), \mu)$, i.e., with the power set as σ -algebra, like the counting measure, is complete. This follows easily from the properties of a measure. Moreover, by construction, the **Lebesgue measure space is complete**, see Corollary 10.17. In fact, every measure space can be “completed” (in the same way as we “completed” the Borel measure space). We omit the details.

The next example is to illustrate that also basic examples can lead to incomplete measure spaces.

Example 10.54. Let us again consider the finite example from Example 10.9, i.e., $\Omega = \{1, 2, 3\}$ with the σ -algebra $\mathcal{A} = \{\emptyset, \{1\}, \{2, 3\}, \{1, 2, 3\}\}$. If we take the measure μ_1 from there, which was defined by $\mu_1(\emptyset) = 0$, $\mu_1(\{1\}) = 1$, $\mu_1(\{2, 3\}) = 0$ and $\mu_1(\{1, 2, 3\}) = 1$, we see that $\{2, 3\}$ is a null set in $(\Omega, \mathcal{A}, \mu_1)$. However, $\{2\}$ and $\{3\}$ are subsets of $\{2, 3\}$ that are not in \mathcal{A} , i.e., not measurable. Hence, $(\Omega, \mathcal{A}, \mu_1)$ is not complete.

Definition 10.55 (Almost everywhere). Let $(\Omega, \mathcal{A}, \mu)$ be a measure space and $f: \Omega \rightarrow \mathbb{R}$ be a measurable function.

If a property of f (e.g. non-negativity, continuity etc.) holds for all $x \in \Omega \setminus N$ for some null set $N \in \mathcal{A}$, i.e., $\mu(N) = 0$, then we say that the property holds **μ -almost everywhere** (μ -a.e.) or for **μ -almost all** $x \in \Omega$.

If μ is clear from the context, we just say “almost everywhere” or “almost all”.

In other words, having a property almost everywhere means that the property holds for all $x \in \Omega$ except for a set of measure 0.

The advantage of complete measure spaces is that we only need to consider functions almost everywhere to verify if they are integrable.

Lemma 10.56. Let $(\Omega, \mathcal{A}, \mu)$ be a complete measure space and g be a measurable function. If $f: \Omega \rightarrow \mathbb{R}$ is such that $f = g$ almost everywhere, i.e., $f(x) = g(x)$ for all $x \in \Omega \setminus N$ for some $N \in \mathcal{A}$ with $\mu(N) = 0$, then, f is measurable.

Moreover, if g is integrable, then f is integrable and

$$\int_E f d\mu = \int_E g d\mu$$

for any measurable $E \in \mathcal{A}$.

In particular, $\int_E f d\mu$ is well-defined if f is only defined almost everywhere.

Proof. We need to show that f is measurable. For this, let $B \in \mathcal{B}(\mathbb{R})$ be a Borel-set. We have

$$f^{-1}(B) = (f^{-1}(B) \cap N) \cup (f^{-1}(B) \cap N^c) = (f^{-1}(B) \cap N) \cup (g^{-1}(B) \cap N^c),$$

where we use that $f = g$ on $N^c = \Omega \setminus N$. Since N , and therefore N^c , and g are measurable, we obtain that $(g^{-1}(B) \cap N^c)$ is measurable. Now, $(f^{-1}(B) \cap N)$ is a subset of a null set, which is measurable since $(\Omega, \mathcal{A}, \mu)$ is complete. Since f is measurable, one can now use the standard machinery to prove the desired equation. (Verify yourself!) \square

We finally obtain the most important result about interchanging limits and integration. This is often called **dominated convergence theorem**, or the **Theorem of Lebesgue**.

Theorem 10.57 (Dominated convergence theorem). Let $(\Omega, \mathcal{A}, \mu)$ be a complete measure space and let $(f_n)_{n \in \mathbb{N}}$ be a sequence of measurable functions such that $f(x) = \lim_n f_n(x)$ exists for almost all $x \in \Omega$. If there exists a integrable function $g: \Omega \rightarrow [0, \infty)$ such that $\forall n \in \mathbb{N}: |f_n| \leq g$ almost everywhere, then all f_n and f are integrable and for any measurable $E \subset \Omega$ it holds

$$\int_E f d\mu = \int_E \lim_{n \rightarrow \infty} f_n d\mu = \lim_{n \rightarrow \infty} \int_E f_n d\mu.$$

We call g the **integrable majorant** of (f_n) .

Remark 10.58. The assumption that the measure space is complete is only necessary because we want to assume the convergence of (f_n) only almost everywhere, which is the assumption that is more suitable for applications. The theorem above is also true for arbitrary measure spaces, but then we need the convergence for all $x \in \Omega$.

Remark 10.59. Recall from Lemma 10.29 that the pointwise limit of measurable functions is measurable. But, under almost everywhere convergence, the function f can be arbitrary on the “set of non-convergence” and is therefore not measurable in general. However, Lemma 10.56 shows that this is not a problem in complete measure spaces.

Proof. First of all, since $\liminf_n f_n$ is measurable and equal to $f := \lim_n f_n$ almost everywhere, we obtain from Lemma 10.56 that f is measurable. Moreover, $|f_n| \leq g$ implies $|f| \leq g$. Hence, f and all f_n are integrable by the monotonicity of the integral, i.e.,

$$\int |f_n| d\mu \leq \int g d\mu < \infty.$$

(Similar for f .) Moreover, since $|f - f_n| \leq |f| + |f_n| \leq 2g$, we also see that $|f - f_n|$ is integrable. By the a.e.-convergence we have $\lim_{n \rightarrow \infty} |f - f_n| = 0$ almost everywhere. The second part of Fatou’s lemma, see Lemma 10.52, and Lemma 10.56 imply

$$0 \leq \liminf_{n \rightarrow \infty} \int_E |f - f_n| d\mu \leq \limsup_{n \rightarrow \infty} \int_E |f - f_n| d\mu \leq \int_E \limsup_{n \rightarrow \infty} |f - f_n| d\mu = \int_E 0 d\mu = 0.$$

Therefore,

$$\lim_{n \rightarrow \infty} \int_E |f - f_n| d\mu = 0.$$

(Recall that the limit exists iff \liminf and \limsup are equal.) Finally, by using linearity and the triangle inequality, we obtain

$$\lim_{n \rightarrow \infty} \left| \int_E f d\mu - \int_E f_n d\mu \right| = \lim_{n \rightarrow \infty} \left| \int_E (f - f_n) d\mu \right| \leq \lim_{n \rightarrow \infty} \int_E |f - f_n| d\mu = 0.$$

□

Let us first recall that we have shown in Example 10.42 that **the Lebesgue integral and the Riemann integral coincide** for non-negative and continuous functions on a bounded interval. Using the dominated convergence theorem, this can be extended to arbitrary Riemann-integrable functions, even on more general domains. Let us fix that in a lemma in a more general form for future reference.

Lemma 10.60. *Let $A \subset \mathbb{R}^d$ be a Jordan-measurable set and $f: A \rightarrow \mathbb{R}$ be Riemann-integrable, see Definition 8.120. Then, f is Lebesgue-integrable and the integrals coincide.*

Sketch of proof. As in Example 10.42 we use that the lower sums L_n are integrals of certain simple functions, say g_n . These g_n converge almost everywhere to f . (We did not show this!) By the dominated convergence theorem, the limit of the L_n equals the Lebesgue integral. Moreover, since f is Riemann-integrable, L_n converge to the Riemann-integral, which shows their equality. □

Note that thereby, we can **use all the calculation rules from Section 8.7** when it comes to such examples. However, recall that the Lebesgue integral is more general as we were also able to integrate functions like $\chi_{\mathbb{Q}}$. Moreover, so far we haven’t presented proofs, e.g., for Lemma 8.122. This will follow from the more general considerations in the next section.

Remark 10.61. Note that there are also Riemann-integrable functions that are not Lebesgue-integrable. However, this can only happen if the domain of integration is infinite. For example, $\frac{\sin(x)}{x}$ is not Lebesgue-integrable, but one can calculate that its improper Riemann-integral over \mathbb{R} equals π . This shows that one must be careful about the ‘chosen’ integral. However, unless otherwise stated, we will work with the Lebesgue integral.

Let us now consider a useful consequence for the computation of integrals of power series.

Example 10.62 (Integrals of series). We consider a continuous function $f: [a, b] \rightarrow \mathbb{R}$ which has the representation

$$f(x) = \sum_{k=0}^{\infty} a_k (x - x_0)^k$$

as an absolutely convergent series, i.e., $\sum_{k=0}^{\infty} |a_k| |x - x_0|^k \leq M$ for some $M < \infty$ and all $x \in [a, b]$. Therefore, $g(x) = \sum_{k=0}^{\infty} |a_k| |x - x_0|^k$ dominates the sequence $f_n(x) = \sum_{k=0}^n |a_k| |x - x_0|^k$, i.e., $f_n \leq g$, and g is integrable, because it is non-negative, measurable and bounded by the integrable function $M \cdot \chi_{[a,b]}$. We therefore obtain from Theorem 10.57 that

$$\int_a^b f(x) dx = \sum_{k=0}^{\infty} a_k \int_a^b (x - x_0)^k dx = \sum_{k=0}^{\infty} \frac{a_k}{k+1} ((b - x_0)^{k+1} - (a - x_0)^{k+1}).$$

(See Lemma 10.46 for the corresponding result for non-negative functions.)

In particular, in the special case $[a, b] = [0, 1]$, $x_0 = 0$ and $a_k = \frac{1}{k+1}$, we obtain

$$\int_0^1 \sum_{k=0}^{\infty} \frac{x^k}{k+1} dx = \frac{\pi^2}{6},$$

where we used Corollary 7.28.

Note that a slight change to $a_0 = 0$ and $a_k = \frac{1}{k}$ for $k \geq 1$ gives the function $f(x) = -\ln(1 - x)$, see Example 5.66, whose integral is $\int_0^1 f(x) dx = 1$, see Example 6.57. However, this integral can be calculated way easier by using dominated convergence and a telescoping trick. (Do it!)

The next example shows that **it is essential to check for an integrable majorant**, i.e., a function g which dominates the sequence $(f_n)_{n \in \mathbb{N}}$ pointwise, if one wants to apply Lebesgue's theorem (Theorem 10.57).

Example 10.63. We have a look at the sequence of measurable functions $f_n: [0, 1] \rightarrow \mathbb{R}$ where

$$f_n(x) = \begin{cases} n & \text{if } x \in [0, \frac{1}{n}] \\ 0 & \text{else.} \end{cases}$$

First note that

$$\int_{[0,1]} f_n(x) dx = \int_{[0, \frac{1}{n}]} n dx = 1,$$

and so $\lim_{n \rightarrow \infty} \int_{[0,1]} f_n(x) dx = 1$. On the other hand $f_n(x) \rightarrow 0$ for all $x \in (0, 1]$, (i.e., for almost all $x \in (0, 1]$) and so

$$\int_{[0,1]} \lim_{n \rightarrow \infty} f_n(x) dx = 0.$$

That is, we can not interchange limit and integral in this case.

The reason is that there is no integrable function g such that $f_n \leq g$ for all $n \in \mathbb{N}$.

To show this, we observe that any function g which dominates the sequence $(f_n)_{n \in \mathbb{N}}$ has to satisfy $g \geq \sup_n f_n$ pointwise. So we can estimate the integral of g from below by

$$\int_{[0,1]} g(x) dx \geq \int_{[\frac{1}{m}, 1]} g(x) dx = \sum_{k=1}^{m-1} \int_{[\frac{1}{k+1}, \frac{1}{k}]} g(x) dx,$$

where $m \in \mathbb{N}$ is arbitrary. Since g dominates the sequence we see that

$$\sum_{k=1}^{m-1} \int_{[\frac{1}{k+1}, \frac{1}{k}]} g(x) dx \geq \sum_{k=1}^{m-1} \int_{[\frac{1}{k+1}, \frac{1}{k}]} f_k(x) dx = \sum_{k=1}^{m-1} \int_{[\frac{1}{k+1}, \frac{1}{k}]} k dx = \sum_{k=1}^{m-1} \frac{1}{k+1}.$$

Since the harmonic series is divergent (to ∞) and m was arbitrary, g cannot have a finite integral, and is therefore not integrable. So $(f_n)_{n \in \mathbb{N}}$ is not dominated by any integrable function.

Finally, let us present one of the most interesting consequences that one can get rather easily based on the above results: the **Leibniz rule for differentiation under the integral sign**.
(Yes, that's the same Leibniz from the convergence test for alternating series.)

Lemma 10.64 (Differentiation under the integral sign). *Let $(\Omega, \mathcal{A}, \mu)$ be a measure space and $I \subset \mathbb{R}$ be an open interval. If a function $f: \Omega \times I \rightarrow \mathbb{R}$ satisfies*

- $x \mapsto f(x, t)$ is integrable for each fixed $t \in I$,
- $t \mapsto f(x, t)$ is differentiable for each fixed $x \in \Omega$, and
- there is some integrable function g with $|\frac{\partial f}{\partial t}(x, t)| \leq g(x)$ for all $x \in \Omega$ and $t \in I$.

Then,

$$\frac{d}{dt} \int_{\Omega} f(x, t) d\mu(x) = \int_{\Omega} \frac{\partial}{\partial t} f(x, t) d\mu(x) \quad \text{for all } t \in I.$$

This lemma is particularly useful when additional parameters are involved, or if we introduce them to compute complicated integrals. (Note that we use “ d ” on the left and “ ∂ ” on the right, because we differentiate a univariate function left, and a multivariate function right.)

Proof. Let $F(t) := \int_{\Omega} f(x, t) d\mu(x)$. Then, we want to prove that $F'(t) = \int_{\Omega} \frac{\partial}{\partial t} f(x, t) d\mu(x)$. For this, let $h_n \rightarrow 0$ be an arbitrary null sequence and define $f_n(x, t) := \frac{1}{h_n} (f(x, t + h_n) - f(x, t))$. We see that $\lim_{n \rightarrow \infty} f_n(x, t) = \frac{\partial}{\partial t} f(x, t)$ for all $x \in \Omega$ and $t \in I$. Moreover, we obtain from the mean value theorem (Theorem 5.34) that $f_n(x, t) = \frac{\partial}{\partial t} f(x, \xi)$ for some $\xi \in (t, t + h_n)$. Therefore, from the assumption, $|f_n(x, t)| \leq g(x)$ for all $x \in \Omega$ and $t \in I$, i.e., (f_n) is dominated by the integrable function g . By using the dominated convergence theorem we obtain for all $t \in I$ that

$$\begin{aligned} \lim_{n \rightarrow \infty} \frac{F(t + h_n) - F(t)}{h_n} &= \lim_{n \rightarrow \infty} \int_{\Omega} \frac{f(x, t + h_n) - f(x, t)}{h_n} d\mu(x) \\ &= \lim_{n \rightarrow \infty} \int_{\Omega} f_n(x, t) d\mu(x) = \int_{\Omega} \lim_{n \rightarrow \infty} f_n(x, t) d\mu(x) \\ &= \int_{\Omega} \frac{\partial}{\partial t} f(x, t) d\mu(x). \end{aligned}$$

Since this holds for arbitrary sequences (h_n) we obtain that F is differentiable and satisfies the desired equation. \square

One particularly fascinating example for this rule is the so-called *Euler integral of the second kind*, which defines the famous **gamma function**.

Example 10.65 (Gamma function). For arbitrary $n \in \mathbb{N}$ we have that

$$\int_0^\infty x^n e^{-x} dx = n!.$$

To see this relation, we introduce a new parameter $t \geq 1$ and use the formula

$$\int_0^\infty e^{-tx} dx = \frac{1}{t}.$$

Now, $f(x, t) = e^{-tx}$ satisfies all assumptions of the lemma above (with $I = [1, \infty)$) so we can just compute the derivative of both sides of the last equation (and interchange integral and derivative) to obtain

$$\int_0^\infty -x e^{-tx} dx = -\frac{1}{t^2},$$

and so $\int_0^\infty x e^{-tx} dx = \frac{1}{t^2}$. Differentiating again leads to $\int_0^\infty x^2 e^{-tx} dx = \frac{2}{t^3}$, and then $\int_0^\infty x^3 e^{-tx} dx = \frac{3!}{t^4}$ and so on. Repeating this n times, we obtain

$$\int_0^\infty x^n e^{-tx} dx = \frac{n!}{t^{n+1}}$$

for all $t \in \mathbb{R}$. This shows the result by using $t = 1$.

Note that this function can also be defined more general as

$$\Gamma(z) := \int_0^\infty x^{z-1} e^{-x} dx,$$

where $z > 0$ (or even $z \in \mathbb{C}$ with $\operatorname{Re}(z) > 0$). This is the *gamma function* which is important to many fields of mathematics. Besides the value computed above, there are not many explicit values. Even more, finding the zeros of this function ($\Gamma(z) = 0$) turned out to be one of the hardest problems in nowadays mathematics, called the **Riemann hypothesis**. This long-standing conjecture is closely related to number theory and the “distribution of prime numbers”. (We clearly can’t go into details here.)

Example 10.66. To see a more practical example, we want to compute

$$\int_0^1 \frac{x^{42}-1}{\ln(x)} dx.$$

(Note that also some computer algebra systems struggle to provide a precise value.)

Since other standard methods do not work, we think about introducing a new variable that makes things easier. For this, note that $f(x, t) = \frac{x^t - 1}{\ln(x)}$ satisfies $\frac{\partial}{\partial t} f(x, t) = x^t$, and therefore all assumptions for differentiation under the integral, if we restrict t to values in a bounded interval, say $t \in [0, 42]$. Denoting $F(t) := \int_0^1 f(x, t) dx$, we see that we are looking for the value $F(42)$, and by the above lemma

$$F'(t) = \int_0^1 \frac{\partial}{\partial t} f(x, t) dx = \int_0^1 x^t dx = \frac{1}{t+1},$$

which holds for all $t > -1$. By computing the antiderivative, we obtain that $F(t) = \ln(t+1) + C$ and, since $F(0) = 0$, we see that $C = 0$ and therefore

$$F(t) = \int_0^1 f(x, t) dx = \ln(t+1).$$

Coming back to our original example, we have $\int_0^1 \frac{x^{42}-1}{\ln(x)} dx = \ln(43)$.

Note that we never needed an antiderivative of $\frac{x^{42}-1}{\ln(x)}$, which actually cannot be given explicitly.

10.4 Product measures and Fubini’s theorem

We now turn (again) to integrals of multivariate functions. Recall that already our ‘old’ definition of an integral allows to use results like Fubini’s theorem (Theorem 8.116) to compute a multivariate integral by computing several univariate integrals sequentially, at least if the function is continuous and defined on a box. (This is very restrictive!) Here, we want to generalize this to rather arbitrary situations. It turns out that *product spaces* are the appropriate tool to do so.

Again, most of the results here hold in rather general situations. But since such a treatment would need quite a lot of tedious considerations, we only consider our most important examples, the Lebesgue measure and the counting measure, and comment on more general situations. We omit most of the proofs.

Definition 10.67. Let $(\Omega_1, \mathcal{A}_1, \mu_1)$, $(\Omega_2, \mathcal{A}_2, \mu_2)$ be measure spaces.

The **product σ -algebra**, denoted by $\mathcal{A}_1 \otimes \mathcal{A}_2$, is the smallest σ -algebra (over $\Omega_1 \times \Omega_2$) that contains all sets of the form $A_1 \times A_2$ with $A_1 \in \mathcal{A}_1$, $A_2 \in \mathcal{A}_2$, i.e., $\mathcal{A}_1 \otimes \mathcal{A}_2 := \sigma(\mathcal{A}_1 \times \mathcal{A}_2)$.

A measure μ on $\mathcal{A}_1 \otimes \mathcal{A}_2$ which satisfies

$$\mu(A_1 \times A_2) = \mu_1(A_1) \cdot \mu_2(A_2)$$

is called **product measure** of μ_1 and μ_2 .

(Note that $\mathcal{A}_1 \otimes \mathcal{A}_2$ does not only contain sets of the form $A_1 \times A_2$.)

It is not clear at all that such a product measure exists and even if, there might be many different such measures. (Note that we only pose restrictions on Cartesian products.) E.g., for the Lebesgue measure on \mathbb{R}^2 and the counting measure on \mathbb{N}^2 , it is clear that the measure of a product set is the product of the 'univariate' measures. So, they are product measures.

Luckily, we will see shortly that such a product measure is unique, and how to compute the measure for general sets, if the two terms (i.e., measures) satisfy a certain technical condition. Note that we have seen this already as assumption 3) of Theorem 10.14.

Definition 10.68. Let $(\Omega, \mathcal{A}, \mu)$ be a measure space.

If $\mu(\Omega) < \infty$, then μ is called a **finite measure** and $(\Omega, \mathcal{A}, \mu)$ is a **finite measure space**.

Moreover, if there exist $E_1 \subset E_2 \subset \dots \in \mathcal{A}$ with $\Omega = \bigcup_{i=1}^{\infty} E_i$ and $\mu(E_i) < \infty$ for all i , then μ is called a **σ -finite measure** and $(\Omega, \mathcal{A}, \mu)$ is called **σ -finite measure space**.

Clearly, finite measures are also σ -finite measures. (Just use $E_i = \Omega$ for all i .)

Example 10.69. The Lebesgue measure on a bounded interval, and the counting measure on a finite set are obviously finite measures.

Moreover, and as we have already seen, the Lebesgue measure on \mathbb{R} and the counting measure on \mathbb{N} or \mathbb{Z} , which are not finite, are σ -finite measures. Use the sets $E_k = [-k, k]$ or $E_k = \{-k, \dots, k\}$, $k \in \mathbb{N}$, respectively.

Example 10.70. A typical example of a measure space $(\Omega, \mathcal{A}, \mu)$ that is not σ -finite, is a counting measure on an uncountable set. For example, let $\Omega = [0, 1]$ (which is uncountable) with an arbitrary σ -algebra \mathcal{A} , and the counting measure μ . That is $\mu(A) = \#A$ (e.g., $\mu(\{0, \frac{1}{2}, 1\}) = 3$ and $\mu([\frac{1}{4}, \frac{1}{2}]) = \infty$). This measure is not σ -finite, since any countable union of finite sets is countable. That's why many results that follow would be false if we consider this measure.

Our goal is to compute multivariate integrals with the help of univariate integrals. For this, let us first consider the corresponding measures and define the **x -section** and **y -section of a set** $A \in \mathcal{A}_1 \otimes \mathcal{A}_2$ by

$$A^x := \{y \in \Omega_2 : (x, y) \in A\} \quad \text{and} \quad A_y := \{x \in \Omega_1 : (x, y) \in A\}.$$

For example, if $A = A_1 \times A_2$, then, we have $A^x = A_2$ for $x \in A_1$ and $A^x = \emptyset$ otherwise.

For definiteness, we clearly need that these sets are measurable. That is, sections of measurable sets are all measurable.

Lemma 10.71. Let $(\Omega_1, \mathcal{A}_1)$ and $(\Omega_2, \mathcal{A}_2)$ be measurable spaces. Then, for any $A \in \mathcal{A}_1 \otimes \mathcal{A}_2$, $x \in \Omega_1$ and $y \in \Omega_2$, we have $A^x \in \mathcal{A}_2$ and $A_y \in \mathcal{A}_1$.

Proof. If $A = A_1 \times A_2$ for some $A_1 \in \mathcal{A}_1$ and $A_2 \in \mathcal{A}_2$ (i.e., A is a 'rectangle'), then, as stated above, A^x is either \emptyset or A_2 , so $A^x \in \mathcal{A}_2$ for every $x \in \Omega_1$.

Now let $\mathcal{F} = \{A \in \mathcal{A}_1 \otimes \mathcal{A}_2 : A^x \in \mathcal{A}_2 \text{ for every } x \in \Omega_1\}$. By the above, all 'rectangles' are in \mathcal{F} . Moreover, we clearly have for every sequence of sets E, E_1, E_2, \dots that

$$\left(\bigcup_{i=1}^{\infty} E_i \right)^x = \bigcup_{i=1}^{\infty} (E_i)^x \quad \text{and} \quad (E^c)^x = (E^x)^c.$$

(Verify that!) This shows that \mathcal{F} is closed under countable unions and complements. It is therefore a σ -algebra. However, $\mathcal{A}_1 \otimes \mathcal{A}_2$ is the smallest σ -algebra that contains all 'rectangles'. So, we have $\mathcal{F} = \mathcal{A}_1 \otimes \mathcal{A}_2$, which proves the claim for A^x . The sets A_y can be treated in the same way. \square

The following result shows that, under some assumptions, the product measure is unique and that the measure of a set can be computed by integrating over the measures of the sections, sometimes called **slices**, of the set.

Theorem 10.72. *Let $(\Omega_1, \mathcal{A}_1, \mu_1), (\Omega_2, \mathcal{A}_2, \mu_2)$ be σ -finite measure spaces. Then, there is a unique product measure of μ_1 and μ_2 on $(\Omega_1 \times \Omega_2, \mathcal{A}_1 \otimes \mathcal{A}_2)$, which is denoted by $\mu_1 \otimes \mu_2$.*

For every $A \in \mathcal{A}_1 \otimes \mathcal{A}_2$, this measure satisfies

$$(\mu_1 \otimes \mu_2)(A) = \int_{\Omega_1} \mu_2(A^x) d\mu_1(x) = \int_{\Omega_2} \mu_1(A_y) d\mu_2(y).$$

Sketch of proof. The uniqueness follows from a variant of Caratheodory's extension theorem (Theorem 10.14). We do not discuss the details here, but note that this requires the σ -finiteness.

However, once uniqueness is established, it is enough to verify that $\nu(A) := \int_{\Omega_1} \mu_2(A^x) d\mu_1(x)$ is a product measure of μ_1 and μ_2 . The uniqueness then implies the first equality $\nu = \mu_1 \otimes \mu_2$.

The proof of the second equality follows the same lines, but with $\nu(A) := \int_{\Omega_2} \mu_2(A_y) d\mu_2(y)$.

Since $A^x \in \mathcal{A}_2$ for every $A \in \mathcal{A}_1 \otimes \mathcal{A}_2$ by Lemma 10.71, we first have that $\mu_2(A^x)$ is well-defined. We need to show that $g_A(x) := \mu_2(A^x)$ is also \mathcal{A}_1 -measurable.

First, for 'rectangles' $A = A_1 \times A_2$ with $A_1 \in \mathcal{A}_1$ and $A_2 \in \mathcal{A}_2$, we have $A^x \in \{\emptyset, A_2\}$ depending on $x \in A_1$ or not. This shows $g_A(x) = \chi_{A_1}(x) \cdot \mu_2(A_2)$, which is a \mathcal{A}_1 -measurable function (in x) as a constant times an indicator function. Therefore, we have $\nu(A) = \mu_1(A_1) \cdot \mu_2(A_2)$ for $A = A_1 \times A_2$ and it remains to show that ν is a measure on $\mathcal{A}_1 \otimes \mathcal{A}_2$.

The set $\mathcal{F} = \{A \in \mathcal{A}_1 \otimes \mathcal{A}_2 : g_A \text{ is measurable}\}$ now contains all 'rectangles'. Using that A^x and B^x are disjoint (in \mathcal{A}_2) for arbitrary disjoint $A, B \in \mathcal{A}_1 \otimes \mathcal{A}_2$, we see that $g_{A \cup B} = g_A + g_B$. With the monotone convergence theorem we obtain that \mathcal{F} is closed under countable disjoint unions and consequently that ν is σ -additive on \mathcal{F} . (Verify this!) To consider complements, let us first assume that μ_1, μ_2 are finite measures. Then, $(A^x)^c = (A^c)^x$ implies $g_{A^c}(x) = \mu_2(\Omega_2) - g_A(x)$, and so that $A^c \in \mathcal{F}$ for $A \in \mathcal{F}$. Hence, \mathcal{F} is a σ -algebra and $\mathcal{F} = \mathcal{A}_1 \otimes \mathcal{A}_2$. If μ_1, μ_2 are 'only' σ -finite, let $E_1 \subset E_2 \subset \dots \in \mathcal{A}_2$ as in Definition 10.68. By the arguments from above, we have that the functions $g_{A \cap E_n}$ are measurable for every $A \in \mathcal{A}_1 \otimes \mathcal{A}_2$. (We only need for the complement that $g_{A^c \cap E_n}(x) = \mu_2(E_n) - g_{A \cap E_n}(x)$ is measurable since $\mu_2(E_n) < \infty$ for all n .) Finally, since $g_{A \cap E_n}$ is increasing in n , we obtain that $g_A = \lim_n g_{A \cap E_n}$ is measurable for all $A \in \mathcal{A}_1 \otimes \mathcal{A}_2$. Overall, we verified that $\nu(A)$, i.e., the integral over the non-negative measurable function $g_A(x) = \mu_2(A^x)$, is well-defined and σ -additive on $\mathcal{A}_1 \otimes \mathcal{A}_2$, and therefore a measure. \square

Let us now verify that our important examples satisfy these assumptions.

Example 10.73. We consider the counting measure μ on \mathbb{N}^2 , which is the product of the counting measure μ_1 on \mathbb{N} with itself.

Let us first show that $\mathcal{P}(\mathbb{N}) \otimes \mathcal{P}(\mathbb{N}) = \mathcal{P}(\mathbb{N}^2) = \mathcal{P}(\mathbb{N} \times \mathbb{N})$. Since any $A \times B$ with $A, B \subset \mathbb{N}$ is contained in $\mathcal{P}(\mathbb{N}^2)$, we have $\mathcal{P}(\mathbb{N}) \otimes \mathcal{P}(\mathbb{N}) \subset \mathcal{P}(\mathbb{N}^2)$. (Recall that $\mathcal{P}(\mathbb{N}) \otimes \mathcal{P}(\mathbb{N})$ is the smallest σ -algebra that contains

all these products.)

Now, any $M \in \mathcal{P}(\mathbb{N}^2)$ is a countable set and can be written as disjoint union of the form

$$M = \bigcup_{i=1}^{\infty} (A_i \times B_i),$$

where each $A_i, B_i \subset \mathbb{N}$ is either a set containing one element or the empty set. Since $\mathcal{P}(\mathbb{N}) \otimes \mathcal{P}(\mathbb{N})$ is a σ -algebra (by definition) and contains all sets of the form $A_i \times B_i \subset \mathbb{N} \times \mathbb{N}$, we see that $M \in \mathcal{P}(\mathbb{N}) \otimes \mathcal{P}(\mathbb{N})$. Hence, $\mathcal{P}(\mathbb{N}^2) \subset \mathcal{P}(\mathbb{N}) \otimes \mathcal{P}(\mathbb{N})$, which shows

$$\mathcal{P}(\mathbb{N}^2) = \mathcal{P}(\mathbb{N}) \otimes \mathcal{P}(\mathbb{N}).$$

To verify that counting measure μ on \mathbb{N}^2 is a product measure of the counting measure on \mathbb{N} with itself, we only need the calculation rule for the cardinality of Cartesian products, i.e.,

$$\mu(A_1 \times A_2) = \#(A_1 \times A_2) = \#A_1 \cdot \#A_2 = \mu_1(A_1) \cdot \mu_1(A_2).$$

By the above theorem, we obtain that this determines μ uniquely, and we can compute the measure of a general set $A \in \mathcal{P}(\mathbb{N}^2)$ by

$$\mu(A) = \sum_{k=1}^{\infty} \mu_1(A^k) = \sum_{k=1}^{\infty} \#\{\ell \in \mathbb{N}: (k, \ell) \in A\}.$$

Moreover, the last theorem shows that this is the only measure with this property.

The situation is a bit more complicated when we talk about the Lebesgue measure, because it is not a product measure. However, it is very close to one and we will see that the following results also holds for the Lebesgue measure. However, let us first show the desired results for product measures.

We therefore clearly need that **sections of functions** are measurable. For this, let us define **x -section** and **y -section of a function** $f: \Omega_1 \times \Omega_2 \rightarrow \mathbb{R}$ by

$$f^x(y) := f(x, y) \quad \text{and} \quad f_y(x) := f(x, y)$$

such that $f^x: \Omega_2 \rightarrow \mathbb{R}$ and $f_y: \Omega_1 \rightarrow \mathbb{R}$ for all $x \in \Omega_1$ and $y \in \Omega_2$. We have the following.

Lemma 10.74. *Let $(\Omega_1, \mathcal{A}_1)$ and $(\Omega_2, \mathcal{A}_2)$ be measurable spaces. Then, for any $x \in \Omega_1$, $y \in \Omega_2$ and any $(\mathcal{A}_1 \otimes \mathcal{A}_2)$ -measurable function $f: \Omega_1 \times \Omega_2 \rightarrow \mathbb{R}$, we have that f^x and f_y are measurable.*

Proof. Since f is measurable, we have that $A = \{(x, y) \in \Omega_1 \times \Omega_2 : f(x, y) > a\}$ is measurable for any $a \in \mathbb{R}$. Therefore, $\{y \in \Omega_2 : f^x(y) > a\} = A^x \in \mathcal{A}_2$ is also measurable for any $a \in \mathbb{R}$ by Lemma 10.71, proving that f^x is measurable. The same works for f_y . \square

If we now consider integration with respect to the (unique) product measure, the above theorem indicates that we can again consider integrals with the two factors iteratively, and in arbitrary order. This is shown in the following theorem. As for Lebesgue's theorem, we start with a statement on non-negative functions first, which is called *Tonelli's theorem*.

Theorem 10.75 (Tonelli). Let $(\Omega_1, \mathcal{A}_1, \mu_1)$, $(\Omega_2, \mathcal{A}_2, \mu_2)$ be σ -finite measure spaces. Let $f: \Omega_1 \times \Omega_2 \rightarrow [0, \infty)$ be measurable (precisely, $(\mathcal{A}_1 \otimes \mathcal{A}_2)$ -measurable). Then,

- $\int_{\Omega_2} f^x d\mu_2$ is \mathcal{A}_1 -measurable (as function of $x \in \Omega_1$),
- $\int_{\Omega_1} f_y d\mu_1$ is \mathcal{A}_2 -measurable (as function of $y \in \Omega_2$),

and we have

$$\int_{\Omega_1 \times \Omega_2} f d(\mu_1 \otimes \mu_2) = \int_{\Omega_1} \left(\int_{\Omega_2} f d\mu_2 \right) d\mu_1 = \int_{\Omega_2} \left(\int_{\Omega_1} f d\mu_1 \right) d\mu_2,$$

or, written out,

$$\int_{\Omega_1 \times \Omega_2} f d(\mu_1 \otimes \mu_2) = \int_{\Omega_1} \left(\int_{\Omega_2} f(x, y) d\mu_2(y) \right) d\mu_1(x) = \int_{\Omega_2} \left(\int_{\Omega_1} f(x, y) d\mu_1(x) \right) d\mu_2(y).$$

Again, all integrals might be ∞ here.

Proof. Note that, for indicator functions $f = \chi_A$ with $A \in \mathcal{A}_1 \otimes \mathcal{A}_2$, this is precisely the statement of Theorem 10.72. By linearity of the integral, it holds for arbitrary simple functions. Using a monotone sequence of simple functions (f_n) with $f_n \rightarrow f$, and note that all limits and integrals can be interchanged due to the monotone convergence theorem, we obtain the result. \square

Tonelli's theorem is a powerful tool as it shows that **one can always interchange the two integrals for a non-negative function**, if it is measurable.

Example 10.76. We can use Tonelli's theorem to compute infinite sums with more than one index. For example, if we want to compute the sum of $\frac{\ell^{k-3}}{(\ell+1)^k}$ over all $k, \ell = 1, 2, \dots$, we can change the order of the summation signs, because all terms are non-negative. We obtain

$$\sum_{k, \ell=1}^{\infty} \frac{\ell^{k-3}}{(\ell+1)^k} = \sum_{k=1}^{\infty} \sum_{\ell=1}^{\infty} \frac{1}{\ell^3} \left(\frac{\ell}{\ell+1} \right)^k = \sum_{\ell=1}^{\infty} \frac{1}{\ell^3} \sum_{k=1}^{\infty} \left(\frac{\ell}{\ell+1} \right)^k = \sum_{\ell=1}^{\infty} \frac{1}{\ell^2} = \frac{\pi^2}{6}.$$

Using the last step of the standard machinery, we finally obtain the result for integrable functions, which is the famous **Theorem of Fubini**. Recall that a function f is integrable if it is measurable and $|f|$ has finite integral.

Theorem 10.77 (Fubini). Let $(\Omega_1, \mathcal{A}_1, \mu_1)$, $(\Omega_2, \mathcal{A}_2, \mu_2)$ be σ -finite measure spaces. Let $f: \Omega_1 \times \Omega_2 \rightarrow \mathbb{R}$ be integrable on $\Omega_1 \times \Omega_2$. Then,

- f^x is μ_2 -integrable for almost every $x \in \Omega_1$,
- f_y is μ_1 -integrable for almost every $y \in \Omega_2$,

and we have

$$\int_{\Omega_1 \times \Omega_2} f d(\mu_1 \otimes \mu_2) = \int_{\Omega_1} \left(\int_{\Omega_2} f d\mu_2 \right) d\mu_1 = \int_{\Omega_2} \left(\int_{\Omega_1} f d\mu_1 \right) d\mu_2.$$

Note that we have the integrability of $f^x(y)$ (w.r.t. y) only for almost every x in general, since integrability of f is a weak assumption. However, this does not affect the integrals.

Proof. Since f is integrable, we can write it by the difference of the non-negative measurable functions $f^+, f^-: \Omega_1 \times \Omega_2 \rightarrow \mathbb{R}$ and decompose the integrals such that $\int_{\Omega_1 \times \Omega_2} f^+ d(\mu_1 \otimes \mu_2) < \infty$ and $\int_{\Omega_1 \times \Omega_2} f^- d(\mu_1 \otimes \mu_2) < \infty$. Applying Tonelli's theorem, we obtain

$$\int_{\Omega_1 \times \Omega_2} f^+ d(\mu_1 \otimes \mu_2) = \int_{\Omega_1} \left(\int_{\Omega_2} f^+ d\mu_2 \right) d\mu_1 = \int_{\Omega_2} \left(\int_{\Omega_1} f^+ d\mu_1 \right) d\mu_2 < \infty,$$

and so for f^- , which implies the stated equality. Moreover, $\int_{\Omega_2} f^+(x, y) d\mu_2(y) < \infty$ and $\int_{\Omega_2} f^-(x, y) d\mu_2(y) < \infty$ for almost every x , and hence that f^x is integrable a.e., and the same works for f_y . \square

Note that Fubini's theorem is unsatisfactory as the assumption that f is integrable is rather general, but it still needs to be verified. Luckily, for this we only need to check if the non-negative function $|f|$ has finite integral, and we can apply Tonelli's theorem. Precisely, Tonelli's theorem shows

$$\int_{\Omega_1 \times \Omega_2} |f| d(\mu_1 \otimes \mu_2) < \infty \iff \int_{\Omega_1} \left(\int_{\Omega_2} |f| d\mu_2 \right) d\mu_1 < \infty \iff \int_{\Omega_2} \left(\int_{\Omega_1} |f| d\mu_1 \right) d\mu_2 < \infty.$$

So, we only need to check finiteness of one of the iterated integrals.

Let us finally turn back to the Lebesgue measure.

It is not hard to show that $\mathcal{B}(\mathbb{R}) \otimes \mathcal{B}(\mathbb{R}) = \mathcal{B}(\mathbb{R}^2)$. A proof follows similar lines as Example 10.73 and is based on the fact, that the generator of $\mathcal{B}(\mathbb{R}^d)$ is of product form, see Definition 10.15.

However, the same does not hold for the Lebesgue- σ -algebra. Recall that we obtained $\mathcal{L}(\mathbb{R}^d)$ by 'completion', i.e., we added all subsets of null sets from $\mathcal{B}(\mathbb{R}^d)$, see Corollary 10.17. In other words, every $A \in \mathcal{L}(\mathbb{R}^d)$ is of the form $A = B \cup N$ with $B \in \mathcal{B}(\mathbb{R}^d)$ and $N \subset M \in \mathcal{B}(\mathbb{R}^d)$ with $\lambda_d(M) = 0$. This does not work together well with the product and we have $\mathcal{L}(\mathbb{R}) \otimes \mathcal{L}(\mathbb{R}) \neq \mathcal{L}(\mathbb{R}^2)$, i.e., $\mathcal{L}(\mathbb{R}^2)$ is not a product- σ -algebra. For example, for $A \notin \mathcal{L}(\mathbb{R})$ we have $\{0\} \times A \notin \mathcal{L}(\mathbb{R}) \otimes \mathcal{L}(\mathbb{R})$, but still $\{0\} \times A \in \mathcal{L}(\mathbb{R}^2)$, because it is a subset of the null set $\{0\} \times \mathbb{R} \in \mathcal{B}(\mathbb{R}^2)$. Therefore, **the Lebesgue measure λ_2 is not a product measure**. (One might call it the 'completion' of a product measure.) Therefore, some of the results above don't hold as stated. In particular, we have just seen that not every section of a measurable set is measurable, in contrast to the case of product measure spaces, see Lemma 10.71.

However, we obtain that this still holds for almost all sections, and thereby also obtain a nice characterization of null sets.

Lemma 10.78. *Let $A \in \mathcal{L}(\mathbb{R}^2)$ be a (Lebesgue-)measurable set.*

Then, the sections $A^x, A_y \subset \mathbb{R}$ are measurable, i.e., $A^x, A_y \in \mathcal{L}(\mathbb{R})$, for almost every $x, y \in \mathbb{R}$.

Moreover, we have that $N \subset \mathbb{R}^2$ is a null set if and only if N^x is a null set for a.e. $x \in \mathbb{R}$. (Similar for y .)

Remark 10.79. The last result would hold for general 'completions of product measures'. However, for incomplete measures we would need to add the assumption that N is measurable.

Proof. By construction of the Lebesgue- σ -algebra we have that every $A \in \mathcal{L}(\mathbb{R}^2)$ is of the form $A = B \cup N$ with $B \in \mathcal{B}(\mathbb{R}^2)$ and $N \subset M \in \mathcal{B}(\mathbb{R}^2)$ with $\lambda_2(M) = 0$. Since $\mathcal{B}(\mathbb{R}^2)$ is a product- σ -algebra, we have that $B^x, M^x \in \mathcal{B}(\mathbb{R})$ for every $x \in \mathbb{R}$. Moreover, by Theorem 10.72 (or Tonelli's theorem) we obtain that M^x is a null set for almost every x is equivalent to M being a null set. (If $\lambda_1(M^x) > 0$ for a set of measure > 0 , then $\lambda_2(M) > 0$, and vice versa.) Since $B^x \subset A^x \subset B^x \cup M^x$, with $B^x, B^x \cup M^x \in \mathcal{B}(\mathbb{R})$, and M^x is a null set for a.e. x , we have that $A^x \in \mathcal{L}(\mathbb{R})$ for a.e. x .

The above already shows that almost all sections of a null set are null sets. For the reverse direction assume that N^x is a null set for a.e. x , i.e., there exist $M^x \in \mathcal{B}(\mathbb{R})$ with $N^x \subset M^x$ and $\lambda_1(M^x) = 0$ for all $x \in \mathbb{R} \setminus N'$, where $N' \in \mathcal{B}(\mathbb{R})$ with $\lambda_1(N') = 0$. Then, the set $B \in \mathcal{B}(\mathbb{R}^2)$ with $B^x = M^x$ for $x \in \mathbb{R} \setminus N'$ and $B^x = \mathbb{R}$ for $x \in N'$ is still a null set (Why?) and satisfies $N \subset B$. This shows that N is a null set. \square

With this we obtain the following result, which combines the above findings.

Theorem 10.80 (Fubini for the Lebesgue measure). *Let $f: \mathbb{R}^2 \rightarrow \mathbb{R}$ be Lebesgue-measurable such that either*

- $\int_{\mathbb{R}} \int_{\mathbb{R}} |f(x, y)| dy dx < \infty$ or
- $\int_{\mathbb{R}} \int_{\mathbb{R}} |f(x, y)| dx dy < \infty$,

then, f is integrable and we have

$$\int_{\mathbb{R}^2} f d\lambda_2 = \int_{\mathbb{R}^2} f(x, y) d(x, y) = \int_{\mathbb{R}} \int_{\mathbb{R}} f(x, y) dy dx = \int_{\mathbb{R}} \int_{\mathbb{R}} f(x, y) dx dy.$$

In particular, we have that f^x and f_y are integrable for almost every x and y .

In addition, for every measurable set $A \subset \mathbb{R}^2$, we have

$$\int_A f d\lambda_2 = \int_{\mathbb{R}} \int_{A^x} f(x, y) dy dx = \int_{\mathbb{R}} \int_{A_y} f(x, y) dx dy$$

if one of the iterated integrals exist.

It is important to note here, that we also obtain measurability only for almost every section, in contrast to the original formulation of Fubini's theorem. This is due to the fact that λ_2 is not a product measure, and is based on the last lemma.

Proof. By construction of the Lebesgue- σ -algebra we have that every $A \in \mathcal{L}(\mathbb{R}^2)$ is of the form $A = B \cup N$ with $B \in \mathcal{B}(\mathbb{R}^2)$ and $N \subset \mathbb{R}^2$ with $\lambda_2(N) = 0$. In particular, $\lambda_2(A) = \lambda_2(B)$. Using this, and the standard machinery, one obtains that every Lebesgue-integrable function f coincides a.e. with a Borel-measurable function g , i.e., $f(x, y) = g(x, y)$ for all $(x, y) \in \mathbb{R}^2 \setminus N$ with $\lambda_2(N) = 0$. Denoting the Borel-measure on \mathbb{R}^d by μ_d and recalling that λ_2 is a complete measure, we obtain from Lemma 10.56 that the integral of f (w.r.t. λ_2) and g (w.r.t. μ_2) are equal, i.e. $\int f d\lambda_2 = \int g d\mu_2$. (Recall that $\lambda_2(B) = \mu_2(B)$ for all $B \in \mathcal{B}(\mathbb{R}^2)$; λ_2 is just defined for more sets.) Since the Borel-measure is a product measure, we can apply Theorem 10.77 to see that $\int g d\mu_2$ can be written by iterated integrals. To see that these iterated integrals equal the iterated (Lebesgue-)integrals of f , we use that the sections f^x and f_y are a.e. equal to the (integrable) g^x and g_y for almost all $x, y \in \mathbb{R}$, respectively. (Verify this! Hint: Use that sections of null sets are null sets.) By Lemma 10.56, we obtain that this implies that f^x and f_y are Lebesgue-integrable for a.e. x and y , and that the iterated integrals coincide with the integrals of g^x and g_y . For the last statement, we just use that $\int_A f d\lambda_2 = \int_{\mathbb{R}^2} \chi_A \cdot f d\lambda_2$. \square

Let us first see an easy non-negative function.

Example 10.81. We see that the integral of the function $f(x, y) = \frac{\sin(x)}{x}$ over the triangle $A := \{(x, y) \in [0, \pi]^2 : x > y\}$ is

$$\int_A f d\lambda_2 = \int_0^\pi \int_{A^x} \frac{\sin(x)}{x} dy dx = \int_0^\pi \int_0^x \frac{\sin(x)}{x} dy dx = \int_0^\pi \sin(x) dx = 2,$$

where we use that $A^x = [0, x)$.

The next example is a bit more involved.

Example 10.82. We want to evaluate the integral of $f(x, y) := x e^{-y} \frac{\sin(y)}{y}$ over the (unbounded) domain $E := \{(x, y) \in \mathbb{R}^2 : 0 \leq x \leq \frac{y}{2}\}$.

To apply Fubini's theorem, we need that f is integrable over E , i.e., that $\chi_E \cdot f$ is integrable.

For this, note that $|f(x, y)| = \left| x e^{-y} \frac{\sin(y)}{y} \right| \leq x e^{-y}$ for all $(x, y) \in E$. To this non-negative function we apply Tonelli's theorem to obtain

$$\begin{aligned} \int_E x e^{-y} d\lambda_2(x, y) &= \int_0^\infty \int_0^{y/2} x e^{-y} dx dy = \int_0^\infty \frac{y^2}{8} e^{-y} dy \\ &= \int_0^\infty \left(\frac{y^2}{8} e^{-y/2} \right) e^{-y/2} dy \leq \int_0^\infty K \cdot e^{-y/2} dy < \infty \end{aligned}$$

for some constant $K < \infty$. (Why?)

Therefore, f is integrable and we can compute

$$\int_E f d\lambda_2 = \int_0^\infty \int_0^{y/2} x e^{-y} \frac{\sin(y)}{y} dx dy = \int_0^\infty \frac{y}{8} e^{-y} \sin(y) dy.$$

Using integration by parts, i.e., $\int uv' dy = uv - \int u'v dy$ with $u(y) = \frac{y}{8}$ and $v'(y) = e^{-y} \sin(y)$, such that $u'(y) = \frac{1}{8}$ and $v(y) = -\frac{e^{-y}}{2}(\cos(y) + \sin(y))$, we obtain

$$\begin{aligned} \int_E f d\lambda_2 &= \left[\frac{y}{8} \frac{-e^{-y}}{2} (\cos(y) + \sin(y)) \right]_0^\infty - \int_0^\infty \frac{-e^{-y}}{16} (\cos(y) + \sin(y)) dy \\ &= \frac{-1}{16} \int_0^\infty -e^{-y} (\cos(y) + \sin(y)) dy = \frac{-1}{16} [e^{-y} \cos(y)]_0^\infty = \frac{1}{16}. \end{aligned}$$

In addition to calculating integrals, one can also **verify if a function is not integrable** using Fubini's theorem.

Example 10.83. Consider the triangles $T_1 := \{(x, y) \in [0, 1]^2 : x < y\}$ and $T_2 := \{(x, y) \in [0, 1]^2 : x > y\}$ and the function $f: [0, 1]^2 \rightarrow \mathbb{R}$

$$f(x, y) := \frac{1}{y^2} \cdot \chi_{T_1}(x, y) - \frac{1}{x^2} \cdot \chi_{T_2}(x, y).$$

(Note that f is only defined almost everywhere.)

If we now compute the integrals over the individual variables we obtain for each $0 < y < 1$ that

$$\int_0^1 f(x, y) dx = \int_0^y \frac{1}{y^2} dx - \int_y^1 \frac{1}{x^2} dx = \frac{1}{y} + \left[\frac{1}{x} \right]_y^1 = 1.$$

However, for fixed $0 < x < 1$ we obtain

$$\int_0^1 f(x, y) dy = - \int_0^x \frac{1}{y^2} dy + \int_x^1 \frac{1}{y^2} dy = -\frac{1}{x} - \left[\frac{1}{y} \right]_x^1 = -1.$$

This implies that $\int_0^1 \int_0^1 f(x, y) dx dy \neq \int_0^1 \int_0^1 f(x, y) dy dx$ and therefore that f is not integrable.

Note that the above procedure can be iterated to obtain **the product of more than two measures**. We only state here the result for the Lebesgue measure.

Theorem 10.84 (Fubini). *Let $f: \mathbb{R}^d \rightarrow \mathbb{R}$ be Lebesgue-integrable. Then, we have*

$$\int_{\mathbb{R}^d} f d\lambda_d = \int_{\mathbb{R}} \left(\int_{\mathbb{R}} \left(\cdots \left(\int_{\mathbb{R}} f(x_1, \dots, x_d) dx_d \right) \cdots \right) dx_2 \right) dx_1$$

with an arbitrary order of integration on the right.

Moreover, f is integrable if one of the iterated integrals over $|f|$ is finite.

This also justifies Theorem 8.116 and Lemma 8.122, even in a more general setting.

Example 10.85. To integrate the (clearly integrable) function $f(x, y, z) = x^2 y e^{xyz}$ over $E = [0, 1]^3$ it is beneficial to first integrate over z . We see that

$$\begin{aligned} \int_{[0,1]^3} f d\lambda_3 &= \int_0^1 \int_0^1 \int_0^1 x^2 y e^{xyz} dz dy dx = \int_0^1 \int_0^1 [xe^{xyz}]_{z=0}^1 dy dx \\ &= \int_0^1 \int_0^1 xe^{xy} - x dy dx = \int_0^1 [e^{xy} - xy]_{y=0}^1 dx \\ &= \int_0^1 e^x - x - 1 dx = e - 1 - \frac{1}{2} - 1 = e - \frac{5}{2}. \end{aligned}$$

Let us finish with a more general, and some cautionary examples.

Example 10.86. One can also consider **the product of two different measures**. For example, one might consider the function $f: \mathbb{N} \times \mathbb{R}_+ \rightarrow \mathbb{R}$, like $f(k, x) = x^{k-1}(1+x)^{-k}e^{-x}$ with $k \in \mathbb{N}$ and $x \in \mathbb{R}_+$. Denoting by ν the product of the counting measure μ (on \mathbb{N}) and the Lebesgue measure λ (on \mathbb{R}_+), we obtain that

$$\int_{\mathbb{N} \times \mathbb{R}} f d\nu = \int_0^\infty \sum_{k=1}^\infty f(k, x) d\lambda(x) = \sum_{k=1}^\infty \int_0^\infty f(k, x) d\lambda(x),$$

which is (for non-negative functions) the statement of Lemma 10.46, see also Example 10.62.

For the specific function f , the integral with respect to x might be difficult to compute. However, the sum over k just leads to a geometric series, and so $\sum_{k=1}^\infty f(k, x) = e^{-x}$. This shows $\int_{\mathbb{N} \times \mathbb{R}} f d\nu = 1$. (Verify that!)

The next example shows that **it is essential to check for integrability or non-negativity**.

Example 10.87. Let us consider summation of the terms $a_{k\ell} = \chi_k(\ell) - \chi_{k+1}(\ell)$ over $k, \ell \in \mathbb{N}$, which might be visualized by

$$\begin{matrix} 1 & -1 & 0 & 0 & \dots \\ 0 & 1 & -1 & 0 & \dots \\ 0 & 0 & 1 & -1 & \dots \\ 0 & 0 & 0 & 1 & \dots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{matrix}$$

Now, depending on the order of summation we obtain

$$\sum_{k=1}^\infty \sum_{\ell=1}^\infty a_{k\ell} = 0 \neq 1 = \sum_{\ell=1}^\infty \sum_{k=1}^\infty a_{k\ell},$$

since for the latter sum we have $\sum_{k=1}^\infty a_{k1} = 1$ and $\sum_{k=1}^\infty a_{k\ell} = 0$ for $\ell \geq 2$. In this example, it is not possible to change the order of summation (aka. integration), but note that also the conditions of Fubini's theorem fail to hold: the function is neither non-negative nor integrable.

The final example shows, that also **σ -finiteness is a necessary assumption**.

Example 10.88. Recall from Example 10.70 that $([0, 1], \mathcal{P}([0, 1]), \mu)$ with the counting measure μ is not σ -finite. The integral with respect to μ is somehow unintuitive as it can only be finite for $g: [0, 1] \rightarrow \mathbb{R}$ that are non-zero at *countably* many points. (If countably infinite, the function values must be summable.) For example, the function $g(x) = \chi_{\{0\}}(x)$ is a (simple) function with integral $\int_{[0,1]} g d\mu = 1$, but indicator functions of infinite sets (countable or uncountable), like the constant function $g(x) = \chi_{[0,1]}(x)$, or $g(x) = \chi_{\mathbb{Q}}(x)$, are not integrable w.r.t. μ . If we consider the product measure of μ and the Lebesgue

measure λ on $[0, 2]$, and the function $f(x, y) = \chi_{\{x\}}(2y)$, i.e., $f(x, y) = 1$ for $x = 2y$ and $f(x, y) = 0$ otherwise, we obtain

$$0 = \int_0^1 \int_0^2 f(x, y) d\lambda(x) d\mu(y) \neq \int_0^2 \int_0^1 f(x, y) d\mu(y) d\lambda(x) = 2.$$

Since f is measurable and non-negative, this shows that σ -finiteness is necessary in general.

10.5 Connection to probability theory

Measure theory is also the basis for (higher) probability theory. Here, we only present the basic concepts and how this relates to our previous considerations. Clearly, this is a huge subject and deserves its own lecture.

Let us start with the special kind of (finite) measures considered here.

Definition 10.89. Let $(\Omega, \mathcal{A}, \mu)$ be a measure space such that

$$\mu(\Omega) = 1.$$

Then, μ is called a **probability measure**, and $(\Omega, \mathcal{A}, \mu)$ is called a **probability space**.

A set $A \in \mathcal{A}$ is usually called an **event**.

A very common notation for probability measures is \mathbb{P} .

Example 10.90. Let $\Omega = \{1, 2, \dots, n\}$ and $\mathcal{A} = \mathcal{P}(\Omega)$. Then $\mathbb{P}: \mathcal{A} \rightarrow [0, 1]$ with

$$\mathbb{P}(A) = \frac{\#A}{n}$$

for $A \in \mathcal{A}$ is a probability measure. Clearly, we have that $\mathbb{P}(\Omega) = \frac{n}{n} = 1$, and the other requirements for \mathbb{P} being a measure are clearly also fulfilled.

In the case $n = 6$, i.e., $\Omega = \{1, 2, 3, 4, 5, 6\}$, this is the classical example to model fair dice. For example, the set $\{5\}$ describes the event that one rolls a 5, and using \mathbb{P} we see that this happens with probability $\frac{1}{6}$. The set $\{1, 2\}$ corresponds to the event that either 1 or 2 is rolled.

Probability measures are in most cases specified by a *density* w.r.t. another measure. For this, let $(\Omega, \mathcal{A}, \mu)$ be a measure space and $\rho: \mathbb{R} \rightarrow [0, \infty]$ be a non-negative measurable function such that

$$\int_{\Omega} \rho d\mu = 1.$$

We can then define a probability measure on the measurable space (Ω, \mathcal{A}) by

$$\mathbb{P}(B) := \int_B \rho d\mu, \quad B \in \mathcal{A}.$$

We call ρ the **probability density function** (p.d.f.) of \mathbb{P} w.r.t. the **reference measure** μ .

We already know that \mathbb{P} is a measure. Moreover, since $\mathbb{P}(\Omega) = 1$, μ is a probability measure.

Note that actually every non-negative integrable function ρ can be made to a probability density function by considering its **normalization** $\tilde{\rho} := \frac{1}{c}\rho$ with $c := \int_{\Omega} \rho d\mu$. (Verify this!)

Example 10.91. Consider $([0, 1], \mathcal{L}([0, 1]), \lambda)$, where λ is the Lebesgue measure on $[0, 1]$. Then, λ is a probability measure, since $\lambda([0, 1]) = 1$. More general, $\rho(x) = \frac{1}{b-a}\chi_{[a, b]}(x)$ for $a, b \in \mathbb{R}$ with $a < b$ is a probability density w.r.t. the Lebesgue measure on \mathbb{R} . The corresponding probability measure, i.e., $\mathbb{P}(A) = \frac{1}{b-a}\lambda(A \cap [a, b])$, is called the **uniform distribution on $[a, b]$** .

In the same way, one can define the uniform distribution on finite sets $\Omega \subset \mathbb{N}$, by using the counting measure μ and $\rho(x) = \frac{1}{\#\Omega}\chi_\Omega(x)$.

Some more standard examples will be shown at the end of this section.

We now want to analyze functions applied to a 'random element', e.g., to model the outcome of a process which depends on some sort of randomness. For this, consider the ground set Ω to be the **set of possible outcomes**, e.g., of an experiment. (Ω is sometimes called *sample space*.) In this context, a measurable function $X : \Omega \rightarrow \mathbb{R}$ on a probability space $(\Omega, \mathcal{A}, \mathbb{P})$ is called a **(real-valued) random variable**. That is, a random variable assigns each outcome $\omega \in \Omega$ a value $X(\omega) \in \mathbb{R}$, and we are interested in the probability that X attains specific values when ω appears according to the probability \mathbb{P} , i.e.,

$$\mathbb{P}(X \in A) := \mathbb{P}(\{\omega : X(\omega) \in A\}) = \int_{X^{-1}(A)} 1 d\mathbb{P}(\omega).$$

Remark 10.92. It is convention to denote random variables by X, Y, Z , and not by f, g, h as in previous sections. However, they are still just measurable functions. Moreover, we usually use ω instead of x or t as input, i.e., we write $X(\omega)$ for $\omega \in \Omega$, but we omit the ω whenever possible.

Note that measurability is precisely the concept we need for random variables, because it assures that all the sets $\{X \in A\}$, i.e., the preimages $X^{-1}(A)$, are measurable for each Borel set $A \in \mathcal{B}(\mathbb{R})$. This makes the probabilities $\mathbb{P}(X \in A)$ well-defined.

We can clearly also consider random variables taking values in another measurable space $(\Omega_2, \mathcal{A}_2)$ and define everything accordingly. We do not discuss that here.

Note that in the same way as \mathbb{P} is in many cases given by a density function, we can also often describe the probabilities $\mathbb{P}_X(A) := \mathbb{P}(X \in A)$ with the help of a density. However, note that A is now a subset of the codomain $X(\Omega)$ and not of Ω . It is therefore natural to use a reference measure that is tailored to the random variable.

For instance, if X is a real-valued random variable and $\rho_X : \mathbb{R} \rightarrow [0, \infty]$ is a Lebesgue-measurable function such that

$$\mathbb{P}(X \in A) = \int_A \rho_X(x) dx \quad \text{for all } A \in \mathcal{B}(\mathbb{R}),$$

then, we call ρ_X a **(continuous) probability density of X** w.r.t. the Lebesgue measure.

Note that for each such measures the probability that $X = x_0$ is 0 for any value of $x_0 \in \mathbb{R}$.

Example 10.93. Consider a *uniformly distributed* random number $\omega \in [0, 2]$, i.e., $\mathbb{P}([a, b]) = \frac{b-a}{2}$ for all $a, b \in [0, 2]$. (A density w.r.t. the Lebesgue measure of this measure is given by $\rho = \frac{1}{2}\chi_{[0,2]}$.)

Now consider the random variable $X(\omega) = \omega^2$. Since

$$\mathbb{P}(X \in [a, b]) = \mathbb{P}(\{\omega^2 \in [a, b]\}) = \mathbb{P}(\{\omega \in [\sqrt{a}, \sqrt{b}]\}) = \frac{\sqrt{b} - \sqrt{a}}{2}$$

for all $a, b \in [0, 4]$, we obtain that $\rho_X(x) = \frac{1}{2\sqrt{x}}\chi_{[0,4]}(x)$ is a density of X . (Verify this!)

Equivalently, we can treat random variables that only take discrete values, i.e., $X(\Omega) \subset \mathbb{Z}$. Such random variables are called **discrete random variables**. (It does not matter if Ω is discrete or not!) Such random variable do always have a probability density w.r.t. the counting measure μ on \mathbb{Z} , which is sometimes called *probability mass function* in this case. To see this, let X be a discrete random variable on some probability space $(\Omega, \mathcal{A}, \mathbb{P})$. We obtain for every $A \subset \mathbb{Z}$ that

$$\mathbb{P}(X \in A) = \sum_{k \in A} \mathbb{P}(X = k) = \int_A \rho_X d\mu$$

for $\rho_X(k) := \mathbb{P}(X = k)$, $k \in \mathbb{Z}$.

Example 10.94. Let us consider again a (fair) die, i.e., we consider the probability space $(\Omega, \mathcal{P}(\Omega), \mathbb{P})$ with $\Omega = \{1, 2, 3, 4, 5, 6\}$, $\mathcal{A} = \mathcal{P}(\Omega)$ and $\mathbb{P}(\{\omega\}) = \frac{1}{6}$ for $\omega = 1, \dots, 6$.

Now consider the *game* that you lose 3 (Euro or whatever) if you roll a 1 or 2, you win 5 if you roll a 6 and nothing happens in any other case. This can be modeled by the discrete random variable $X := -3\chi_{\{1,2\}} + 5\chi_{\{6\}}$. We then obtain

$$\mathbb{P}(X = -3) = \mathbb{P}(\{1, 2\}) = \frac{2}{6} = \frac{1}{3} \quad \text{and} \quad \mathbb{P}(X = 5) = \frac{1}{6}.$$

Moreover, $\mathbb{P}(X = 0) = 1 - \mathbb{P}(X = -3) - \mathbb{P}(X = 5) = \frac{1}{2}$. (Why?) Since X only attains the values 0, -3 and 5 from \mathbb{Z} , we have that all probabilities over sets that do not contain these values are just zero, e.g., $\mathbb{P}(X = 1) = 0$. We obtain that $\rho_X := \frac{1}{2}\chi_{\{0\}} + \frac{1}{3}\chi_{\{-3\}} + \frac{1}{6}\chi_{\{5\}}$ is a density of X w.r.t. the counting measure on \mathbb{Z} .

Clearly, we again want to extract certain *characteristic values* from a random variable. Here, we only discuss the two most important quantities, i.e., the expected value and the variance.

Definition 10.95. Let $(\Omega, \mathcal{A}, \mathbb{P})$ be a probability space and X a random variable.

We define the **expected value** (or **expectation** or **mean**) of X as

$$\mathbb{E}X := \mathbb{E}(X) := \int_{\Omega} X d\mathbb{P} = \int_{\Omega} X(\omega) d\mathbb{P}(\omega).$$

The **variance** of X is defined as

$$\text{Var}(X) := \mathbb{E}(|X - \mathbb{E}(X)|^2).$$

By writing $\mathbb{E}X$ and $\text{Var}(X)$ we generally assume that the expectations (aka. integrals) exist.

Remark 10.96. To compute the variance of a random variable one can also use the formula

$$\text{Var}(X) = \mathbb{E}(X^2) - \mathbb{E}(X)^2,$$

which follows from $\mathbb{E}(|X - \mathbb{E}(X)|^2) = \mathbb{E}(X^2) - 2\mathbb{E}(X \cdot \mathbb{E}(X)) + \mathbb{E}(X)^2 = \mathbb{E}(X^2) - \mathbb{E}(X)^2$. Here $\mathbb{E}(X^2) = \int_{\Omega} X^2 d\mathbb{P}$.

The variance is used to measure how far the outcome of an experiment varies from its expectation, which is of large interest because we want to know how close a single trial is to the 'expected behavior' of a random variable. Let us discuss an easy example, before we make this precise.

Example 10.97. We consider a fair dice, which can be modeled by $\Omega = \{1, 2, 3, 4, 5, 6\}$, $\mathcal{A} = \mathcal{P}(\Omega)$ and $\mathbb{P}(A) = \frac{\#A}{6}$ for $A \in \mathcal{A}$.

If we want to study the outcome of rolling dice directly, this can be modeled by considering the random variable $X(i) := i$ for $i = 1, \dots, 6$. Expectation and variance can be computed by

$$\mathbb{E}(X) = \int_{\Omega} X d\mathbb{P} = \sum_{i=1}^6 X(i) \mathbb{P}(\{i\}) = \sum_{i=1}^6 i \frac{1}{6} = \frac{7}{2}$$

(Note that the expectation is in general not an element of Ω .) and

$$\text{Var}(X) = \int_{\Omega} \left(X(\omega) - \frac{7}{2} \right)^2 d\mathbb{P}(\omega) = \sum_{i=1}^6 \left(i - \frac{7}{2} \right)^2 \frac{1}{6} = \frac{35}{12}.$$

If we are interested in another experiment, e.g., the *game* from Example 10.94, which was given by a random variable $Y := -3\chi_{\{1,2\}} + 5\chi_{\{6\}}$, we see that

$$\mathbb{E}(Y) = -3\mathbb{P}(\{1, 2\}) + 5\mathbb{P}(\{6\}) = -\frac{1}{6}.$$

(This shows that you would *lose in expectation* if you play this game.) Moreover,

$$\begin{aligned}\text{Var}(Y) &= \int_{\Omega} (Y(\omega) - \mathbb{E}(Y))^2 d\mathbb{P}(\omega) = \sum_{i=1}^6 \left(Y(i) + \frac{1}{6} \right)^2 \frac{1}{6} \\ &= \frac{1}{6} \left(\left(-3 + \frac{1}{6} \right)^2 + \left(-3 + \frac{1}{6} \right)^2 + \left(0 + \frac{1}{6} \right)^2 + \left(0 + \frac{1}{6} \right)^2 + \left(0 + \frac{1}{6} \right)^2 + \left(5 + \frac{1}{6} \right)^2 \right) \\ &= \frac{771}{108} \approx 7.14.\end{aligned}$$

This shows that the outcomes of this experiment have a large fluctuation (aka. variance).

There are thousands of (useful) results, like probability bounds, that are beneficial for theoretical and applied purposes. We do not discuss that here, as we have another scope. However, let us state the most fundamental result in this direction and an important application. This is, using only the variance, we can show bounds on the probability that the outcome of an experiment is far away from the expectation. In its most classical form, this is **Chebyshev's inequality**.

Lemma 10.98 (Chebyshev's inequality). *Let $(\Omega, \mathcal{A}, \mathbb{P})$ be a probability space and X a random variable such that expectation $\mathbb{E}X$ and variance $\text{Var}(X)$ exist. Then,*

$$\mathbb{P}(|X - \mathbb{E}X| > a) \leq \frac{\text{Var}(X)}{a^2} \quad \text{for all } a > 0.$$

Proof. Let $A := \{\omega: |X(\omega) - \mathbb{E}X| > a\}$, so the left hand side above is $\mathbb{P}(A)$. Then,

$$\text{Var}(X) = \int_{\Omega} (X(\omega) - \mathbb{E}(X))^2 d\mathbb{P}(\omega) \geq \int_A (X(\omega) - \mathbb{E}(X))^2 d\mathbb{P}(\omega) \geq a^2 \int_A 1 d\mathbb{P} \geq a^2 \mathbb{P}(A).$$

Dividing by a^2 proves the claim. \square

One particularly interesting application of all the things we've discussed in this section is the estimation of the expectation by repeating the experiment, say n times, and taking the mean

$$\bar{X} := \frac{1}{n} \sum_{i=1}^n X_i,$$

where the X_1, \dots, X_n are *independent copies* of a random variable X .

Remark 10.99. A precise definition of *independence* goes beyond this lecture. It means, roughly speaking, that the outcome of the i -th experiment does not depend on all the other (independent) experiments. This can be expressed precisely by some properties of their probabilities.

One possible substitute property is that the random variables are **uncorrelated**, which means that $\mathbb{E}(X_i \cdot X_j) = \mathbb{E}(X_i)\mathbb{E}(X_j)$ for all $i \neq j$. That is, the random variables might depend on each other, but do not show any *correlation*. We omit the details.

That the averages \bar{X} can lead to quite precise estimates, is based on the following properties.

Lemma 10.100. *Let $(\Omega, \mathcal{A}, \mathbb{P})$ be a probability space and X_1, \dots, X_n be random variables. Then,*

$$\mathbb{E}(\bar{X}) = \frac{1}{n} \sum_{i=1}^n \mathbb{E}(X_i)$$

and if we have additionally that $\mathbb{E}(X_i \cdot X_j) = \mathbb{E}(X_i)\mathbb{E}(X_j)$ for all $i \neq j$, then

$$\text{Var}(\bar{X}) = \frac{1}{n^2} \sum_{i=1}^n \text{Var}(X_i).$$

Proof. The first equality is just the linearity of the expectation (aka. integral). For the second, we multiply out and use the first part to obtain

$$\begin{aligned}\text{Var}(\bar{X}) &= \mathbb{E}\left(\left(\frac{1}{n} \sum_{i=1}^n (X_i - \mathbb{E}(X_i))\right)^2\right) = \frac{1}{n^2} \sum_{i,j=1}^n \mathbb{E}((X_i - \mathbb{E}(X_i))(X_j - \mathbb{E}(X_j))) \\ &= \frac{1}{n^2} \sum_{i,j=1}^n (\mathbb{E}(X_i X_j) - \mathbb{E}(X_i)\mathbb{E}(X_j)) = \frac{1}{n^2} \sum_{i=1}^n (\mathbb{E}(X_i^2) - \mathbb{E}(X_i)^2) = \frac{1}{n^2} \sum_{i=1}^n \text{Var}(X_i),\end{aligned}$$

where we use the assumption $\mathbb{E}(X_i \cdot X_j) = \mathbb{E}(X_i)\mathbb{E}(X_j)$ for $i \neq j$. \square

This finally implies the following bound on the **error of a Monte Carlo method**, which is a name for numerical methods that use averages over independent random variables.

Corollary 10.101. Let $(\Omega, \mathcal{A}, \mathbb{P})$ be a probability space and X_1, \dots, X_n be uncorrelated random variables with $\mathbb{E}X_i =: Z \in \mathbb{R}$ and $\text{Var}(X_i) =: \sigma^2$ for all $i = 1, \dots, n$. Then,

$$\mathbb{P}\left(\left|\frac{1}{n} \sum_{i=1}^n X_i - Z\right| > \frac{\sigma}{\sqrt{\delta n}}\right) \leq \delta \quad \text{for all } \delta > 0.$$

This shows that we need at most $n \approx \frac{\sigma^2}{\delta \cdot \varepsilon}$ independent trials of an experiment to achieve an *error* at most $\varepsilon > 0$ with probability at least $1 - \delta$. We omit the (rather straightforward) proof.

10.5.1 Some special distributions

In this subsection we shortly introduce some of the most important special distributions and present their expectations and variances. All of them are given by their densities.

Let us start with popular continuous random variables.

Example 10.102. The most well-known example for a distribution is the **standard normal distribution** which is given by the density

$$\rho(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}.$$

If a random variable X has this density, i.e.,

$$\mathbb{P}(X \in [a, b]) = \mathbb{P}(a \leq X \leq b) = \frac{1}{\sqrt{2\pi}} \int_a^b e^{-\frac{1}{2}x^2} dx,$$

we use the notation $X \sim \mathcal{N}(0, 1)$.

To compute the expected value and variance of normally distributed random variables, we use the formulas

$$\mathbb{E}(X) = \int_{\mathbb{R}} x\rho(x) dx \quad \text{and} \quad \mathbb{E}(X^2) = \int_{\mathbb{R}} x^2\rho(x) dx$$

which hold for any real random variable X with density $\rho: \mathbb{R} \rightarrow \mathbb{R}^+$.

We have that, for $X \sim \mathcal{N}(0, 1)$,

$$\mathbb{E}(X) = \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} x e^{-\frac{1}{2}x^2} dx = 0,$$

since the integrand is an odd function. (The density is symmetric.) For the variance we compute

$$\text{Var}(X) = \mathbb{E}(X^2) - \mathbb{E}(X)^2 = \mathbb{E}(X^2) = \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} x^2 e^{-\frac{1}{2}x^2} dx.$$

Setting $f'(x) = xe^{-\frac{1}{2}x^2}$ and $g(x) = x$ it follows by integration by parts and the property that a density integrates to 1 that

$$\text{Var}(X) = 1.$$

Moreover, if we consider a random variable $Y := \sigma \cdot X + \mu$ for $X \sim \mathcal{N}(0, 1)$ and $\sigma, \mu \in \mathbb{R}$, we have that

$$\mathbb{E}(Y) = \mu \quad \text{and} \quad \text{Var}(Y) = \text{Var}(\sigma X) = \sigma^2.$$

The density of Y is given by

$$\rho(x) = \frac{1}{\sqrt{2\pi}\sigma^2} e^{-\left(\frac{x-\mu}{\sqrt{2}\sigma}\right)^2}$$

and we call Y a **normally distributed random variable** with **mean** μ and **variance** σ^2 , and write $Y \sim \mathcal{N}(\mu, \sigma)$.

Example 10.103. Another important example is the **exponential distribution** with parameter $\lambda > 0$, which is given by the density

$$\rho(x) = \begin{cases} \lambda e^{-\lambda x} & \text{if } x \geq 0, \\ 0 & \text{else,} \end{cases} \quad x \in \mathbb{R}.$$

We write $X \sim \text{Exp}(\lambda)$ if X has this density. In physics, this distribution is used to model radioactive decay, where λ is related to the half-life period, i.e., the time an atom needs (on average) to break up. More practical, one can model the appearance of a new customer at a waiting line. We obtain

$$\mathbb{E}(X) = \int_{\mathbb{R}^+} x \lambda e^{-\lambda x} dx = \frac{1}{\lambda},$$

which follows by partial integration. Similar arguments can be used to compute that

$$\text{Var}(X) = \frac{1}{\lambda^2}.$$

Let us now discuss some discrete random variables.

Example 10.104. First of all, let us present the general form of expected value and variance for discrete random variables, i.e., random variables X on some probability space $(\Omega, \mathcal{A}, \mathbb{P})$ with $X(\omega) \in \mathbb{Z}$ for every $\omega \in \Omega$. For this, we let

$$p_k := \mathbb{P}(\{X = k\}), \quad k \in \mathbb{Z},$$

and observe that $\rho(k) := p_k$, $k \in \mathbb{Z}$, is the density of X w.r.t. the counting measure.

By plugging in the definition we obtain

$$\mathbb{E}(X) = \sum_{k \in \mathbb{Z}} k \cdot p_k \quad \text{and} \quad \text{Var}(X) = \sum_{k \in \mathbb{Z}} k^2 p_k - \left(\sum_{k \in \mathbb{Z}} k p_k \right)^2.$$

The last discrete distribution we want to discuss is the *Poisson distribution*, which is closely connected to the exponential distribution.

Example 10.105. The **Poisson distribution** with parameter $\lambda > 0$ is defined by its density

$$\rho(k) := \frac{\lambda^k}{k!} e^{-\lambda}$$

w.r.t. the counting measure on \mathbb{N}_0 . That is, a random variable X is Poisson-distributed with parameter λ , written $X \sim \text{Poi}(\lambda)$, if

$$\mathbb{P}(X \in A) = \sum_{k \in A} \frac{\lambda^k}{k!} e^{-\lambda}$$

for every $A \subset \mathbb{N}_0$. The Poisson distribution is used to model the number of events happening in a fixed period of time, if each event happens independently after a exponentially distributed time. Again, one might think of decaying atoms or appearing customers, but this time we are interested in their number in a fixed period. With the above, we obtain

$$\mathbb{E}(X) = e^{-\lambda} \sum_{k \in \mathbb{N}_0} k \cdot \frac{\lambda^k}{k!} = \lambda \quad \text{and} \quad \text{Var}(X) = \lambda.$$

(Verify yourself!)

Remark 10.106. Let us finally add a comment on programming issues.

We've already learned the set $\{x_0\}$ has (Lebesgue-)measure 0 for any $x_0 \in [0, 1]$. Thus, we obtain,

$$\mathbb{P}(\mathbb{Q} \cap [0, 1]) = 0,$$

This means that the probability to receive a rational number by randomly picking a number out of the unit interval is 0. However, this is an issue in computer science, because computers are limited to rational numbers. (A computer can not even save an irrational number, like π , due to its infinite 'length'.) Thus, each 'random number' in $[0, 1]$ you receive by using a computer is not really a randomly chosen number, as it is produced by some 'deterministic' algorithm. (We do not discuss the possibilities of quantum computers.) In fact it is an important topic in computer science/mathematics how to generate numbers that 'behave' like random numbers, called the theory of *pseudo-random numbers*.

11 Basic functional analysis

The branch of functional analysis is concerned with the transformation of functions (or other objects), in the way like functions transform numbers or vectors to other numbers or vectors.

For this, we first need to introduce a 'structure' that allows to treat classes of functions in the same way as vectors, the so-called *vector spaces*. Then, we will discuss *norms* of functions (or other mappings) in a more general fashion and introduce the corresponding *normed spaces*. This allows to assign a 'size' also to functions and, therefore, leads to some interesting consequences, like certain inequalities. However, the main intention here is to introduce an universal notation that allows to write down *quantitative mathematical results* not only for individual functions (like for integration or differentiation in the last chapters) but also for whole classes of functions. Again, this provides an important piece of mathematical language.

11.1 Vector spaces

In this section we start by introducing *vector spaces*, also called *linear spaces*, which are one of the fundamental structures in mathematics. These are generalizations of the 'spaces of vectors' \mathbb{R}^d , and we will see that vector spaces share many characteristics with them. Therefore, they lead to a rather intuitive way of treating certain classes of objects, like functions. In particular, by discussing the *basis* and the *dimension* of a vector space and (*linear*) *mappings* between them, we will learn that rather different objects (functions, vectors etc.) may be mapped one-on-one onto each other, which may allow for an easier representation and handling of them. Here, we concentrate on the most important concepts and present several examples.

Let us start by recalling the **field axioms**. We say that a set \mathbb{F} with an *addition* $+: \mathbb{F} \times \mathbb{F} \rightarrow \mathbb{F}$ and a *multiplication* $\cdot: \mathbb{F} \times \mathbb{F} \rightarrow \mathbb{F}$ is a field, if the following properties hold for arbitrary $a, b, c \in \mathbb{F}$:

- $+$ and \cdot are associative: $(a + b) + c = a + (b + c)$ and $a \cdot (b \cdot c) = (a \cdot b) \cdot c$.
- $+$ and \cdot are commutative: $a + b = b + a$ and $a \cdot b = b \cdot a$.
- Neutral elements: there exist $0, 1 \in \mathbb{F}$ such that $a + 0 = 0 + a = a$ and $a \cdot 1 = 1 \cdot a = a$.
- Inverse elements: there exists $-a \in \mathbb{F}$ such that $a - a = -a + a = 0$ and, if $a \neq 0$, then there exists $a^{-1} \in \mathbb{F}$ such that $a \cdot a^{-1} = a^{-1} \cdot a = 1$.
- Distributivity: we have $a \cdot (b + c) = a \cdot b + a \cdot c$.

(Precisely, one has to say that $(\mathbb{F}, +, \cdot)$ is a field, but we usually omit $+/\cdot$.)

In most cases, we just consider $\mathbb{F} = \mathbb{R}$ or $\mathbb{F} = \mathbb{C}$, i.e., the field of real or complex numbers, but recall that this notion is quite flexible and there are also *finite fields*, like $\{0, 1\}$ with addition and multiplication modulo 2, which are of some interest in computer science and numerical analysis.

A vector space is now a set with similar properties, but we only require that some operations are well-defined between elements of the set and a given field. For example, one might think on a set of vectors, for which a multiplication is not allowed.

Definition 11.1. Let \mathbb{F} be a field and V be a non-empty set.

We call V a **vector space over \mathbb{F}** (or **linear space over \mathbb{F}** or **\mathbb{F} -vector space**), if the following properties hold:

There exists a **vector addition** $+: V \times V \rightarrow V$ such that

- 1) **Associativity:** For all $x, y, z \in V$ there holds $x + (y + z) = (x + y) + z$
- 2) **Commutativity:** For any vectors $x, y \in V$ we have $x + y = y + x$
- 3) **Neutral element:** there exists an element $0 \in V$ such that $x + 0 = x$ for all $x \in V$
- 4) **Inverse element:** For any $x \in V$ there exists an element $-x \in V$ such that $x - x = 0$.

Moreover, there is a **scalar multiplication** $\cdot: \mathbb{F} \times V \rightarrow V$ such that for all $x, y \in V$ and $\mu, \lambda \in \mathbb{F}$ the following properties are valid:

- 5) **Distributivity w.r.t. vector addition:** $\lambda \cdot (x + y) = \lambda \cdot x + \lambda \cdot y$
- 6) **Distributivity w.r.t. field addition:** $(\lambda + \mu) \cdot x = \lambda \cdot x + \mu \cdot x$
- 7) **Associativity:** $(\lambda \cdot \mu) \cdot x = \lambda \cdot (\mu \cdot x)$
- 8) **Neutral element:** $1 \cdot x = x$

The elements $x \in V$ are called **vectors** and the elements of \mathbb{F} are called **scalars**.

In the case $\mathbb{F} = \mathbb{R}$ or $\mathbb{F} = \mathbb{C}$, we call V a **real or complex vector space**, respectively.

Remark 11.2 (Notation). We usually use the convention that Latin letters denote elements of a vector space V , whereas Greek letters are used for scalars, i.e., elements in \mathbb{F} .

Remark 11.3. The phrase 'there is a vector addition' might sound elusive, but in most cases we consider it is just the *natural* (point-wise) addition of functions. This formulation above is still needed because, formally, the addition of two real numbers is not the same as the addition of vectors or functions or binary digits.

Note that the definition of a vector space states that the **sum of two vectors and the multiplication of a vector with a scalar have to be elements of V** . That is, for any $x, y \in V$ and $\lambda \in F$ one has to show that $x + y \in V$ and $\lambda x \in V$. In many cases these are actually the only requirements to check, while the properties 1) - 8) are more or less clear by definition of the vector addition and the scalar multiplication.

The most canonical example for a \mathbb{R} -vector space is the Euclidean space, which is also the reason for the name *vector space*.

Example 11.4 (Euclidean space). Let $d \in \mathbb{N}$ and define $\mathbb{R}^d = \mathbb{R} \times \mathbb{R} \times \cdots \times \mathbb{R}$, i.e. the d -fold Cartesian product of \mathbb{R} . Then, \mathbb{R}^d together with the component-wise addition and the component-wise multiplication builds a (real) vector space. (Double-check the conditions above!)

In the same way, \mathbb{C}^d is a \mathbb{C} -vector space.

The next example shows that one also needs to be careful with the underlying field.

Example 11.5 (Different fields). We have that \mathbb{Q} together with the usual addition and multiplication is a field, and we easily see that \mathbb{R}^d is a \mathbb{Q} -vector space. However, considered as a \mathbb{Q} -vector space, we are only allowed to multiply a vector in \mathbb{R}^d by rational numbers.

The complex numbers \mathbb{C} are also a field (with the corresponding rules for addition and multiplication), but \mathbb{R} cannot be a \mathbb{C} -vector space because $i \cdot x = (ix_1, ix_2, \dots, ix_d) \notin \mathbb{R}^d$, where i is the imaginary unit.

However, also more complicated sets than just the 'sets of vectors' are vector spaces. First of all, we can consider vectors of *infinite length*, i.e., sequences.

Example 11.6 (Space of sequences). Let

$$\mathbb{R}^{\mathbb{N}} := \{(x_1, x_2, \dots) : x_k \in \mathbb{R}, k \in \mathbb{N}\},$$

i.e., the set of all real sequences, and define $+$ and \cdot component-wise (one can see this as vectors with ‘infinite length’). Then, $\mathbb{R}^{\mathbb{N}}$ is a real vector space.

One defines analogously $\mathbb{C}^{\mathbb{N}}$ as the complex vector space of complex sequences.

Moreover, and this is one of our main purposes, we can also consider vector spaces of functions, which are usually called **function spaces**.

Example 11.7 (Function spaces). Let $\Omega \subset \mathbb{R}^d$ and define the set

$$\mathbb{R}^\Omega := \{f: \Omega \rightarrow \mathbb{R}\},$$

i.e., the **set of all real-valued functions**. The addition of functions $f, g \in \mathbb{R}^\Omega$ is defined point-wise (as usual) by $(f + g)(x) = f(x) + g(x)$ for any $x \in \Omega$. Moreover, for $\lambda \in \mathbb{R}$ and $f \in \mathbb{R}^\Omega$ we set (also as usual) $(\lambda \cdot f)(x) = \lambda \cdot f(x)$ for any $x \in \Omega$. It is not hard to show that this is indeed a vector space. (Verify, in particular, the points 1)-8) of Definition 11.1!)

Another important vector space is the set

$$C(\Omega) := \{f: \Omega \rightarrow \mathbb{R} : f \text{ is continuous}\}$$

of continuous functions on Ω with point-wise addition and scalar multiplication. Since we know that the sum of continuous functions, and the multiplication by a constant, results in a continuous function, see e.g. Lemma 8.11, we obtain that $C(\Omega)$ is a **real vector space**.

Similarly, for an open set Ω and $k \in \mathbb{N}$, we define the set

$$C^k(\Omega) := \left\{ f: \Omega \rightarrow \mathbb{R} : D^\alpha f \text{ exists and is continuous for all } \alpha \in \mathbb{N}_0^d \text{ with } |\alpha| \leq k \right\},$$

where $D^\alpha f$ denotes the partial derivative of f in multi-index notation as introduced in Section 8.6. (One usually writes $C^0(\Omega) := C(\Omega)$.) Using that the derivative is linear, i.e., $D^\alpha(f + g) = D^\alpha f + D^\alpha g$ from Theorem 5.11, we also obtain that $C^k(\Omega)$ is a **real vector space**.

(If Ω is closed, then we consider the one-sided derivatives at the boundary.)

Remark 11.8. We will omit what we mean by addition and multiplication in the sequel if it is clear from the context. This is the case, in particular, if we consider function spaces with point-wise addition and scalar multiplication.

There are also some important function spaces of specific functions.

Example 11.9 (Space of polynomials). Let \mathcal{P}_n denote the set of all (algebraic) polynomials $p: \mathbb{R} \rightarrow \mathbb{R}$ which have degree at most n , i.e.,

$$\mathcal{P}_n := \left\{ a_n x^n + a_{n-1} x^{n-1} + \cdots + a_1 x + a_0 : a_0, \dots, a_n \in \mathbb{R} \right\}.$$

With the usual point-wise addition and scalar multiplication this forms a \mathbb{R} -vector space.

To see this, first note that each polynomial $p \in \mathcal{P}_n$ of degree at most n has the form $p(x) := a_n x^n + a_{n-1} x^{n-1} + \cdots + a_1 x + a_0$. Clearly, λp is again a polynomial of degree at most n . Moreover, with another polynomial $q(x) := b_n x^n + b_{n-1} x^{n-1} + \cdots + b_1 x + b_0$, we obtain

$$(p + q)(x) = (a_n + b_n)x^n + (a_{n-1} + b_{n-1})x^{n-1} + \cdots + (a_1 + b_1)x + (a_0 + b_0) \in \mathcal{P}_n,$$

since $a_k + b_k \in \mathbb{R}$ for $k = 0, \dots, n$. The properties 1) to 8) from Definition 11.1 can be shown easily. (Verify this!)

However, specific classes of functions are **often not closed under addition** and are therefore not a vector space.

Example 11.10 (Polynomials of fixed degree). Let \mathbf{P}_n be the set of all polynomials of degree exactly n , i.e. polynomials of the form

$$p(x) = a_n x^n + a_{n-1} x^{n-1} + \cdots + a_1 x + a_0$$

with $a_n \neq 0$. This looks very similar to the last example. However, this is not a vector space.

This can be seen by considering the two polynomials $p(x) = x^n$ and $q(x) = -x^n + 1$, which are both contained in \mathbf{P}_n . However, $p + q = 1$ is not a polynomial of degree n , but has degree 0, so the sum of two elements from \mathbf{P}_n is not in \mathbf{P}_n . Therefore, \mathbf{P}_n is not a vector space.

Alternatively, one might already see that \mathbf{P}_n does not contain a neutral element for the addition, as the zero polynomial is not in \mathbf{P}_n . Hence, \mathbf{P}_n also violates point 3) from Definition 11.1.

The next example shows a very important example of a function space consisting of complex-valued functions, namely it consists of **trigonometric polynomials** with bounded degree. This is a particularly important function space when it comes to numerical approximation by Fourier series, see Section 7.

Example 11.11 (Trigonometric polynomials). Let $I := [0, 1]$ be the unit interval and \mathcal{T}_n be the set of all trigonometric polynomials of degree at most n on I , i.e., functions of the form

$$p(x) = \sum_{k=-n}^n c_k e^{2\pi i k x}$$

with $c_{-n}, \dots, c_n \in \mathbb{C}$, see Definition 7.6. Recall from Section 7 that every $p \in \mathcal{T}_n$ is a periodic, complex-valued function, and that we have the Euler identity $e^{2\pi i k x} = \cos(2\pi k x) + i \sin(2\pi k x)$.

One can easily prove, as for algebraic polynomials, that \mathcal{T}_n is a complex function space (i.e., a \mathbb{C} -vector space consisting of functions).

A more interesting appearance of vector spaces are the **sets of solutions** of certain equations.

Example 11.12. Let $A \in \mathbb{R}^{m \times n}$ be a real $m \times n$ matrix, then the set $N(A) = \{x \in \mathbb{R}^n : Ax = 0\}$ together with the usual vector addition and scalar multiplication is a vector space. (Verify this!) This space is called **nullspace** of A .

To see a specific example, consider

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

which has the nullspace

$$N(A) = \left\{ \begin{pmatrix} 0 \\ 0 \\ \lambda \end{pmatrix} : \lambda \in \mathbb{R} \right\}.$$

It is easy to verify that $N(A)$ satisfies all requirements of a vector space.

Let us consider a more involved example that demonstrates, again, that solutions of certain equations build a vector space.

Example 11.13. Let us assume that we want to find functions $y: [0, \infty) \rightarrow \mathbb{R}$ that satisfy

$$y''(t) = -\omega^2 y(t).$$

for all $t \geq 0$. This is an *ordinary differential equation*, i.e., we know the solution only by a relation of the function and its derivatives. Moreover, this equation has a nice interpretation: In physics, this equation is used to model **Hooke's law** (invented around 1676 by *Robert Hooke*), which describes the oscillating position of an object that is attached to a helical spring. Here, ω is a parameter depending on the spring

and the weight of the object. (Other applications of this law are, e.g., deformations of elastic bodies like wind blowing on a tall building.)

Although we didn't learn so far how to solve such equations systematically, it is easy to check that

$$y_1(t) := \sin(\omega t) \quad \text{and} \quad y_2(t) := \cos(\omega t)$$

solve the above equation. (Just compute their second-order derivative.)

Moreover, since differentiation is a linear operation, it follows that also linear combinations of y_1 and y_2 , i.e., functions of the form $c_1 y_1(t) + c_2 y_2(t)$ with $c_1, c_2 \in \mathbb{R}$, solve the equation, which shows that every function from the function space

$$S := \left\{ y(t) = c_1 \sin(\omega t) + c_2 \cos(\omega t) : c_1, c_2 \in \mathbb{R} \right\}$$

are solutions to the equation. (One can even show that every solution is of this form.)

To find the explicit solution, we need some more info: For example, if we know the *initial* position $y(0) = a$ and the initial velocity $y'(0) = b$, we obtain the solution $y(t) = \frac{b}{\omega} \sin(\omega t) + a \cos(\omega t)$.

In the previous example we have seen that in some cases we consider vector spaces, e.g. $N(A)$, that are contained in another larger vector space, e.g. \mathbb{R}^3 . Actually, this is mostly the case and motivates the definition of a subspace.

Definition 11.14 (Subspaces). Let V be a \mathbb{F} -vector space and suppose that $U \subset V$.

If U is a \mathbb{F} -vector space, then we call U a **(linear) subspace** of V .

If additionally $U \subsetneq V$, then we say it is a **proper subspace**.

Remark 11.15. By writing out the conditions of being a vector space, we obtain that $U \subset V$ is a subspace of V if and only if

- 1) $U \neq \emptyset$,
- 2) $u_1 + u_2 \in U$ and
- 3) $\alpha u_1 \in U$,

for all $u_1, u_2 \in U$ and $\alpha \in \mathbb{F}$.

Example 11.16. We already saw, that for any matrix $A \in \mathbb{R}^{m \times n}$ we have that $N(A) \subset \mathbb{R}^n$ is a subspace of \mathbb{R}^n , where n is the number of columns.

Example 11.17. A vector space may clearly contain several different subspaces. For example, there are the following proper subspaces of \mathbb{R}^3 :

$$U_1 = \left\{ \begin{pmatrix} 0 \\ 0 \\ \lambda \end{pmatrix} : \lambda \in \mathbb{R} \right\}, \quad U_2 = \left\{ \begin{pmatrix} 0 \\ \lambda \\ \lambda \end{pmatrix} : \lambda \in \mathbb{R} \right\} \quad \text{and} \quad U_3 = \left\{ \begin{pmatrix} 0 \\ 0 \\ \lambda \end{pmatrix} + \begin{pmatrix} 0 \\ \mu \\ 0 \end{pmatrix} : \lambda, \mu \in \mathbb{R} \right\}.$$

One can also verify that U_1 and U_2 are also subspaces of U_3 .

Example 11.18. Typical proper subspaces arise when we restrict elements of a given vector space to satisfy a given property. It is then still necessary to check if this property is ‘linear’ in the sense that it leads to a vector space. For example, $C([0, 1])$ is a subspace of $\mathbb{R}^{\mathbb{N}}$. Moreover, the set

$$\mathcal{F} := \{f \in C([0, 1]): f(0) = 0\},$$

i.e., all continuous functions on $[0, 1]$ with a fixed function value, is clearly a subset of $C([0, 1])$. Since the sum of two functions from \mathcal{F} are still continuous and have function value 0 at 0 (same for scalar multiplication), we obtain that \mathcal{F} is a subspace of $C([0, 1])$, and therefore also a subspace of \mathbb{R}^{Ω} . It is even a proper subspace, since $C([0, 1])$ contains functions with $f(0) \neq 0$.

However, note that $\mathcal{G} := \{f \in C([0, 1]): f(0) = 1\}$ is still a subset, but not a subspace of $C([0, 1])$. (Why?)

We have already seen that vector spaces have the nice property that sums and multiples of elements are still in the space. We will now turn to an even more powerful insight: In a rather general setting, **every element of a vector space can be written as a linear combination of only a few fixed elements**. These elements will be called a *generating system* or *basis* of a vector space, and play a fundamental role, in particular, for working with function spaces.

Let us again start with some formalism.

Definition 11.19. A set of vectors $\{v_1, v_2, \dots, v_n\} \subset V$, where V is a \mathbb{F} -vector space is called **linearly dependent** if there exist $\alpha_1, \dots, \alpha_n \in \mathbb{F}$, not all zero, such that

$$\alpha_1 v_1 + \alpha_2 v_2 + \cdots + \alpha_n v_n = 0.$$

Otherwise, i.e., if

$$\alpha_1 v_1 + \alpha_2 v_2 + \cdots + \alpha_n v_n = 0 \iff \alpha_1 = \alpha_2 = \cdots = \alpha_n = 0,$$

then we call the set $\{v_1, v_2, \dots, v_n\}$ **linearly independent**.

A vector of the form

$$y = \alpha_1 v_1 + \alpha_2 v_2 + \cdots + \alpha_n v_n,$$

for some $\alpha_k \in \mathbb{F}$ is called **linear combination** (in V) of the vectors v_1, \dots, v_n .

In words, a set of vectors is linearly dependent if and only if one of its elements is a linear combination of the other elements. Although this is a rather elementary reformulation of the definition, it is still quite useful, so we will state it as lemma.

Lemma 11.20. Let V be a vector space and $\{v_1, \dots, v_n\} \subset V$ be linearly dependent if and only if at least one of the v_k ’s is a linear combination of the other vectors.

Proof. By definition of linear dependence, we know that there exist some scalars $\alpha_1, \dots, \alpha_n \in \mathbb{F}$ such that at least one of them is non-zero, say α_k , and

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

By regrouping we obtain

$$v_k = -\frac{1}{\alpha_k} \sum_{\substack{i=1 \\ i \neq k}}^n \alpha_i v_i$$

□

Remark 11.21. The case of infinitely many vectors is more involved and postponed to later.

Note that the '0' in $\sum_{i=1}^n \alpha_i v_i = 0$ is not only a number but the zero element in V (i.e., a vector or a function), and that the equality should therefore be understood in the right sense.

For example, if V is a function space, e.g. $V = C(\Omega)$, then the 'vectors' are functions and all operations have to be understood point-wise. This means that a set of functions $\{f_1, \dots, f_n\} \subset V$ on a set Ω is linearly independent if

$$\alpha_1 f_1(x) + \alpha_2 f_2(x) + \cdots + \alpha_n f_n(x) = 0 \quad \forall x \in \Omega$$

implies that $\alpha_i = 0$ for all $1 \leq i \leq n$.

Let's see some examples.

Example 11.22. The easiest case is to consider a set of just two elements $v_1, v_2 \neq 0$. (Note that, by definition, every set that contains a zero vector is linearly dependent.) We obtain that

$$\{v_1, v_2\} \text{ is linearly dependent} \iff v_1 = \alpha v_2 \quad \text{for some } \alpha \in \mathbb{F}.$$

That is, two 'vectors' are either multiples of each other, or independent.

Example 11.23. The vectors

$$v_1 = (1, 2, 3) \quad v_2 = (5, 4, 3) \quad v_3 = (7, 8, 9)$$

are linearly dependent. This follows by observing that $2v_1 + v_2 = v_3$. In particular, v_3 is a linear combination of v_1 and v_2 . Moreover, by this equality, we see that every v_k is a linear combination of the other two.

Example 11.24. Let us consider linear independence of the three functions $v_1(x) = 1$, $v_2(x) = \sin(x)$ and $v_3(x) = \cos(x)$ on the interval $[0, 2\pi]$. We have to check that the equation

$$\alpha_1 + \alpha_2 \sin(x) + \alpha_3 \cos(x) = 0$$

can only be satisfied (for arbitrary $x \in [0, 2\pi]$) if $\alpha_1 = \alpha_2 = \alpha_3 = 0$.

For this we just try different inputs x . We start by plugging in $x_1 = 0$ and observe that in this case α_2 can be chosen freely since $\sin 0 = 0$, but we obtain $\alpha_1 + \alpha_3 \cos 0 = 0 \iff \alpha_1 = -\alpha_3$.

Now, analogously, for $x_2 = \pi$ we have $\sin(\pi) = 0$ and $\cos(\pi) = -1$, leading to $\alpha_1 = \alpha_3$. The last two equations together show that $\alpha_1 = -\alpha_1$ and $\alpha_3 = -\alpha_3$, which is only possible if $\alpha_1 = \alpha_3 = 0$. Finally, using any $x_3 \in (0, \pi) \cup (\pi, 2\pi)$ and $\alpha_1 = \alpha_3 = 0$, we conclude that also $\alpha_2 = 0$ is required. Thus the only possible choice for $\alpha_1, \alpha_2, \alpha_3$ is $\alpha_1 = \alpha_2 = \alpha_3 = 0$ if we want $\alpha_1 + \alpha_2 \sin(x) + \alpha_3 \cos(x) = 0$ for all $x \in [0, 2\pi]$. This implies that the functions are linearly independent.

Another important set of independent vectors are the **unit vectors**.

Example 11.25 (Unit vectors). Consider $e_1, e_2, \dots, e_d \in \mathbb{R}^d$ such that each e_k has a 1 in the k -th component and zeros in all other components, i.e., e_k is the k -th unit vector.

These vectors are clearly linearly independent, because two different unit vectors never have a one in the same position. However, any vector $x \in \mathbb{R}^d$ can be written as linear combination of e_1, e_2, \dots, e_d . To see this, we denote the k -th component of x by x_k (which is a real number) and obtain

$$x = \sum_{k=1}^d x_k \cdot e_k.$$

Therefore, there is no other vector $x \in \mathbb{R}^d$ such that $\{e_1, e_2, \dots, e_d, x\}$ is linearly independent.

The above example suggests that by finding 'large enough' linearly independent sets we can write each element of a vector space as linear combination of elements of this set. This motivates the concept of a *basis* and *dimension* of a vector space.

Definition 11.26 (Basis). Let V be a vector space over some field \mathbb{F} and $n \in \mathbb{N}$.

For a set $G := \{b_1, \dots, b_n\} \subset V$ we define the **(linear) span of G** (in V) by

$$\text{span}(G) := \left\{ v \in V : v = \sum_{k=1}^n c_k b_k \quad \text{for some } c_1, \dots, c_n \in \mathbb{F} \right\},$$

i.e., the set of all linear combinations of elements of G .

We call a finite set of vectors $G := \{b_1, \dots, b_n\} \subset V$ a **generating set of V** if each element in V is a linear combination of elements in G , i.e., $V = \text{span}(G)$.

A set $B \subset V$ is called a **basis of V** if B is a generating system of V and all elements in B are linearly independent.

Another common notation for the span of G is $\langle G \rangle$. Moreover, note the trivial fact that, if a set G contains only linearly independent vectors, then G is a basis of $\text{span}(G)$.

Remark 11.27. The case of infinitely many vectors has to be treated a bit more carefully, and there exist different concepts of a basis in such a case. Here, we will only consider the case that is closest to the above and is anyhow the most common concept, which is called a *Schauder basis*. That is, we call $B := \{b_1, b_2, \dots\} \subset V$ a **(Schauder) basis of V** if all elements of B are linearly independent and each element in V can be written as a **unique convergent series** of elements in B , i.e., for all $v \in V$ there are unique $c_1, c_2, \dots \in \mathbb{F}$ such that

$$v = \sum_{k=1}^{\infty} c_k b_k.$$

Clearly, this concept makes only sense if there are infinitely many linearly independent vectors. Moreover, we need to be careful what mean by this convergence of such series. (Note that the b_k are possibly functions.) Therefore, we present more on this once we introduce normed spaces.

Example 11.28. We saw above that the unit vectors e_1, e_2, \dots, e_d are a generating system of \mathbb{R}^d , which is linearly independent. Therefore, $B := \{e_1, \dots, e_d\}$ is a basis of \mathbb{R}^d , which is called the **standard basis** of \mathbb{R}^d .

Example 11.29. The space \mathcal{P}_n of real polynomials with degree at most n is the set of all polynomials $p: \mathbb{R} \rightarrow \mathbb{R}$ of the form

$$p(x) = a_n x^n + a_{n-1} x^{n-1} + \dots + a_1 x + a_0, \quad a_0, \dots, a_n \in \mathbb{R}.$$

These are all linear combinations of functions from $B = \{1, x, x^2, \dots, x^n\}$, i.e., the *monomials*, i.e., $\mathcal{P}_n = \text{span}\{1, x, x^2, \dots, x^n\}$. By plugging in some different function values ($n+1$ are enough), we obtain that $\{1, x, x^2, \dots, x^n\}$ is linearly independent, and therefore that B is a basis of \mathcal{P}_n .

An important characterization of finite bases is given by the following theorem.

Theorem 11.30. *Let V be a vector space and $B = \{b_1, b_2, \dots, b_n\}$ be any (finite) basis of V . Then,*

- 1) *for every $v \in V$ there are unique $c_1, \dots, c_n \in F$ such that $v = \sum_{i=1}^n c_i b_i$, these c_i are called the **coordinates of v w.r.t. the basis B** .*
- 2) *If $B' \subsetneq B$ then B' is no longer a generating system of V .*
- 3) *For any $x \in V \setminus B$ we have that $B \cup \{x\}$ is linearly dependent.*

Proof. We start by proving the first point. By definition every basis is a generating system of V , so we know at least one representation of the above form has to exist. To show that it is unique, we assume that there exist two different representations, i.e., there exist $\alpha_1, \dots, \alpha_n$ and β_1, \dots, β_n such that

$$\sum_{i=1}^n \alpha_i b_i = v = \sum_{i=1}^n \beta_i b_i.$$

Thus we get

$$0 = \sum_{i=1}^n \alpha_i b_i - \sum_{i=1}^n \beta_i b_i = \sum_{i=1}^n (\alpha_i - \beta_i) b_i.$$

(Note that we only used allowed operations in the vector space.)

Since B is a basis we know that its elements are linearly independent which implies that the above equation holds only if $\alpha_i = \beta_i$ for every $1 \leq i \leq n$. So there can exist only one representation.

For the second point we have a look at the set $\{b_1, b_2, \dots, b_{k-1}, b_{k+1}, \dots, b_n\} = B' \subsetneq B$. Clearly, the elements in B' have to be linearly independent, since B is a basis. Thus $b_k \notin \text{span}(B')$, which implies that $\text{span}(B') \subsetneq V$. Therefore, by removing an arbitrary vector of B we do no longer have a basis, i.e., no proper subset of B is a basis of V .

The third point follows by observing that for any $x \in V$ we can find a unique representation like we did in the fist part of the proof. So $B \cup \{x\}$ cannot be linearly independent, since x is a linear combination of the basis vectors. \square

It is not always easy to find a basis of a vector space. However, if V is **finitely generated**, i.e., it is the span of finitely many vectors, then one always finds a basis among the generating vectors.

Lemma 11.31. *Let V be a vector space which is generated by a finite set $\{b_1, b_2, \dots, b_m\}$. Then, $\{b_1, b_2, \dots, b_m\}$ contains a basis. In particular, every such V has a basis.*

Proof. If the set $\{b_1, b_2, \dots, b_m\}$ is linearly independent we are done. Otherwise we can find a element which a linear combination of the other elements in this set. W.l.o.g. we may assume that b_m is a linear combination of the other vectors. Thus $\{b_1, b_2, \dots, b_{m-1}\}$ is still a generating set of V . We repeat the above procedure until we finally end up with a linearly independent set B which is still a generating set of V . Hence B is a basis. \square

The following lemma, sometimes called *Steinitz exchange lemma*, shows that elements of a basis can be replaced by other linear independent vectors, and we still have a basis.

Lemma 11.32 (Steinitz). *Let V be a vector space which is generated by $\{b_1, b_2, \dots, b_n\} \subset V$, i.e., $V = \text{span}\{b_1, \dots, b_n\}$. ($n = \infty$ is allowed.)*

If $\{a_1, a_2, \dots, a_k\} \subset V$ are linearly independent, then $k \leq n$.

Moreover, there is a renumbering of the b_i 's, such that the set $\{a_1, \dots, a_k, b_{k+1}, \dots, b_n\}$ is a generating system of V .

Proof. We show by induction on l , that the set $\{a_1, \dots, a_l, b_{l+1}, \dots, b_n\}$ is a generating set of V ($l \leq k$). For $l = 0$ the statement is true by our assumptions. If we know that the statement is true for $l - 1$, then we can write a_l as

$$a_l = \sum_{i=1}^{l-1} \alpha_i a_i + \sum_{i=l}^n \beta_i b_i.$$

We know that the a_i 's are linearly independent, so at least one of the β_i 's ($l \leq i \leq m$) has to be non-zero. This implies also that $k \leq n$, otherwise the a_i 's would not be linearly independent. After renumbering we may assume that this index is l and we get

$$b_l = \frac{1}{\beta_l} \left(a_l - \sum_{i=1}^{l-1} \alpha_i a_i - \sum_{i=l+1}^m \beta_i b_i \right).$$

Hence the set $\{a_1, \dots, a_l, b_{l+1}, \dots, b_m\}$ is a generating system as claimed. \square

The last result has a nice consequence. Namely, that **every basis of a vector space has the same cardinality**. We only consider the case finitely generated vector spaces. (The other case is again postponed.)

Corollary 11.33. *Let V be a vector space which is generated by a finite set.*

If $\{a_1, a_2, \dots, a_n\}$ and $\{b_1, b_2, \dots, b_m\}$ are bases of V , then $m = n$.

This implies that a set $\{x_1, x_2, \dots, x_n\} \subset V$, where n is the length of one basis, forms a basis, if one of the following is fulfilled:

- 1) $\{x_1, x_2, \dots, x_n\}$ is linear independent, or
- 2) $\{x_1, x_2, \dots, x_n\}$ is a generating set.

Proof. By the Lemma of Steinitz, see Lemma 11.32, we immediately see that for two bases $\{a_1, a_2, \dots, a_n\}$ and $\{b_1, b_2, \dots, b_m\}$ it holds $n \leq m$ and $m \leq n$ at the same time. Thus every basis has the same cardinality. For the first point of second part we use once more the Steinitz lemma and see that we are allowed to exchange linearly independent elements and elements of a generating system. Therefore, the set $\{x_1, x_2, \dots, x_n\}$ also forms a basis. The last point follows since a generating set must have at least as many independent elements as the basis. Therefore, $\{x_1, x_2, \dots, x_n\}$ is linearly independent. With the first points, this proves that it is a basis. \square

The previous results showed that all bases of a vector space have the same cardinality, which allows to define this as a **characteristic of a vector space**, called its *dimension*.

Definition 11.34. Let V be a vector space with basis $\{b_1, b_2, \dots, b_d\}$.

Then, we define the **dimension of V** by $\dim(V) := d = \#\{b_1, b_2, \dots, b_d\}$.

If $\dim(V) < \infty$, then we call V a **finite-dimensional** vector/function space.

In the case that there are infinitely many linearly independent vectors, we write $\dim(V) = \infty$, and call V an **infinite-dimensional** vector/function space.

Let us first see some examples of finite dimensional spaces.

Example 11.35. The Euclidian space \mathbb{R}^d , where $d \in \mathbb{N}$ has dimension d .

Example 11.36. We consider the set $\mathbb{R}^{m \times n}$ which is the set of all $m \times n$ -matrices. Together with the usual matrix addition and the scalar multiplication this is a vector space. (Verify this!) Moreover, the matrices E_{ij} which have a 1 in the i,j -th component from a basis. Hence, the vector space $\mathbb{R}^{m \times n}$ has dimension $m \cdot n$.

Example 11.37. From Example 11.29 we know that $\{1, x, x^2, \dots, x^n\}$ is a basis of the space \mathcal{P}_n of real polynomials with degree at most n . Therefore, $\dim(\mathcal{P}_n) = n + 1$.

In the same way, the space \mathcal{T}_n of trigonometric polynomials of degree at most n , has the basis $B := \{e^{-2\pi i k \cdot} : -n \leq k \leq n\}$ and therefore $\dim(\mathcal{T}_n) = 2n + 1$.

Example 11.38. As in the previous example, the monomials $1, x, \dots, x^n$ are linearly independent functions for any $n \in \mathbb{N}$. So, the vector space of all polynomials, denoted by $\mathcal{P} := \bigcup_{n=0}^{\infty} \mathcal{P}_n$, cannot have finite dimension, as we can find arbitrary large sets of linear independent functions.

The same is true for the spaces $C^k(\Omega)$ with $k \in \mathbb{N}_0$ and an open set $\Omega \subset \mathbb{R}^d$, because the (univariate) monomials are clearly contained in these sets, and so,

$$\dim(C^k(\Omega)) = \infty.$$

11.2 Normed spaces

So far, we introduced vector spaces as kind of a generalization of the (well-understood) spaces of vectors \mathbb{R}^d , and discussed how elements of general vector spaces may be written with the help of a basis. However, this was restricted to finite-dimensional vector spaces. The main reason for that is that, while a representation with a basis in finite-dimensional spaces was just a linear combination, in infinite-dimensional spaces it is necessary to consider series (i.e., infinite sums) of the basis elements. However, as always when it comes to series, we need to be careful what it means that a series converges.

For this we need a substitute for all the concepts, like absolute value of a number or norm of a vector. Let us summarize the essential properties we require. In what follows, we only consider real or complex vector spaces, i.e., we have $\mathbb{F} = \mathbb{R}$ or $\mathbb{F} = \mathbb{C}$.

Definition 11.39. Let X be a \mathbb{F} -vector space and $\|\cdot\|: X \rightarrow \mathbb{R}$ be a real-valued mapping.

We say that $\|\cdot\|$ is a **norm (on X)** if the following holds:

- | | |
|---|--------------------------------|
| 1) For any $x \in X$ with $x \neq 0$ we have $\ x\ > 0$. | (positive definiteness) |
| 2) For $x \in X, \lambda \in \mathbb{F}$ it holds $\ \lambda x\ = \lambda \ x\ $. | (homogeneity) |
| 3) If $x, y \in X$ then we have $\ x + y\ \leq \ x\ + \ y\ $. | (triangle inequality) |

The tuple $(X, \|\cdot\|)$ is then called a **normed space**.

Note that the homogeneity of a norm (property 2) implies that $x = 0 \implies \|x\| = 0$. Therefore, the positive definiteness (property 1) is equivalent to

$$1') \quad \|x\| \geq 0 \text{ for all } x \in X \quad \text{and} \quad x = 0 \iff \|x\| = 0.$$

That is, every mapping $\|\cdot\|: X \rightarrow \mathbb{R}$ satisfying 1', 2 and 3 is a norm.

Let us start by mentioning that the classes we discussed in earlier chapters are in indeed special cases of the above definition.

Example 11.40. $(\mathbb{R}, |\cdot|)$ is a normed space.

The following norms on \mathbb{R}^d were already discussed in Section 8.1.

Example 11.41. We consider the vector space \mathbb{R}^d , $d \in \mathbb{N}$, and

$$\begin{aligned}\|x\|_1 &:= \sum_{i=1}^d |x_i|, \\ \|x\|_2 &:= \sqrt{\sum_{i=1}^d |x_i|^2} \quad \text{and} \\ \|x\|_\infty &:= \max\{|x_1|, |x_2|, \dots, |x_d|\}.\end{aligned}$$

We have seen in Section 8.1 that all these mappings define norms on \mathbb{R}^d . Therefore, the tuples $(\mathbb{R}^d, \|\cdot\|_p)$ are normed spaces, which are denoted by

$$\begin{aligned}\ell_1^d &:= (\mathbb{R}^d, \|\cdot\|_1), \\ \ell_2^d &:= (\mathbb{R}^d, \|\cdot\|_2) \quad \text{and} \\ \ell_\infty^d &:= (\mathbb{R}^d, \|\cdot\|_\infty).\end{aligned}$$

That is, by saying $x \in \ell_p^d$, we mean that $x \in \mathbb{R}^d$ and that we consider the 'length' of x to be measured in the norm $\|\cdot\|_p$, which is sometimes just called ℓ_p -norm. Due to their importance, we will discuss this type of normed spaces more detailed (including a definition for arbitrary $p > 0$) in the next section.

Let us turn to more interesting examples.

Example 11.42. Let $\Omega \subset \mathbb{R}^d$ be closed and bounded. Recall from Example 11.7 that the set

$$C(\Omega) := \{f: \Omega \rightarrow \mathbb{R}: f \text{ is continuous}\}$$

with the usual point-wise addition and scalar multiplication is a vector space. We define

$$\|f\|_\infty := \sup_{x \in \Omega} |f(x)|,$$

and claim that this is a norm on $C(\Omega)$.

First note that we need that Ω is closed and bounded to obtain that $\|f\|_\infty < \infty$ for any $f \in C(\Omega)$. (Continuous functions on open sets might be unbounded.) Positive definiteness and homogeneity are clear, so we only need to show the triangle inequality. We compute

$$\|f + g\|_\infty = \sup_{x \in \Omega} |(f + g)(x)| \leq \sup_{x \in \Omega} (|f(x)| + |g(x)|) \leq \sup_{x \in \Omega} |(f)(x)| + \sup_{x \in \Omega} |g(x)| = \|f\|_\infty + \|g\|_\infty.$$

So indeed we have that $(C(\Omega), \|\cdot\|_\infty)$ is a normed space.

One may also consider much more difficult norms which are, e.g., combinations of known norms. Note that a norm is used to **measure the 'size'** of an element of 'our' vector space, and that this 'size' is usually **specified by the application** in mind.

Example 11.43. For $x = (x_1, x_2, x_3) \in \mathbb{R}^3$ consider the mapping

$$\|x\| := |x_2| + \sqrt{|x_1|^2 + (|x_1| + |x_3|)^2}.$$

It is not hard to verify that $\|\cdot\|$ is a norm on \mathbb{R}^3 .

Just note that $x \neq 0$ implies $x_k \neq 0$ for some $k = 1, 2, 3$, and so $\|x\| > 0$ (positive definiteness). For the homogeneity observe that the squares and square roots cancel each other, and so $\|\lambda x\| = |\lambda| \|x\|$ for every $\lambda \in \mathbb{R}$. The hardest part is usually the triangle inequality. For this, let $x, y \in \mathbb{R}^3$ and compute

$$\begin{aligned}\|x + y\| &= |x_2 + y_2| + \sqrt{|x_1 + y_1|^2 + (|x_1 + y_1| + |x_3 + y_3|)^2} \\ &\leq |x_2| + |y_2| + \sqrt{|x_1 + y_1|^2 + (|x_1| + |y_1| + |x_3| + |y_3|)^2} \\ &\leq |x_2| + |y_2| + \sqrt{|x_1|^2 + (|x_1| + |x_3|)^2} + \sqrt{|y_1|^2 + (|y_1| + |y_3|)^2},\end{aligned}$$

where we used the triangle inequality for real numbers (twice) in the first inequality, and the triangle inequality for the Euclidean norm in the second. To be precise, we've used

$$\begin{aligned} \left\| \begin{pmatrix} x_1 + y_1 \\ |x_1| + |y_1| + |x_3| + |y_3| \end{pmatrix} \right\|_2 &= \left\| \begin{pmatrix} x_1 \\ |x_1| + |x_3| \end{pmatrix} + \begin{pmatrix} y_1 \\ |y_1| + |y_3| \end{pmatrix} \right\|_2 \\ &\leq \left\| \begin{pmatrix} x_1 \\ |x_1| + |x_3| \end{pmatrix} \right\|_2 + \left\| \begin{pmatrix} y_1 \\ |y_1| + |y_3| \end{pmatrix} \right\|_2. \end{aligned}$$

Example 11.44. A rather general way to obtain norms on \mathbb{R}^d is to fix a matrix $A \in \mathbb{R}^{d \times d}$ and some $p = 1, 2, \infty$, and define

$$\|x\|_A := \|Ax\|_p \quad \text{for } x \in \mathbb{R}^d.$$

Non-negativity, homogeneity and triangle inequality follow directly from the linearity of matrix multiplication and the corresponding properties for the ℓ_p -norm.

We also obtain that $\|x\|_A = 0 \iff Ax = 0$. However, this is only equivalent to $x = 0$ if A is invertible. Therefore, $\|\cdot\|_A$ is a norm on \mathbb{R}^d if and only if A is invertible.

Note that it is really **important to check the properties a norm**, because some of results in the following do not hold otherwise. Moreover, it is sometimes not so easy to see whether it is a norm or not, which makes it necessary to be quite precise.

Let us discuss some illustrative counterexamples.

Example 11.45. One typical mistake is to forget to check that a norm on a vector space is indeed **defined (and non-negative) for all elements**. For example, the mapping

$$\|x\| := \sqrt{|x_1|^2 - (|x_1| + |x_2|)^2} \quad \text{for } x \in \mathbb{R}^2$$

is, in contrast to the norm considered in Example 11.43, not a norm on \mathbb{R}^2 , because it is not non-negative for all $x \in \mathbb{R}^2$. (It is not even real-valued for $x_2 \neq 0$.)

All this is particularly important when we talk about function spaces and norms that involve derivatives. For example, if we consider $f \in C(I)$ for some closed interval I and the mapping

$$\|f\| := \sup_{x \in I} (|f(x)| + |f'(x)|),$$

then we see that $\|f\|$ is not even defined for all $f \in C(I)$ as not every continuous function is differentiable. However, if we restrict the space to functions such that the norm is defined, e.g., to the differentiable functions $C^1(I)$, where we consider one-sided derivatives at the boundary, then we see that $\|\cdot\|$ is a norm on $C^1(I)$. (Verify yourself!)

Remark 11.46. Note that functions from $C^1(I)$ are additionally required to have a continuous derivative, which was not needed in the last example. Moreover, we may clearly consider also the multivariate situation, when we replace the absolute value of the univariate derivative $|f'(x)|$ by a multivariate analog like a norm of the gradient. (Recall that the gradient is a vector.)

Once we checked that a norm is well-defined and non-negative, it is still required to check all the other properties since all of them might be violated separately, as the following examples show. First, to show positive definiteness of a norm when we already know that its non-negative, it remains to show that, **if an element has norm zero, then it is the zero element** (which is unique in a vector space).

Example 11.47. Consider the norms $\|x\|_A = \|Ax\|_p$ for some $p = 1, 2, \infty$ and $A \in \mathbb{R}^{d \times d}$ from Example 11.44. If A is not invertible, then there must exist a vector $x^* \in \mathbb{R}^d$ with $x^* \neq 0$ and $Ax^* = 0$. (Why? Hint: There are many correct answers.) But this readily implies that $\|x^*\| = 0$, and therefore that $\|\cdot\|_A$ is not a norm.

Example 11.48. In the case of function spaces, e.g., $C^1([0, 1])$ (meaning functions whose derivative is continuous on $[0, 1]$, where we consider one-sided derivatives at the endpoints), then we might want to measure the 'size' of a function solely by means of its derivative. That is, we consider the mapping

$$\|f\| := \|f'\|_\infty = \sup_{x \in (0,1)} |f'(x)| \quad \text{for } f \in C^1([0, 1]).$$

Although this mapping is well-defined, non-negative, homogeneous and satisfies the triangle inequality, it is not positive definite. This is seen by considering a constant function, say $f(x) := 1$, on $[0, 1]$, which is continuously differentiable with $f'(x) = 0$ on $(0, 1)$. Therefore, $\|f\| = 0$ but $f \neq 0$, and so, $\|\cdot\|$ is **not a norm**. This is due to the fact that the mapping 'kills' some important information (here the constant part).

However, if we consider such a norm on a subspace, then it might be a norm. For example, let $\mathcal{F} = \{f \in C^1([0, 1]) : f(0) = 0\}$, which is a subspace of $C^1([0, 1])$, see Example 11.18. Now, only constant functions f satisfy $\|f\| = 0$ and the only constant function in \mathcal{F} is the zero function $f = 0$. Therefore, and since all other properties are not affected by restriction to a subspace, we obtain that $(\mathcal{F}, \|\cdot\|)$ is a normed space.

Remark 11.49. One way to verify the positive definiteness of a norm is to show that it is **bounded from below by another norm**. That is, a mapping $\|\cdot\|_* : X \rightarrow \mathbb{R}$ is positive definite if there is a norm $\|\cdot\|$ on X with $\|x\|_* \geq c \cdot \|x\|$ for any $x \in X$ and some fixed $c > 0$. This clearly implies that $\|x\|_* > 0$ for all $x \neq 0$.

If $\|\cdot\|_*$ is additionally homogeneous and satisfies the triangle inequality, then it is also a norm.

The second property, i.e., the homogeneity, of a norm is usually rather easy to verify. One just needs to multiply an element by a number and see that all the powers of the number, if there even exist some, cancel each other. In the same way one can easily 'see' that a mapping is not a norm.

Example 11.50. For $x = (x_1, x_2, x_3) \in \mathbb{R}^3$ consider the mapping $\|x\| := |x_1| + |x_2| + |x_3|^2$. We see that this mapping is not homogeneous in the last component. So, for a vector $x = \lambda \cdot e_3 = (0, 0, \lambda)$, $\lambda \in \mathbb{R}$, we see that $\|\lambda \cdot e_3\| = \lambda^2 \neq |\lambda| = |\lambda| \cdot \|e_3\|$ for $\lambda \notin \{0, 1\}$.

The last property to verify, i.e., the triangle inequality, is often the hardest. However, it can usually be reduced to a series of applications of 'known' triangle inequalities, as in the example presented above. To 'see' that a mapping is not a norm because it violates the triangle inequality, one may **look for products** of components which are usually not compatible with norms.

Example 11.51. Let us consider the mapping $\|x\| := |x_1| + |x_2| + \sqrt{|x_1 x_2|}$ for $x = (x_1, x_2) \in \mathbb{R}^2$. It is not hard to see that this mapping is positive definite and homogeneous. However, as there is a product of components one might be careful. (Note that this is not a general rule!) To check if this is a problem, one might just input two vectors that are non-zero only for one of the components and try to verify the triangle inequality. Here, we see $\|e_1\| = 1$, $\|e_2\| = 1$, but $\|e_1 + e_2\| = \|(1, 1)\| = 3$, which violates the triangle inequality. Hence, $\|\cdot\|$ is not a norm on \mathbb{R}^2 .

Note that verifying the properties of a norm is nothing you need to do frequently. It is mostly done once in the beginning of a study to see if the proposed setting is 'well-suited'. One can then use all the results that follow for *all* elements of the defined normed space.

However, some spaces are so important that they build the foundation of most theoretical work. This is due some very useful properties that hold for elements of these spaces, which we will discuss in the following section.

11.2.1 The L_p -spaces

In this section we have a look at the probably most important family of function spaces, namely the L_p -spaces, where $p > 0$ is a real parameter. These function spaces a fundamental tool, e.g., in stochastics,

theory of ODEs, PDEs, etc., in particular because of their flexibility with respect to the setting. That is, by employing the measure theory from the last chapter, we are able to tackle functions on discrete sets (like sequences) or rather arbitrary sets at once. Let us start with the following sets, which we need to 'modify' a bit later to make them normed spaces.

Definition 11.52. Let $(\Omega, \mathcal{A}, \mu)$ be measure space and let $0 < p < \infty$. Then, we define

$$\mathcal{L}_p(\Omega, \mu) := \left\{ f: \Omega \rightarrow \mathbb{C} : f \text{ is measurable and } \int_{\Omega} |f|^p d\mu < \infty \right\}.$$

So the set $\mathcal{L}_p(\Omega, \mu)$ is the set of all functions such that $|f|^p$ is an integrable function w.r.t. μ . Note that sometimes, the spaces are defined to consist only of real-valued functions. This would not make a difference in the following and so we directly introduce the larger space.

(The case $p = \infty$ will be discussed below.)

Remark 11.53. Let us already mention here that only the case $p \geq 1$ will lead to a normed space. However, also the case $p \in (0, 1)$ is of large practical relevance, as it leads to a framework for *sparse approximation*. We do not discuss that here in detail and refer to the literature.

Let us discuss some classes of functions that are contained in $\mathcal{L}_p(\Omega, \mu)$.

Example 11.54. Let $f: \Omega \rightarrow \{0, 1\}$ be an indicator function of some measurable $A \subset \Omega$ such that $\mu(A) < \infty$. Then, for any $0 < p < \infty$ it follows

$$\int_{\Omega} |f|^p d\mu = \int_A |1|^p d\mu = \mu(A) < \infty.$$

This shows that all such indicator functions are in $\mathcal{L}_p(\Omega, \mu)$ for all $0 < p < \infty$.

For a set $\Omega \subset \mathbb{R}^d$ equipped with the **Lebesgue measure** we write $\mathcal{L}_p(\Omega)$ instead of $\mathcal{L}_p(\Omega, \lambda)$.

Example 11.55. Let $\Omega = [0, 1]$ and let $f: [0, 1] \rightarrow \mathbb{R}$ be a continuous function. Then $|f|^p$ is still a continuous function for any $0 < p < \infty$. Thus we have that $c = \max_{x \in [0, 1]} |f(x)|^p < \infty$ exists. This implies that

$$\int_0^1 |f(x)|^p dx \leq \int_0^1 c dx = c < \infty.$$

So any continuous function on $[0, 1]$ is contained in $\mathcal{L}_p([0, 1])$, i.e., $C([0, 1]) \subset \mathcal{L}_p([0, 1])$ for any $p \in (0, \infty)$. A similar result holds if we consider arbitrary closed intervals.

However, note that **the same is not true for unbounded intervals**. If we consider, e.g., $\mathcal{L}_p(\mathbb{R})$, then we see that the constant function $\chi_{\mathbb{R}}$, which is clearly continuous, is not integrable. Therefore, $C(\mathbb{R}) \not\subset \mathcal{L}_p(\mathbb{R})$ for $p \in (0, \infty)$.

Moreover, we can also treat the **sequence spaces**.

Example 11.56. We have a look at the measure space $(\mathbb{N}, \mathcal{P}(\mathbb{N}), \mu)$, where μ is the counting measure. In this case we write $\ell_p := \ell_p(\mathbb{N}) := \mathcal{L}_p(\mathbb{N}, \mu)$. We observe that $f \in \ell_p$ if and only if

$$\int_{\mathbb{N}} |f|^p d\mu = \sum_{k=1}^{\infty} |f(k)|^p < \infty.$$

Using the identification $a_n := f(n)$, we see that ℓ_p is **the space of all p -summable sequences**, i.e., $(a_n)_{n \in \mathbb{N}} \subset \mathbb{C}$ with $\sum_{k=1}^{\infty} |a_k|^p < \infty$.

Let us first verify that $\mathcal{L}_p(\Omega, \mu)$ is a vector space.

Lemma 11.57. *Let $(\Omega, \mathcal{A}, \mu)$ be measure space and let $0 < p < \infty$.*

Then, $\mathcal{L}_p(\Omega, \mu)$ together with the point-wise addition and the usual scalar multiplication is a vector space.

Proof. We only need to show linearity. The other requirements follow since $\mathcal{L}_p(\Omega, \mu)$ is then a subspace of \mathbb{C}^Ω . Thereto, let $f, g \in \mathcal{L}_p(\Omega, \mu)$ and $\lambda \in \mathbb{C}$. Measurability of λf and $f + g$ follow from the measurability of f and g , respectively. We compute that

$$\int_{\Omega} |\lambda f|^p d\mu = |\lambda|^p \int_{\Omega} |f|^p d\mu < \infty.$$

Thus $\lambda f \in \mathcal{L}_p(\Omega, \mu)$. To see that $f + g$ is again an element of $\mathcal{L}_p(\Omega, \mu)$ we use the inequality

$$|f + g| \leq |f| + |g| \leq 2 \max\{|f|, |g|\},$$

which implies

$$\begin{aligned} \int_{\Omega} |f + g|^p d\mu &\leq \int_{\Omega} 2^p \max\{|f|, |g|\}^p d\mu \\ &\leq 2^p \int_{\Omega} |f|^p + |g|^p d\mu \\ &= 2^p \left(\int_{\Omega} |f|^p d\mu + \int_{\Omega} |g|^p d\mu \right) < \infty. \end{aligned}$$

This shows that $f + g \in \mathcal{L}_p(\Omega, \mu)$ as claimed. \square

To make $\mathcal{L}_p(\Omega, \mu)$ a normed space, we need a norm. Therefore, we define the mapping

$$\|\cdot\|_p : \mathcal{L}_p(\Omega, \mu) \rightarrow [0, \infty)$$

for with

$$\|f\|_p := \left(\int_{\Omega} |f|^p d\mu \right)^{\frac{1}{p}}.$$

We now want to show whether $\|\cdot\|_p$ satisfies all requirements of a norm. Note that it is clearly defined and non-negative for all $f \in \mathcal{L}_p(\Omega, \mu)$, and homogeneous. However, to prove the triangle inequality we need some auxiliary results, which are of independent interest. Afterwards we discuss the positive definiteness.

Lemma 11.58. *Let $a, b \geq 0$ and $r \in (0, 1)$. Then*

$$a^r b^{1-r} \leq r a + (1 - r)b.$$

Proof. The inequality is clearly true if either $a = 0$ or $b = 0$. If $a, b > 0$ then the inequality is equivalent to the following estimate

$$\log(a^r b^{1-r}) = r \log a + (1 - r) \log b \leq \log(ra + (1 - r)b).$$

Since the second derivative of \log is negative, the logarithm is a concave function. Thus, the above inequality is true, which concludes the proof. \square

We are now able to state one of the most important inequalities in all of mathematics. This inequality was found by *Otto Hölder* (1859–1937) in 1889.

Theorem 11.59 (Hölder's inequality). *Let $1 < p, q < \infty$ such that $\frac{1}{p} + \frac{1}{q} = 1$. If $f \in \mathcal{L}_p(\Omega, \mu)$ and $g \in \mathcal{L}_q(\Omega, \mu)$, then $fg \in \mathcal{L}_1(\Omega, \mu)$ and*

$$\|fg\|_1 \leq \|f\|_p \|g\|_q.$$

Two numbers $p, q \in (1, \infty)$ with $\frac{1}{p} + \frac{1}{q} = 1$ are called **(Hölder) conjugates**. Note that it is essential to consider $p \geq 1$ to have a Hölder conjugate. ($p = 1$ will be considered soon.) Therefore, we do not have Hölder's inequality for $p < 1$.

Proof. Clearly, fg is measurable. So it remains to show the above stated inequality which also implies that $fg \in \mathcal{L}_1(\Omega, \mu)$.

To show this inequality, let $A := \|f\|_p^p$ and $B := \|g\|_q^q$. If either A or B is zero then there is nothing to show. Therefore, we assume that both of them are strictly positive.

For arbitrary but fixed $x \in \Omega$ we define

$$a = \frac{|f(x)|^p}{A} \quad \text{and} \quad b = \frac{|g(x)|^q}{B}.$$

Setting $r = \frac{1}{p}$, which implies that $1 - r = \frac{1}{q}$, we can use Lemma 11.58 to see that

$$\frac{|f(x)|}{A^{\frac{1}{p}}} \frac{|g(x)|}{B^{\frac{1}{q}}} = a^r b^{1-r} \leq ra + (1-r)b = \frac{1}{p} \frac{|f(x)|^p}{A} + \frac{1}{q} \frac{|g(x)|^q}{B}.$$

Since x was arbitrary this is valid for any $x \in \Omega$. In combination with the definition of A, B this yields

$$\int_{\Omega} \frac{|f(x)|}{A^{\frac{1}{p}}} \frac{|g(x)|}{B^{\frac{1}{q}}} d\mu(x) \leq \frac{1}{p} \int_{\Omega} \frac{|f(x)|^p}{A} d\mu(x) + \frac{1}{q} \int_{\Omega} \frac{|g(x)|^q}{B} d\mu(x) = \frac{1}{p} \frac{A}{A} + \frac{1}{q} \frac{B}{B} = \frac{1}{p} + \frac{1}{q} = 1.$$

This implies that

$$\int_{\Omega} |fg| d\mu \leq A^{\frac{1}{p}} B^{\frac{1}{q}} = \|f\|_p \|g\|_q.$$

□

Note that the last result holds for general measure spaces. In particular, it is often used for sequences. Therefore, we formulate it for this case as a corollary.

Corollary 11.60. *Let $1 < p, q < \infty$ such that $\frac{1}{p} + \frac{1}{q} = 1$.*

Then, for all sequences $(x_n)_{n \in \mathbb{N}}, (y_n)_{n \in \mathbb{N}} \in \mathbb{C}^{\mathbb{N}}$, we have

$$\sum_{k=1}^{\infty} |x_k y_k| \leq \left(\sum_{k=1}^{\infty} |x_k|^p \right)^{\frac{1}{p}} \left(\sum_{k=1}^{\infty} |y_k|^q \right)^{\frac{1}{q}}$$

For $p = q = 2$, this is the **Cauchy-Schwarz inequality**, which we already proved in Lemma 1.76 for finite sums.

With the above, we can now prove that $\|\cdot\|_p$ satisfies the triangle inequality on $\mathcal{L}_p(\Omega, \mu)$ for every $p \in [1, \infty)$. This inequality is usually attributed to *Hermann Minkowski* (1864–1909).

Theorem 11.61 (Minkowski's inequality). *Let $f, g \in \mathcal{L}_p(\Omega, \mu)$ with $p \in [1, \infty)$.*

Then,

$$\|f + g\|_p \leq \|f\|_p + \|g\|_p.$$

Proof. The case $p = 1$ just follows by using $|f + g| \leq |f| + |g|$ point-wise. For $p > 1$ we set $q = \frac{p}{p-1}$ which implies that $\frac{1}{p} + \frac{1}{q} = 1$. We observe that

$$\begin{aligned}\int_{\Omega} |f + g|^p d\mu &= \int_{\Omega} |f + g| \cdot |f + g|^{p-1} d\mu \\ &\leq \int_{\Omega} |f| \cdot |f + g|^{p-1} d\mu + \int_{\Omega} |g| \cdot |f + g|^{p-1} d\mu.\end{aligned}$$

Furthermore,

$$\| |f + g|^{p-1} \|_q^q = \int_{\Omega} (|f + g|^{p-1})^q d\mu = \int_{\Omega} |f + g|^p d\mu = \| f + g \|_p^p < \infty,$$

which shows that $|f + g|^{p-1} \in \mathcal{L}_q(\Omega, \mu)$. (Here, we only used the definition of p and q .) In conclusion,

$$\begin{aligned}\| f + g \|_p^p &\leq \| f \cdot |f + g|^{p-1} \|_1 + \| g \cdot |f + g|^{p-1} \|_1 \\ &\leq \| f \|_p \| |f + g|^{p-1} \|_q + \| g \|_p \| |f + g|^{p-1} \|_q \\ &= (\| f \|_p + \| g \|_p) \| |f + g|^{p-1} \|_q \\ &= (\| f \|_p + \| g \|_p) \| f + g \|_q^{p/q} \\ &= (\| f \|_p + \| g \|_p) \| f + g \|_p^{p-1}.\end{aligned}$$

Note that in the last step we only used that $p/q = p - 1$. Dividing by $\| f + g \|_p^{p-1}$ implies the desired result. \square

It only remains to verify whether the mapping $\|\cdot\|_p$ is positive definite or not. As it is clearly non-negative, we only need to check that $\|f\|_p = 0$ implies $f = 0$. In some cases, like for the sequence spaces $\ell_p = \ell_p(\mathbb{N})$, we easily see that this is the case. Just note that for $a = (a_n)_{n \in \mathbb{N}} \in \ell_p$ we obtain that $\|a\|_p^p = \sum_{k=1}^{\infty} |a_k|^p = 0$ implies $a_k = 0$ for all $k \in \mathbb{N}$, and therefore $a = 0$. Note that the same holds for the counting measure on arbitrary countable sets. Let us fix this finding in the following proposition.

Proposition 11.62. *Let I be a countable set and $p \in [1, \infty)$. Then,*

$$\|a\|_p := \left(\sum_{k \in I} |a_k|^p \right)^{1/p}$$

defines a norm on $\ell_p(I) := \{a = (a_i)_{i \in I} : \|a\|_p < \infty\}$.

(We use the letter ' I ' for *index set*.)

However, what about other measures, like the Lebesgue measure? Unfortunately, the mapping $\|\cdot\|_p$ is not a norm on $\mathcal{L}_p(\Omega, \mu)$ in general. This is due to the existence of null sets in the underlying measure space. Precisely, we have $\chi_N \in \mathcal{L}_p(\Omega, \mu)$ for any measurable null set N , as well as $\chi_N^p = \chi_N$ for any $p \in [1, \infty)$. Therefore, $\|\chi_N\| = 0$ although $\chi_N \neq 0$ for $N \neq \emptyset$.

For example, considering $\mathcal{L}_p([0, 1])$, we see that $\|\chi_{\{0\}}\|_p = 0$, or even $\|\chi_{\mathbb{Q}}\|_p = 0$ for all $p \in [1, \infty)$ (as we have learned in the last chapter), but these functions are clearly not the zero function.

However, there is a way to 'construct' normed spaces consisting of $\mathcal{L}_p(\Omega, \mu)$ -functions (independent of the measure we are using) which are normed w.r.t. $\|\cdot\|_p$. Thereto, we have to weaken our notion of equality of two functions. This means that we wish to say two functions are just the same if they agree almost everywhere. For example, we would assume $\chi_{\mathbb{Q}}$ equals the zero function. There are many ways to make this precise, but the probably most elegant is by using *equivalence classes* of functions. That is, instead of considering individual functions, we work directly with sets of functions that are equal almost everywhere.

This leads to the following definition.

Definition 11.63. Let $(\Omega, \mathcal{A}, \mu)$ be a measure space and $0 < p < \infty$.

Moreover, for each $f \in \mathcal{L}_p(\Omega, \mu)$, we define the set

$$[f] := \left\{ g \in \mathcal{L}_p(\Omega, \mu) : f = g \text{ almost everywhere} \right\},$$

which is called the **equivalence class of f** .

Then, we define the L_p -spaces

$$L_p(\Omega, \mu) := \left\{ [f] : f \in \mathcal{L}_p(\Omega, \mu) \right\}.$$

We write $L_p(\Omega)$ for $L_p(\Omega, \mu)$ if $\Omega \subset \mathbb{R}^d$ and μ is the Lebesgue measure.

First of all observe that, formally, $L^p(\Omega, \mu)$ is **not a function space**. It consists of sets of functions, which might seem a bit confusing first. However, this is just a way of getting rid of the problem, that there are elements of norm zero, which are not the zero element.

Still, it is convention to treat the elements of $L_p(\Omega, \mu)$ like functions. That is, we can still write $\int f d\mu$, $\|f\|_p$ etc. for $f \in L_p(\Omega, \mu)$, although these mappings are only defined for functions. The reason is that these quantities do not depend on the specific element of an equivalence class. Precisely, we have $\int_{\Omega} g d\mu = \int_{\Omega} f d\mu$ for every $g \in [f]$, and we therefore omit the notational difficulties when we are working with integrals.

However, when it comes to function evaluations at certain points $x \in \Omega$, then we need to be careful. In many cases, sets of the form $\{x\}$, i.e., *singletons*, are sets of measure zero, e.g., for the Lebesgue measure. In such a case, functions in the same equivalence class can have different values at x , but are still considered equal. Since $f \in L_p(\Omega, \mu)$ is (formally) a set of functions we cannot speak about a function value $f(x)$. For example, we have $\chi_{\mathbb{Q}} \in \mathcal{L}_p(\mathbb{R})$ with $\chi_{\mathbb{Q}}(0) = 1$, but $\chi_{\mathbb{Q}} \in [0]$ since $\|\chi_{\mathbb{Q}}\|_p = 0$ for all $p \geq 1$. That is, we consider ' $0 = \chi_{\mathbb{Q}}$ ' in $L_p(\mathbb{R})$. (Formally, we have $[0] = [\chi_{\mathbb{Q}}]$. Verify that!) This shows that **function evaluations are not well defined in general**. That is, we need additional considerations (of other spaces of functions) when we are interested in point-wise statements. This might seem not very useful at first but these spaces obey a rich structure and many important function spaces build upon these concepts.

We will see later that there are function spaces (i.e., those with a *reproducing kernel*) that are subspaces of these $L_p(\Omega, \mu)$ such that function evaluation is well-defined.

Remark 11.64. Note that it would be more precise to write $[f] \in L_p(\Omega, \mu)$ for $f \in \mathcal{L}_p(\Omega, \mu)$ to indicate that elements of $L_p(\Omega, \mu)$ are equivalence classes. For notational convenience, we often omit this technicality and write $f \in L_p(\Omega, \mu)$ for $f \in [f] \in L_p(\Omega, \mu)$, i.e., we assume that f is a function, but keep in mind that it is undetermined on null sets.

Remark 11.65. The sets $[f]$ are called equivalence classes, because they are defined via an *equivalence relation*. (Recall the notion of an equivalence relation from Definition 1.23.) An equivalence relation \sim is somehow a substitute for the usual equality $=$, as it is reflexive, symmetric and transitive. If we now define $f \sim g : \iff f = g$ almost everywhere, then we have $[f] = \{g \in \mathcal{L}_p(\Omega, \mu) : g \sim f\}$. Note that, by the properties of the norm, we also have $[f] = \{g \in \mathcal{L}_p(\Omega, \mu) : \|f - g\|_p = 0\}$. Hence, we use equivalence classes to **partition the space into elements that can be distinguished**.

Remark 11.66. Note that the procedure applied in the last definition works in general, to 'construct' a vector space such that a given non-negative, homogeneous mapping $\|\cdot\|$ that satisfies the triangle inequality, is positive definite on it.

For this, let X be the given vector space and define $N := \{x \in X : \|x\| = 0\}$. We then define the equivalence class of $x \in X$ by $[x] := x + N$, i.e., all elements of X that differ from x only by an element

of norm zero. Clearly, we can also write $[x] = \{y \in X : \|y - x\| = 0\}$. Since, by construction, all elements from $[x]$ have the same norm, we just define $\|[x]\| := \|x\|$. The space $X/N := \{[x] : x \in X\}$ is then called the **quotient space of X and N** , and we see that $[0]$ is the only element of X/N with norm 0, making $\|\cdot\|$ a norm on X/N .

We now show that this definition of space of 'almost everywhere defined functions' leads to a normed space in general. First of all, to make L_p a vector space, we need to say what we mean by addition and scalar multiplication for equivalence classes. Therefore, we use the straightforward definitions

$$[f] + [g] := [f + g]$$

and

$$\lambda \cdot [f] := [\lambda \cdot f]$$

for all $f, g \in \mathcal{L}_p(\Omega, \mu)$ and $\lambda \in \mathbb{F}$. (Verify that $L_p(\Omega, \mu)$ is indeed a vector space!)

We are now able to prove that $\|\cdot\|_p$ defines a norm on $L_p(\Omega, \mu)$.

Theorem 11.67. *For any $1 \leq p < \infty$ we have that $L_p(\Omega, \mu)$ together with*

$$\|f\|_p := \left(\int_{\Omega} |f|^p d\mu \right)^{\frac{1}{p}}$$

is a normed space.

Proof. Since $L_p(\Omega, \mu)$ is a vector space, it remains to verify that $\|\cdot\|_p$ is a norm.

First of all, note that $\|f\|_p$ does not depend on a specific element of the equivalence class of f , i.e., $\|f\|_p = \|g\|_p$ for all $g \in [f]$. Hence, $\|\cdot\|$ is well-defined for equivalence classes, and we use the identification $\|[f]\|_p = \|f\|_p$.

Moreover, $\|\cdot\|_p$ is non-negative and homogeneous, and satisfies the triangle inequality due to Minkowski's inequality. It remains to show that $\|\cdot\|_p$ is positive definite, i.e., that $\|f\|_p = 0$ implies $f = 0$ a.e. (i.e., $[f] = [0]$). For this, assume that $f = 0$ a.e. is not true, which is equivalent to $\mu(\{x : |f(x)| > c\}) > 0$ for some $c > 0$. We obtain that $\|f\|_p^p > c^p \cdot \mu(\{x : |f(x)| > c\}) > 0$, and so $\|f\|_p > 0$. Taking the reverse statement, we get that $\|f\|_p = 0$ implies $f = 0$ a.e. \square

So far we only had a look at $1 \leq p < \infty$. Now we discuss the case $p = \infty$.

That is, we want to define something like the supremum-norm for continuous functions (i.e., $\|f\|_{\infty} = \sup_x |f(x)|$), but on vector spaces similar to the above L_p -spaces. In particular, as function values are not defined, the supremum-norm would not be defined for "L_p-functions" (which are actually sets). Let us therefore introduce a notion of supremum and infimum that is compatible with this concept.

Recall that, for a measurable function $f : \Omega \rightarrow \mathbb{R}$, we have that all the sets

$$\{f > a\} := \{x \in \Omega : f(x) > a\} = f^{-1}(a, \infty),$$

i.e., those $x \in \Omega$ such that $f(x)$ is larger than a , are measurable sets. Clearly, if there is some $a \in (0, \infty)$ such that $\{f > a\} = \emptyset$, then f is bounded from above (in the 'classical' sense). However, we now consider two functions to be equal if they only disagree on a null set. Hence, we should also consider a version of the supremum that 'ignores' the values on a null set. This quantity is called the **essential supremum of f** and is defined by

$$\text{ess sup}_{\Omega} f := \inf \left\{ a \in \mathbb{R} : \mu(\{f > a\}) = 0 \right\}.$$

(We write $\text{ess sup } f$ and omit the Ω if the underlying set is clear.)

In words $\text{ess sup } f$ is the smallest number which can be used to bound f almost everywhere. Similarly, we can define the the **essential infimum of f** by

$$\text{ess inf}_{\Omega} f := \sup \left\{ a \in \mathbb{R} : \mu(\{f < a\}) = 0 \right\}.$$

We call f **essentially bounded** if

$$\|f\|_{\infty} := \operatorname{ess\,sup}_{x \in \Omega} |f| < \infty.$$

Note that $\|\cdot\|_{\infty}$ is often also called **supremum norm**. Therefore, it is essential to check the setting you are working in to not getting confused with this notation overload.

Remark 11.68. There is another equivalent definition of the supremum norm essential supremum that is sometimes useful. Namely, instead of taking the infimum over the function values (i.e., a), we can also take the infimum over all null sets that we can exclude, i.e.,

$$\operatorname{ess\,sup} f = \inf \left\{ \sup \left\{ f(x) : x \in \Omega \setminus N \right\} : \mu(N) = 0 \right\}.$$

That is, we take the supremum of f on $\Omega \setminus N$ for all null sets N , and then take the infimum of all these values.

Similar as before we can define the following spaces of essentially bounded functions.

Definition 11.69. Let $(\Omega, \mathcal{A}, \mu)$ be a measure space. Then, we define

$$\mathcal{L}_{\infty}(\Omega, \mu) := \left\{ f : \Omega \rightarrow \mathbb{R} : f \text{ is measurable and } \|f\|_{\infty} < \infty \right\}.$$

If we consider again the equivalence classes $[f] = \{g \in \mathcal{L}_{\infty}(\Omega, \mu) : f = g \text{ a.e.}\}$, then we define

$$L_{\infty}(\Omega, \mu) := \left\{ [f] : f \in \mathcal{L}_{\infty}(\Omega, \mu) \right\}.$$

We write $L_{\infty}(\Omega)$ for $L_{\infty}(\Omega, \mu)$ if $\Omega \subset \mathbb{R}^d$ and μ is the Lebesgue measure, and $\ell_{\infty}(I)$ if I is a countable set and μ is the counting measure.

Let us note, again, that $g \in [f]$, i.e., $f = g$ a.e., holds if and only if $\|f - g\|_{\infty} = 0$. We also follow the usual convention that we speak of “ L_{∞} -functions” instead of equivalence classes of functions. However, keep in mind that elements of L_{∞} are not always defined point-wise.

Example 11.70. For every measurable set A , we obviously have that $\|\chi_A\|_{\infty} \leq 1$. Indeed, we have that

$$\|\chi_A\|_{\infty} = \begin{cases} 1, & \mu(A) > 0, \\ 0, & \mu(A) = 0. \end{cases}$$

Example 11.71. Note that in measure spaces without a non-empty null set, we have that all equivalence classes contain only one element. In such a case, the construction of L_{∞} is again not needed and we can just use the identification $f = [f]$.

The most prominent example are the sequence spaces $\ell_{\infty} = \ell_{\infty}(\mathbb{N})$ of all bounded sequences, i.e., $\ell_{\infty} = \{(a_n)_{n \in \mathbb{N}} : \sup_n |a_n| < \infty\}$. For these normed spaces we clearly have $(a_n)_{n \in \mathbb{N}} = (b_n)_{n \in \mathbb{N}}$ almost everywhere if and only if $a_n = b_n$ for all $n \in \mathbb{N}$. (If $a_k \neq b_k$ for some k , then (a_n) and (b_n) differ on a set of measure at least one.)

Let us finally verify that $\|\cdot\|_{\infty}$ defines a norm on $L_{\infty}(\Omega, \mu)$.

Theorem 11.72. $L_{\infty}(\Omega, \mu)$ together with $\|\cdot\|_{\infty}$ is a normed space.

Proof. The proof is very similar to the above ones. Let us first assume that $f, g \in \mathcal{L}_\infty(\Omega, \mu)$. Then, there exist null sets N, M such that $|f(x)| \leq \|f\|_\infty$ for all $x \notin N$, and $|g(x)| \leq \|g\|_\infty$ for all $x \notin M$. Using the triangle inequality point-wise it follows that

$$|f(x) + g(x)| \leq |f(x)| + |g(x)| \leq \|f\|_\infty + \|g\|_\infty \quad \text{for all } x \notin N \cup M,$$

which shows that $\sup\{|f(x) + g(x)| : x \in \Omega \setminus (N \cup M)\} \leq \|f\|_\infty + \|g\|_\infty$. However, since $N \cup M$ is still a null set, we obtain that

$$\|f + g\|_\infty = \operatorname{ess\,sup}_{\Omega} |f + g| \leq \sup\{|f(x) + g(x)| : x \in \Omega \setminus (N \cup M)\} \leq \|f\|_\infty + \|g\|_\infty < \infty.$$

Using this we easily prove that $\mathcal{L}_\infty(\Omega, \mu)$, and therefore $L_\infty(\Omega, \mu)$, are vector spaces. This also shows the triangle inequality, and the other properties of a norm are also straightforward. Let us just recall the fact that $\|f\|_\infty = 0$ if and only if $|f(x)| = 0$ almost everywhere, which clearly implies that $f = 0$ a.e. \square

Let us finish this section with the final statement of the Hölder inequality for the L_p -spaces, including the limiting cases $p = 1, \infty$.

Theorem 11.73 (Hoelder's inequality continued). *Let $1 \leq p, q \leq \infty$ such that $\frac{1}{p} + \frac{1}{q} = 1$. If $f \in L_p(\Omega, \mu)$ and $g \in L_q(\Omega, \mu)$, then $fg \in L_1(\Omega, \mu)$ and*

$$\|fg\|_1 \leq \|f\|_p \|g\|_q.$$

For $p = \infty$ we have $q = 1$ and therefore $\|fg\|_1 \leq \|f\|_\infty \|g\|_1$.

Proof. The case $p, q \in (1, \infty)$ was already proven in Theorem 11.59. Moreover, the cases $p = 1$ and $p = \infty$ work analogous. We therefore only prove the case $p = \infty$. For this, observe that

$$|f(x)g(x)| \leq \|f\|_\infty |g(x)|$$

for almost all $x \in \Omega$. Thus,

$$\int_{\Omega} |fg| d\mu \leq \|f\|_\infty \int_{\Omega} |g| d\mu.$$

This implies the desired result. \square

11.2.2 Sequences in normed spaces and Banach spaces

Note that many (advanced) concepts of mathematics are intrinsically concerned with sequences and their limits. Besides quantities (like the derivative or an integral) which are actually defined using limits, there are also many other situations where we want to employ that a sequence or series converges to something. One may think about Taylor series (Sections 5.5 & 8.6) or Fourier series (Chapter 7), where the goal was to write down (or even approximate) a function as a sum of other functions. Although we have learned many results on *point-wise convergence* of sequences of functions, and also some for *uniform convergence* (see Section 7.3), we still don't have a general theory at hand that allows to classify 'how' a sequence converges. However, this is essential to deduce properties of the *limit function* without calculating it explicitly, which is impossible (or at least impractical) in general. In particular, **we want to know if the limit of a sequence of functions with a certain property has the same property**. Ideally, we would even like to say that this is true for all sequences in a normed space. We will see shortly that this is indeed the case for many important normed spaces, and these spaces will be denoted by a special name, i.e., *Banach spaces*. However, such a powerful statement is clearly not true for all normed spaces and we will also discuss some counterexamples to illustrate typical pitfalls.

Let us start with the general definition of convergence, which is again based on the underlying normed space, and is therefore a rather flexible notion.

Definition 11.74 (Convergence in normed spaces.). Let $(F, \|\cdot\|)$ be a normed space. A sequence $(f_n)_{n \in \mathbb{N}}$ is called **convergent in** $(F, \|\cdot\|)$ if there exists some $f \in F$ such that

$$\lim_{n \rightarrow \infty} \|f - f_n\| = 0.$$

If this is the case we write $f = \lim_{n \rightarrow \infty} f_n$ or $f_n \rightarrow f$ in $(F, \|\cdot\|)$.

If the norm is fixed, we just say $(f_n)_{n \in \mathbb{N}}$ is **convergent in** F and write $f_n \rightarrow f$ in F .

(We use F for the underlying vector space, which will mostly be function spaces, to reduce confusion between the elements of the space and the inputs.)

Note that it is important here to say “convergent in $(F, \|\cdot\|)$ ” to indicate the norm that is considered for the convergence and that the limit is in F .

Let us start with a very basic example.

Example 11.75. We already know that $(\mathbb{R}, |\cdot|)$ is a normed space. Now, convergence in \mathbb{R} is just the usual convergence of sequences of numbers. The same holds for the convergence of vectors, i.e., convergence in $(\mathbb{R}^d, \|\cdot\|_p)$ for $d \in \mathbb{N}$, see Section 8.1.

However, already when talking about sequences, the specific norm of a normed space is crucial for the convergence.

Example 11.76. We consider the space $(\ell_2, \|\cdot\|_2)$ and have a look at the sequence $(x_n)_{n \in \mathbb{N}} \subset \ell_2$ with

$$x_n = \left(1, \frac{1}{2}, \frac{1}{3}, \dots, \frac{1}{n}, 0, 0, \dots\right).$$

(Note that a sequence $(x_n)_{n \in \mathbb{N}} \subset \ell_2$ is a sequence of sequences!)

First we note that any x_n is contained in ℓ_2 since we only have to consider finite sums. Moreover, we have that

$$\sum_{k=1}^{\infty} \frac{1}{k^2} = \frac{\pi^2}{6},$$

see Corollary 7.28. Thus, $x = (1, 1/2, 1/3, \dots) \in \ell_2$, because its elements are *square-summable*. Moreover, we see that

$$\|x - x_n\|_2^2 = \sum_{k=n+1}^{\infty} \frac{1}{k^2} \leq \frac{1}{n},$$

which follows from Lemma 7.34 and the example below this lemma. Hence, $x_n \rightarrow x$ in ℓ_2 .

We now consider the same sequence but as sequence in ℓ_1 , i.e., $(x_n)_{n \in \mathbb{N}} \subset \ell_1$. Note that $x_n \in \ell_1$ follows again from the fact that it has only finitely many non-zero entries. However, we have $\|x_n\|_1 = \sum_{k=1}^n \frac{1}{k}$, which is the harmonic series, and so $\|x_n\|_1 \rightarrow \infty$. Therefore, there cannot be any $x \in \ell_1$ (which implies $\|x\|_1 < \infty$) such that $\|x - x_n\|_1 \rightarrow 0$, i.e., $(x_n)_{n \in \mathbb{N}}$ is not convergent in ℓ_1 .

Remark 11.77. It is worth mentioning that the choice of a norm is also a choice for a specific ‘strength of convergence’, and this is also related to the ‘size’ of the corresponding normed space.

For example, it is easy to verify that $\|x\|_\infty \leq \|x\|_p$ for any $x \in \ell_p$, for any $1 \leq p < \infty$. This also shows that $\ell_p \subset \ell_\infty$, i.e., every summable sequence is bounded. Therefore, we obtain that every sequence $(x_n)_{n \in \mathbb{N}} \subset \ell_p$ that converges to some $x \in \ell_p$, i.e., $\|x - x_n\|_p \rightarrow 0$, also satisfies $\|x - x_n\|_\infty \rightarrow 0$, i.e., it is also convergent in ℓ_∞ .

The converse is not true in general: The sequence $(x_n)_{n \in \mathbb{N}} \subset \ell_\infty$ with $x_n = (\frac{1}{n}, \frac{1}{n}, \dots)$ satisfies $\|x_n\|_\infty = \frac{1}{n} \rightarrow 0$ and hence, converges in ℓ_∞ (to 0), but it is not even a sequence in ℓ_p for $p < \infty$.

Remark 11.78. Note that the dependence on the norm for the question whether a sequence converges or not, cannot appear for sequences of finite-dimensional vectors. We have already seen in Lemma 8.6 that in these cases convergence w.r.t. the ℓ_2 -norm is equivalent to component-wise convergence and the same result could be proven for any norm. (See also Lemma 8.8.) Therefore, convergence in finite dimensional normed spaces is independent of the norm. We omit a formal proof.

We now turn to function spaces, which are the main motivation for most of the upcoming concepts. Recall from Section 7.3 that there is a **crucial difference between point-wise and uniform convergence** of a sequence of functions. Here, we will discuss some more concepts of convergence, but let us first repeat an example from there with our new notation.

Example 11.79. We consider a closed interval I and the normed space $(C(I), \|\cdot\|_\infty)$. A sequence of functions $(f_n)_{n \in \mathbb{N}}$ is convergent in $C(I)$ if there exists some function $f \in C(I)$ such that

$$\|f - f_n\|_\infty = \sup_{x \in I} |f(x) - f_n(x)| \rightarrow 0.$$

(Note that we called that *uniform convergence* earlier.)

If we now consider the function $f_n(x) = x^n$ on set $I = [0, 1/2]$, we first note that $f_n \rightarrow 0$ point-wise, i.e., $f_n(x) \rightarrow 0$ for every fixed x . Moreover, by observing that $\|f_n\|_\infty = \sup_{x \in I} |x^n| = \frac{1}{2^n} \rightarrow 0$, we also obtain that $f_n \rightarrow 0$ in $C(I)$.

However, if we consider the same functions f_n on $I = [0, 1]$, then we see that the point-wise limit is the Dirac delta $f = \delta_{\{1\}}$, i.e., $f_n(x) \rightarrow 0$ for $x \in [0, 1)$ and $f_n(1) \rightarrow 1$. Since the (point-wise) limit f is not a continuous function, we see that $(f_n)_{n \in \mathbb{N}}$ is not convergent in $C([0, 1])$. One might also argue that $\|f - f_n\|_\infty = 1$ for all $n \in \mathbb{N}$ and therefore, that $\|f - f_n\|_\infty \not\rightarrow 0$.

The main disadvantage of the above concept is (as we already encountered earlier) that we need to 'know the limit' to verify the type of convergence. Fortunately, this is actually not needed in many cases and we can define another type of sequences to prevent this. Recall that we have already discussed the same in Section 3.5 for real-valued sequences, i.e., sequences in $(\mathbb{R}, |\cdot|)$.

Definition 11.80 (Cauchy sequences). Let $(F, \|\cdot\|)$ be a normed space and let $(f_n)_{n \in \mathbb{N}} \subset F$.

If for any $\varepsilon > 0$ there exists some $n_0 \in \mathbb{N}$ such that for any $n, m \geq n_0$ we have that

$$\|f_n - f_m\| < \varepsilon,$$

then we say $(f_n)_{n \in \mathbb{N}}$ is a **Cauchy sequence (w.r.t. $\|\cdot\|$)**.

If all Cauchy sequences of a normed space $(F, \|\cdot\|)$ are convergent sequences, then we call $(F, \|\cdot\|)$ a **complete normed space** or a **Banach space**.

Remark 11.81. First note that every convergent sequence $(f_n)_{n \in \mathbb{N}}$ in normed space $(F, \|\cdot\|)$ is a Cauchy sequence. This follows from the triangle inequality $\|f_n - f_m\| \leq \|f_n - f\| + \|f - f_m\|$ and the definition of convergence to $f := \lim_{n \rightarrow \infty} f_n$. See Section 3.5 for details. This shows that, **in Banach spaces, Cauchy sequences and convergent sequences are just the same**.

There was one particularly helpful tool for Cauchy sequences, which we state now again in our more general setting.

Lemma 11.82. Let $(F, \|\cdot\|)$ be a normed space. If $(f_n)_{n \in \mathbb{N}}$ is a Cauchy sequence, which has a convergent subsequence, then $(f_n)_{n \in \mathbb{N}}$ is a convergent sequence.

Proof. Denote the convergent subsequence of (f_n) by (f_{n_k}) and assume that it has the limit $f \in F$. It follows that for arbitrary $\varepsilon > 0$ and large enough n_k, m that

$$\|f_m - f\| \leq \|f_m - f_{n_k}\| + \|f_{n_k} - f\| < 2\varepsilon,$$

which proves the claim. \square

We have a look at some examples. This first shows the earlier result on real-valued sequences.

Example 11.83. The normed space $(\mathbb{R}, |\cdot|)$ is a Banach space since all Cauchy sequences are convergent in \mathbb{R} , see Theorem 3.56.

The same also applies to finite-dimensional vectors.

Example 11.84. The normed spaces $(\mathbb{R}^d, \|\cdot\|_p)$ with $1 \leq p \leq \infty$, i.e., the spaces ℓ_p^d , are Banach spaces. This follows from the fact that each component of a convergent or Cauchy sequence in \mathbb{R}^d is a convergent or Cauchy sequence in \mathbb{R} , respectively. (Verify this precisely!)

Let us now discuss the most important examples, starting with the space of continuous functions $(C(I), \|\cdot\|_\infty)$ on closed intervals. One may prove a similar result for arbitrary bounded and closed sets.

Theorem 11.85. Let $I = [a, b]$ be a closed interval. Then, $(C(I), \|\cdot\|_\infty)$ is a Banach space.

In particular, this implies that whenever a sequence of functions in $C(I)$ converges w.r.t. $\|\cdot\|_\infty$ to an arbitrary function f , then this is a continuous function. So, all possible limits of sequences in $C(\Omega)$ are contained in $C(\Omega)$. Note that we basically proved that already in Lemma 7.24.

Proof. We consider a Cauchy sequence $(f_n)_{n \in \mathbb{N}}$, i.e., we have for arbitrary $\varepsilon > 0$ that

$$\sup_{x \in I} |f_n(x) - f_m(x)| < \varepsilon$$

for n, m large enough. Thus, for any fixed $x_0 \in I$ the sequence $(f_n(x_0))_{n \in \mathbb{N}}$ is a Cauchy sequence in \mathbb{R} , hence convergent. Thus, the mapping $f: I \rightarrow \mathbb{R}$ with

$$f(x) := \lim_{n \rightarrow \infty} f_n(x)$$

is well-defined. It remains to prove that $f_n \rightarrow f$ in $C(I)$ and that $f \in C(I)$.

Again, since $(f_n)_{n \in \mathbb{N}}$ is a Cauchy sequence, there exists a number $N = N(\varepsilon)$ such that

$$\|f_m - f_n\|_\infty < \frac{\varepsilon}{2}$$

for any $n, m \geq N$. This means that $|f_m(x) - f_n(x)| < \frac{\varepsilon}{2}$ for all $x \in I$ and $n, m \geq N$. If we let $m \rightarrow \infty$, we obtain $|f(x) - f_n(x)| < \frac{\varepsilon}{2}$ for all $x \in I$ and $n \geq N$ and therefore, $\|f - f_n\|_\infty \leq \frac{\varepsilon}{2} < \varepsilon$. This shows that the sequence $(f_n)_{n \in \mathbb{N}}$ converges also uniformly (i.e., in $C(I)$) to f .

To prove that $f \in C(I)$, i.e., that f is continuous, we fix some arbitrary $x_0 \in I$ and show that f has to be continuous in x_0 . This implies the result.

Thereto, let $\varepsilon > 0$ be arbitrary. We choose $k \in \mathbb{N}$ (depending on ε) such that

$$\|f - f_k\|_\infty < \frac{\varepsilon}{3}.$$

Now, since f_k is continuous at each $x_0 \in I$ there exists some $\delta > 0$ such that for any x with $|x - x_0| < \delta$ it follows

$$|f_k(x) - f_k(x_0)| < \frac{\varepsilon}{3}.$$

Hence, for x such that $|x - x_0| < \delta$ we have

$$\begin{aligned} |f(x) - f(x_0)| &\leq |f(x) - f_k(x)| + |f_k(x) - f_k(x_0)| + |f_k(x_0) - f(x_0)| \\ &\leq 2\|f_k - f\|_\infty + |f_k(x) - f_k(x_0)| \\ &< 2\frac{\varepsilon}{3} + \frac{\varepsilon}{3} = \varepsilon. \end{aligned}$$

This implies that f is continuous at x_0 . \square

The other important example here are the L_p -spaces.

Theorem 11.86. *For any $1 \leq p \leq \infty$, the space $L_p(\Omega, \mu)$ is a Banach space.*

That is, any Cauchy sequence in $(L_p(\Omega, \mu), \|\cdot\|_p)$ is convergent in this space.

Proof of Theorem 11.86. We only prove the case $p < \infty$. The case $p = \infty$ is very similar to the case $C(I)$ above. We show that any Cauchy sequence in $L_p(\Omega, \mu)$ converges to some function $f \in L_p(\Omega, \mu)$. (Recall that elements of L_p are equivalence classes and all statements have to be understood almost everywhere. Specific functions are arbitrary elements from a class.)

Since $(f_n)_{n \in \mathbb{N}}$ is a Cauchy sequence, it follows that there exists a subsequence $(f_{n_k})_{k \in \mathbb{N}}$ with the property

$$\|f_{n_{k+1}} - f_{n_k}\|_p < 2^{-k}.$$

Define

$$g_m = \sum_{k=1}^m |f_{n_{k+1}} - f_{n_k}|,$$

and

$$g = \sum_{k=1}^{\infty} |f_{n_{k+1}} - f_{n_k}|.$$

We have that

$$\|g_m\|_p \leq \sum_{k=1}^m \|f_{n_{k+1}} - f_{n_k}\|_p \leq \sum_{k=1}^m 2^{-k} \leq 1.$$

Moreover, we have that g_m is an increasing sequence of functions which converges a.e. to g . The monotone convergence theorem (Theorem 10.43) then implies that

$$\int_{\Omega} g^p d\mu = \lim_{m \rightarrow \infty} \int_{\Omega} g_m^p d\mu \leq 1.$$

(Note that this requires that g_m^p is integrable, i.e., $g_m \in L_p(\Omega, \mu)$.) Moreover, this shows that g is finite μ -almost everywhere.

We conclude that the series given by $\sum_{k=1}^{\infty} (f_{n_{k+1}}(x) - f_{n_k}(x))$ converges absolutely for almost every $x \in \Omega$. That is, the functions

$$f_{n_m} = f_{n_1} + \sum_{k=1}^{m-1} (f_{n_{k+1}} - f_{n_k}),$$

(Note the telescoping trick.)

converge as $m \rightarrow \infty$ almost everywhere to

$$f := f_{n_1} + \sum_{k=1}^{\infty} (f_{n_{k+1}} - f_{n_k}).$$

Moreover, it follows that $|f|, |f_{n_m}| \leq |f_{n_1}| + g$ and hence $f, f_{n_m} \in L_p(\Omega, \mu)$ for every $m \in \mathbb{N}$.

It remains to prove that $f_{n_m} \rightarrow f$ (as $m \rightarrow \infty$) in $L_p(\Omega, \mu)$. Therefore, we use the dominated convergence

theorem (which is allowed since $f - f_{n_m}$ is dominated by $2g$) to interchange the norm and the limit, and obtain

$$\|f - f_{n_m}\|_p \leq \sum_{k=m}^{\infty} \|f_{n_{k+1}} - f_{n_k}\|_p \leq 2^{-m+1} \rightarrow 0.$$

Hence, $(f_{n_m})_{m \in \mathbb{N}}$ is a convergent sequence in $L_p(\Omega, \mu)$, and therefore a convergent subsequence of $(f_n)_{n \in \mathbb{N}}$. By Lemma 11.82, the sequence $(f_n)_{n \in \mathbb{N}}$ is also convergent, since $(f_n)_{n \in \mathbb{N}}$ is a Cauchy sequence. \square

Corollary 11.87. *The spaces $(\ell_p, \|\cdot\|_p)$ are Banach spaces for $1 \leq p \leq \infty$.*

Recall that $\|x\|_p = (\sum_{i=1}^{\infty} |x_i|^p)^{1/p}$ for $p < \infty$, and $\|x\|_{\infty} = \sup_k |x_k|$.

To get a bit of intuition what Banach spaces are about, we finish with some examples of *non-Banach spaces*. Recall that we defined a norm on a vector space to quantify a certain property (a 'size') of its elements. Positive definiteness somehow **required the space to be small enough** such that the only zero-norm element is indeed zero. The property of being complete goes in the other direction and requires the space to be large enough such that all its sequences converge. This can be phrased as " $(F, \|\cdot\|)$ is a Banach space if its norm 'fits' to it".

Let us demonstrate how this can fail with the example of continuous functions in L_p .

Example 11.88. Let us consider the space $C([-1, 1])$ of continuous functions equipped with the L_p -norm (w.r.t. the Lebesgue measure on $[0, 1]$). Since $C([-1, 1])$ is a subspace of $L_p([-1, 1])$, see Example 11.55, we obtain that $(C([-1, 1]), \|\cdot\|_p)$ is a normed space. (Here, we can really identify equivalence classes with functions, as any class contains only one continuous function. Verify all this!)

However, we will now see that $C([-1, 1])$ is 'not large enough' to be a Banach space w.r.t. $\|\cdot\|$.

For this, we consider the sequence of functions given by

$$f_n(x) := \begin{cases} 0 & \text{if } -1 \leq x < 0 \\ n \cdot x & \text{if } 0 \leq x \leq \frac{1}{n} \\ 1 & \text{else.} \end{cases}$$

It is clear that this is a sequence of continuous functions. Moreover, (f_n) is a Cauchy sequence w.r.t. $\|\cdot\|_p$ since for $n > m$ we see that

$$\|f_n - f_m\|_p^p = \int_0^{\frac{1}{m}} |f_n - f_m|^p dx \leq \frac{2^p}{m},$$

i.e., $\|f_n - f_m\|_p < \varepsilon$ for all large enough n, m , where we use that $\|f_n\|_{\infty} \leq 1$ for all $n \in \mathbb{N}$. However, $(f_n)_{n \in \mathbb{N}}$ is not a convergent sequence in $(C([-1, 1]), \|\cdot\|_p)$ since we see that the limit would be the Heaviside function which is not continuous.

Next we have a look at an example that involves derivatives.

Example 11.89. The space $C^1([-1, 1])$ of continuously differentiable functions on $[-1, 1]$ equipped with the supremum-norm $\|f\|_{\infty}$ is a normed spaces. (Why?) However, one might already guess that this is not a Banach space, because the norm does not involve derivatives. Hence, we might not expect that convergence w.r.t. this norm preserves this property.

To verify this, we are looking for a sequence of differentiable functions whose limit is not differentiable. For example, we can consider

$$f_n(x) := |x|^{1+\frac{1}{n}}.$$

First note that f_n is clearly differentiable at any point different from 0. Moreover,

$$\frac{f_n(h)}{h} = \frac{|h|^{1+\frac{1}{n}}}{h} \rightarrow 0,$$

as $h \rightarrow 0$ shows that for any $n \in \mathbb{N}$ the function f_n is continuously differentiable in 0 and so, $f_n \in C^1([-1, 1])$ for all $n \in \mathbb{N}$.

However, we see that $f_n(x) \rightarrow f(x) := |x|$ point-wise and also

$$\|f - f_n\|_\infty = \sup_{x \in [-1, 1]} \left| |x| - |x|^{1+\frac{1}{n}} \right| = \sup_{x \in I} \left| |x| \left(1 - |x|^{\frac{1}{n}} \right) \right|.$$

Since this is zero for $x = 0$, and the term $|x|^{\frac{1}{n}}$ converges to 1 for $x \neq 0$ as $n \rightarrow \infty$, we obtain that $\|f - f_n\|_\infty \rightarrow 0$. Therefore, (f_n) is convergent w.r.t. $\|\cdot\|_\infty$ and hence a Cauchy sequence w.r.t. $\|\cdot\|_\infty$. Hence, the Cauchy sequence w.r.t. $\|\cdot\|_\infty$ (f_n) converges to the function $f \notin C^1([-1, 1])$, which shows that $(C^1([-1, 1]), \|\cdot\|_\infty)$ is not a Banach space.

11.2.3 Metric spaces

Let us shortly mention that there is also a more general concept of *spaces* that have a certain structure.

Definition 11.90. Let X be a set and $d: X \times X \rightarrow \mathbb{R}$ with the properties

- 1) For any $x, y \in X$ it holds $d(x, y) \geq 0$ and $d(x, y) = 0 \iff x = y$.
- 2) $d(x, y) = d(y, x)$ holds for all $x, y \in X$.
- 3) If $x, y, z \in X$ then we have $d(x, z) \leq d(x, y) + d(y, z)$.

The properties 1), 2) and 3) are called **positive definiteness**, **symmetry** and **triangle inequality**, and the tuple (X, d) is called **metric space**.

Remark 11.91. Note that in this context a metric space does not need to be a vector space.

A normed space is also a metric space. To see this we assume that $(X, \|\cdot\|)$ is a normed space and define $d: X \times X \rightarrow \mathbb{R}$ where

$$d(x, y) = \|x - y\|.$$

This metric is called the **induced metric** of $\|\cdot\|$.

It is easy to show that conditions 1) and 2) (for a metric space) are satisfied for d . To see the triangle inequality we estimate

$$d(x, z) = \|x - z\| = \|x - y + y - z\| \leq \|x - y\| + \|y - z\| = d(x, y) + d(y, z).$$

So each normed space is also a metric space, so this is really a generalization. The other direction does not hold in general.

Let us only mention that many concepts, like continuity, can be generalized to arbitrary metric spaces, others can not. We do not discuss the details here.

11.3 Inner products and Hilbert spaces

We now turn to a somehow very special but still quite general class of vector spaces, which are called Hilbert spaces. What is special about this kind of vector spaces is that we do not only have a norm but also an inner product.

Definition 11.92. Let X be a \mathbb{R} -vector space. A mapping $\langle \cdot, \cdot \rangle: X \times X \rightarrow \mathbb{R}$ which is

- 1) **linear in both arguments**, i.e., $\langle \lambda x + \mu y, z \rangle = \lambda \langle x, z \rangle + \mu \langle y, z \rangle$ and $\langle x, \lambda y + \mu z \rangle = \lambda \langle x, y \rangle + \mu \langle x, z \rangle$,
- 2) **symmetric**, i.e., $\langle x, y \rangle = \langle y, x \rangle$, and
- 3) **positive definite**, i.e., $\langle x, x \rangle \geq 0$ and $\langle x, x \rangle = 0 \iff x = 0$,

where $x, y, z \in X$ and $\lambda, \mu \in \mathbb{R}$ are arbitrary, is called **(real) inner product**.

The tuple $(X, \langle \cdot, \cdot \rangle)$ is called **inner product space**.

Remark 11.93. In German, an inner product is called *Skalarprodukt* and inner product spaces are called *Prähilberträume*. Note that we will not use the terminology *scalar product* (in English) in order to not get confused with the scalar multiplication.

Remark 11.94. Inner products for complex vector spaces are a bit different. In this case, we let $\langle \cdot, \cdot \rangle: X \times X \rightarrow \mathbb{C}$ be an inner product if it satisfies

$$\langle \lambda x + \mu y, z \rangle = \lambda \langle x, z \rangle + \mu \langle y, z \rangle$$

and

$$\langle x, \lambda y + \mu z \rangle = \bar{\lambda} \langle x, y \rangle + \bar{\mu} \langle x, z \rangle,$$

where $\bar{\lambda}$ denotes the complex conjugate of λ . One says, the inner product is *sesquilinear*. Symmetry changes to *conjugate symmetry*

$$\langle x, y \rangle = \overline{\langle y, x \rangle}.$$

However, the third condition is the same for complex inner products. Note that, due to 2), we have $\langle x, x \rangle = \langle x, x \rangle$ and hence that $\langle x, x \rangle \in \mathbb{R}$ for any x .

Let us start with a very basic example that we have already seen earlier.

Example 11.95. \mathbb{R}^d together with $\langle x, y \rangle := \sum_{k=1}^d x_k y_k$ is an inner product space. $\langle \cdot, \cdot \rangle$ is usually called the *standard inner product* of \mathbb{R}^d . When concerned with complex vectors, the standard inner product is given by

$$\langle x, y \rangle := \sum_{k=1}^d x_k \bar{y}_k$$

for $x, y \in \mathbb{C}^d$. (Verify yourself!)

The prototype of a inner product space is the space L_2 .

Example 11.96. The function space $L_2(\Omega, \mu)$, i.e., the normed space of (complex-valued) functions on Ω with $\|f\|_2 = \sqrt{\int_{\Omega} |f|^2 d\mu} < \infty$, is an inner product space w.r.t. the inner product

$$\langle f, g \rangle_{L_2} = \int_{\Omega} f \cdot \bar{g} d\mu.$$

(Note that we sometimes write a subscript to $\langle \cdot, \cdot \rangle$ to indicate which inner product we are using.) Linearity and symmetry follow immediately from the corresponding properties of the integral, and positive definiteness is easily obtained from $\langle f, f \rangle_{L_2} = \int_{\Omega} |f|^2 d\mu = \|f\|_2^2$ and the properties of a norm.

As a direct corollary, we can again consider sequences.

Example 11.97. The space ℓ_2 with $\langle x, y \rangle = \sum_{k=1}^{\infty} x_k \bar{y}_k$ is an inner product space.

Linearity and symmetry follow again by the representation as an integral, and the corresponding properties. Positive definiteness follows again by observing that

$$\sum_{k=1}^{\infty} x_k \bar{x}_k = \sum_{k=1}^{\infty} |x_k|^2 = \|x\|_2^2$$

and the positive definiteness of the ℓ_2 -norm.

In the above examples we saw that we always had $\langle x, x \rangle = \|x\|^2$ (for some norm). This is no coincidence as we will see now. However, we first need an important inequality.

Theorem 11.98 (Cauchy-Schwarz). *Let $(X, \langle \cdot, \cdot \rangle)$ be an inner product space. Then, for all $x, y \in X$ it holds*

$$|\langle x, y \rangle|^2 \leq \langle x, x \rangle \cdot \langle y, y \rangle.$$

Note that we have proven the Cauchy-Schwarz inequality already for \mathbb{R}^d with the standard inner product. Here, we present the general result for arbitrary inner product spaces.

Proof. We observe that for any $x \in X$ we have that $\langle 0, x \rangle = \langle 0 \cdot x, x \rangle = 0 \langle 0, x \rangle = 0$. This means that for the case $y = 0$ there is nothing to show. So assume that $y \neq 0$ from here on.

Due to the properties of the inner product we have that for any $x, y \in X$ ($y \neq 0$) and $\lambda \in \mathbb{C}$ it holds

$$0 \leq \langle x - \lambda y, x - \lambda y \rangle = \langle x, x \rangle - \bar{\lambda} \langle x, y \rangle - \lambda \overline{\langle x, y \rangle} + |\lambda|^2 \langle y, y \rangle.$$

Setting $\lambda = \frac{\langle x, y \rangle}{\langle y, y \rangle}$ we obtain

$$0 \leq \langle x, x \rangle - \frac{2|\langle x, y \rangle|^2}{\langle y, y \rangle} + \frac{|\langle x, y \rangle|^2}{\langle y, y \rangle} = \langle x, x \rangle - \frac{|\langle x, y \rangle|^2}{\langle y, y \rangle}.$$

Thus

$$|\langle x, y \rangle|^2 \leq \langle x, x \rangle \langle y, y \rangle.$$

□

From this result we can directly deduce the triangle inequality for the mapping $x \mapsto \sqrt{\langle x, x \rangle}$, which then implies that it indeed defines a norm.

Theorem 11.99. *Let $(X, \langle \cdot, \cdot \rangle)$ be an inner product space.*

*Then, the mapping $\|\cdot\|: X \rightarrow \mathbb{R}$ with $\|x\| := \sqrt{\langle x, x \rangle}$ is a norm, called the **induced norm**.*

In other words, every inner product space is a normed space.

Proof. Homogeneity and positive definiteness of $\|\cdot\|$ follow from the corresponding properties of $\langle \cdot, \cdot \rangle$. For the triangle inequality we first compute

$$\begin{aligned} \|x + y\|^2 &= \langle x + y, x + y \rangle \\ &= \langle x, x \rangle + \langle x, y \rangle + \langle y, x \rangle + \langle y, y \rangle \\ &= \|x\|^2 + \langle x, y \rangle + \langle y, x \rangle + \|y\|^2 \\ &\leq \|x\|^2 + 2|\langle x, y \rangle| + \|y\|^2. \end{aligned}$$

We apply the Cauchy-Schwarz inequality to estimate $|\langle x, y \rangle| \leq \langle x, x \rangle^{\frac{1}{2}} \langle y, y \rangle^{\frac{1}{2}} = \|x\| \cdot \|y\|$. Hence

$$\|x + y\|^2 \leq \|x\|^2 + 2\|x\| \cdot \|y\| + \|y\|^2 = (\|x\| + \|y\|)^2.$$

□

This result allows to reformulate the Cauchy-Schwarz inequality by using the induced norm.

Corollary 11.100. *Let $(X, \langle \cdot, \cdot \rangle)$ be an inner product space. Then for all $x, y \in X$ we have*

$$\langle x, y \rangle \leq \|x\| \cdot \|y\|,$$

where $\|\cdot\| = \sqrt{\langle x, x \rangle}$ is the induced norm.

Hence, we can employ all the concepts already introduced for normed spaces. Most importantly, we can 'naturally' define convergence of sequences in an inner product space by means of its induced norm. In particular, this leads to the notion of a **complete inner product space**, which are called *Hilbert spaces*, named after *David Hilbert* (1862–1943).

Recall that a normed space is complete, i.e., a Banach space, if all Cauchy sequences in it converge (w.r.t. its norm).

Definition 11.101 (Hilbert spaces). Let $(H, \langle \cdot, \cdot \rangle)$ be an inner product space.

If $(H, \|\cdot\|)$ is a Banach space, where $\|f\| := \sqrt{\langle f, f \rangle}$ is the induced norm, then we call $(H, \|\cdot\|)$ (or just H) a **Hilbert space**.

(H is the typical letter to denote Hilbert spaces.)

First, recall that all Cauchy sequences in \mathbb{R}^d are convergent, independent of the norm. Hence, equipping it with the standard inner product, and the corresponding induced norm, we see that this is a Hilbert space. Actually, Hilbert spaces were precisely invented to generalize these spaces to an infinite-dimensional setting.

Example 11.102. Let us have a look at \mathbb{R}^d together with the standard inner product

$$\langle x, y \rangle = \sum_{k=1}^d x_k y_k.$$

The induced norm is given by the Euclidean norm in this case, i.e.,

$$\|x\|_2 = \sqrt{\sum_{k=1}^d x_k^2},$$

and we have already showed that this defines a Banach space.

The next example is the one originally introduced by Hilbert.

Example 11.103. Consider the space ℓ_2 with the inner product

$$\langle x, y \rangle = \sum_{k=1}^{\infty} x_k \bar{y}_k.$$

We know from Corollary 11.87 that ℓ_2 is a Banach space w.r.t. the norm

$$\|x\| = \sqrt{\sum_{k=1}^{\infty} |x_k|^2}.$$

Observing that this is exactly the induced norm we obtain the result.

Considering general measures leads to the general concept of **Hilbert spaces of functions**.

Example 11.104. We have a look at the space $L_2(\Omega, \mu)$, where $(\Omega, \mathcal{A}, \mu)$ is an arbitrary measure space. We already know that

$$\langle f, g \rangle = \int_{\Omega} f \bar{g} d\mu$$

is an inner product. The induced norm therefore is given by

$$\|f\| = \sqrt{\int_{\Omega} |f|^2 d\mu}.$$

It was already shown in Theorem 11.86 that this is a Banach space w.r.t. this norm. Hence, $L_2(\Omega, \mu)$ is a **Hilbert space**.

Remark 11.105. As discussed earlier, the spaces $L_p(\Omega, \mu)$ are defined for arbitrary measure space $(\Omega, \mathcal{A}, \mu)$. Note that we omit the σ -algebra \mathcal{A} in this notation, but it still needs to be defined. However, note that \mathcal{A} is often clear from the context.

Remark 11.106. Let us already mention here that L_2 is the only Hilbert space among the L_p -spaces. That is, one cannot define an inner product that makes L_p with $p \neq 2$ complete.

The main advantage of Hilbert spaces is that we are able to define **what it means to be orthogonal**, in the same way as we did for vectors.

Definition 11.107. Let $(H, \langle \cdot, \cdot \rangle)$ be an inner product space.

We say that $x, y \in H$ are **orthogonal** if

$$\langle x, y \rangle = 0.$$

If x and y additionally satisfy $\langle x, x \rangle = \langle y, y \rangle = 1$, then we call them **orthonormal**.

Moreover, we say a set $A \subset H$ is **orthogonal/orthonormal**, if arbitrary $x, y \in H$ with $x \neq y$ are orthogonal/orthonormal.

For a set $A \subset H$ we define the **orthogonal complement** of A by

$$A^{\perp} := \left\{ x \in H : \langle x, y \rangle = 0 \text{ for all } y \in A \right\}.$$

(Some authors write $x \perp y$ to say that $\langle x, y \rangle = 0$.)

Let us see a example.

Example 11.108. We consider the space \mathbb{R}^3 with the standard inner product. Having a look at the vector

$$e_1 = (1, 0, 0),$$

we see that

$$\langle e_1, x \rangle = x_1$$

for $x = (x_1, x_2, x_3) \in \mathbb{R}^3$. Thus, e_1 is orthogonal to any vector whose first component is zero. Thus

$$\{e_1\}^{\perp} = \left\{ x \in \mathbb{R}^3 : x_1 = 0 \right\}.$$

The vector $e_2 = (0, 1, 0)$ is orthogonal to all vectors which have a zero in their second component, and we obtain

$$\{e_1, e_2\}^{\perp} = \left\{ x \in \mathbb{R}^3 : x_1 = x_2 = 0 \right\}.$$

Moreover, note that $\{e_1, e_2\}$ is orthonormal.

Example 11.109 (Complex trigonometric functions). Let us recall that from the section about Fourier analysis we know that

$$\int_0^1 e^{2\pi i kx} dx = 0$$

whenever $k \in \mathbb{Z} \setminus \{0\}$. Since $\overline{e^{2\pi i xk}} = e^{-2\pi i xk}$ (this follows immediately by Euler's identity $e^{ix} = \cos x + i \sin x$) we obtain that

$$\int_0^1 e^{2\pi i n x} \overline{e^{2\pi i m x}} dx = \int_0^1 e^{2\pi i (n-m)x} dx = \begin{cases} 1 & \text{if } m = n \\ 0 & \text{else.} \end{cases}$$

Therefore, denoting $e_k(x) := e^{2\pi i kx}$, we obtain that the set $\{e_k : k \in \mathbb{Z}\}$ is orthonormal w.r.t. the inner product

$$\langle f, g \rangle = \int_{[0,1]} f(x) \overline{g(x)} dx.$$

We will now see that **orthonormal sets can take the role of a basis** as considered for finite-dimensional spaces. Hence, we can, in a rather general context, write a function as a combination of other (simpler) elements of the space.

However, we also need that the orthonormal set is rich enough to 'reach' every element. This leads to the concept of an *orthonormal bases*.

Definition 11.110. Let H be a Hilbert space.

A countable orthonormal set $\{e_1, e_2, e_3, \dots\}$ is called **orthonormal basis (ONB) of H** if each element in $x \in H$ can be written as

$$x = \sum_{k=1}^{\infty} \alpha_k e_k$$

for some $\alpha_k \in \mathbb{R}$ (or \mathbb{C}). This means that, for some sequence $(\alpha_k)_{k \in \mathbb{N}}$, we have

$$\lim_{n \rightarrow \infty} \left\| \sum_{k=1}^n \alpha_k e_k - x \right\| = 0.$$

A very powerful result is the following, which we state without proof.

Theorem 11.111. Let H be a Hilbert space and $\{e_1, e_2, e_3, \dots\}$ be an orthonormal set.

Then, the following statements are equivalent:

- 1) If for $x \in H$ we have that $\langle x, e_k \rangle = 0$ for all $k \in \mathbb{N}$, then $x = 0$.
- 2) The set $\{e_1, e_2, e_3, \dots\}$ is an orthonormal basis of H .

3) For any $x \in H$ we have $x = \sum_{k=1}^{\infty} \langle x, e_k \rangle e_k$.

4) For any $x, y \in H$ we have $\langle x, y \rangle = \sum_{k=1}^{\infty} \langle x, e_k \rangle \langle e_k, y \rangle$.

5) For any $x \in H$ we have $\|x\|^2 = \sum_{k=1}^{\infty} |\langle x, e_k \rangle|^2$

Remark 11.112. If H is a finite-dimensional Hilbert space then the above theorem is still true. We then just set $e_k = 0$ for all $k > \dim(H)$, and all series become finite sums.

We have a look at some examples.

Example 11.113. Consider $H = \mathbb{R}^d$ together with the standard inner product. If we now consider the set of unit vectors $\mathcal{B} := \{e_1, \dots, e_d\}$, then, clearly, property 1 above is fulfilled and we obtain that \mathcal{B} is an orthonormal basis. We actually knew already before that every vector $x = (x_1, x_2, \dots, x_d)$ has the representation

$$x = \sum_{k=1}^d x_k e_k.$$

With this we obtain (even without Theorem 11.111) that

$$x = \sum_{k=1}^d x_k e_k = \sum_{k=1}^d \langle x_k, e_k \rangle e_k$$

and

$$\|x\|_2^2 = \sum_{k=1}^d |\langle x, e_k \rangle|^2.$$

Although these statements are rather obvious, note that they would also hold with \mathcal{B} replaced by any other orthonormal basis of \mathbb{R}^d .

For example, if we consider the case $d = 2$, it is not hard to verify that $\{v_1, v_2\}$ with $v_1 = (\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}})$ and $v_2 = (\frac{1}{\sqrt{2}}, -\frac{1}{\sqrt{2}})$ is also an orthonormal basis of \mathbb{R}^2 . We obtain, in particular, that $\|x\|_2^2 = \sum_{k=1}^2 |\langle x, v_k \rangle|^2$. Just plugging in the definition, this equality shows that

$$\|x\|_2^2 = x_1^2 + x_2^2 = \sum_{k=1}^2 |\langle x, v_k \rangle|^2 = \frac{(x_1 + x_2)^2 + (x_1 - x_2)^2}{2}$$

for $x = (x_1, x_2) \in \mathbb{R}^2$.

Next we turn to some infinite-dimensional spaces.

Example 11.114. Consider the Hilbert space ℓ_2 , and consider the unit vectors e_1, e_2, e_3, \dots , i.e., e_k contains a 1 at position k and zeros on all other positions. Clearly, $\langle e_i, e_j \rangle = 0$ if $i \neq j$ and $\|e_k\|^2 = \langle e_k, e_k \rangle = 1$ for any $k \in \mathbb{N}$. Again, we note that

$$\langle x, e_i \rangle = x_i = 0 \iff x_i = 0$$

for any $x = (x_1, x_2, x_3, \dots) \in \ell_2$. So, if $\langle x, e_k \rangle = 0$ for any $k \in \mathbb{N}$, then we have that $x = 0$. Theorem 11.111 now implies that $\{e_1, e_2, e_3, \dots\}$ is an orthonormal basis of ℓ_2 , with all the other properties stated there.

Example 11.115. Let us consider $L_2([0, 1])$ and the set of complex-valued functions $\{e^{2\pi i k \cdot}\}_{k \in \mathbb{Z}}$. Denoting $e_k(x) := e^{2\pi i k x}$, we proved in Lemma 7.26 that if

$$\hat{f}(k) := \langle f, e_k \rangle = \int_0^1 f(x) e^{-2\pi i k x} dx = 0$$

for all $k \in \mathbb{Z}$, then it follows that $f = 0$. (We stated the result only for continuous functions, but it also holds for $f \in L_2$. Note that we did not prove this important result.)

So $\{e^{2\pi i k \cdot}\}_{k \in \mathbb{Z}}$ is an ONB of $L_2([0, 1])$. In particular, we obtain that

$$f = \sum_{k \in \mathbb{Z}} \langle f, e_k \rangle e_k = \sum_{k \in \mathbb{Z}} \hat{f}(k) e^{2\pi i k \cdot}$$

for every $f \in L_2([0, 1])$. Note that this equality (which is actually a convergence statement) only holds “in the L^2 -sense”, i.e., almost everywhere. This should be compared to the (point-wise) convergence of Fourier series we discussed in Section 7.

11.3.1 Reproducing kernel Hilbert spaces

Let us finally discuss a special kind of Hilbert spaces which are, in contrast to general Hilbert spaces, **compatible with function values**. This means that these Hilbert spaces, which are called *reproducing kernel Hilbert spaces*, consists of functions whose function values are well-defined at every point in the domain. This is particularly useful as we can use the powerful techniques we have available for Hilbert spaces, for function spaces that really consists of functions (rather than equivalence classes).

Note that most Hilbert spaces that appear in applications are actually of this type, and that these spaces were originally introduced by *Stanisław Zaremba* (1863–1942) in 1907 in an attempt to solve certain *boundary value problems*.

Definition 11.116. Let H be a Hilbert space of functions on some set Ω , i.e., any $f \in H$ has the form $f: \Omega \rightarrow \mathbb{C}$.

If there exists some function $k: \Omega \times \Omega \rightarrow \mathbb{C}$ with the properties

- 1) for any $x \in \Omega$ we have that $k(x, \cdot) \in H$, and
- 2) for any $x \in \Omega$ and any $f \in H$ we have that $f(x) = \langle f, k(x, \cdot) \rangle$,

then we call k a **reproducing kernel** and the space H is called a **reproducing kernel Hilbert space (RKHS)** on Ω .

(Property 2 is sometimes called the **reproducing property**.)

In particular, this implies that $f(x)$ is well-defined for any $f \in H$ and $x \in \Omega$. Applying the Cauchy-Schwarz inequality to the second property, we can even make a more quantitative statement. If we define $C_x := \|k(x, \cdot)\|_H < \infty$ for $x \in \Omega$, i.e., the norm of the function $k(x, \cdot)$ in H , then we obtain

$$|f(x)| = |\langle f, k(x, \cdot) \rangle| \leq C_x \|f\|_H.$$

(Here, we use the notation $\|\cdot\|_H$ for the norm that is induced by the inner product in H .)

This means that for functions from a reproducing kernel Hilbert space H , we have that **a small norm in H implies small function values**. Note that such a statement is not true for general Hilbert spaces, like $L_2(\Omega)$. To see this, note that we can easily define functions with arbitrary small norm but constant function value at a given point. (For example?)

Following these lines, we can prove the powerful result that convergence in a RKHS also implies point-wise convergence at every point.

Lemma 11.117. Let H be a RKHS on Ω with kernel k .

If a sequence $(f_n)_{n \in \mathbb{N}}$ converges in H to some f , that is

$$\lim_{n \rightarrow \infty} \|f_n - f\| = 0,$$

then $f_n(x) \rightarrow f(x)$ as $n \rightarrow \infty$ for every $x \in \Omega$.

Note that we do not claim uniform convergence, and this would be false in general. Moreover, one might even prove that the property in the last lemma is indeed equivalent to the existence of a reproducing kernel. We omit the details.

Proof. Note that

$$|f_n(x) - f(x)| = |\langle f_n - f, k(x, \cdot) \rangle| \leq \|k(x, \cdot)\| \cdot \|f_n - f\|,$$

which implies the result. □

Based on this, we can give a useful representation of the reproducing kernel based on an orthonormal basis in the Hilbert space.

Theorem 11.118. *Let H be a RKHS on Ω with kernel k .*

If $\{e_1, e_2, \dots\}$ is an orthonormal basis in H , then we have that

$$k(x, y) = \sum_{i=1}^{\infty} \overline{e_i(x)} \cdot e_i(y),$$

where the series converges point-wise for every $x, y \in \Omega$.

Remark 11.119. This result is also of huge practical interest as it allows to prove bounds on the *error* that appears if a kernel is approximated by a finite sum of functions (similar as for Fourier or Taylor series). That is, it is a typical (numerical) approach to work with the *truncated kernel* $k_n(x, y) := \sum_{i=1}^n e_i(x)e_i(y)$ and use later that k_n is close to k , i.e., that $\|k - k_n\|$ is small. It is worth noting that $\langle f, k_n(x, \cdot) \rangle$ is then an approximation of the function value $f(x) = \langle f, k(x, \cdot) \rangle$. Actually, one can prove that

$$\langle f, k_n(x, \cdot) \rangle = \sum_{i=1}^n \langle f, e_i \rangle \cdot e_i(x).$$

Proof. Due to Theorem 11.111 and the fact that $k(x, \cdot) \in H$, we obtain

$$k(x, y) = \sum_{i=1}^{\infty} \langle k(x, \cdot), e_i \rangle \cdot e_i(y),$$

with convergence of the series in H . (We used the representation as series for $k(x, y)$, considered as function in y .) By Lemma 11.117, we obtain that the convergence is also point-wise, which implies that the above equality holds for every $x, y \in \Omega$. The reproducing property finally implies that $\langle k(x, \cdot), e_i \rangle = \langle e_i, k(x, \cdot) \rangle = \overline{e_i(x)}$ for all $x \in \Omega$, which proves the result. \square

Let us now discuss the most important properties of a reproducing kernel. In particular, we obtain that reproducing kernels have certain properties and, moreover, that every function with these properties is the reproducing kernel of some RKHS. This turns out to be a nice characterization of reproducing kernels.

Lemma 11.120. *Let H be a RKHS on Ω with kernel k . Then for any $x, y \in \Omega$*

- 1) $k(x, y) = \langle k(x, \cdot), k(y, \cdot) \rangle$
- 2) k is conjugate symmetric, i.e., $k(x, y) = \overline{k(y, x)}$
- 3) k is unique, i.e., there is not other reproducing kernel in H .

Proof. The first point immediately follows from the definition of k , and the second point from the first and the symmetry of the inner product.

To show that there is no other reproducing kernel we observe that if there would be another such kernel k_1 we would have that

$$k(x, y) = \langle k(\cdot, y), k_1(\cdot, x) \rangle = \overline{\langle k_1(\cdot, x), k(\cdot, y) \rangle} = \overline{k_1(y, x)}.$$

However, due to the second part of the proof, k_1 is conjugate symmetric and so

$$k(x, y) = k_1(x, y).$$

\square

Another interesting property is that a reproducing kernel is a positive definite mapping. Recall that we said that a real matrix $A = (a_{ij}) \in \mathbb{R}^{n \times n}$ is positive semi-definite iff for every $c = (c_1, \dots, c_n)^T \in \mathbb{R}^n$ we have

$$c^T A c = \sum_{i,j=1}^n c_i c_j a_{ij} \geq 0.$$

Here, we assume the same for the reproducing kernel, when evaluated at arbitrary n points.

Definition 11.121. Let Ω be a non-empty set.

We say that a symmetric function $k: \Omega \times \Omega \rightarrow \mathbb{R}$ is **positive definite** if the matrix $K := (k(x_i, x_j))_{i,j=1}^n \in \mathbb{R}^{n \times n}$ is positive semi-definite for any $n \in \mathbb{N}$ and any choice $x_1, \dots, x_n \in \Omega$.

That is, if we have for any $c_1, \dots, c_n \in \mathbb{R}$ that

$$\sum_{i=1}^n \sum_{j=1}^n c_i c_j k(x_i, x_j) \geq 0.$$

Remark 11.122. Note that, in analogy to the matrix notation, it would be better to call a kernel with the above property *positive semi-definite*, and some authors prefer to do so. However, in the context of kernels, this distinction is in general not necessary, and would just complicate the definition. (Note that K above is not positive definite if $x_i = x_j$ for some $i \neq j$.)

The following result shows that every reproducing kernel is positive definite. For simplicity, we consider only real-valued kernels here.

Lemma 11.123. Let H be a RKHS on Ω with kernel k . Then, k is positive definite.

Proof. We already know that k is symmetric. Now, let $c_1, \dots, c_n \in \mathbb{R}$ and $x_1, \dots, x_n \in \Omega$ be arbitrary. We compute that

$$\begin{aligned} \sum_{i=1}^n \sum_{j=1}^n c_i c_j k(x_i, x_j) &= \sum_{i=1}^n \sum_{j=1}^n c_i c_j \langle k(x_i, \cdot), k(x_j, \cdot) \rangle = \sum_{i=1}^n \sum_{j=1}^n \langle c_i k(x_i, \cdot), c_j k(x_j, \cdot) \rangle \\ &= \left\langle \sum_{i=1}^n c_i k(x_i, \cdot), \sum_{j=1}^n c_j k(x_j, \cdot) \right\rangle \\ &= \left\| \sum_{i=1}^n c_i k(x_i, \cdot) \right\|^2 \geq 0. \end{aligned}$$

□

Remarkably, it is true that every positive definite kernel corresponds to some RKHS. This result was first published by *Nachman Aronszajn* (1907–1980) around 1950, but he himself attributed it to *Eliakim H. Moore* (1862–1932). Note that Aronszajn, together with *Stefan Bergmann* (1895–1977), carried out a systematic study on RKHS only in the early 1950s, although these spaces were known for almost 50 years. The fascinating properties of these spaces were not realized at this time, because the important applications from *machine learning*, like *empirical risk minimization*, were just not known.

Theorem 11.124 (Moore-Aronszajn). Let Ω be a non-empty set and let $k: \Omega \times \Omega \rightarrow \mathbb{R}$ be a positive definite function on Ω .

Then, there exists a unique Hilbert space H such that k is its reproducing kernel.

The proof is quite technical so we will omit a detailed proof, however, we will state the main ideas as they give information about what H 'looks like'.

Sketch of proof. The main idea is to consider the vector space of functions which have the following form

$$f(x) = \sum_{i=1}^n \alpha_i k(y_i, x),$$

for some points $y_1, \dots, y_n \in \Omega$ and $\alpha_1, \dots, \alpha_n \in \mathbb{R}$. One can define the mapping

$$\left\langle \sum_{i=1}^n \alpha_i k(y_i, \cdot), \sum_{j=1}^m \beta_j k(z_j, \cdot) \right\rangle := \sum_{i=1}^n \sum_{j=1}^m \alpha_i \beta_j k(y_i, z_j).$$

Using the positive definiteness of k , one can show that this defines an inner product.

To obtain the (unique) Hilbert space for the given kernel, we now need to consider arbitrary limits of sequences of such functions.

□

Let us finally note that one can equip almost arbitrary sets (of data) with a Hilbert space structure. This procedure is often called **kernel trick**, and is one of the most fundamental tool when working in machine learning, especially with *artificial neural networks*. The basic idea behind this is, that kernels can be used as kind of a **measure of similarity**: The larger $k(x, y)$ the more 'similar' are x and y . To find similarities in large data sets, which is (one of) the main subject of *artificial intelligence*, it is therefore of interest to consider such 'similarity measures', as they also come with many (theoretical) advances. Clearly, all this is subject of its own lecture and we can only state the basic idea here.

Example 11.125 (Feature maps). Given an arbitrary set Ω (i.e., the data) and a mapping $\varphi: \Omega \rightarrow H$, where H is some Hilbert space. The mapping φ is called **feature map** and must be predefined to apply the "kernel trick". (Finding appropriate feature maps is its own area of research, but note that there exist whole libraries of feature maps for common applications.) We define the function $k: \Omega \times \Omega \rightarrow \mathbb{R}$ as

$$k(x, y) := \langle \varphi(x), \varphi(y) \rangle_H.$$

First observe that for every $n \in \mathbb{N}$, $a_1, \dots, a_n \in \mathbb{R}$ and $x_1, \dots, x_n \in \Omega$ we have

$$\begin{aligned} \sum_{i=1}^n \sum_{j=1}^n a_i a_j k(x_i, x_j) &= \sum_{i=1}^n \sum_{j=1}^n \langle a_i \varphi(x_i), a_j \varphi(x_j) \rangle_H = \left\langle \sum_{i=1}^n a_i \varphi(x_i), \sum_{j=1}^n a_j \varphi(x_j) \right\rangle_H \\ &= \left\| \sum_{i=1}^n a_i \varphi(x_i) \right\|_H^2 \geq 0. \end{aligned}$$

Hence, k is a positive definite kernel and therefore the reproducing kernel of some RKHS, see Theorem 11.124.

This shows that **feature maps lead to some RKHS** of functions on Ω . However, note that we do not 'know' the Hilbert space (or its inner product) in general, and this is also not necessary. Although it is fundamental that we construct a Hilbert space here to employ some of the deep theory introduced earlier, it turns out that, in practical applications, it is completely enough to know the kernel at specific points (i.e., the data) to find good approximations. This is the main insight for the success of reproducing kernel Hilbert spaces in a machine learning context.