Foundations of Higher Mathematics

Winter Semester 2025/2026

University of Bayreuth

Lars Grüne and Michael Stoll

Contents

Version of March 5, 2026, 8:44.

1. Introduction

Date:
March 5, 2026

These are the lecture notes for the course *Foundations of Higher Mathematics*. This is the basic introductory mathematics module for the study course *Data, Science, and AI*. Its intended outcome is to enable you to work with concepts of "higher" mathematics, in particular from Linear Algebra and Analysis. One important aspect of this is that you are supposed to learn how to carry out mathematical *proofs* of statements from these subjects. More generally, the goal is that you *understand* what is going on; this goes quite a bit beyond the ability to *reproduce* definitions and theorems. The point is that you are able to *combine* them in suitable ways to solve *new problems*. To get there, you need *practice*, which you are supposed to acquire by working on the weekly Problem Sheets. Don't take the easy way out when you think you are stuck at a problem (i.e., copy a solution from somebody else or an AI)! Rather, find a group of fellow students and work together, explaining your thoughts to one another. One good venue for doing this is the *Maths Support Centre*, located in the seminar room S 79 (building NW II) and open every afternoon Monday to Friday.

The structure of this course is slightly unusual (for German standards, at least), in that it combines material from both Linear Algebra and Analysis. (The usual set-up is to have two separate courses that one takes in parallel during the first two semesters.) After a first part that introduces basic notions from logic and set theory, some algebraic structures like groups and fields, and the real and complex numbers, the second part teaches you about sequences and series of real numbers. Then we switch to Linear Algebra and introduce vector spaces, linear subspaces, the notions of linear independence, basis, and dimension, and we discuss linear maps. Following this, we return to Analysis and study functions of one real variable and their properties like continuity and differentiability. Finally, in the last part of the course, we return to Linear Algebra to introduce matrices and to discuss how to use them for computations, for example to solve linear systems of equations.

We hope that this way of doing things will work well for you and for us. Feedback is always welcome!

Your lecturers: Lars Grüne and Michael Stoll

2. The language of mathematics

Date:
March 5, 2026

What is mathematics all about?

Wikipedia (← text in this color is a *link;* you can click on it!) says (as of October 2025):

> *Mathematics* involves the description and manipulation of abstract objects that consist of either abstractions from nature or—in modern mathematics—purely abstract entities that are stipulated to have certain properties, called axioms. Mathematics uses pure reason to prove properties of objects, a *proof* consisting of a succession of applications of deductive rules to already established results. These results include previously proved theorems, axioms, and—in case of abstraction from nature—some basic properties that are considered true starting points of the theory under consideration.

So one important part is *abstraction*. This means that one aims at determining the important common features of various different situations, and then one tries to deduce as many statements as possible based on these features that will then hold in *all* situations that exhibit these features. This is done via the key component of all mathematical activity, which is carrying out a *mathematical proof*.

- The most important learning goal in this basic course on higher mathematics is to be able to *produce mathematical proofs*.

Of course, you are also supposed to learn about results and methods of Linear Algebra and Analysis, but without the essential proving skills, they would be of limited use.

Before we can begin, we need to learn the vocabulary and grammar of the language of mathematics. Mathematical statements and proofs are formulated in the language of *logic*; the objects being discussed are described in the language of *set theory*. We will introduce both here (or review them, depending on how much you already know). These are the "tools of the trade" that you will be dealing with every day, so pay close attention!

2.A Propositional logic.

A *proposition* is a (mathematical) statement for which it makes sense to say that it is true or false.

DEF
proposition

The word "proposition" is frequently also used to denote a mathematical theorem of minor importance compared to the main results.

Propositional logic deals with propositions and how they can be combined, and in particular how the truth or falsehood of the combined proposition depends on the truth or falsehood of their components.

We give a definition of the common combinations below. (The color green marks definitions in these notes.)

2.1. **Definition.** In the following, A and B stand for arbitrary propositions.

DEF
$\neg A$
$A \wedge B$
$A \vee B$
$A \Rightarrow B$
$A \Leftrightarrow B$

(1) *Negation*: we write "not A" or "$\neg A$" for the negation of the proposition A. $\neg A$ is true if and only if A is false, and conversely.

(2) *Conjunction*: we write "A and B" or "$A \wedge B$" for the conjunction of A and B. This proposition is true if and only if both A and B are true.

(3) *Disjunction*: we write "A or B" or "$A \vee B$" for the disjunction of A and B. This proposition is true if and only if at least one of A and B is true. (Note that this includes the case that both A and B are true.)

(4) *Implication*: we write "B follows from A", "A implies B" or "$A \Rightarrow B$". This proposition is true if and only if A is false or B is true (or both).

(5) *Equivalence*: we write "A if and only if B" (also shortened to "A iff B"), "A and B are equivalent" or "$A \Leftrightarrow B$". This proposition is true if and only if either A and B are both true or A and B are both false. ♢

All notations mentioned above are possible and allowed; writing "$A \wedge B$" is not better or worse than writing "A and B" (only slightly shorter). When writing mathematical text, you should not try to put your arguments in the most compact form possible, rather you should strive to make it easy for the reader to follow them!

When combinations are nested, we use parentheses to make clear what we mean: "A and B or C" could mean either "$(A \wedge B) \vee C$" or "$A \wedge (B \vee C)$", and these are not equivalent.

The most important, and at the same time hardest to understand, of these combinations is the implication. It is important, because most mathematical theorems have the form of an implication: if some assumptions A hold, then we can conclude that some statement B is true. It is a bit difficult, because people are dealing with it frequently in an imprecise or even wrong way in daily life. In particular, the distinction between "B follows from A" and "A follows from B" is often blurred. To understand how implications work is your first important task. Here are the most important facts.

- $A \Rightarrow B$ is certainly always true *when* A *is false*.
- $A \Rightarrow B$ is also always true when B is true.
- $A \Rightarrow B$ can *only* be false, if A is true, but B is false.

We can also express that $A \Rightarrow B$ is true by saying that "A is sufficient for B" or "B is necessary for A".

We will sometimes use the notation "$\bot$" for the *falsum*, the proposition that is always false. Analogously, there is "$\top$", the always true proposition. Using this, we can write

$$\bot \Rightarrow B \qquad \text{and} \qquad A \Rightarrow \top \qquad \text{are always true.}$$

(The first one of these is known by its Latin name as *ex falso quodlibet*, also known as the *principle of explosion*.)

The following is *not* a legal way of concluding.

> We want to show A. So we assume that A is true. Then it follows that B must be true. But we know that B is true, so A must also be true.

(The color blue marks examples.)

2.2. **Example.** Let's show $0 = 1$. We do this by manipulating the equality. Duplicating both sides, we obtain $0 = 2$, and then subtracting 1 from both sides gives $-1 = 1$. Now we square both sides, which gives $1 = 1$, which is clearly true. Therefore our original equation $0 = 1$ must also be true. In symbols:

EXAMPLE wrong argument

$$0 = 1 \implies 0 = 2 \implies -1 = 1 \implies 1 = 1 \checkmark$$

Everything here is correct, except the last sentence: it is not possible to conclude A from $A \Rightarrow B$ and B. The mistake is that we need the *reverse* implications to conclude:

$$0 = 1 \impliedby 0 = 2 \impliedby -1 = 1 \overset{\lightning}{\impliedby} 1 = 1 \checkmark$$

but the last of these does not hold (the first two are equivalences).

The rule is that the implication arrows always need to point *to the desired conclusion.* ♣

On the contrary, it is definitely possible to conclude B from $A \Rightarrow B$ and A. This is one of the basic logical rules used in mathematical proofs. In many cases "$A \Rightarrow B$" is some mathematical theorem that we want to apply. We show that its assumption A holds, and then we can conclude that B must also be true.

Carrying out a mathematical proof can be seen as playing a game. At each step, we have one or several "proof states" that contain a "goal" statement to be proved together with the assumptions that are currently valid. There are deduction rules that tell us how to change these proof states. One important rule allows us to remove a proof state when the goal is among the assumptions; this means that its goal statement has been proved. We begin with one proof state that only contains the goal we want to prove, and we win the game when we reach the position with zero proof states (then nothing is left to prove).

A general rule is that we can add any proposition that has already been proved to the assumptions. This also means that we can add additional assumptions if we are able to prove them from the assumptions that we already have. (We can also always remove assumptions to keep things nice and tidy, but this is less important).

Each type of combination of propositions comes with two rules. One of them (the *elimination rule*) tells us how to make use of an assumption that has this form. The other one (the *introduction rule*) tells us how to obtain a proposition of this form from others.

DEF elimination rule

DEF introduction rule

The rule mentioned above (we can conclude B from $A \Rightarrow B$ and A, also called *modus ponens*) is the elimination rule of **implication**; it tells us how to use the implication $A \Rightarrow B$. The other elimination rules are as follows.

- **Conjunction:** given $A \wedge B$, we obtain A and B.
- **Disjunction:** if we know $A \vee B$, then we can make a case distinction. This means that we replace a proof state that has $A \vee B$ among its assumptions by *two* proof states, in which we replace $A \vee B$ by A and by B, respectively.
- **Falsum:** if $\bot$ is among the assumptions, then we can add any proposition (in particular the goal, so we can finish the proof immediately). This is because the implication $\bot \Rightarrow A$ is always true, so this is in fact a special case of the elimination rule of implication.
- **Negation:** $\neg A$ is equivalent to $A \Rightarrow \bot$. Again using the elimination rule of implication, this means that if we have both A and $\neg A$ as assumptions, then we can add $\bot$ (and thus finish the proof).

- **Equivalence:** $A \Leftrightarrow B$ is equivalent to $(A \Rightarrow B) \wedge (B \Rightarrow A)$, so, given $A \Leftrightarrow B$, we obtain $A \Rightarrow B$ and $B \Rightarrow A$.

The introduction rules allow us to add combinations of propositions to our assumptions. They are as follows.

- **Conjunction:** given A and B, we obtain $A \wedge B$.
  To prove a goal of the form $A \wedge B$, we can split the proof into two parts, one where we replace the goal by A and one where we replace the goal by B.
- **Disjunction:** given A, we obtain $A \vee B$; similarly, given B, we obtain $A \vee B$.
  To prove a goal of the form $A \vee B$, we can replace it by A or we can replace it by B. (We may have to be careful here to not take a wrong turn!)
- **Implication:** to prove $A \Rightarrow B$, we replace the goal by B and add A to the assumptions (we prove B under the assumption that A holds).
- **Equivalence:** similarly as for the elimination rule, we obtain $A \Leftrightarrow B$ from the two implications $A \Rightarrow B$ and $B \Rightarrow A$.
  So to prove a goal of the form $A \Leftrightarrow B$, we instead prove the two implications ("directions") separately. This is a proof step that is used quite frequently.
- **Negation:** since $\neg A$ is equivalent to $A \Rightarrow \bot$, this is a special case of the introduction rule of implication. To prove $\neg A$, we add A to the assumptions and replace the goal with $\bot$. This means that we want to derive a contradiction from the assumption that A is true.

We can combine these to obtain further rules. For example, since $A \vee \neg A$ is always true, we are free to add a statement of this form to our assumptions at any time. If we follow this by the elimination rule of disjunction, we obtain the *proof by cases*: To prove some statement, it is sufficient to, on the one hand, prove it under the assumption that A is true, and, on the other hand, prove it under the assumption that $\neg A$ is true (i.e., A is false). (The statement that $A \vee \neg A$ is always true is known as the *law of excluded middle*—there are no other possibilities than A is true or A is false.)

A special case of this is the *proof by contradiction*. This rule says that, in order to prove A, it is sufficient to show that the assumption of $\neg A$ leads to a contradiction (which is actually a proof of $\neg\neg A$). To see that this rule follows from the rule of excluded middle, note that we can split the proof of A into a proof that A is true assuming A (which is clear) and a proof that A is true assuming $\neg A$. But if we can deduce a contradiction from $\neg A$, then we can prove anything, including A.

EXAMPLES
proofs

2.3. **Examples.** We give some examples of proofs using the rules above. They are written as text; we add in parentheses which rule is used at each step, in the form "elim($\wedge$)" for the elimination rule of conjunction or "intro($\Rightarrow$)" for the introduction rule of implication.

(1) We show that $((A \Rightarrow B) \wedge (B \Rightarrow C)) \Rightarrow (A \Rightarrow C)$ holds for arbitrary propositions A, B and C (if B follows from A and C follows from B, then C follows from A).

We assume $(A \Rightarrow B) \wedge (B \Rightarrow C)$ and have to show $A \Rightarrow C$ (intro($\Rightarrow$)). We assume in addition A and have to show C (intro($\Rightarrow$)). From $(A \Rightarrow B) \wedge (B \Rightarrow C)$ we deduce $A \Rightarrow B$ and $B \Rightarrow C$ (elim($\wedge$)). Now from A and $A \Rightarrow B$, we obtain B (elim($\Rightarrow$)), and from B and $B \Rightarrow C$, we obtain C (elim($\Rightarrow$)), so C is proved.

(2) The conjunction $A \wedge B$ is false if and only if at least one of A and B is false: $\neg(A \wedge B) \Leftrightarrow (\neg A \vee \neg B)$.

We split the proof into a proof of $\neg(A \wedge B) \Rightarrow (\neg A \vee \neg B)$ and a proof of $(\neg A \vee \neg B) \Rightarrow \neg(A \wedge B)$ (intro($\Leftrightarrow$)).

To prove $\neg(A \wedge B) \Rightarrow (\neg A \vee \neg B)$, we assume $\neg(A \wedge B)$ and have to show $\neg A \vee \neg B$ (intro($\Rightarrow$)). Here we need to proceed by cases. We first assume $\neg A$. Then $\neg A \vee \neg B$ follows (intro($\vee$)), and we are done. Now we assume A instead. We again proceed by cases. We first assume $\neg B$. Then $\neg A \vee \neg B$ follows again (intro($\vee$)), and we are done. Now we assume B instead. Then from A and B we obtain $A \wedge B$ (intro($\wedge$)), and from $\neg(A \wedge B)$ and $A \wedge B$ we obtain a contradiction (elim($\neg$)), so we can conclude anything we like, which finishes the proof.

To prove the reverse implication $(\neg A \vee \neg B) \Rightarrow \neg(A \wedge B)$, we assume $\neg A \vee \neg B$ and have to show $\neg(A \wedge B)$ (intro($\Rightarrow$)). To do that, we assume $A \wedge B$ and have to produce a contradiction (intro($\neg$)). It suffices to do that once assuming $\neg A$ and once assuming $\neg B$ (elim($\vee$)). In both cases, we obtain A and B from $A \wedge B$ (elim($\wedge$)); the contradiction follows from $\neg A$ and A in the first case and from $\neg B$ and B in the second case (elim($\neg$)).

(3) Another important rule follows from the fact that $(A \Rightarrow B) \Leftrightarrow (\neg B \Rightarrow \neg A)$ is always true. This means that we can replace the implication $A \Rightarrow B$ by the negated-and-reversed version $\neg B \Rightarrow \neg A$, the *contrapositive* of $A \Rightarrow B$. Using this is called *proof by contraposition*.

DEF contrapositive

But first we need to prove the equivalence. As usual, we split into the two implications (intro($\Leftrightarrow$)).

"$\Rightarrow$" (we will use this shorthand to indicate that we are going to prove the "forward" implication "from left to right"): We assume $A \Rightarrow B$ and have to show $\neg B \Rightarrow \neg A$ (intro($\Rightarrow$)). We then in addition assume $\neg B$ and have to show $\neg A$ (intro($\Rightarrow$)). Finally, we further assume A and have to produce a contradiction (intro($\neg$)). Now from $A \Rightarrow B$ and A, we obtain B (elim($\Rightarrow$)), and then from $\neg B$ and B, we obtain the desired contradiction (elim($\neg$)).

"$\Leftarrow$" (this indicates that we now prove the "reverse" implication "from right to left"): We assume $\neg B \Rightarrow \neg A$ and have to show $A \Rightarrow B$ (intro($\Rightarrow$)). So we also assume A and have to show B (intro($\Rightarrow$)). Now we argue by cases. Either B is true, then we are done. Or $\neg B$ is true, then from $\neg B$ and $\neg B \Rightarrow \neg A$, we obtain $\neg A$ (elim($\Rightarrow$)), and then from $\neg A$ and A, we obtain a contradiction (elim($\neg$)), which finishes the proof. ♣

In practice (and very soon also in these lectures) one obviously does not mention the rules that are used in every step; also, as you progress in your studies, more and more steps or sequences of steps in proofs are considered "obvious" and are not spelled out. But it must still be possible to expand each proof into a sequence of applications of the rules; otherwise it is not a correct proof (which means that it is not a proof at all—there is no such thing as a "partial proof": it is all or nothing).

Some parts of these notes are in smaller print (and a different color) like this one. They contain material that goes beyond the core contents of this course, but may be of some interest.

What I want to point out here is that one can actually play proving as a game. There are computer programs that allow to interactively construct mathematical proofs and that check whether these proofs are correct. This can then be turned into an actual

game. One such system is called Lean; you can find some such games on the Lean Game Server. Try it out!

2.B Sets.

We assume that you have learned at school how to work with *sets*. Therefore we will content ourselves here with recalling the most important facts and definitions; we also recall/introduce some notation.

Finite sets can be specified by enumerating their elements, e.g.,

$$\{1, 2, 4, 8\}\,, \qquad \{\{1\}, \{1, 2\}\}\,.$$

Note that the elements of a set can themselves be sets as in the second example above. One of the most important sets is the *empty set*, which can be written $\{\}$, but in higher mathematics is usually denoted $\emptyset$, which is the notation we will also use. We write $x \in M$ for the statement that x is an element of the set M and $x \notin M$ for its negation. The statement $x \in \emptyset$ is always false, for example, as the empty set has no elements. Two sets are *equal* if and only if they have the same elements. In particular, it is irrelevant in which order and how often the elements are listed:

DEF
$\emptyset$
DEF
$x \in M$
$x \notin M$
DEF
$M = N$

$$\{1, 3, 2, 1, 2, 2\} = \{1, 2, 3\}\,.$$

We can also describe sets by specifying the properties that their elements have. For example,

$$\{n \mid n \text{ is a prime number}\}\,.$$

(Instead of the vertical bar "|" one often writes a colon ":" or sometimes a semicolon ";".)

There are standard symbols for certain sets that occur frequently, like

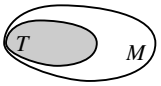
- the set $\mathbb{N} = \{0, 1, 2, 3, \ldots\}$ of *natural numbers*,
- the set $\mathbb{Z} = \{\ldots, -3, -2, -1, 0, 1, 2, 3, \ldots\}$ of *integers*,
- the set $\mathbb{Q} = \{\frac{a}{b} \mid a \in \mathbb{Z}, b \in \mathbb{N}, b \neq 0\}$ of *rational numbers*, and
- the set $\mathbb{R}$ of *real numbers*.

DEF
$\mathbb{N}, \mathbb{Z}, \mathbb{Q}, \mathbb{R}$

Warning. The definition of the set $\mathbb{N}$ of natural numbers is not consistent in the literature; frequently one defines $\mathbb{N} = \{1, 2, 3, \ldots\}$ (without zero). This is not a question of right or wrong; this is basically a matter of taste (or, perhaps rather, convenience). My take is that the natural numbers are the cardinalities (i.e., number of elements) of finite set, and since the empty set is clearly a finite set, its cardinality zero should be a natural number.

2.4. **Definition.** A set T is a *subset* of the set M, written $T \subset M$, if every element of T is also an element of M. Note that $T = M$ is allowed here. If we want to express that T is a *proper subset* of M (meaning a subset that is not M itself), we write $T \subsetneq M$. Instead of $T \subset M$, one can also write $M \supset T$ and say that M is a *superset* of T. The subset relation is also called *inclusion*. ♢

DEF
subset



Warning. The notation is not used consistently in the literature. Another fairly usual convention is to write $T \subset M$ for proper subsets ("strict inclusion") and $T \subseteq M$ for arbitrary subsets. It is important to be aware of such potential differences when working with a textbook!

The empty set is a subset of *every* set, $\emptyset \subset M$, and every set is a subset of itself, $M \subset M$. For the sets of numbers introduced above we have the inclusions $\mathbb{N} \subset \mathbb{Z} \subset \mathbb{Q} \subset \mathbb{R}$.

2.5. **Definition.** The set of all subsets of M is the *power set* of M, written

$$\mathcal{P}(M) = \{T \mid T \subset M\}\,. \qquad \diamondsuit$$

DEF
power set
$\mathcal{P}(M)$

2.6. **Example.** We have for example

$$\mathcal{P}(\emptyset) = \{\emptyset\}\,, \quad \mathcal{P}(\{\emptyset\}) = \{\emptyset, \{\emptyset\}\} \quad \text{and} \quad \mathcal{P}(\{1,2\}) = \{\emptyset, \{1\}, \{2\}, \{1,2\}\}\,. \qquad \clubsuit$$

EXAMPLE
power set

An important remark:

- It is very important to carefully distinguish between sets and their elements. For example, "$a \in M$" and "$a \subset M$" have completely different meanings.
- It is particularly hard to distinguish between the element a and the one-element set $\{a\}$. Make sure you understand the difference!

On the other hand, $a \in M$ is equivalent to $\{a\} \subset M$, so making both possible mistakes at the same time leads to something correct again.

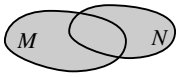
We can produce new sets from given sets.

2.7. **Definition.** Let M and N be two sets.

DEF
$M \cup N$

(1) The *union* of M and N is

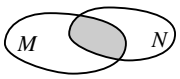
$$M \cup N = \{x \mid x \in M \vee x \in N\}\,.$$



$M \cap N$

(2) The *intersection* of M and N is

$$M \cap N = \{x \mid x \in M \wedge x \in N\}\,.$$



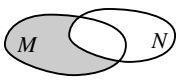
Two sets whose intersection is empty ($M \cap N = \emptyset$) are *disjoint*.

disjoint

$M \setminus N$

(3) The *(set) difference* of M and N is

$$M \setminus N = \{x \mid x \in M \wedge x \notin N\}\,.$$



It contains the elements of M that are not elements of N. $\diamondsuit$

These combinations obey certain rules, and they are related to the notion of subset. Below, we will state and prove some of them; it is a good exercise for you to come up with further ones and prove them yourselves, to get fluent in carrying out simple proofs. You will find some exercises like this on the problem sheet.

(The color red marks mathematical theorems, which are sometimes also called "lemmas", when they are of an auxiliary nature, or "corollaries", when they are fairly immediate consequences of another theorem.)

2.8. **Theorem.** THM set properties

(1) *For all sets M and N, M is a subset of N if and only if the sets $M \cup N$ and N are the same:*

$$M \subset N \iff M \cup N = N\,.$$

(2) *For any two sets X and Y, we have the "absorption law"*

$$(X \cap Y) \cup Y = Y\,.$$

(3) *If A, B, C are arbitrary sets, then we have the "distributive law"*

$$(A \cap B) \cup C = (A \cup C) \cap (B \cup C)\,.$$

The following analogous statements are also true, but we will not give proofs here.

(1) For all sets M and N: $\quad M \subset N \iff M \cap N = M$.

(2) For all sets X and Y: $\quad (X \cup Y) \cap Y = Y$.

(3) For all sets A, B, C: $\quad (A \cup B) \cap C = (A \cap C) \cup (B \cap C)$.

(The color dark red marks proofs.)

Can you figure out which proof rules are used in each step in the proofs below?

Proof.

(1) We want to show an equivalence $M \subset N \iff M \cup N = N$. We split the proof into the two directions.
"$\Rightarrow$": We assume that M is a subset of N and have to show that $M \cup N = N$. By definition, this is again an equivalence, namely $x \in M \cup N \iff x \in N$ (where x is any potential element). We split this equivalence again. Note that we can write it as "$(M \cup N \subset N) \wedge (N \subset M \cup N)$".
"$M \cup N \subset N$": Assume $x \in M \cup N$. By definition, this means $x \in M$ or $x \in N$. In the first case ($x \in M$), we obtain $x \in N$ since by assumption M is a subset of N. In the second case $x \in N$ is true by assumption.
"$N \subset M \cup N$": This is always true; note that $x \in N$ implies $x \in M \vee x \in N$.
We have shown the equality $M \cup N = N$, which finishes the proof of "$\Rightarrow$".
"$\Leftarrow$": We assume that $M \cup N = N$ and have to show that $M \subset N$. So we assume that $x \in M$ (and have to show that $x \in N$). From $x \in M$ we obtain $x \in M \cup N$. But $M \cup N = N$ by assumption, so $x \in N$ as desired.

(2) We have to show an equality of sets. We have just seen that we can split the proof into showing two inclusions. In analogy with "$\Rightarrow$" and "$\Leftarrow$", we use the shorthand "$\subset$" and "$\supset$" to indicate the two parts of the proof.
"$\subset$": (Here we prove $(X \cap Y) \cup Y \subset Y$.) Assume that $a \in (X \cap Y) \cup Y$. This means $a \in X \cap Y$ or $a \in Y$. We want to show $a \in Y$. In the second case, this is true by assumption; in the first case, $a \in X \cap Y$ implies ($a \in X$ and) $a \in Y$.
"$\supset$": (Now we prove $(X \cap Y) \cup Y \supset Y$, i.e., $Y \subset (X \cap Y) \cup Y$.) Assume that $a \in Y$. This implies "$a \in X \cap Y$ or $a \in Y$", which exactly means that $a \in (X \cap Y) \cup Y$.
An alternative argument uses the already proved statement (1). We can use it for $M = X \cap Y$ and $N = Y$; then the inclusion $X \cap Y \subset Y$ on the left of the equivalence is always true, so the equality on the right must also be true.

(3) "$\subset$": Assume $s \in (A \cap B) \cup C$. This means that $s \in A \cap B$ or $s \in C$. In the first case, we obtain $s \in A$ and $s \in B$, which imply $s \in A \cup C$ and $s \in B \cup C$, respectively. These together give us $s \in (A \cup C) \cap (B \cup C)$. In the second case ($s \in C$), we also obtain $s \in A \cup C$ and $s \in B \cup C$, and can then conclude in the same way.

"'$\supset$"': Assume $s \in (A \cup C) \cap (B \cup C)$. This means that $s \in A \cup C$ and $s \in B \cup C$. The first of these means $s \in A$ or $s \in C$. If $s \in C$, then we obtain $s \in (A \cap B) \cup C$. If $s \notin C$, then necessarily $s \in A$ and $s \in B$, so $s \in A \cap B$, which again implies $s \in (A \cap B) \cup C$. ❑

2.C Predicate logic.

In mathematics, we usually need to deal not just with simple statements that are combined in some way, rather, in most cases, the relevant statements depend on certain parameters, called *variables*. A typical example is the statement "$x \in M$", whose truth value depends on what x and M are. Such a statement that can contain variables is called a *predicate* (whereas "proposition" refers to a statement that does not depend on variables and thus has a fixed truth value).

DEF predicate

There are then essentially two ways to turn a predicate depending on a variable x into one that does not depend on x, by *quantifying* over x. These are as follows.

2.9. **Definition.** Let $A(x)$ denote a predicate that depends on the variable x (and potentially on further variables).

DEF quantifiers $\forall$, $\exists$

(1) We can state "for all x, $A(x)$ is true", also written "$\forall x\colon A(x)$".

(2) We can state "there is some x such that $A(x)$ is true", also written "$\exists x\colon A(x)$".

The symbols $\forall$ and $\exists$ are called *quantifiers*, more precisely, the *universal quantifier* and the *existential quantifier*, respectively. ♢

In practice, we always consider values for the variables that are elements of some given set. This means that we basically only see the combinations

$$\forall x\colon x \in M \Rightarrow A(x) \qquad \text{and} \qquad \exists x\colon x \in M \wedge A(x)\,,$$

which we abbreviate to

$$\forall x \in M\colon A(x) \qquad \text{"for all } x \in M,\ A(x) \text{ holds"}$$

and

$$\exists x \in M\colon A(x) \qquad \text{"there exists an } x \in M \text{ such that } A(x) \text{ holds"}\,.$$

Note that

$$\forall x \in \emptyset\colon A(x)$$

is *always true* ("vacuously true"), because the assumption (or "antecedent") in the implication "$x \in \emptyset \Rightarrow A(x)$" is false. similarly,

$$\exists x \in \emptyset\colon A(x)$$

is never true: the empty set has no element, and so in particular no element with some additional property.

We list a few important rules for dealing with quantified statements.

- The name of the variable in a quantified statement is arbitrary. If y does not appear in $A(x)$, then
$$\forall x \in M\colon A(x) \qquad \text{and} \qquad \forall y \in M\colon A(y)$$
are equivalent ("renaming of bound variables"), and similarly for the existential quantifier. You may be familiar with this in the context of integrals:
$$\int_a^b f(x)\,dx = \int_a^b f(y)\,dy\,.$$

- It is very important to understand that the *ordering* of quantifiers of different types matters. The statements
$$\forall x \in M\ \exists y \in N\colon A(x,y) \qquad \text{and} \qquad \exists y \in N\ \forall x \in M\colon A(x,y)$$
mean rather different things: in the first statement, the y whose existence is asserted is allowed to depend on x, whereas in the second statement, there must be one (fixed) y that works for every x. See the example below.

  Therefore, it is good practice to *always* put the quantifiers *before* the statement they are applied to. Writing something like
$$\exists a \in M\colon A(a,x), \forall x \in N$$
must be avoided at all costs! This applies in the same way to statements written out in words (however, when there is only one quantifier around, we sometimes put it at the end: "blah holds for all $x \in M$").

- On the other hand,
$$\forall x \in M\ \forall y \in N\colon A(x,y) \qquad \text{and} \qquad \forall y \in N\ \forall x \in M\colon A(x,y)$$
are equivalent; therefore one also abbreviates this to
$$\forall x \in M, y \in N\colon A(x,y)$$
or, if $M = N$,
$$\forall x, y \in M\colon A(x,y)\,.$$
Similarly for existential quantifiers.

EXAMPLE
∀∃ and ∃∀

2.10. **Example.** The statement
$$\forall x \in \mathbb{N}\ \exists y \in \mathbb{N}\colon y > x$$
("for every natural number x there exists a larger natural number y") is certainly true, whereas
$$\exists y \in \mathbb{N}\ \forall x \in \mathbb{N}\colon y > x$$
("there is a natural number y that is larger than every natural number x") is false. The variant (obtained by switching the variables)
$$\exists x \in \mathbb{N}\ \forall y \in \mathbb{N}\colon y > x$$
("there is a natural number x that is smaller than every natural number y") is also false, but it is a "near miss" in a sense: replacing "$>$" with "$\geq$" makes it true (we can then take $x = 0$). ♣

Most statements that we will deal with in higher mathematics do involve quantifiers. So how does one prove such statements?

In the same way as with the logical connectives ($\wedge$, $\vee$, $\Rightarrow$, $\Leftrightarrow$, $\neg$) in propositional logic, there are elimination and introduction rules for the universal and the existential quantifiers.

- The elimination rule for $\forall$ says that if we have $\forall x \in M\colon A(x)$ among our assumptions and m is any element of M, then we can deduce $A(m)$. (If $A(x)$ holds for *all* $x \in M$, then $A(m)$ must hold for any *specific* $m \in M$. This is also called *specialization*.)

DEF specialization

- The elimination rule for $\exists$ says that if we have $\exists x \in M\colon A(x)$ among our assumptions, then we may add an unspecified element $m \in M$ to our context together with the statement $A(m)$. (m is a *witness* for the existential statement.)

DEF witness

- The introduction rule for $\forall$ says that to prove $\forall x \in M\colon A(x)$, we introduce an arbitrary $x \in M$ into the context (this is usually indicated by writing "let $x \in M$ be arbitrary" or just "let $x \in M$" in a proof) and replace the goal by $A(x)$. (To prove that $A(x)$ holds for all $x \in M$, we pick an arbitrary $x \in M$ and prove $A(x)$ for that x.)
- The introduction rule for $\exists$ says that to prove $\exists x \in M\colon A(x)$, we exhibit some specific $m \in M$ and prove $A(m)$. (Given a concrete witness, we can deduce that some witness exists.)

Note that similarly to the introduction rule of disjunction, we need to be careful when deciding on the witness we want to use when proving an existential statement (indeed, the existential statement $\exists x \in M\colon A(x)$ can be considered as a disjunction over $A(m)$ for each $m \in M$)—if we pick a wrong witness m, we may not be able to prove the statement $A(m)$.

We also have the following rules that tell us how negation interacts with our quantifiers.

- Negation of a universal statement:

$$\neg\forall x \in M\colon A(x) \qquad \text{is equivalent to} \qquad \exists x \in M\colon \neg A(x)\,.$$

  This shows how to *refute* a universal statement: one gives a *counterexample*, i.e., an element m of M that does *not* satisfy $A(m)$.

- Negation of an existential statement:

$$\neg\exists x \in M\colon A(x) \qquad \text{is equivalent to} \qquad \forall x \in M\colon \neg A(x)\,.$$

  Combined with a proof by contradiction, this shows how we can prove an existential statement *without* having to produce a witness: we show that its negation $\forall x \in M\colon \neg A(x)$ leads to a contradiction.

Here is an example how one can prove an existential statement without producing a witness. Recall that a natural number n is *composite* if $n > 1$ and n can be written as a product $n = k \cdot l$ of natural numbers k and l with $k > 1$ and $l > 1$. This is an existential statement:

$$\exists k, l \in \mathbb{N}\colon k > 1 \wedge l > 1 \wedge n = k \cdot l\,.$$

A natural number $n > 1$ that is not composite is *prime*. One can show ("Fermat's Little Theorem") that when p is a prime number and a is any natural number, then p divides $a^p - a$. Let now $n > 1$ be some natural number. If we can find some $a \in \mathbb{N}$ such that n does *not* divide $a^n - a$, then this implies that n is not a prime number, and so it must be composite. But this does *not* tell us what the factors k and l are! (Indeed, finding a nontrivial factorization efficiently on a classical computer is a famous open problem, whereas there are efficient algorithms that decide whether n is a prime or not.)

More formally, we assume that Fermat's Little Theorem

$$\forall p \in \mathbb{N}\colon \big(p \text{ is prime} \Rightarrow \forall a \in \mathbb{N}\colon p \mid a^p - a\big)$$

is a known fact ("$a \mid b$" means "a divides b"). We specialize it for n (elim($\forall$)) and obtain

$$n \text{ is prime} \Rightarrow \forall a \in \mathbb{N}\colon n \mid a^n - a\,.$$

To show that n is composite, we use a proof by contradiction, so we assume that n is not composite and deduce a contradiction. Since we also assume that $n > 1$, we then obtain that n is prime (by definition). Then we can use the implication above to deduce $\forall a \in \mathbb{N}\colon n \mid a^n - a$ (elim($\Rightarrow$)). Now, say we know that n does not divide $2^n - 2$. Then (intro($\exists$)) we can conclude $\exists a \in \mathbb{N}\colon n \nmid a^n - a$, which is equivalent to $\neg\forall a \in \mathbb{N}\colon n \mid a^n - a$. This gives the desired contradiction.

Here is a question to think about. Is

$$\big(\forall x \in M\colon A(x)\big) \Rightarrow \big(\exists x \in M\colon A(x)\big)$$

always true? If not, when is it true?

To illustrate the proof rules for quantifiers, we'll give some examples.

EXAMPLES
proofs

2.11. **Examples.**

(1) We show the statement

$$\forall x \in \mathbb{N}\ \exists y \in \mathbb{N}\colon y > x$$

from the example above.

Let $x \in \mathbb{N}$ be arbitrary; we have to show $\exists y \in \mathbb{N}\colon y > x$ (intro($\forall$)). Now we have to pick a suitable y. Note that y can depend on x, so we use $x+1$ as our witness and have to show $x + 1 > x$ (intro($\exists$)). This is a known fact about natural numbers, so the proof is complete. (More precisely, $\forall n \in \mathbb{N}\colon n+1 > n$ is a known fact, which we can add to our assumptions. Then we can *specialize* it for $x \in \mathbb{N}$ (elim($\forall$)).)

(2) We show the "drinker's paradox": in a bar there are people, some of which drink alcohol and some of which don't. If there is at least one person in the bar, then there is someone in the bar such that if this person drinks alcohol, then everybody drinks alcohol. We write "$A(x)$" to denote "person x drinks alcohol", and we write B for the set of people in the bar. Then the statement can be written as

$$B \neq \emptyset \Rightarrow \exists p \in B\colon \big(A(p) \Rightarrow \forall q \in B\colon A(q)\big)\,.$$

We assume $B \neq \emptyset$ and have to show the existential statement (intro($\Rightarrow$)). The statement "$B \neq \emptyset$" means "$\exists b \in B$", so let $b \in B$ be some person in the bar (elim($\exists$)). We argue by cases. If $\forall q \in B\colon A(q)$ is true, then we pick b as our witness and have to show $A(b) \Rightarrow \forall q \in B\colon A(q)$ (intro($\exists$)). This implication holds since its conclusion is true by assumption. Now assume that $\forall q \in B\colon A(q)$ is false, i.e., $\neg\forall q \in B\colon A(q)$ is true. This is equivalent with $\exists q \in B\colon \neg A(q)$. So there is some person x in the bar such that $A(x)$ is false (elim($\exists$)). We take x as our witness and have to show $A(x) \Rightarrow \forall q \in B\colon A(q)$ (intro($\exists$)). This implication holds since its assumption is false. ♣

You will see many more (examples of) proofs in the course of these lectures.

One thing that you might have noticed is that to a large extent, the *structure* of the statement to be proved dictates the rules that we apply. But there are some points where we can get stuck unless we know what we are doing. One such situation is when we have to decide whether we want to prove the left or right part of a disjunction or which witness to use to prove an existential statement. Another situation that may require some creativity is when we need to figure out according to which statement we want to argue by cases.

2.D Ordered pairs and Cartesian products.

We will frequently need to work with two (or even more) elements of possibly different sets together, where the ordering of theses elements is important. (When the ordering does not matter, so we deal with *unordered* pairs, we can use two-element sets $\{a, b\}$.) For this purpose, we introduce ordered pairs.

DEF ordered pair (a, b)

2.12. **Definition.** If a and b are elements of some sets, then (a, b) denotes the *ordered pair* formed by them. The characterizing property of these ordered pairs is that two ordered pairs are equal if and only if both components are the same:

$$(a, b) = (x, y) \iff \big(a = x \text{ and } b = y\big)\,. \qquad \diamondsuit$$

It is possible to define ordered pairs within set theory, for example by setting

$$(a, b) = \{\{a\}, \{a, b\}\}\,.$$

(Note the special case $(a, a) = \{\{a\}\}$.) One then has to show that the pairs defined in this way have the characterizing property from the definition above. Try to prove it!

DEF $M \times N$

2.13. **Definition.** If M and N are two sets, then we write

$$M \times N = \{(m, n) \mid m \in M, n \in N\}$$

for the set of all ordered pairs whose first component is from M and whose second component is from N. This set $M \times N$ is the *Cartesian product* of the sets M and N. $\diamondsuit$

R. Descartes (1596–1650)

("Cartesian" is derived from the latinized name *Cartesius* of the mathematician and philosopher ("cogito, ergo sum": "I think, therefore I am") René Descartes, who introduced coordinates of points in the plane (say) into mathematics.)

DEF triple, ..., n-tuple $M_1 \times \cdots \times M_n$ M^n

2.14. **Definition.** In a similar way as for pairs, we can also introduce (ordered) *triples* (a, b, c), *quadruples* (a, b, c, d), *quintuples* (a, b, c, d, e), and quite generally *n-tuples* $(a_1, a_2, \ldots, a_n)$ (for $n \in \mathbb{N}$) and define Cartesian products with more than two factors, for example,

$$A \times B \times C \times D = \{(a, b, c, d) \mid a \in A, b \in B, c \in C, d \in D\}\,.$$

The equality of two n-tuples $(a_1, a_2, \ldots, a_n)$ and $(b_1, b_2, \ldots, b_n)$ is defined analogously as for pairs:

$$(a_1, a_2, \ldots, a_n) = (b_1, b_2, \ldots, b_n) \iff a_1 = b_1 \wedge a_2 = b_2 \wedge \ldots \wedge a_n = b_n\,.$$

We obtain an important special case of the general Cartesian product when all the sets involved are the same. In this case, we write M^2 for $M \times M$, M^3 for $M \times M \times M$, and generally

$$M^n = \{(m_1, m_2, \ldots, m_n) \mid m_1, m_2, \ldots, m_n \in M\}$$

for the set of n-tuples whose components are taken from the set M. When $n = 1$, we have "1-tuples"; we set $(a) = a$, which gives us $M^1 = M$. When $n = 0$, there is the unique zero-tuple $()$ (that does not have any components), and $M^0 = \{()\}$ is a one-element set. $\diamondsuit$

EXAMPLES Cartesian products

2.15. **Examples.** For example, $\mathbb{R}^2$ is the set of all pairs of real numbers. If we interpret the components as the x and y coordinates, then we can think of $\mathbb{R}^2$ as the set of all points in the plane. Similarly, $\mathbb{R}^3$ is the set of all points in (three-dimensional) space. These sets and also their more general form $\mathbb{R}^n$ will soon show up again as standard examples of "vector spaces". $\clubsuit$

2.E Maps and families.

The last (for now) important notion that we need to introduce is that of a map (or mapping) between two sets.

2.16. **Definition.** Let M and N be two sets. A *map* f *from* M *to* N is a prescription that assigns to every $x \in M$ a uniquely determined $y \in N$; we write $f(x)$ for this y. We denote the fact that f is a map from M to N by

$$f\colon M \longrightarrow N$$

or, if we also want to specify the prescription,

$$f\colon M \longrightarrow N, \quad x \longmapsto f(x)\,,$$

where we usually write some concrete formula or something similar in place of "$f(x)$". Note the two different arrows "$\to$" and "$\mapsto$"! The first one is placed between the *sets* M and N, whereas the second one sits between the *elements* $x \in M$ and $f(x) \in N$. The element $f(x)$ of N is then the *image of* x *under* f. If $f(x) = y$ for some $y \in N$, then x is a *preimage of* y *under* f. Note that an element $y \in N$ can have no or several (even infinitely many) preimages, but each $x \in M$ always has a unique image under f.

Frequently (in particular in Real or Complex Analysis and related fields) one also uses the word *function* for a map (mainly when its target set is $\mathbb{R}$ or the set of complex numbers $\mathbb{C}$, which we will introduce soon). (This explains why maps are frequently called "f".) ♢

DEF
map

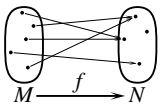


image
preimage

The term "prescription" in the definition above is *not* meant to imply that $f(x)$ must be given by a computational formula. The *only* relevant thing is that we obtain *precisely one* $f(x) \in N$ for *every* $x \in M$. One can think of f as a "black box", which eats an $x \in M$ and spits out an $f(x) \in N$ (which is always the same when we input the same x). We can visualize this as follows.

$$M \ni x \longrightarrow \boxed{f} \longrightarrow f(x) \in N$$

2.17. **Definition.** If $f\colon M \to N$ is a map, then M is the *domain* or *source* of f and N is the *codomain* or *target* of f. Note that the specification of the domain and codomain is part of the map; the specification of the rule $x \mapsto f(x)$ is not sufficient.

Two maps f and g are equal, written $f = g$, if and only if they have the same domains and codomains, and for all elements x of their common domain, we have $f(x) = g(x)$. This says that (assuming common domains and codomains) maps are determined by their values.

We write $\mathrm{Map}(M, N)$ for the set of all maps with domain M and codomain N. ♢

DEF
domain
codomain
$f = g$
$\mathrm{Map}(M, N)$

The codomain or target N of f should not be confused with the *range* or *image* of f, which is the set of all values of f,

$$\{f(x) \mid x \in M\} \subset N\,.$$

The range can be a proper subset of the codomain.

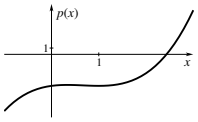
DEF
range
image

2.18. **Examples.** Here are some examples of maps.

EXAMPLES
maps

$$z\colon \mathbb{R} \longrightarrow \mathbb{R}, \quad x \longmapsto 0$$

(this is the zero map; we have $z(x) = 0 \in \mathbb{R}$ for all $x \in \mathbb{R}$),



$$p\colon \mathbb{R} \longrightarrow \mathbb{R}, \quad x \longmapsto x^3 - 2x^2 + x - 5$$

(a polynomial function; for example, we have $p(1) = p(0) = -5$),

$$s\colon \mathbb{R} \longrightarrow \{-1, 0, 1\}, \quad x \longmapsto \begin{cases} 1 & \text{if } x > 0, \\ 0 & \text{if } x = 0, \\ -1 & \text{if } x < 0 \end{cases}$$

(the sign function).

For an arbitrary set M we can define the "singleton set map"

$$e\colon M \longrightarrow \mathcal{P}(M), \quad x \longmapsto \{x\}\,.$$

The Cartesian product $M \times N$ comes with two *projection maps*

DEF
projection

$$\mathrm{pr}_1\colon M \times N \longrightarrow M, \quad (a, b) \longmapsto a \qquad \text{and} \qquad \mathrm{pr}_2\colon M \times N \to N, \quad (a, b) \longmapsto b\,.$$

If T is a subset of S, then we have the *inclusion map*

DEF
inclusion map

$$i\colon T \longrightarrow S, \quad x \longmapsto x\,.$$

For every set X there is (as a special case of the inclusion map) the *identity map* or simply *identity*

DEF
identity id_X

$$\mathrm{id}_X\colon X \longrightarrow X, \quad x \longmapsto x$$

that maps every element of X to itself.

For each set X there is exactly one map $\emptyset \to X$ (the inclusion map). A map $X \to \emptyset$ exists only when X is itself empty—if X has an element x, then x cannot be mapped to any value in $\emptyset$. ♣

If the term "prescription" that we have used above is too vague for your taste, then you will learn here how one can put the notion of "map" on a solid basis within set theory. To do this, we identify a map $f\colon M \to N$ with its *graph*

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

(This generalizes the graphs of functions $\mathbb{R} \to \mathbb{R}$ that you probably know from school.) Then one can say that a subset $F \subset M \times N$ corresponds to a map $f\colon M \to N$ if and only if the conditions

$$\forall x \in M\ \exists y \in N\colon (x, y) \in F$$

and

$$\forall x \in M\ \forall y_1, y_2 \in N\colon\ ((x, y_1) \in F \wedge (x, y_2) \in F) \Rightarrow y_1 = y_2$$

are satisfied. The first of these says that *every* $x \in M$ must be mapped to some element of N, and the second one says that there is *at most one* such element of N.

There are some important properties that a map can have or not have.

2.19. **Definition.** Let $f\colon M \to N$ be a map.

DEF
injective
surjective
bijective

(1) f is *injective*, *one-to-one* or an *injection*, if f does not map two distinct elements of M to the same element of N,

$$\forall x_1, x_2 \in M\colon f(x_1) = f(x_2) \Rightarrow x_1 = x_2$$

(So to show that f is injective, we take two elements of M and assume that they have the same image under f; then we show that these two elements must actually be equal.)

(2) f is *surjective*, *onto* or a *surjection* ("sur" = "on" in French), if every element of N occurs as the image of some element of M,

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

(3) f is *bijective* or a *bijection*, if f is both injective and surjective. ♢

Alternatively, we can characterize these properties in the following way.

- f is injective if and only if every element of N has *at most* one preimage under f.
- f is surjective if and only if every element of N has *at least* one preimage under f.
- f is bijective if and only if every element of N has *exactly* one preimage under f.

2.20. **Definition.** Let $f\colon M \to N$ be a bijective map. We can then define a map $f^{-1}\colon N \to M$ by setting $f^{-1}(y)$ to be the uniquely determined $x \in M$ such that $f(x) = y$. This map f^{-1} is the *inverse map* (or just *inverse*) of f.

DEF
inverse map
permutation

A bijective map $f\colon X \to X$ is also called a *permutation* of X. ♢

2.21. **Examples.** We write $\mathbb{R}_{\geq 0}$ for the set $\{x \in \mathbb{R} \mid x \geq 0\}$ of nonnegative real numbers.

EXAMPLES
injective
surjective

(1) $f_1\colon \mathbb{R} \to \mathbb{R}$, $x \mapsto x^2$, is neither injective nor surjective.
For example, $f_1(1) = f_1(-1) = 1$, and $-1 \in \mathbb{R}$ has no preimage.

(2) $f_2\colon \mathbb{R}_{\geq 0} \to \mathbb{R}$, $x \mapsto x^2$, is injective, but not surjective.

(3) $f_3\colon \mathbb{R} \to \mathbb{R}_{\geq 0}$, $x \mapsto x^2$, is surjective, but not injective.

(4) $f_4\colon \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$, $x \mapsto x^2$, is bijective.
The inverse map is $f_4^{-1}\colon \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$, $x \mapsto \sqrt{x}$.

This also illustrates nicely that the domain and codomain are essential for defining a map. Some further general examples are as follows.

(5) For every set M, the identity map id_M is bijective.

(6) For every set M, the "empty map" $\emptyset \to M$ is injective.

(7) Every map $\{a\} \to M$ is injective.

(8) A map $M \to \{a\}$ is surjective if and only if M is not empty.

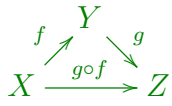
(9) The singleton set map $e\colon M \to \mathcal{P}(M)$ is injective, but not surjective (for example, the empty set has no preimage). ♣

We can compose maps by applying them one after the other.

2.22. **Definition.** If $f\colon X \to Y$ and $g\colon Y \to Z$ are maps (such that the codomain of f is the same as the domain of g), then we can form the composed map $g \circ f\colon X \to Z$ that maps $x \in X$ to $g(f(x)) \in Z$:

DEF composition

$$x \longrightarrow \boxed{f} \longrightarrow f(x) \longrightarrow \boxed{g} \longrightarrow g(f(x)) \qquad \diamondsuit$$



One has to remember that in $g \circ f$ the map f is applied *first*, even though it comes after g in the notation. We pronounce "$g \circ f$" as "g after f", which hopefully helps.

This composition of maps has some important properties.

2.23. **Theorem.**

THM properties of maps

(1) *If $f\colon W \to X$, $g\colon X \to Y$ and $h\colon Y \to Z$ are maps, then we have $(h \circ g) \circ f = h \circ (g \circ f)$. Therefore one usually omits the parentheses and writes $h \circ g \circ f$.*

(2) *If $f\colon X \to Y$ is a map, then we have*

$$f \circ \mathrm{id}_X = f \qquad \text{and} \qquad \mathrm{id}_Y \circ f = f\,.$$

(3) *If $f\colon X \to Y$ and $g\colon Y \to Z$ are both injective, then $g \circ f\colon X \to Z$ is also injective.*

(4) *If $f\colon X \to Y$ and $g\colon Y \to Z$ are both surjective, then $g \circ f\colon X \to Z$ is also surjective.*

(5) *If $f\colon X \to Y$ is bijective with inverse $f^{-1}\colon Y \to X$, then*

$$f^{-1} \circ f = \mathrm{id}_X \qquad \text{and} \qquad f \circ f^{-1} = \mathrm{id}_Y\,.$$

(6) *If $f\colon X \to Y$ and $g\colon Y \to Z$ are two maps, then we have the implications*

$$g \circ f \text{ injective} \implies f \text{ injective} \qquad \text{and} \qquad g \circ f \text{ surjective} \implies g \text{ surjective.}$$

(7) *If $f\colon X \to Y$ is a map, then f is injective if and only if X is empty or there is a map $g\colon Y \to X$ such that $g \circ f = \mathrm{id}_X$.*

(8) *If $f\colon X \to Y$ is a map, then f is surjective if and only if there is a map $g\colon Y \to X$ such that $f \circ g = \mathrm{id}_Y$.*

(9) *If $f\colon X \to Y$ is a map, then f is bijective if and only if there is a map $g\colon Y \to X$ such that $g \circ f = \mathrm{id}_X$ and $f \circ g = \mathrm{id}_Y$.*

There are (at least) two different ways of proving that a given map $f\colon X \to Y$ is bijective:

- we show that f is injective and surjective, or
- we find a candidate g for the inverse map and verify that $g \circ f = \mathrm{id}_X$ and $f \circ g = \mathrm{id}_Y$.

In many cases the second method is easier to carry out.

Proof.

(1) First of all, it is clear that both mas have the same domain W and the same codomain Z. It remains to show

$$\forall w \in W\colon \big((h \circ g) \circ f\big)(w) = \big(h \circ (g \circ f)\big)(w)\,.$$

So let $w \in W$ be arbitrary. Then we have

$$\big((h \circ g) \circ f\big)(w) = (h \circ g)(f(w)) = h\big(g(f(w))\big)$$

and similarly

$$\big(h \circ (g \circ f)\big)(w) = h\big((g \circ f)(w)\big) = h\big(g(f(w))\big)\,.$$

$$w \longrightarrow \boxed{f} \longrightarrow f(w) \longrightarrow \boxed{g} \longrightarrow g(f(w)) \longrightarrow \boxed{h} \longrightarrow h(g(f(w)))$$

(2) In both cases the two maps in the claimed equality have the same domain X and the same codomain Y. For $x \in X$ we have

$$(f \circ \mathrm{id}_X)(x) = f(\mathrm{id}_X(x)) = f(x)\,,$$

which shows $f \circ \mathrm{id}_X = f$, and

$$(\mathrm{id}_Y \circ f)(x) = \mathrm{id}_Y(f(x)) = f(x)\,,$$

which shows $\mathrm{id}_Y \circ f = f$.

(3) Exercise.

(4) Exercise.

(5) The domains and codomains on both sides are the same for both equalities. For $x \in X$ we have $f^{-1}(f(x)) = x = \mathrm{id}_X(x)$ by definition of the inverse map, which implies that $f^{-1} \circ f = \mathrm{id}_X$. For $y \in Y$ we have $f(f^{-1}(y)) = y = \mathrm{id}_Y(y)$ again by definition of the inverse map, which implies that $f \circ f^{-1} = \mathrm{id}_Y$.

(6) To show the first implication, we assume that $g \circ f$ is injective; we must show that f is also injective. So let $x_1, x_2 \in X$ be such that $f(x_1) = f(x_2)$. Then

$$(g \circ f)(x_1) = g(f(x_1)) = g(f(x_2)) = (g \circ f)(x_2)\,,$$

and because $g \circ f$ is injective by assumption, it follows that $x_1 = x_2$. This shows that f is injective.
To show the second implication, we assume that $g \circ f$ is surjective; we must show that g is also surjective. Let $z \in Z$ be arbitrary. Since by assumption, $g \circ f$ is surjective, there is some $x \in X$ such that $g(f(x)) = (g \circ f)(x) = z$. Setting $y = g(x)$, we then obtain $g(y) = z$. So g is indeed surjective.

(7) We want to show the equivalence

$$f\colon X \to Y \text{ injective} \iff (X = \emptyset \ \vee\ \exists g \in \mathrm{Map}(Y, X)\colon g \circ f = \mathrm{id}_X)\,.$$

"$\Rightarrow$": We assume that f is injective. If X is empty, the right hand side holds (by the introduction rule of disjunction). If X is nonempty, let $x_0 \in X$ be some element. We construct a suitable map $g\colon Y \to X$ as follows. Let $y \in Y$. If there exists an $x \in X$ such that $f(x) = y$, then we set $g(y) = x$. Since f is injective, there is exactly one such x, so $g(y)$ is uniquely defined. If there is no $x \in X$ with $f(x) = y$, then we set $g(y) = x_0$. Now we need to check that g satisfies the desired property $g \circ f = \mathrm{id}_X$. Both sides have the same domain and codomain, and for $x \in X$ we have by our definition of g that $(g \circ f)(x) = g(f(x)) = x = \mathrm{id}_X(x)$. This shows that the two maps are equal.
"$\Leftarrow$": If $X = \emptyset$, then f is injective. If there is a map $g\colon Y \to X$ such that $g \circ f = \mathrm{id}_X$, then f is also injective by part (6) since id_X is injective.

(8) "$\Rightarrow$": If f is surjective, then for each $y \in Y$ we can pick an $x_y \in X$ such that $f(x_y) = y$ (by definition there is always at least one preimage). We then set $g(y) = x_y$, and obtain $f \circ g = \mathrm{id}_Y$.
"$\Leftarrow$": This follows from part (6) since id_Y is surjective.

(9) "⇒": If f is bijective, then $g = f^{-1}$ has the required property.
"⇐": By part (7), f is injective, and by part (8), f is surjective, so f is bijective. ❑

If one wants to define maps that depend on two (or more) elements of possibly different sets, then one can do this using Cartesian products. If we want to map an element of M_1 and an element of M_2 to an element of N, then this can be done by defining a map $M_1 \times M_2 \to N$. For example, we can interpret the addition of real numbers as a map $\mathbb{R} \times \mathbb{R} \to \mathbb{R}$, $(x, y) \mapsto x + y$. If $f\colon M_1 \times M_2 \to N$ is a map, then we usually write $f(m_1, m_2)$ instead of $f\big((m_1, m_2)\big)$. A map of the form $M \times M \to M$ is sometimes called a *binary operation* on M.

DEF binary operation

To conclude this section, we introduce an alternative interpretation and notation for maps that will occur frequently later.

2.24. **Definition.** If $a\colon I \to M$ is a map, then one also writes this as $(a_i)_{i\in I}$ and calls it a *family* of elements of M *indexed* by (the *index set*) I. Here $a_i = a(i)$ is the value of the map a on the element $i \in I$. You may be familiar with this when $I = \mathbb{N}$; then $(a_n)_{n\in\mathbb{N}}$ is a sequence of elements of M.

DEF family M^I subfamily

The n-tuples introduced earlier can be seen as the special case $I = \{1, 2, \dots, n\}$ of such a family. Generalizing the notation M^n for the set of all n-tuples with components in M, we also write M^I for the set of all families of elements of M indexed by I. This is the same (up to the notation) as the set $\mathrm{Map}(I, M)$ of all maps from I to M.

If $J \subset I$ is a subset, then $(a_i)_{i\in J}$ is a *subfamily* of $(a_i)_{i\in I}$. ♢

The difference between the *map* $a\colon I \to M$ and the *family* $(a_i)_{i\in I}$ is just in the interpretation.

- In the map version, the focus is more on the act of *mapping* elements.
- In the family version, the focus is more on the *values* that occur.

B. Russell (1872–1970)

We use the opportunity to say a bit more about set theory. What we do here is "naive" set theory; we do not consider which constructions with sets are really possible or allowed. Usually, this does not lead to problems. However, you should know that set theory is not as harmless as it may at first appear. If one allows too much freedom in the construction of sets, one gets into difficulties, as is demonstrated by the famous *Russell's Paradox*. If one can construct the "set of all sets that do not contain themselves as an element", $M = \{x \mid x \notin x\}$, then the question whether M is an element of itself leads to a contradiction that cannot be resolved. (In recreational mathematics, the following version is popular. In a village there is a barber who shaves exactly those men in the village that do not shave themselves. Does the barber shave himself or not?) To avoid this contradiction, one needs to formulate precise rules how one can construct sets from other sets (and that in particular do not allow the construction of M above). This leads to *axiomatic set theory*.

Most of the axioms are rather harmless; for example, they assert the existence of the empty set, that one can construct sets with one or two elements, that one can always form subsets, and that unions and power sets exist. However, there is one axiom, the *Axiom of Choice*, which was rejected by some mathematicians. It says that "for every family of nonempty sets, there exists a choice function". More precisely: if $(X_i)_{i\in I}$ is a family of sets such that $X_i \neq \emptyset$ for all $i \in I$, then there exists a *choice function* $f\colon I \to X$, where $X = \{x \mid \exists i \in I\colon x \in X_i\}$ is the union of all sets X_i (whose existence is given by one of the harmless axioms), such that for every $i \in I$ the image $f(i)$ is an element of X_i. In other words, the choice function selects one element in each set X_i.

We have actually used the Axiom of Choice in the proof of part (8) of Theorem 2.23 when we picked a preimage x_y for each $y \in Y$.

The reason for rejecting the Axiom of Choice is that it is not "constructive": it makes an existential statement ("there is a choice function"), but does not say *how* such a choice function can be defined. Nowadays most mathematicians tend to accept the Axiom of Choice, because it is useful. In particular in Analysis, where one frequently deals with uncountable objects, one cannot get very far without it. It is known that adding the Axiom of Choice to the other axioms of set theory does not lead to a contradiction if the other axioms are consistent (which is not known). It is also true that adding the negation of the Axiom of Choice also does not lead to new contradictions: the Axiom of Choice is independent of the other axioms of set theory.

3. Algebraic structures: groups, rings, fields

Date: March 5, 2026

We now introduce the most important algebraic structures: groups, rings, and fields. Groups are important in many contexts within mathematics, although this will not be very apparent in these basic lectures. Most important in the following will be fields (these are structures in which we can use the basic arithmetic operations addition, subtraction, multiplication and division, and they obey the usual rules): in Linear Algebra the domain that provides the "scalars" in a linear structure (a vector space) is a field, and in Analysis, we work with the fields of real and of complex numbers. Rings are "fields without division"; they are an intermediate step on the way to defining fields and are also important in algebra.

In general, an algebraic structure is a set with one or more binary operations on it that satisfy some rules. The following defines an in a certain way "minimally interesting" algebraic structure.

DEF semigroup

3.1. **Definition.** A *semigroup* is a pair $(S, *)$, where S is a set and $*$ is a binary operation $*\colon S \times S \to S$, $(a, b) \mapsto a * b$, which satisfies the *associative law*,

$$\forall a, b, c \in S\colon (a * b) * c = a * (b * c)\,.$$

The semigroup is *commutative*, if in addition the *commutative law* holds:

$$\forall a, b \in S\colon a * b = b * a\,.$$

When the binary operation $*$ is clear from the context, one usually just writes "the semigroup S". ♢

A remark on notation: binary operations in algebraic structures are usually written in "infix notation" like "$a * b$" instead of "$*(a, b)$".

A consequence of the associative law is that it does not matter how we set parentheses when combining three or more elements using the binary operation. For example, we have for $a, b, c, d, e \in S$ that

$$\begin{aligned} a * ((b * c) * d) &= a * (b * (c * d)) = (a * b) * (c * d) \\ &= ((a * b) * c) * d = (a * (b * c)) * d \qquad \text{and} \\ a * (b * (c * (d * e))) &= (a * b) * (c * (d * e)) = ((a * b) * (c * d)) * e = \dots\,. \end{aligned}$$

We can therefore simply write $a * b * c * d$, $a * b * c * d * e$ and so on.

This leads to an interesting combinatorial question, namely, *how many* different ways there are to fully parenthesize a combination of n elements. Let C_n denote this number. Then we see from the above that $C_1 = C_2 = 1$, $C_3 = 2$, and $C_4 = 5$. From the consideration that we obtain a combination of n elements by applying our binary operation to a combination of k elements and a combination of $n - k$ elements (for some k such that $1 \le k < n$), we obtain the following *recurrence* for C_n:

$$C_n = \sum_{k=1}^{n-1} C_k C_{n-k} = C_1 C_{n-1} + C_2 C_{n-2} + \dots + C_{n-2} C_2 + C_{n-1} C_1 \qquad \text{for all } n \ge 2.$$

(We will introduce the summation sign $\sum$ soon. In general, a recurrence expresses the value of a function $\mathbb{N} \to X$ at n in terms of its values at numbers less than n.) This allows us to compute $C_5 = 1 \cdot 5 + 1 \cdot 2 + 2 \cdot 1 + 5 \cdot 1 = 14$, $C_6 = 42$, $C_7 = 132$, and so on. There is actually a closed formula for C_n in terms of binomial coefficients (see later),

$$C_n = \frac{1}{n}\binom{2n-2}{n-1} = \frac{1}{2n-1}\binom{2n-1}{n-1} = \frac{(2n-2)!}{(n-1)!\,n!}\,,$$

which, however, is not so easy to prove directly. (You are welcome to try!) These numbers C_n are known as the *Catalan numbers* (which explains the "C"); frequently (e.g., in Wikipedia) the index is shifted by 1 and one starts with $C_0 = C_1 = 1$. They show up in combinatorics in many different contexts.

If our semigroup is commutative, then the ordering of the elements does not matter:

$$a * b * c = b * a * c = b * c * a = c * b * a = c * a * b = a * c * b\,.$$

In the following we write $\mathbb{N}_{>0} = \{n \in \mathbb{N} \mid n > 0\} = \{1, 2, 3, \ldots\}$ for the set of strictly positive natural numbers. **DEF** $\mathbb{N}_{>0}$

3.2. **Examples.** The trivial example of a semigroup is $(\emptyset, *)$, where $*\colon \emptyset \times \emptyset \to \emptyset$ is the empty map (note that $\emptyset \times \emptyset = \emptyset$). **EXAMPLES** semigroups

Examples of commutative semigroups are $(\mathbb{N}_{>0}, +)$, $(\mathbb{N}, +)$, $(\mathbb{Z}, +)$, $(\mathbb{N}_{>0}, \cdot)$, $(\mathbb{N}, \cdot)$, $(\mathbb{Z}, \cdot)$. The semigroup $(\mathrm{Map}(X, X), \circ)$, where X is an arbitrary set and the binary operation is composition of maps, is not commutative in general (in fact, it is commutative if and only if X has at most one element; this is an exercise). ♣

A bare semigroup is not yet very useful. So we impose some more structure.

3.3. **Definition.** A *monoid* is a triple $(M, *, e)$, where M is a set, $*\colon M \times M \to M$ is a binary operation, and $e \in M$ is an element, such that $(M, *)$ is a semigroup with *neutral element* e, **DEF** monoid

$$\forall a \in M\colon e * a = a = a * e\,.$$

The monoid is *commutative* if the semigroup $(M, *)$ is commutative. ♢

If there exists a neutral element, then it is uniquely determined. This is shown in the following lemma (a *lemma* is an auxiliary statement or a less important mathematical theorem).

3.4. **Lemma.** *Let $(S, *)$ be a semigroup. If $e \in S$ is a left neutral element and $e' \in S$ is a right neutral element in this semigroup, meaning that* **LEMMA** uniqueness of neutral elements

$$\forall a \in S\colon e * a = a \text{ and } a * e' = a\,,$$

then $e = e'$.

Proof. Since e is left neutral, we have $e * e' = e'$. Since e' is right neutral, we have $e * e' = e$. Both together imply $e = e'$. ❑

Because of this, one usually leaves out the specification of the neutral element and talks about the "monoid $(M, *)$" or even the "monoid M" when the binary operation is clear from the context.

Note that it is indeed possible that a semigroup has (for example) several left neutral elements (and then clearly no right neutral element). If M is an arbitrary set and we choose pr_2 for the binary operation (so that $a * b = b$), then we obtain a semigroup in which *all* elements are left neutral.

EXAMPLES
monoids

3.5. **Examples.** Since the definition of "monoid" requires a neutral element, the empty set cannot be a monoid. The "trivial" monoid is then $(\{e\}, *, e)$, where $*$ is the only possible map $\{e\} \times \{e\} \to \{e\}$ (so that $e * e = e$).

With the exception of $(\mathbb{N}_{>0}, +)$, which has no neutral element, all the examples of semigroups in 3.2 can be considered as monoids $(\mathbb{N}, +, 0)$, $(\mathbb{Z}, +, 0)$, $(\mathbb{N}_{>0}, \cdot, 1)$, $(\mathbb{N}, \cdot, 1)$, $(\mathbb{Z}, \cdot, 1)$, and $(\mathrm{Map}(X, X), \circ, \mathrm{id}_X)$. ♣

It is even better when we can undo the binary operation with some other element by operating with some element again. This leads to the notion of a group.

DEF
group

3.6. **Definition.** A *group* is s quadruple $(G, *, e, i)$, where G is a set, $*$ is a binary operation on G, $e \in G$, and i is a map $i\colon G \to G$, such that $(G, *, e)$ is a monoid and for every $g \in G$ the element $i(g) \in G$ is an *inverse* of g,

$$\forall g \in G\colon i(g) * g = e = g * i(g)\,.$$

The group is *commutative* or *abelian* if the monoid $(G, *, e)$ is commutative. ♢

N.H. Abel
1802–1829

The name "abelian" refers to the Norwegian mathematician Niels Henrik Abel, after whom also the *Abel Prize* is named. This is a prize for mathematical achievements that is comparable to a Nobel Prize and is awarded yearly, beginning in 2003.

Inverses are also uniquely determined, as the following lemma shows.

LEMMA
uniqueness of inverses

3.7. **Lemma.** *Let $(M, *, e)$ be a monoid and let $a \in M$. If $b \in M$ is a left inverse and $c \in M$ is a right inverse of a,*

$$b * a = e = a * c\,,$$

then $b = c$.

Proof. We have

$$b = b * e = b * (a * c) = (b * a) * c = e * c = c\,.$$ ❑

In the same way as for monoids, we therefore usually simply talk about the "group $(G, *)$" or even the "group G" when the binary operation is clear from the context.

There are two main notations used for binary operations in groups.

- The "multiplicative" notation. One writes the operation as $a \cdot b$ or also ab, the neutral element is denoted 1 and the inverse of a is written a^{-1}.
- The "additive" notation. This is almost exclusively used for commutative groups. The operation is written $a+b$, the neutral element is denoted 0, and the inverse of a is written $-a$ (and called the *negative* of a). One also abbreviates $a + (-b)$ to $a - b$.

EXAMPLES
groups

3.8. **Examples.** The trivial monoid can also be seen as a group: its only element e is its own inverse.

Among the other examples of monoids in 3.5 only $(\mathbb{Z}, +, 0, -)$ can be "promoted" to a group (and the last example $\mathrm{Map}(X, X)$ when X has at most one element; this gives a trivial group). Another example of a commutative group is given by $(\mathbb{R}_{>0}, \cdot, 1, x \mapsto 1/x)$, where $\mathbb{R}_{>0}$ is the set of all positive real numbers.

DEF
symmetric group $S(X)$

If we only consider the *bijective* maps (permutations) $X \to X$, then we obtain a group $(S(X), \circ, \mathrm{id}_X, f \mapsto f^{-1})$, which is known as the *symmetric group* of X.

Here we set

$$S(X) = \{f \colon X \to X \mid f \text{ bijective}\}\,.$$

This group is commutative if and only if X has at most two elements (exercise). ♣

For a semigroup to "be" a group, it is actually sufficient to require only the existence of a left neutral element e and for every element x the existence of a left inverse $i(x)$. Then it follows that e is also right neutral:

$$x*e = e*x*e = i(i(x))*i(x)*x*e = i(i(x))*e*e = i(i(x))*e = i(i(x))*i(x)*x = e*x = x\,.$$

Using this, we also obtain $i(i(x)) = i(i(x)) * e = x * e = x$, which allows us to show that $i(x)$ is also a right inverse of x:

$$x * i(x) = i(i(x)) * i(x) = e\,.$$

We can obviously replace "left" by "right" in the claim above, and it will also work. On the other hand, there exist semigroups that have left neutral and right inverse elements, but are not groups. Find an example!

Groups are nice, because we can always solve certain types of equations in them. Before we prove that, we show a cancellation rule.

3.9. **Lemma.** *Let $(G, *, e, i)$ be a group and let $a, b, c \in G$. Then we have*

$$a * c = b * c \iff a = b \iff c * a = c * b\,.$$

LEMMA cancellation in groups

Proof. We prove the first equivalence; the second is done in the same way.

"$\Leftarrow$" is clear. For "$\Rightarrow$" we proceed as follows.

$$\begin{aligned} a * c = b * c &\implies (a * c) * i(c) = (b * c) * i(c) \implies a * (c * i(c)) = b * (c * i(c)) \\ &\implies a * e = b * e \implies a = b\,. \end{aligned}$$

❑

3.10. **Lemma.** *Let $(G, *, e, i)$ be a group and let $a, b \in G$. Then the equations*

$$a * x = b \qquad \text{and} \qquad x * a = b$$

each have a unique solution $x \in G$, namely $x = i(a)*b$ and $x = b*i(a)$, respectively.

LEMMA equations in groups

Proof. We carry out the proof for the first equation; the second one is similar.

$$a * x = b \iff i(a) * a * x = i(a) * b \iff e * x = i(a) * b \iff x = i(a) * b\,.$$

We have used Lemma 3.9 for the first equivalence. ❑

Now we consider structures that have two binary operations.

3.11. **Definition.** A *ring* is a sextuple $(R, +, 0, -, \cdot, 1)$, where R is a set, $+, \cdot\colon R \times R \to R$ are binary operations, $0, 1 \in R$ are elements, and $-\colon R \to R$ is a map, such that $(R, +, 0, -)$ is a commutative group, $(R, \cdot, 1)$ is a monoid, and the *distributive laws* hold,

$$\forall a, b, c \in R\colon\ a \cdot (b + c) = a \cdot b + a \cdot c \quad \text{and} \quad (a + b) \cdot c = a \cdot c + b \cdot c\,.$$

The ring is *commutative* if the monoid $(R, \cdot, 1)$ is commutative. ♢

DEF ring

Since the neutral and inverse elements are uniquely determined, we usually just talk about the "ring $(R, +, \cdot)$" or even the "ring R" when the operations are clear from the context.

If the ring is commutative, then it suffices to require only one of the two distributive laws (the other then follows). The product $a \cdot b$ of two elements is also written ab.

The definition tells us that we can add, subtract, and multiply in a ring, and the usual rules are valid, for example $0 \cdot a = a \cdot 0 = 0$, $-(a + b) = -a - b$, $(-a) \cdot (-b) = a \cdot b$ (exercise). The implication $a \cdot b = 0 \Rightarrow a = 0 \vee b = 0$ is *not* necessarily true, however!

In terms that involve addition and multiplication, the convention is that multiplication is "stronger", so that

$$a \cdot b + c \cdot d = (a \cdot b) + (c \cdot d) \qquad \text{rather than} \qquad a \cdot (b + c) \cdot d\,.$$

3.12. **Examples.** The trivial example of a ring is the *zero ring* $(\{0\}, +, 0, -, \cdot, 0)$, where $0 = 1$ and $0 + 0 = -0 = 0 \cdot 0 = 0$. Every ring R in which the zero element 0_R is the same as the unit element 1_R is a zero ring: for all $r \in R$ we have that $r = 1_R \cdot r = 0_R \cdot r = 0_R$, so 0_R is the only element of R.

The standard example of a (commutative) ring is the ring $\mathbb{Z}$ of integers with the usual addition and multiplication as operations.

If you already know matrices and how to add and multiply them, then you can verify that the set of all 2×2 matrices with entries in $\mathbb{R}$ together with the addition and multiplication of matrices forms a non-commutative ring. ♣

EXAMPLES rings

Finally, we consider fields.

3.13. **Definition.** A *field* is a septuple $(K, +, 0, -, \cdot, 1, i)$, where K is a set, $+, \cdot\colon K \times K \to K$ are binary operations, $0, 1 \in K$ are elements, and $-\colon K \to K$, $i\colon K \setminus \{0\} \to K \setminus \{0\}$ are maps, such that $(K, +, 0, -, \cdot, 1)$ is a commutative ring and $(K \setminus \{0\}, \cdot, 1, i)$ is a (commutative) group. One writes a^{-1} for $i(a)$. ♢

DEF field

As usual, we talk about the "field $(K, +, \cdot)$" or the "field K". The definition implies that 0 and 1 must be distinct elements in any field since 1 is required to be the neutral element of the group $K \setminus \{0\}$.

The group $(K \setminus \{0\}, \cdot)$ is also denoted $K^\times$ and is called the *multiplicative group* of K. (Another common notation is K^*.)

DEF multiplicative group $K^\times$

It is of course also possible to define what a field is without mentioning rings and groups. Given the "data" $+$, $\cdot$, 0, 1, $-$, and i, the following "field axioms" have to be satisfied:

$$\begin{array}{ll} (a + b) + c = a + (b + c)\,, & a + b = b + a \\ a + 0 = a\,, & a + (-a) = 0 \\ (a \cdot b) \cdot c = a \cdot (b \cdot c) & a \cdot b = b \cdot a \\ a \cdot 1 = a\,, & a \neq 0 \Rightarrow a \cdot a^{-1} = 1 \\ 0 \neq 1\,, & a \cdot (b + c) = a \cdot b + a \cdot c \end{array}$$

For two elements $a, b \in K$ with $b \neq 0$ we can define division by $a/b = a \cdot b^{-1}$. Then we have all four basic arithmetic operations at our disposal, and the usual rules are valid (they can be deduced from the field axioms). For example, in a field it is always true that $a \cdot b = 0$ implies that $a = 0$ or $b = 0$. (If $a \neq 0$, then we obtain $0 = a^{-1} \cdot 0 = a^{-1} \cdot a \cdot b = 1 \cdot b = b$.)

EXAMPLES
fields

3.14. **Examples.** The smallest example of a field has only the two necessary elements 0 and 1. The rules $0 + 0 = 0$, $0 + 1 = 1 + 0 = 1$, $0 \cdot 0 = 0 \cdot 1 = 1 \cdot 0 = 0$, and $1 \cdot 1 = 1$ follow directly from the definition; the remaining sum $1 + 1$ must then give 0 since the equation $a + 1 = 0$ must be solvable. One can then show (this is easy, but somewhat tedious) that this structure, which is denoted $\mathbb{F}_2$, is indeed a field.

Standard examples of fields are given by the field $\mathbb{Q}$ of rational numbers and the field $\mathbb{R}$ of real numbers (see next section), with the usual addition and multiplication. We will soon construct another field, the field $\mathbb{C}$ of complex numbers. ♣

4. Proof by induction

Date:
March 5, 2026

4.A An Introductory Example.

In this section, we introduce a proof technique that allows us to prove mathematical statements for all natural numbers. As motivation, let us consider the following problem:

Calculate the sum of the first n positive natural numbers:

$$1 + 2 + 3 + \ldots + (n-1) + n.$$

To write this sum more concisely, we express it using the summation symbol $\sum$, which is defined as follows:

DEF Sum

4.1. **Definition.** Let m, n be natural numbers with $m \leq n$. For each natural number $k = m, \ldots, n$, let $a_k \in G$, where G is a commutative group. Then we define the sum of the numbers a_m to a_n as

$$\sum_{k=m}^{n} a_k := a_m + a_{m+1} + \ldots + a_n.$$

Additionally, for $m \geq 1$, we define the *empty sum* as

$$\sum_{k=m}^{m-1} a_k := 0$$

independent of the values of the a_k. ♢

The symbol ":=" indicates that the object on the left-hand side is defined to be equal to the expression on the right-hand side.

C.F. Gauss
1777–1855

Using this summation notation, the above task becomes:

$$\text{Calculate } \sum_{k=1}^{n} k.$$

There is an anecdote about the mathematician Carl Friedrich Gauss (1777–1855), whose teacher, when Gauss was a child, once assigned the class the task of adding the first 100 numbers. This is precisely our problem with $n = 100$. Instead of adding the numbers one by one, Gauss proceeded as follows:

$$\begin{aligned}\sum_{k=1}^{100} k &= (1 + 100) + (2 + 99) + \ldots + (49 + 52) + (50 + 51) \\ &= \underbrace{101 + 101 + \ldots + 101 + 101}_{\text{a total of 50 terms}} = 50 \cdot 101 = 5050\end{aligned}$$

and was thus able to calculate the result much more quickly.

This idea can be generalized to any n. In the case where n is even, we can write:

$$\begin{aligned}\sum_{k=1}^{n} k &= (1 + n) + (2 + (n-1)) + \ldots + \left(\left(\frac{n}{2} - 1\right) + \left(\frac{n}{2} + 2\right)\right) + \left(\frac{n}{2} + \left(\frac{n}{2} + 1\right)\right) \\ &= \underbrace{(n+1) + (n+1) + \ldots + (n+1) + (n+1)}_{\text{a total of } n/2 \text{ terms}} = \frac{n}{2} \cdot (n+1) = \frac{n(n+1)}{2}.\end{aligned}$$

In the case where n is odd, one can suitably modify the trick and obtain the same result (exercise).

One might now think the problem is solved. However, in the strict mathematical sense (which we will always apply in this course), this is not yet a valid proof. While the above calculation makes it plausible that the equation

$$\sum_{k=1}^{n} k = \frac{n(n+1)}{2} \tag{4.1}$$

holds for all $n \in \mathbb{N}$, it is not formally proven! The issue lies in the ellipsis "..." in the sum $(1+n)+(2+(n-1))+\ldots+((n/2-1)+(n/2+2))+(n/2+(n/2+1))$. Although it is intuitively clear what is meant by the ellipsis, a rigorous mathematical proof requires us to explicitly write out all the missing terms[1]. To do this, we would need to know the value of n, so we can determine how many terms to write in place of the ellipsis.

Thus, while we could expand the above calculation into a correct proof for each specific $n \in \mathbb{N}$, we would have to do so individually for each n. Since there are infinitely many natural numbers, this approach cannot provide a proof that establishes the formula for all $n \in \mathbb{N}$.

4.B The Induction Principle.

The proof by induction provides an elegant solution to the problem just discussed. For a given $n_0 \in \mathbb{N}$, the principle is used to prove a statement $A(n)$ for all $n \in \mathbb{N}$ with $n \geq n_0$. The principle consists of the following two steps:

(1) **Induction anchor ($n = n_0$)**: Prove that $A(n_0)$ is a true statement.
(2) **Induction step ($n \to n+1$)**: Prove that for all $n \geq n_0$, the following holds: If $A(n)$ is true, then $A(n+1)$ is also true.

That this implies $A(n)$ is true for all $n \geq n_0$ is easy to see: That $A(n_0)$ holds follows immediately from (1). Repeated application of (2) then yields $A(n_0+1)$, $A(n_0+2)$, $A(n_0+3)$, and so on. Importantly, one does not actually have to carry out this repeated application of (2): what matters is that the proof of (2) guarantees the induction step can be performed *as many times as needed*.

We illustrate the use of the principle with the problem from the previous section and formulate the result as a theorem.

THM
Sum of 1 to n

4.2. **Theorem.** *For all $n \in \mathbb{N}$, the following holds:*

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

Proof. We apply the principle of mathematical induction with $n_0 = 0$ to the statement

$$A(n) := \left[\sum_{k=1}^{n} k = \frac{n(n+1)}{2}\right].$$

Induction anchor: $n = n_0 = 0$. For $n = 0$, the sum is the empty sum, so

$$\sum_{k=1}^{n} k = \sum_{k=1}^{0} k = 0.$$

[1] This may seem excessively fussy for this task, but as soon as we consider more complex problems (some of which will be presented later in this section), intuition quickly fails, and it is no longer clear what the ellipsis actually stands for.

On the other hand,

$$\frac{n(n+1)}{2} = \frac{0 \cdot 1}{2} = \frac{0}{2} = 0.$$

Thus, $A(0)$ is proven.

Induction step: Assume $A(n)$ holds for some $n \in \mathbb{N}$. This assumption is called the *induction hypothesis*. We need to show that $A(n+1)$ also holds, i.e.,

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

From the induction hypothesis, we know

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

We aim to show:

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

Expanding the numerators, this is equivalent to:

$$\frac{n^2+n}{2} + (n+1) = \frac{n^2+3n+2}{2}.$$

Since

$$\frac{n^2+3n+2}{2} = \frac{n^2+n+2n+2}{2} = \frac{n^2+n}{2} + \frac{2n+2}{2} = \frac{n^2+n}{2} + (n+1),$$

the equation is indeed satisfied. ❑

Using the notation we introduced at the beginning of this chapter, the implication proven in the proof of Theorem 4.2 can then be concisely expressed as

$$\text{For all } n \geq n_0: \quad \sum_{k=1}^{n} k = \frac{n(n+1)}{2} \quad \Rightarrow \quad \sum_{k=1}^{n+1} k = \frac{(n+1)(n+2)}{2}.$$

In general, the inductive step can be concisely written as

$$\text{For all } n \geq n_0: \quad A(n) \;\Rightarrow\; A(n+1).$$

4.C Examples for Induction Proofs.

In this section, we will illustrate the principle of mathematical induction with a few example statements, while also proving some results that will be useful later in the course.

We begin with the question of a formula for the sum

$$\sum_{k=0}^{n} x^k,$$

where $x \in \mathbb{K}$, $\mathbb{K}$ an arbitrary field, and we use the convention that $x^0 = 1$. This sum is called a *geometric series* and will be used several times throughout this course. We will consider the case $x \neq 1$—the reason for this will become clear shortly. This is not a significant restriction, since we do not need an induction proof for $x = 1$: in this case, $x^k = 1$ for all k, and we directly obtain

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

To get an idea of what the solution for $x \neq 1$ might look like (since such an "idea" is needed for a proof by induction), let us test small values of k:

$$\sum_{k=0}^{0} x^k = 1, \quad \sum_{k=0}^{1} x^k = 1+x, \quad \sum_{k=0}^{2} x^k = 1+x+x^2, \quad \sum_{k=0}^{3} x^k = 1+x+x^2+x^3, \dots$$

This is not yet very enlightening. The key idea is to expand the right-hand side expressions by multiplying numerator and denominator with $1-x$, which leads to

$$1 \cdot \frac{1-x}{1-x} = \frac{1-x}{1-x}, \qquad (1+x) \cdot \frac{1-x}{1-x} = \frac{1+x-x-x^2}{1-x} = \frac{1-x^2}{1-x},$$

$$(1+x+x^2) \cdot \frac{1-x}{1-x} = \frac{1+x+x^2-x-x^2-x^3}{1-x} = \frac{1-x^3}{1-x},$$

$$(1+x+x^2+x^3) \cdot \frac{1-x}{1-x} = \frac{1+x+x^2+x^3-x-x^2-x^3-x^4}{1-x} = \frac{1-x^4}{1-x}.$$

This suggests the conjecture:

$$\sum_{k=0}^{n} x^k = \frac{1-x^{n+1}}{1-x}$$

(which also explains why $x = 1$ was excluded, since we cannot divide by $1-1=0$). We will now prove this conjecture using induction.

4.3. **Theorem.** *For the geometric series, the following formula holds for any $x \neq 1$ and any $n \in \mathbb{N}$:*

THM Geometric Series

$$\sum_{k=0}^{n} x^k = \frac{1-x^{n+1}}{1-x}.$$

Proof. We prove the statement[2]

$$A(n) := \left[\sum_{k=0}^{n} x^k = \frac{1-x^{n+1}}{1-x} \right]$$

by induction starting at $n_0 = 0$. For this,

$$\sum_{k=0}^{0} x^k = 1 = \frac{1-x}{1-x} = \frac{1-x^1}{1-x} \quad \Rightarrow \quad A(n_0).$$

For the induction step $n \to n+1$, assume $A(n)$ holds for a given $n \in \mathbb{N}$. Then:

$$\sum_{k=0}^{n+1} x^k = \underbrace{\sum_{k=0}^{n} x^k}_{= \frac{1-x^{n+1}}{1-x} \text{ (ind. hyp.)}} + x^{n+1} = \frac{1-x^{n+1}}{1-x} + x^{n+1}$$

$$= \frac{1-x^{n+1}+x^{n+1}(1-x)}{1-x} = \frac{1-x^{n+2}}{1-x},$$

which is exactly $A(n+1)$. ❑

[2] If one has acquired some practice in using mathematical induction, the detailed definition of $A(n)$ in the proof can be omitted, what we will also do in the following. For the start it is nevertheless helpful to write down as precisely as possible what one is aiming to prove.

Our next example is the expansion of the expression

$$(x+y)^n$$

for real numbers $x, y \in \mathbb{R}$ and arbitrary $n \in \mathbb{N}$. The result is known as the *Binomial Theorem*.

To formulate a conjecture that we can use as the basis of an induction proof, we compute for small values of n:

$$\begin{array}{rcl}
(x+y)^0 & = & 1 \\
(x+y)^1 & = & x+y \\
(x+y)^2 & = & x^2+2xy+y^2 \\
(x+y)^3 & = & x^3+3x^2y+3xy^2+y^3 \\
(x+y)^4 & = & x^4+4x^3y+6x^2y^2+4xy^3+y^4
\end{array}$$

If we write the factors in front of the terms (the so-called *coefficients*) without the x's and y's, we obtain

$$\begin{array}{ccccccccc}
 & & & & 1 & & & & \\
 & & & 1 & & 1 & & & \\
 & & 1 & & 2 & & 1 & & \\
 & 1 & & 3 & & 3 & & 1 & \\
1 & & 4 & & 6 & & 4 & & 1 \\
\vdots & & \vdots & & \vdots & & \vdots & & \vdots
\end{array}$$

B. Pascal
1623–1662

The pattern that can now be observed is that the entries on the edges always have the value 1, and each number not on the edge is exactly the sum of the two numbers directly above it. The resulting figure is called *Pascal's Triangle*, named after Blaise Pascal (1623–1662), although it was actually known before Pascal[3].

The entries in Pascal's Triangle are called *binomial coefficients*. We now label the entries in Pascal's Triangle from top to bottom with n and from left to right with k (each starting from 0), and denote the respective binomial coefficient by $\binom{n}{k}$, for example:

$$\binom{2}{1} := 2, \quad \binom{4}{3} := 4,$$

or more generally:

$$\begin{array}{ccccccccc}
 & & & & \binom{0}{0} & & & & \\
 & & & \binom{1}{0} & & \binom{1}{1} & & & \\
 & & \binom{2}{0} & & \binom{2}{1} & & \binom{2}{2} & & \\
 & \binom{3}{0} & & \binom{3}{1} & & \binom{3}{2} & & \binom{3}{3} & \\
\binom{4}{0} & & \binom{4}{1} & & \binom{4}{2} & & \binom{4}{3} & & \binom{4}{4} \\
\vdots & & \vdots & & \vdots & & \vdots & & \vdots
\end{array}$$

Based on the pattern we observe in Pascal's Triangle, we define the binomial coefficients by means of the following the relationships:

$$\binom{n+1}{k+1} = \binom{n}{k} + \binom{n}{k+1} \tag{4.2}$$

for $n \geq 0$ and $k = 0, \ldots, n-1$, and

$$\binom{n}{0} = \binom{n}{n} = 1. \tag{4.3}$$

[3] Wikipedia contains many interesting facts about this, which should of course be viewed with the usual caution applied to Wikipedia entries.

These relationships uniquely determine the binomial coefficients for all $n \geq 0$ and all $k = 0, \ldots, n$.

In the few examples of the binomial formula shown above, the coefficient in front of the term $x_{n-k}y^k$ in the binomial formula is given by $\binom{n}{k}$. The question now is whether this pattern continues for arbitrary large n. Among the examples we had so far, this is the one where it is probably most obvious that an induction proof is needed as there is no simple way to write the desired result with dots. Without a formal proof, there is no way to check whether all powers really satisfy this rule or whether for some n a different rule emerges. Fortunately, the following theorem, the so-called Binomial Theorem, shows that the rule holds for all n.

4.4. **Theorem.** *For all real numbers $x, y \in \mathbb{R}$ and every natural number $n \in \mathbb{N}$:* **THM** Binomial Formula

$$(x+y)^n = \sum_{k=0}^{n} \binom{n}{k} x^{n-k} y^k \tag{4.4}$$

Proof. By induction over n. For $n = 0$, we have

$$(x+y)^0 = 1 \quad \text{and} \quad \sum_{k=0}^{0} \binom{0}{k} x^{n-k} y^k = \binom{0}{0} x^0 y^0 = 1.$$

Thus, the induction anchor is proven.

To prove the induction step, we compute:

$$\begin{aligned}
(x+y)^{n+1} &= (x+y)^n (x+y) \\
\left(\begin{array}{c}\text{Induction}\\ \text{hypothesis}\end{array}\right) &= \left(\sum_{k=0}^{n} \binom{n}{k} x^{n-k} y^k\right)(x+y) \\
&= \sum_{k=0}^{n} \binom{n}{k} x^{n+1-k} y^k + \sum_{k=0}^{n} \binom{n}{k} x^{n-k} y^{k+1} \\
&= \left(x^{n+1} + \sum_{k=1}^{n} \binom{n}{k} x^{n+1-k} y^k\right) + \left(\sum_{k=0}^{n-1} \binom{n}{k} x^{n-k} y^{k+1} + y^{n+1}\right) \\
&= x^{n+1} + \sum_{k=1}^{n} \binom{n}{k} x^{n+1-k} y^k + \sum_{k=1}^{n} \binom{n}{k-1} x^{n+1-k} y^k + y^{n+1} \\
&= x^{n+1} + \sum_{k=1}^{n} \left[\binom{n}{k} + \binom{n}{k-1}\right] x^{n+1-k} y^k + y^{n+1} \\
\text{(by (4.2))} \quad &= \binom{n+1}{0} x^{n+1} + \sum_{k=1}^{n} \binom{n+1}{k} x^{n+1-k} y^k + \binom{n+1}{n+1} y^{n+1} \\
&= \sum_{k=0}^{n+1} \binom{n+1}{k} x^{n+1-k} y^k,
\end{aligned}$$

which proves the statement for $n+1$ and completes the induction step. ❑

What is still missing is a closed formula for the binomial coefficients. To this end, the so-called *factorial* of a natural number n, defined as

$$n! = 1 \cdot 2 \cdot 3 \cdot \ldots \cdot n$$

is helpful.If we introduce a product notation analogous to the summation symbol by defining

$$\prod_{k=m}^{n} a_k := a_m \cdot a_{m+1} \cdot \ldots \cdot a_n,$$

and define the *empty product* as

$$\prod_{k=m}^{m-1} a_k := 1,$$

then we can write the factorial as

$$n! = \prod_{k=1}^{n} k.$$

Using combinatorial arguments, which we omit here for the sake of brevity, one arrives at the conjecture that the binomial coefficients are given by the formula

$$\prod_{j=1}^{k} \frac{n-j+1}{j} = \frac{n \cdot (n-1) \cdot \ldots \cdot (n-k+1)}{1 \cdot 2 \cdot \ldots \cdot k}. \tag{4.5}$$

For $k = 0, \ldots, n$ this can also be written as

$$\prod_{j=1}^{k} \frac{n-j+1}{j} = \frac{n!}{k!(n-k)!},$$

if we use the convention $0! = 1$. In the interest of time, the following induction proof of the correctness of Formula (4.5) is left to the interested reader.

Formula (4.5) yields in particular

$$\frac{n!}{0!(n-0)!} = \frac{n!}{n!} = 1 \quad \text{and} \quad \frac{n!}{n!(n-n)!} = \frac{n!}{n!} = 1 \quad \text{for all } n \in \mathbb{N}, \tag{4.6}$$

i.e., because of (4.3) the formula yields the correct value for the numbers at the edges of Pascal's triangle. We now prove by induction that (4.5) yields the correct value for all entries of Pascal's triangle.

LEMMA
Formula for binomial coefficient

4.5. **Lemma.** [4] *For all $n \in \mathbb{N}$ and $k = 0, \ldots, n$,*

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

Proof. We prove the statement by induction over n. For $n = 0$, there is only the coefficient $\binom{0}{0}$, for which the formula holds because

$$\binom{0}{0} = 1 \quad \text{and} \quad \prod_{j=1}^{0} \frac{n-j+1}{j} = 1$$

(note that we are again using the empty product here).

For the induction step $n \to n+1$, we assume that the formula holds for a given $n \in \mathbb{N}$ and all $k = 0, \ldots, n$. For the elements on the edge, we have already observed that the formula holds due to (4.6). In the induction step, we therefore only need to show that the formula also holds for the non-edge elements of row $n+1$. That is, we must

[4] A lemma in mathematics is an auxiliary statement, which is typically used in the proof of a larger theorem.

prove that (4.2) holds, i.e., that the formula for $n+1$ is exactly equal to the sum of the formulas for the two coefficients directly above it in row n, namely:

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

Indeed, we compute:

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

which confirms the identity and completes the induction step. ❑

As a final example for the use of induction, we show the most well known use of the factorial. It describes, in how many different ways can we arrange the elements of an n-element set $\{E_1, E_2, \ldots, E_n\}$.

We begin by observing: for the first position we have n choices, for the second $n-1$, etc., leading to

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

ways. This is a fact that we can again prove by induction.

4.6. **Theorem.** *The number of possible arrangements of a set with n-elements $\{E_1, E_2, \ldots, E_n\}$ with $n \in \mathbb{N}, n \neq 0$, is $n!$.*

THM
Arrangements of a finite set

Proof. By induction with anchor $n_0 = 1$. For $n = 1$, there's exactly one arrangement: $1! = 1$. Assume the statement holds for n. For a set with $n+1$ elements, there are $n+1$ choices for the first element, and for the remaining n elements, there are $n!$ arrangements (by the induction hypothesis). Thus, in total we obtain

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

which proves the statement for $n+1$. ❑

4.D Further Remarks on the Principle of Induction.

Looking back at the previous sections, one might rightly observe that some arguments are not entirely consistent. We went to great lengths to eliminate the "dots" from the proof of (4.1) using complete induction. However, these dots already appear in the definition of the sum $\sum_{k=m}^{n} a_k$ and even in the definition of the natural numbers earlier in this chapter.

In fact, by using inductive definitions, one can completely avoid the dots. For example, the sum (with m and n as in Definition 4.1) can alternatively be defined inductively as:

$$\sum_{k=m}^{m-1} a_k := 0 \quad \text{and} \quad \sum_{k=m}^{j} a_k := \sum_{k=m}^{j-1} a_k + a_j \quad \text{for } j = m, \ldots, n.$$

The reason we did not use this form from the beginning is simply that the form in Definition 4.1 is unlikely to cause mathematical misunderstandings in this simple case, and is significantly more intuitive.

Similarly, the natural numbers themselves can be defined without "dots" using the following conditions:

(N0) The natural numbers form a set $\mathbb{N}$ that contains a distinguished element "0".
(N1) There is a function ν on $\mathbb{N}$ that assigns to each number $n \in \mathbb{N}$ a number $\nu(n) \in \mathbb{N}$ with $\nu(n) \neq 0$. This function satisfies: if $n_1 \neq n_2$, then $\nu(n_1) \neq \nu(n_2)$ for all $n_1, n_2 \in \mathbb{N}$.
(N2) For every subset N of $\mathbb{N}$ that contains 0 and also contains $\nu(n)$ for every $n \in N$, it follows that $N = \mathbb{N}$.

The number $\nu(n)$ is called the successor of n. If we use the familiar addition from school instead of the abstract function ν, we get $\nu(n) = n + 1$.

The statements (N0)–(N2) are called the *Peano axioms*, named after the Italian mathematician Giuseppe Peano (1858–1932). An *axiom* is generally a condition that cannot be derived from other conditions and thus serves as a building block in the definition of mathematical objects.

It takes some thought to be convinced that the natural numbers we have known since elementary school are uniquely determined by the Peano axioms—and indeed they are, though we will not delve into this aspect for reasons of time. A detailed treatment can be found, for example, in Chapter I.5 of the book Analysis I by Amman and Escher. It is worth noting, however, that the axioms determine only the structure of the natural numbers, not their notation.

In the usual notation of natural numbers with Arabic numerals, we have:

$$\nu(0) = 1, \quad \nu(\nu(0)) = 2, \quad \nu(\nu(\nu(0))) = 3, \quad \text{etc.}$$

But one could also choose entirely different symbols, such as I, II, III, IV, V, ... as in ancient Rome (where zero was not yet known), or 0, 1, 10, 11, 100, ... as in the binary representation of numbers in computers. What matters is that the structure of the set $\mathbb{N}$ is uniquely determined by (N0)–(N2), regardless of the notation. This means that each number in one notation can be uniquely mapped to a number in any other such notation, such that their respective successors also correspond. Axiom (N2) forms the formal foundation for the principle of complete induction, as it ensures that we can indeed "reach" every $n \in \mathbb{N}$ using the induction principle.

Date: March 5, 2026

5.A The order axiom.

Some of the fields that we use in higher mathematics have more properties than only being a field. For instance, well known properties of the rational numbers $\mathbb{K} = \mathbb{Q}$ are the following:

(A1) For every number $a \in \mathbb{K}$, exactly one of the three relations holds:

$$a > 0, \qquad a = 0, \qquad -a > 0.$$

(A2) From $a > 0$ and $b > 0$ it follows that $a + b > 0$ and $ab > 0$.

(A3) For every number $a \in \mathbb{K}$ there exists a natural number $n \in \mathbb{N}$ such that $n - a > 0$.

(A1) and (A2) are called the *order axioms*, (A3) is called the Archimedian axiom. A field that satisfies the axioms (A1), (A2), and (A3) is called an *Archimedean ordered* field.

Numbers with $a > 0$ are called *positive*, numbers with $-a > 0$ are called *negative*

Axiom (A1) can then also be formulated as follows: every number $a \in \mathbb{K}$ is either positive, negative, or equal to zero. "Either ... or" means that a cannot have two of these properties at the same time. Axiom (A2) states in words that sums and products of positive numbers are again positive.

The set of positive numbers is denoted by $\mathbb{K}_+$, that of negative numbers by $\mathbb{K}_-$. Moreover, for $a, b \in \mathbb{K}$ we write:

$a > b$,	if $a - b > 0$	"a is greater than b"
$a \geq b$,	if $a > b$ or $a = b$	"a is greater than or equal to b"
$a < b$,	if $b > a$	"a is less than b"
$a \leq b$,	if $b \geq a$	"a is less than or equal to b"

The requirement in Axiom (A3) can then be rewritten as $n > a$. Thus, the axiom states that for every number $a \in \mathbb{K}$ there must exist a natural number n that is greater than a.

For the rational numbers $\mathbb{K} = \mathbb{Q}$, the positive numbers are

$$\mathbb{K}^+ = \mathbb{Q}^+ := \{p/q \in \mathbb{Q} \,|\, p, q \in \mathbb{N} \setminus \{0\}\}.$$

All the familiar rules of calculation for inequalities can be derived from the axioms. The following theorem gives some examples of statements that can be deduced from (A1) and (A2) alone.

THM Consequences of (A1) and (A2)

5.1. **Theorem.** *For a field $\mathbb{K}$ that satisfies (A1) and (A2), the following holds for arbitrary $a, b, c \in \mathbb{K}$:*

(1) $a > b$ *and* $b > c \Rightarrow a > c$.
(2) $a > b \Leftrightarrow a + c > b + c$.
(3) *If* $c > 0$ *then* $a > b \Rightarrow ac > bc$.
(4) $a > 0 \Rightarrow -a < 0$ *and* $a < 0 \Rightarrow -a > 0$.
(5) *Exactly one of the following statements holds:* $a > b$, $a = b$, $a < b$.
(6) $a > 0 \Rightarrow a + c > c$.
(7) $a \neq 0 \Rightarrow a^2 = a \cdot a > 0$. *In particular,* $1 = 1^2 > 0$.
(8) *For all* $n \in \mathbb{N}$: $a > 0 \Rightarrow a^n := \underbrace{a \cdot a \cdot \ldots \cdot a}_{n \text{ times}} > 0$.
(9) *If* $a > 0$ *and* $b \geq 0$, *then for all* $n \in \mathbb{N} \setminus \{0\}$: $a > b \Leftrightarrow a^n > b^n$.

(10) *If $a \geq -1$, then for all $n \in \mathbb{N}$ the* Bernoulli inequality *holds:*

$$(1+a)^n \geq 1 + na.$$

Proof. (1) From the assumption we have by definition $a - b > 0$ and $b - c > 0$. Thus,

$$a - c = (a - b) + (b - c) > 0$$

by (A2).

(2) By definition of equivalence "$\Leftrightarrow$", the statement to be proved means

$$a > b \;\Rightarrow\; a + c > b + c \quad \text{and} \quad a > b \;\Leftarrow\; a + c > b + c.$$

We prove both implications separately.

"$\Rightarrow$": Assume $a > b$. Then

$$a - b > 0 \;\Rightarrow\; (a - b) + (c - c) = a - b > 0.$$

But $(a - b) + (c - c) = (a + c) - (b + c)$, so $(a + c) - (b + c) > 0$, i.e. $a + c > b + c$.

"$\Leftarrow$": Assume $a + c > b + c$. Then also $a + c + d > b + c + d$ for all $d \in \mathbb{K}$. Choosing $d = -c$, we get $a > b$.

(3) From $a > b$ we have $a - b > 0$. Since $c > 0$, (A2) gives $(a - b)c > 0$, i.e. $ac > bc$.

(4) From (2) with $b = 0$ and $c = -a$, we get $a > 0 \Rightarrow -a < 0$. The second claim follows similarly.

(5) By (A1), exactly one of (i) $a - b > 0$, (ii) $a - b = 0$, or (iii) $a - b < 0$ holds. These correspond respectively to $a > b$, $a = b$, and $a < b$.

(6) This follows from (2) with $b = 0$.

(7) If $a > 0$, then $a^2 > 0$ by (A2). If $a < 0$, then $(-a) > 0$, hence $(-a)^2 > 0$. But $a^2 = (-a)^2$, so $a^2 > 0$.

(8) Proof by induction over n: clear for $n = 0$ since $a^0 = 1 > 0$. For $n \to n + 1$, assume $a^n > 0$, then $a^{n+1} = a^n \cdot a > 0$.

(9) If $b = 0$, then the claim follows from (8). It remains to show the claim for $b > 0$.

"$\Rightarrow$": **Proof Method 1, using the Binomial Theorem:**
For $c := a - b > 0$, we have $a = b + c$, and thus $a^n = (b + c)^n$. By the Binomial Theorem (Theorem 4.4) it follows that

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

From $b > 0$ and $c > 0$, it follows by (A2) and (8) that all terms in the last sum are > 0, and hence, by (A2), the entire sum is > 0. Using (6), we obtain

$$a^n = (b + c)^n = \underbrace{b^n + c^n}_{c \text{ in (6)}} + \underbrace{\sum_{k=1}^{n-1} \binom{n}{k} b^{n-k} c^k}_{a \text{ in (6)}} > b^n + c^n.$$

Since $c > 0$, it follows by (6) that $b^n + c^n > b^n$, and thus the statement holds.

Proof Method 2, by Induction over n:
For $n = 1$, the statement is obvious. Assume it holds for n, i.e. $a^n > b^n$, hence $a^n - b^n > 0$. By assumption $a > b$, and from (8) it follows that $a^n > 0$. Therefore,

$$a^{n+1} = a^n a \overset{(2a)}{>} a^n b.$$

Thus,

$$a^{n+1} - b^{n+1} = a^{n+1} - b^n b \overset{(2)}{>} a^n b - b^n b = (a^n - b^n) b \overset{(A2)}{>} 0,$$

which by definition is equivalent to the inequality $a^{n+1} > b^{n+1}$, as required.

"$\Leftarrow$": For this part of the proof, we use for the first time a proof by contraposition. Instead of proving $a^n > b^n \Rightarrow a > b$, we show the contrapositive statement "$a > b$ does not hold $\Rightarrow$ $a^n > b^n$ does not hold". According to (5), "$a > b$ does not hold" is equivalent to "$a \leq b$", and "$a^n > b^n$ does not hold" is equivalent to "$a^n \leq b^n$". Hence, we must prove the implication

$$a \leq b \;\Rightarrow\; a^n \leq b^n.$$

We distinguish two cases:

In the case $a = b$, it follows that $a^n = b^n$.

In the case $a < b$, it follows from the first part of the proof (with a and b interchanged) that $a^n < b^n$. Together, these two cases establish the desired implication.

(10) Induction over n. Induction anchor $n = 0$: $(1 + a)^0 = 1 = 1 + 0 \cdot a$. For $n \to n + 1$, we use $1 + a \geq 0$ and the induction hypothesis:

$$(1+a)^{n+1} = (1+a)^n(1+a) \geq (1+na)(1+a) = 1+(n+1)a+na^2 \geq 1+(n+1)a.$$

❑

All these properties could be proved using only (A1) and (A2). For the next theorem, we also need (A3).

5.2. **Theorem.** *For a field $\mathbb{K}$ that satisfies (A1)–(A3), the following holds:*

THM
Growth of powers

(a) *For every $b \in \mathbb{K}$ with $b > 1$ and every $M \in \mathbb{K}$, there exists $n \in \mathbb{N}$ with $b^n > M$.*
(b) *For every $c \in \mathbb{K}$ with $0 < c < 1$ and every $\varepsilon \in \mathbb{K}$ with $\varepsilon > 0$, there exists $n \in \mathbb{N}$ with $c^n < \varepsilon$.*

Proof. (a) Let $a = b - 1$, so $b = 1 + a$. By Bernoulli's inequality,

$$b^n = (1 + a)^n \geq 1 + na.$$

By (A3), we can find n such that $n > M/a$, hence $b^n > na > M$.

(b) First we show $c^{-1} > 0$. Suppose not. Then $c^{-1} < 0$, so $-c^{-1} > 0$. Then $(-c^{-1})c > 0$ by (A2), i.e. $-1 > 0$, a contradiction since we know $1 > 0$ and hence $-1 < 0$. Thus $c^{-1} > 0$.

Next, since $c < 1$, we get $c^{-1} > 1$. Now set $b = c^{-1}$ and $K = \varepsilon^{-1}$, and apply part (a). Then $(c^{-1})^n > \varepsilon^{-1}$. Multiplying by $\varepsilon c^n > 0$ gives

$$c^n < \varepsilon.$$

❑

Intuitively, every Archimedean ordered field can be represented by the familiar number line, on which larger numbers are placed further to the right and smaller numbers further to the left. Axiom (A3) then states that no matter how far to the right you go, there will always still be natural numbers. Without the order axioms, such a graphical visualization would not be possible.

5.3. **Remark.** Now that we have introduced all axioms necessary for the rules of calculation, from now on we will use the usual calculation rules without explicitly referring to the individual axioms each time. At some points, however, we will, for completeness, refer back to statements from the theorems just proved, even if they could easily be derived from the usual calculation rules. ♠

REMARK
Use of axioms

To conclude this section, we want to introduce an important concept, the so-called *absolute value*, also called the *modulus*.

5.4. **Definition.** Let $\mathbb{K}$ be an Archimedean ordered field. The absolute value or modulus of a number $a \in \mathbb{K}$ is defined as

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

♢

DEF
Absolute value

Obviously, the absolute value always satisfies $|a| \geq 0$ and $|a| = |-a|$. In addition, the following theorem holds.

5.5. **Theorem.** *In an Archimedean ordered field, the absolute value satisfies for all $a, b \in \mathbb{K}$:*

$$\begin{aligned} |ab| &= |a| \cdot |b|, \\ |a+b| &\leq |a| + |b|, \\ \big||a| - |b|\big| &\leq |a-b|. \end{aligned}$$

The second statement is called the triangle inequality*, the third is called the* reverse triangle inequality*.*

THM
Properties of the absolute value

Proof. For the first statement, one considers all combinations of the cases $a \geq 0$, $a < 0$, and $b \geq 0$, $b < 0$ individually (exercise).

Also, by case distinction, one checks that for every $a \in \mathbb{K}$, always $a \leq |a|$ and $-a \leq |a|$. Hence

$$a + b \leq |a| + |b|, \qquad -(a+b) = (-a) + (-b) \leq |a| + |b|.$$

Since either $|a+b| = a+b$ or $|a+b| = -(a+b)$, the triangle inequality follows.

From the triangle inequality we get

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

and by subtracting $|b|$ from both sides,

$$|a| - |b| \leq |a-b|.$$

The same inequality holds if we swap a and b, giving $|b| - |a| \leq |a-b|$. Since either $\big||a| - |b|\big| = |a| - |b|$ or $\big||a| - |b|\big| = |b| - |a|$, the claimed inequality follows. ❑

The absolute value has a useful interpretation on the number line: $|a - b|$ is precisely the distance between the numbers a and b. We use this for the following definition.

5.6. **Definition.** For two numbers $a, b \in \mathbb{K}$ we define the *distance*

$$d(a,b) := |a-b|. \qquad \diamondsuit$$

DEF
Distance

We will mostly use the more intuitive notation $d(a,b)$ for the distance. From Definition 5.6 and Theorem 5.5 it follows immediately that the distance satisfies, for all $a, b, c \in \mathbb{K}$, the three properties

$$\begin{array}{lrcl} \text{Positive definiteness:} & d(a,b) & \geq & 0 \text{ and } d(a,b) = 0 \text{ iff } a = b, \\ \text{Symmetry:} & d(a,b) & = & d(b,a), \\ \text{Triangle inequality:} & d(a,c) & \leq & d(a,b) + d(b,c). \end{array}$$

A mapping with these three properties is called a *metric*, and a set on which a metric is defined is called a *metric space*. Thus, an Archimedean ordered field with $d(a,b) = |a-b|$ is a metric space. In Analysis II, we will encounter a number of further metric spaces (and another metric on the exercise sheet).

From the triangle inequality, we also obtain the reverse triangle inequality:

$$d(a,b) - d(b,c) \leq d(a,c).$$

5.B The completeness axiom.

In mathematical analysis we use the real numbers, denoted by $\mathbb{R}$, as our basic set of numbers. So far, we have not yet defined them rigorously. The main feature of this set of numbers is that it is *complete*. Actually, the rigorous definition of "completeness" is far from trivial and we will only sketch its meaning here, rather than giving a completely rigorous definition. However, we will come back to this topic later in the lecture, when we have more mathematical tools at hand.

The first thing to realize when talking about completeness is to convince ourselves that the rational numbers, i.e.,

$$\mathbb{Q} = \left\{ \frac{p}{q} \,\middle|\, p \in \mathbb{Z},\, q \in \mathbb{N} \setminus \{0\} \right\}$$

have "holes", i.e., that there important numbers are missing. This may seem paradox at a first glance, because between any two rational numbers p_1/q_1 and p_2/q_2 we can always find another rational number "in the middle", given by

$$\frac{1}{2}\left(\frac{p_1}{q_1} + \frac{p_2}{q_2}\right) = \frac{p_1 q_2 + p_2 q_1}{2 q_1 q_2}.$$

This could lead to the conjecture that between the elements of $\mathbb{Q}$ there is no space for further numbers on the number line. Yet, this conjecture is wrong, as the following example shows.

We want to find a number $x \in \mathbb{Q}$ with $x > 0$, which solves the equation $x^2 = 2$. Let us assume that there exists rational number with this property, i.e., a fraction p/q with $p \in \mathbb{N}$, $q \in \mathbb{N}$, $q \neq 0$ such that

$$\left(\frac{p}{q}\right)^2 = 2$$

holds. We may assume that one of the two number is odd, because otherwise we could divide both p and q by 2 as long as one of the two numbers is not even, anymore. Clearly, this does not change the value of p/q.[5]

The equation above can be rewritten equivalently as

$$\frac{p^2}{q^2} = 2 \quad \Leftrightarrow \quad p^2 = 2q^2.$$

This representation immediately implies that p^2 must be divisible by 2 without remainder, i.e., that p^2 is an even number. This implies that p must be an even number, as well, because the square of an odd number is always an odd number (formally, this follows from the fact that any odd integer number can be written as $2k+1$ for a $k \in \mathbb{Z}$ and that $(2k+1)^2 = 4k^2+4k+1$ is odd, since it is one more than the even number $4k^2+2k = 2(2k^2+k)$).

Euclid
$\sim$ 300 BC

Hence, $p/2$ is an integer number and consequently $(p/2)^2 = p^2/4$ is also an integer number. Since $q^2 = p^2/2$, the number $q^2/2 = p^2/4$ must also be an integer, i.e., q^2 can be divided by 2 without remainder and is thus an even number. With the same argument as for p^2, above, q must be an even number, too. Hence, both p and q are even, which contradicts our assumption that at least one of the two numbers is odd. Hence, an $x \in \mathbb{Q}$ with $x^2 = 2$ cannot exist.

This proof goes back to the ancient Greek mathematician Euclid. Thus, the insight that $x^2 = 2$ is not solvable in $\mathbb{Q}$ is more than 2000 years old.

The idea of the following completeness axiom is to define the field of real numbers $\mathbb{R}$ by "filling in" the gaps in $\mathbb{Q}$, that is, by completing the rational numbers. There are several different ways to formulate this axiom. Here we present an intuitive method going back to Karl Weierstrass (1815–1897). For this we need a few more definitions.

K. Weierstrass
1815–1897

DEF
Intervals

5.7. **Definition.** Let $\mathbb{K}$ be an Archimedean ordered field. For $a, b \in \mathbb{K}$ with $a < b$, we define

$$\begin{aligned}
[a,b] &:= \{x \in \mathbb{K} \,|\, a \le x \le b\} && \text{compact (or also closed) interval} \\
(a,b) &:= \{x \in \mathbb{K} \,|\, a < x < b\} && \text{open interval} \\
[a,b) &:= \{x \in \mathbb{K} \,|\, a \le x < b\} && \text{(right) half-open interval} \\
(a,b] &:= \{x \in \mathbb{K} \,|\, a < x \le b\} && \text{(left) half-open interval}
\end{aligned}$$

The *length* of an interval is defined as $|I| := b - a$. ♢

The difference between the terms "compact" and "closed" is that compact intervals are always of the form $[a,b]$, whereas closed intervals may also take other forms (more on this later). "Compact" is thus the more precise term.

DEF
interval
nesting

5.8. **Definition.** An *interval nesting* is an infinite sequence of compact intervals $I_0, I_1, I_2, I_3, \ldots$, denoted briefly as I_n, $n \in \mathbb{N}$, which satisfies the following two properties:

(I1) $I_{n+1} \subset I_n$ for all $n \in \mathbb{N}$.
(I2) For every $\varepsilon > 0$ there exists an interval I_n with length $|I_n| < \varepsilon$. ♢

[5] In mathematics we often use assumptions that can be met if the problem formulation or the problem data (here the two integers p and q) is suitably modified. One says that such assumptions hold *without loss of generality* (sometimes abbreviated as w.l.o.g.).

With the help of these interval nestings, we can now formulate our final axiom. Here $\mathbb{K}$ is an Archimedean ordered field.

Completeness Axiom: For every interval nesting in $\mathbb{K}$, there exists a number $a \in \mathbb{K}$ which lies in all intervals I_n.

Note that this number a is unique for every interval nesting: Suppose there were two different numbers a and $\tilde{a}$ lying in all intervals. Without loss of generality, let $a < \tilde{a}$. Set $\varepsilon := d(\tilde{a}, a) = |\tilde{a} - a| = \tilde{a} - a > 0$ and consider an interval I_n with $|I_n| < \varepsilon$. Let a_n and b_n denote the lower and upper bounds of I_n. If $a \in I_n$, then $a \geq a_n$ and thus $b_n - a \leq b_n - a_n < \varepsilon$, i.e. $b_n < a + \varepsilon$. It follows that $\tilde{a} = a + \varepsilon > b_n$, so $\tilde{a}$ cannot lie in I_n. Similarly, assuming $\tilde{a} \in I_n$ leads to $a \notin I_n$. Hence, no two distinct numbers can lie in all intervals of an interval nesting.

The set of real numbers is now the field obtained by adjoining precisely those numbers to $\mathbb{Q}$ so that the completeness axiom holds. In other words, we associate a number with each interval nesting (where interval nestings that lead to the same number are of course assigned the same number) and add these to $\mathbb{R}$. The resulting "new" numbers can no longer be written as fractions, since they do not belong to $\mathbb{Q}$. On the other hand, it would be cumbersome to always write down the interval nesting from which a new number arises. One can prove that every real number can be written as a decimal fraction,

$$c_l c_{l-1} \dots c_1, d_1 d_2 d_3 \dots$$

with (possibly) infinitely many digits d_i after the decimal point, where $c_i, d_i \in \{0, \dots, 9\}$ (for the interested students a proof of this fact is given at the end of Section 7). Since this representation is also not particularly convenient to write down, many real numbers are given their own symbols. The unique positive solution of the equation $x^2 = 2$ is denoted (as is well known) by $\sqrt{2}$, just as in general the positive solution of $x^k = y$ is denoted by $\sqrt[k]{y}$. The circle constant π, which also does not lie in $\mathbb{Q}$, has its own symbol due to its great importance in mathematics.

That the above process of completing $\mathbb{Q}$ actually leads to a meaningful field, i.e. that the real numbers $\mathbb{R}$ actually exist and that the familiar calculation rules hold within them, must of course be proved formally. This would, however, require much more time than we have available in this course. A much more detailed—though still not entirely gap-free—presentation can be found, for example, in the book by Amman and Escher. From now on, we will work with the real numbers without explicitly carrying out this proof. The completeness axiom will, however, appear again at several points.

Before we finish this section, we want to answer a question that we passed over silently above, even though it motivated the entire section: do we actually know that by adding numbers according to the completeness axiom we have indeed adjoined a number $x > 0$ with $x^2 = 2$? The following theorem shows that this is indeed the case.

THM
$\mathbb{R}$ contains x with $x^2 = 2$

5.9. **Theorem.** *There exists an interval nesting in $\mathbb{R}$ such that the unique number x that lies in all intervals I_n, $n \in \mathbb{N}$, is positive and satisfies $x^2 = 2$. In particular, there exists $x \in \mathbb{R}$ with $x > 0$ and $x^2 = 2$.*

Proof. We construct an interval nesting $I_n = [a_n, b_n]$ with $a_n, b_n \in \mathbb{Q} \subset \mathbb{R}$ as follows:

(1) Set $a_0 := 1$, $b_0 := 2$, $n := 0$.

(2) Define $c_n := \dfrac{a_n + b_n}{2}$ (this is the midpoint of the interval I_n).
(3) (i) If $c_n^2 \geq 2$, set $a_{n+1} := a_n$ and $b_{n+1} := c_n$;
(ii) otherwise set $a_{n+1} := c_n$ and $b_{n+1} := b_n$.
(4) Set $n := n + 1$ and go to (2).

From the construction we obtain $I_{n+1} \subseteq I_n$, $a_n^2 \leq 2$ and $b_n^2 \geq 2$. Since the length of the intervals is halved at each step and $|I_0| = 2 - 1 = 1$, we furthermore have $|I_n| = \left(\frac{1}{2}\right)^n$. Hence I_n is an interval nesting, because given any $\varepsilon > 0$ we can choose $n \in \mathbb{N}$ with $\left(\frac{1}{2}\right)^n < \varepsilon$ (see Theorem 5.2(b)), which yields $|I_n| < \varepsilon$.

By the completeness axiom there therefore exists a (unique) $x \in \mathbb{R}$ that lies in all I_n, i.e. $a_n \leq x \leq b_n$ for all n. Since $x \geq a_0 = 1$, x is certainly positive. It remains to show that $x^2 = 2$.

From the inequalities $b_n - a_n = |I_n|$, $a_n < b_n \leq b_0 = 2$, $a_n^2 \leq 2 \leq b_n^2$ and $a_n^2 \leq x^2 \leq b_n^2$ we obtain for all $n \in \mathbb{N}$ the estimate

$$d(x^2, 2) = |x^2 - 2| \leq b_n^2 - a_n^2 = (b_n + a_n)(b_n - a_n) \leq 4\,|I_n|. \tag{5.1}$$

It follows that $x^2 = 2$. Indeed, if $x^2 \neq 2$ then $d(x^2, 2) > 0$ and setting $\varepsilon := d(x^2, 2)/4 > 0$ we can choose $n \in \mathbb{N}$ with $|I_n| < \varepsilon$. But then

$$d(x^2, 2) = 4\varepsilon > 4|I_n|,$$

which contradicts (5.1). Therefore $x^2 = 2$. ❑

6. The field of complex numbers

Date:
March 5, 2026

We have seen that the field $\mathbb{R}$ of real numbers has many nice properties. But we cannot solve all polynomial equations

$$x^n + a_{n-1}x^{n-1} + \ldots + a_1 x + a_0 = 0$$

(with $n \geq 1$ and $a_0, a_1, \ldots, a_{n-1} \in \mathbb{R}$) in $\mathbb{R}$.

If n is odd, the *Intermediate Value Theorem* implies that there must always be a solution. The Intermediate Value Theorem will be discussed later in these lectures.

The simplest equation of this kind that has no solution is $x^2 + 1 = 0$: the left hand side is always ≥ 1 and so can never be zero. We will now construct a field extending $\mathbb{R}$, in which this equation has a solution.

In order to see how we can proceed, we assume for a moment that we already have such a field; we denote it $\mathbb{C}$. Then there is a solution $\boldsymbol{i} \in \mathbb{C}$ of the equation above, i.e., we must have $\boldsymbol{i}^2 = -1$. Since $\mathbb{C}$ extends $\mathbb{R}$, we also have all real numbers in $\mathbb{C}$. Given $a, b \in \mathbb{R}$, we can then form the element $a + b\boldsymbol{i} \in \mathbb{C}$. Do we need further elements? To find out, we need to check if the basic arithmetic operations in $\mathbb{C}$ can generate elements that are not of this form. So let $a, b, a', b' \in \mathbb{R}$. Since $\mathbb{C}$ is a field by assumption, we then have

$$(a + b\boldsymbol{i}) \pm (a' + b'\boldsymbol{i}) = (a + a') \pm (b + b')\boldsymbol{i} \qquad \text{and}$$
$$(a + b\boldsymbol{i}) \cdot (a' + b'\boldsymbol{i}) = aa' + ab'\boldsymbol{i} + ba'\boldsymbol{i} + bb'\boldsymbol{i}^2 = (aa' - bb') + (ab' + ba')\boldsymbol{i}\,.$$

Note that we have used $\boldsymbol{i}^2 = -1$ for the multiplication. What about division? It suffices to consider multiplicative inverses. We first show that $a + b\boldsymbol{i} = 0$ if and only if $a = b = 0$. The direction "$\Leftarrow$" is clear. For the converse, assume that $a + b\boldsymbol{i} = 0$. Then we have

$$0 = (a - b\boldsymbol{i}) \cdot 0 = (a - b\boldsymbol{i}) \cdot (a + b\boldsymbol{i}) = a^2 + b^2\,.$$

Since a and b are real numbers, this is only possible when $a = b = 0$. So we can assume that a and b are not both zero. Then we should have (using the old trick how one gets rid of square roots in the denominator; note that $\boldsymbol{i}$ is a square root of -1)

$$\frac{1}{a + b\boldsymbol{i}} = \frac{a - b\boldsymbol{i}}{(a - b\boldsymbol{i})(a + b\boldsymbol{i})} = \frac{a - b\boldsymbol{i}}{a^2 + b^2} = \frac{a}{a^2 + b^2} + \frac{-b}{a^2 + b^2}\boldsymbol{i}\,,$$

which is again of the form $x + y\boldsymbol{i}$. So it looks like we do not need further elements.

For a rigorous construction of the field $\mathbb{C}$ we replace the expression $a + b\boldsymbol{i}$ (with $a, b \in \mathbb{R}$) by the pair $(a, b) \in \mathbb{R} \times \mathbb{R}$. So we declare $\mathbb{C}$ to be the set $\mathbb{R} \times \mathbb{R}$, and we define the further data of a field as follows-

$$+_{\mathbb{C}} \colon \mathbb{C} \times \mathbb{C} \longrightarrow \mathbb{C}, \qquad \big((a, b), (a', b')\big) \longmapsto (a + a', b + b')$$
$$\cdot_{\mathbb{C}} \colon \mathbb{C} \times \mathbb{C} \longrightarrow \mathbb{C}, \qquad \big((a, b), (a', b')\big) \longmapsto (aa' - bb', ab' + ba')$$
$$0_{\mathbb{C}} = (0, 0)$$
$$1_{\mathbb{C}} = (1, 0)$$
$$-_{\mathbb{C}} \colon \mathbb{C} \longrightarrow \mathbb{C}, \qquad (a, b) \longmapsto (-a, -b)$$
$$i_{\mathbb{C}} \colon \mathbb{C} \setminus \{(0, 0)\} \longrightarrow \mathbb{C} \setminus \{(0, 0)\}, \qquad (a, b) \longmapsto \left(\frac{a}{a^2 + b^2}, \frac{-b}{a^2 + b^2}\right)$$

6.1. **Theorem.** *The set $\mathbb{C} = \mathbb{R} \times \mathbb{R}$ together with the maps and elements defined above forms a field.*

THM
the field $\mathbb{C}$

Proof. We need to verify the various axioms. To show that $(\mathbb{C}, +_{\mathbb{C}}, 0_{\mathbb{C}}, -_{\mathbb{C}})$ is a commutative group is rather easy; we leave that part as an exercise. We have to verify the associativity and commutativity of multiplication. To do this, we use that $\mathbb{R}$ is already known to be a field, so that we can use the usual rules with real numbers.

$$\begin{aligned}\big((a,b) \cdot_{\mathbb{C}} (a',b')\big) \cdot_{\mathbb{C}} (a'',b'') &= (aa' - bb', ab' + a'b) \cdot_{\mathbb{C}} (a'',b'') \\ &= \big((aa' - bb')a'' - (ab' + ba')b'', (aa' - bb')b'' + (ab' + ba')a''\big) \\ &= (aa'a'' - ab'b'' - ba'b'' - bb'a'', aa'b'' + ab'a'' + ba'a'' - bb'b'')\end{aligned}$$

and we obtain the same result from $(a,b) \cdot_{\mathbb{C}} \big((a',b') \cdot_{\mathbb{C}} (a'',b'')\big)$. Similarly,

$$(a,b) \cdot_{\mathbb{C}} (a',b') = (aa' - bb', ab' + ba') = (a'a - b'b, ba' + ab') = (a',b') \cdot_{\mathbb{C}} (a,b)\,.$$

We easily see that $1_{\mathbb{C}} = (1,0)$ is the neutral element of multiplication:

$$(1,0) \cdot_{\mathbb{C}} (a,b) = (1 \cdot a - 0 \cdot b, 1 \cdot b + 0 \cdot a) = (a,b)\,.$$

Now we check that $i_{\mathbb{C}}\big((a,b)\big)$ is the multiplicative inverse of $(a,b) \neq (0,0)$:

$$\begin{aligned}(a,b) \cdot_{\mathbb{C}} i_{\mathbb{C}}\big((a,b)\big) &= (a,b) \cdot_{\mathbb{C}} \left(\frac{a}{a^2+b^2}, \frac{-b}{a^2+b^2}\right) \\ &= \left(\frac{a^2}{a^2+b^2} - \frac{-b^2}{a^2+b^2}, \frac{-ab}{a^2+b^2} + \frac{ba}{a^2+b^2}\right) \\ &= (1,0) = 1_{\mathbb{C}}\,.\end{aligned}$$

$0_{\mathbb{C}} \neq 1_{\mathbb{C}}$ is clear. It remains to verify the distributive law:

$$\begin{aligned}(a,b) \cdot_{\mathbb{C}} \big((a',b') +_{\mathbb{C}} (a'',b'')\big) &= (a,b) \cdot_{\mathbb{C}} (a' + a'', b' + b'') \\ &= \big(a(a' + a'') - b(b' + b''), a(b' + b'') + b(a' + a'')\big) \\ &= (aa' + aa'' - bb' - bb'', ab' + ab'' + ba' + ba'') \\ &= (aa' - bb' + aa'' - bb'', ab' + ba' + ab'' + ba'') \\ &= (aa' - bb', ab' + ba') +_{\mathbb{C}} (aa'' - bb'', ab'' + ba'') \\ &= (a,b) \cdot_{\mathbb{C}} (a',b') +_{\mathbb{C}} (a,b) \cdot_{\mathbb{C}} (a'',b'')\,.\end{aligned}$$

❑

For a real number a we have the element $a_{\mathbb{C}} = (a,0) \in \mathbb{C}$. For $a, b \in \mathbb{R}$ we then have

$$a = b \iff a_{\mathbb{C}} = b_{\mathbb{C}}\,, \qquad (a+b)_{\mathbb{C}} = a_{\mathbb{C}} +_{\mathbb{C}} b_{\mathbb{C}} \qquad \text{and} \qquad (ab)_{\mathbb{C}} = a_{\mathbb{C}} \cdot_{\mathbb{C}} b_{\mathbb{C}}\,.$$

This means that we compute with the elements $a_{\mathbb{C}}$ in exactly the same way as with the corresponding real numbers a. We therefore do not distinguish between a and $a_{\mathbb{C}}$ and thus consider $\mathbb{R}$ as a subset of $\mathbb{C}$. So we simply write a for $a_{\mathbb{C}} = (a,0) \in \mathbb{C}$. Also, we will (mostly) just write $+$, $\cdot$ etc. instead of $+_{\mathbb{C}}$, $\cdot_{\mathbb{C}}$ etc.

6.2. **Definition.** The field $\mathbb{C}$ introduced in Theorem 6.1 is the *field of complex numbers*. We write $\boldsymbol{i}$ for the element $(0,1) \in \mathbb{C}$. Then we have $\boldsymbol{i}^2 = -1$, and every element $z = (a,b) \in \mathbb{C}$ can be written as $z = a + b\boldsymbol{i}$ (or $a + \boldsymbol{i}b$) with $a, b \in \mathbb{R}$. Then a is the *real part* $\operatorname{Re} z$ and b is the *imaginary part* $\operatorname{Im} z$ of z. If $\operatorname{Re} z = 0$, then z is *purely imaginary*. ♢

DEF field of complex numbers

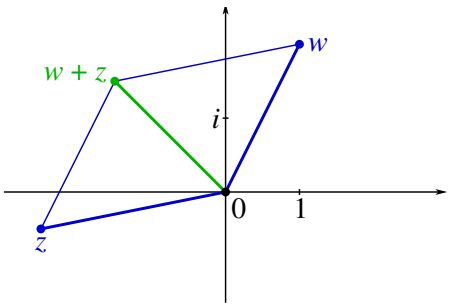
We should check the claims made in the definition:

$$\boldsymbol{i}^2 = (0,1) \cdot_{\mathbb{C}} (0,1) = (0 \cdot 0 - 1 \cdot 1, 0 \cdot 1 + 1 \cdot 0) = (-1,0) = (-1)_{\mathbb{C}} = -1$$

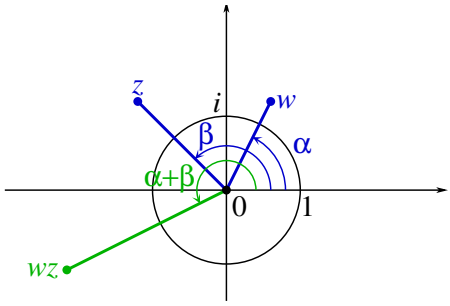
and

$$a + b\boldsymbol{i} = (a,0) +_{\mathbb{C}} (b,0) \cdot_{\mathbb{C}} (0,1) = (a,0) +_{\mathbb{C}} (0,b) = (a,b)\,.$$

We obtain a way to visualize complex numbers by recalling that $\mathbb{C} = \mathbb{R} \times \mathbb{R} = \mathbb{R}^2$ can be seen as the set of points in the plane. Then every point in the plane represents a complex number (in this context, one calls it the "complex plane"). The addition of complex numbers then corresponds to the "parallelogram of force" known from physics.



Addition $w + z$



Multiplication $w \cdot z$

Multiplication of complex numbers can also be interpreted in a geometric way. We consider $z = a + b\boldsymbol{i} \in \mathbb{C}$. Then $a^2 + b^2 \geq 0$; one defines $|z| = \sqrt{a^2 + b^2}$ and calls it the *absolute value* of z. It gives the distance of the point z in the complex plane from the origin $0 \in \mathbb{C}$. If $z \in \mathbb{R}$ (so $b = 0$), this recovers the usual absolute value on $\mathbb{R}$. If $z \neq 0$, then $w = z/|z|$ has absolute value 1. Writing $w = u + v\boldsymbol{i}$, we then have $u^2 + v^2 = 1$: the point (u,v) lies on the unit circle. There is then an angle $\alpha \in \mathbb{R}$ such that $u = \cos\alpha$, $v = \sin\alpha$. This angle α is called the *argument* of w and of z. (Note that the argument is only defined up to multiples of $2\pi \mathrel{\hat{=}} 360°$.) We have the relation

DEF absolute value $|z|$

DEF argument

$$(\cos\alpha + \boldsymbol{i}\sin\alpha) \cdot (\cos\beta + \boldsymbol{i}\sin\beta) = \cos(\alpha+\beta) + \boldsymbol{i}\sin(\alpha+\beta)\,;$$

this is equivalent to the addition identities for sine and cosine,

$$\cos(\alpha+\beta) = \cos\alpha\cos\beta - \sin\alpha\sin\beta, \qquad \sin(\alpha+\beta) = \cos\alpha\sin\beta + \sin\alpha\cos\beta\,.$$

This shows that the arguments (angles) add under multiplication. We see that multiplication by $z \neq 0$ effects a *spiral similarity* of the complex plane with angle of rotation α and dilation factor $|z|$.

We will later see (when discussing the complex exponential function) that

$$\cos\alpha + \boldsymbol{i}\sin\alpha = e^{\alpha\boldsymbol{i}}\,.$$

The relation above is then a consequence of $e^{x+y} = e^x \cdot e^y$.

We can now at least prove that one can solve all quadratic equations in $\mathbb{C}$.

6.3. **Theorem.** *Let $a, b, c \in \mathbb{C}$ such that $a \neq 0$. Then the equation*

$$az^2 + bz + c = 0$$

has at least one solution $z \in \mathbb{C}$.

THM quadratic equations in $\mathbb{C}$

Proof. The equation is equivalent with $(2az + b)^2 = b^2 - 4ac$. It is therefore sufficient to show that every complex number has a square root in $\mathbb{C}$. (If we have $z' \in \mathbb{C}$ such that $z'^2 = b^2 - 4ac$, then $z = (-b + z')/(2a)$ is a solution of the equation. This is the well-known solution formula for quadratic equations.) So let $w \in \mathbb{C}$. We want to find $z \in \mathbb{C}$ such that $z^2 = w$. If $w = 0$, then $z = 0$ is a solution. Otherwise we can write w as $w = |w|(\cos\alpha + \boldsymbol{i}\sin\alpha)$, where α is the argument of w. Then $z = \sqrt{|w|}\big(\cos\frac{\alpha}{2} + \boldsymbol{i}\sin\frac{\alpha}{2}\big)$ is a solution. ❑

In basically the same way, one can show that for every complex number w and every natural number $n \geq 1$ there is an nth root $z \in \mathbb{C}$ of w, i.e., a solution of the equation $z^n = w$. In fact, much more is true.

6.4. **Theorem.** *Every equation*

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

with $n \geq 1$ and $a_0, a_1, \ldots, a_{n-1} \in \mathbb{C}$ has at least one solution $z \in \mathbb{C}$.

THM Fundamental Theorem of Algebra

We don't have the prerequisites to be able to give a proof at this point. There are various different proofs; see the Wikipedia entry.

Equations of the general form above with $n \leq 4$ can be solved by extracting square roots and cube roots (this was already discovered in the 16th century by del Ferro, Tartaglia and Ferrari); equations with $n \geq 5$, however, can no longer solved using the basic operations of arithmetic and the extraction of arbitrary mth roots (Abel-Ruffini Theorem; the first complete proof is due to Abel in 1824). The statement of Theorem 6.4 is therefore much stronger than the existence of nth roots in $\mathbb{C}$.

A field K with the property that every equation

$$x^n + a_{n-1}x^{n-1} + \ldots + a_1 x + a_0 = 0$$

with $n \geq 1$ and $a_0, a_1, \ldots, a_{n-1} \in K$ has a solution $x \in K$ is called *algebraically closed*. The "Fundamental Theorem of Algebra" can therefore also phrased as

DEF algebraically closed

The field of complex numbers is algebraically closed.

In contrast, the field of real numbers is *not* algebraically closed as we have seen. In this respect, $\mathbb{C}$ is clearly "better" than $\mathbb{R}$. On the other hand, $\mathbb{C}$ is not an *ordered* field (and cannot be; see below), so we also lose something when passing from $\mathbb{R}$ to $\mathbb{C}$. (Note that in an ordered field K we have $x^2 \geq 0$ for every $x \in K$. In $\mathbb{C}$ this would imply $-1 = \boldsymbol{i}^2 \geq 0$, which contradicts $1 > 0$.)

In addition to $\boldsymbol{i}$, the equation $x^2 + 1 = 0$ has also the solution $-\boldsymbol{i}$. We can therefore replace $\boldsymbol{i}$ by $-\boldsymbol{i}$ everywhere, and everything would work in the same way. For $z = a + b\boldsymbol{i} \in \mathbb{C}$ we therefore set $\bar{z} = a - b\boldsymbol{i}$; the map $\mathbb{C} \to \mathbb{C}$, $z \mapsto \bar{z}$, is called *complex conjugation*. We have $\overline{w + z} = \bar{w} + \bar{z}$ and $\overline{wz} = \bar{w} \cdot \bar{z}$ (easy exercise); also $z\bar{z} = a^2 + b^2 = |z|^2$ (we already used that). This gives the formula $z^{-1} = \bar{z}/|z|^2$ for the reciprocal of a complex number $z \neq 0$; it is the same expression that we derived earlier, but in a more compact form. Complex conjugation has the following additional properties.

$$z \in \mathbb{R} \iff z = \bar{z}\,, \qquad \operatorname{Re} z = \frac{z + \bar{z}}{2}\,, \qquad \operatorname{Im} z = \frac{z - \bar{z}}{2\boldsymbol{i}}\,.$$

7. Countability

Date:
March 5, 2026

To conclude this chapter, we want to briefly address the question of how many real numbers actually exist. The obvious answer "infinitely many" is of course correct, but we want to investigate this more precisely here. In particular, we want to compare the size of the set $\mathbb{R}$ with the size of the sets $\mathbb{N}$, $\mathbb{Z}$, and $\mathbb{Q}$. To do so, we define a notion of size for infinite sets.

DEF
Countability

7.1. **Definition.** A set A with infinitely many elements is called *countable* if we can assign a natural number $n \in \mathbb{N}$ to each element $a \in A$ such that no n occurs more than once.

If this is not possible, it is called *uncountable*. ♢

This definition formalizes something quite intuitive, namely the process of numbering or counting the elements of the set A. It is obvious that $\mathbb{N}$ is countable, since each $n \in \mathbb{N}$ can simply be assigned to itself. It is somewhat less obvious that $\mathbb{Z}$ is also countable, because at first glance one might think that $\mathbb{Z}$ has about twice as many elements as $\mathbb{N}$. However, this argument only applies to finite sets; with infinite sets, one can use certain tricks that are not possible in the finite case. For $\mathbb{Z}$, this trick consists of arranging the elements as follows:

$$0, -1, 1, -2, 2, -3, 3, \ldots$$

In this arrangement, every $z \in \mathbb{Z}$ eventually appears, and if we number the elements from left to right, we obtain exactly the desired assignment.

For the rational numbers $\mathbb{Q}$, it seems even more surprising that a counting can be found, since for every natural number n there are infinitely many rational numbers (which follows from the simple fact that between any two natural numbers there are infinitely many rationals). Nevertheless, it is possible if we arrange the rational numbers p/q as follows and number them according to the directions of the arrows:

$$\begin{array}{ccccccccc}
\frac{0}{1} & \rightarrow & \frac{-1}{1} & & \frac{1}{1} & \rightarrow & \frac{-2}{1} & & \frac{2}{1} & \rightarrow & \cdots \\
 & \swarrow & & \nearrow & & \swarrow & & \nearrow & & \swarrow & \\
\frac{0}{2} & & \frac{-1}{2} & & \frac{1}{2} & & \frac{-2}{2} & & \frac{2}{2} & & \cdots \\
\downarrow & \nearrow & & \swarrow & & \nearrow & & \swarrow & & \nearrow & \\
\frac{0}{3} & & \frac{-1}{3} & & \frac{1}{3} & & \frac{-2}{3} & & \frac{2}{3} & & \cdots \\
 & \swarrow & & \nearrow & & \swarrow & & \nearrow & & \swarrow & \\
\frac{0}{4} & & \frac{-1}{4} & & \frac{1}{4} & & \frac{-2}{4} & & \frac{2}{4} & & \cdots \\
\downarrow & \nearrow & & \swarrow & & \nearrow & & \swarrow & & \nearrow & \\
\frac{0}{5} & & \frac{-1}{5} & & \frac{1}{5} & & \frac{-2}{5} & & \frac{2}{5} & & \cdots \\
\vdots & & \vdots & & \vdots & & \vdots & & \vdots & & \ddots
\end{array}$$

Here, from left to right, all possible numerators $p \in \mathbb{Z}$ are arranged, and from top to bottom all possible denominators $q \in \mathbb{N} \setminus \{0\}$. In this way, all possible rational numbers appear, so none are missing from the table. The enumeration is then obtained by counting along the diagonals. The fact that many of the fractions that appear have the same value does not matter; if we were to omit these, the set to be counted would only become smaller.

The sets $\mathbb{N}$, $\mathbb{Z}$, and $\mathbb{Q}$ are therefore countable — but not $\mathbb{R}$. More precisely, the following theorem holds.

THM
$\mathbb{R}$ is uncountable

7.2. **Theorem.** *Every compact interval $[a,b] \subset \mathbb{R}$ with $a < b$ contains uncountably many numbers.*

Proof. We proceed by contradiction: assume there exists an enumeration $[a,b] = \{x_1, x_2, x_3, \ldots\}$. We now construct the following nested sequence of intervals I_n.

We begin with the interval $I_0 = [a,b]$ and define the intervals I_n for $n \geq 1$ inductively as follows: for each $n \in \mathbb{N}$ we subdivide $I_n = [a_n, b_n]$ into three subintervals of equal length:

$$\left[a_n, a_n + \tfrac{1}{3}(b_n - a_n)\right] \quad \left[a_n + \tfrac{1}{3}(b_n - a_n), a_n + \tfrac{2}{3}(b_n - a_n)\right] \quad \left[a_n + \tfrac{2}{3}(b_n - a_n), b_n\right].$$

If the number x_{n+1} is less than or equal to $(a_n + b_n)/2 = a_n + (b_n - a_n)/2$, then it is not contained in the third subinterval, and we choose I_{n+1} to be precisely the third subinterval. Otherwise, x_{n+1} does not lie in the first subinterval, and we set I_{n+1} to be the first subinterval.

In this way, we obtain a nested sequence of intervals, since each interval is contained in the previous one, and the length of the n-th interval I_n is $(1/3)^n$, which becomes smaller than any $\varepsilon > 0$. By the completeness axiom, there exists an $s \in \mathbb{R}$ that is contained in all intervals I_n. By construction, however, we have $x_1 \notin I_1$, $x_2 \notin I_2$, and so on. Thus $s \neq x_n$ for all $n \in \mathbb{N}$. Hence the above enumeration is not complete, giving the desired contradiction. ❑

Therefore, even in the smallest subinterval of $\mathbb{R}$, there are strictly more numbers than in the entire set $\mathbb{Q}$. In fact, the set $\mathbb{R} \setminus \mathbb{Q}$ is also uncountable, because if it were countable, then so would be the union $\mathbb{R} = \mathbb{Q} \cup (\mathbb{R} \setminus \mathbb{Q})$.

Can we conclude from this that there are regions on the real line in which no rational numbers lie? That would certainly contradict intuition, since as we observed at the beginning of Section 5.B, between any two rational numbers there is another rational number. And indeed, such regions cannot exist, for between any two real numbers there is always a rational number, as the following theorem shows.

THM
$\mathbb{Q}$ is dense in $\mathbb{R}$

7.3. **Theorem.** *For every $x \in \mathbb{R}$ and every $\varepsilon > 0$ there exists an $r \in \mathbb{Q}$ with $r \in (x - \varepsilon, x]$, i.e. in particular $d(x,r) < \varepsilon$.*

Proof. Choose $n \in \mathbb{N}$ such that $1/n < \varepsilon$. Let A be the set of integers $> nx$. This set is bounded from below by $s = nx$ and therefore has a minimal element m (we will learn more about minimal elements later). It follows that $m > nx$, and by minimality also $m - 1 \leq nx$ (otherwise $m - 1 \in A$ and m would not be minimal). It follows that $m/n > x$ and $(m-1)/n \leq x$, and thus

$$x \geq \frac{m-1}{n} = \frac{m}{n} - \frac{1}{n} > x - \varepsilon.$$

Hence for $r = (m-1)/n \in \mathbb{Q}$ the inequalities $r \leq x$ and $r > x - \varepsilon$ hold, i.e. $r \in (x - \varepsilon, x]$, and $d(x,r) = |x - r| = x - r < \varepsilon$. ❑

This theorem shows that there is always a rational number arbitrarily close to any real number. One says that $\mathbb{Q}$ is *dense* in $\mathbb{R}$. This property justifies, for example, the fact that computers often work only with rational numbers (and not even all of them), since any real number can be approximated arbitrarily well by a rational number.

With a small variation of the above proof one can also show that every real number can be written as a decimal fraction[6]

$$\pm c_l c_{l-1} \dots c_1 . d_1 d_2 d_3 \dots$$

with (possibly) infinitely many decimal places d_i, where $c_i, d_i \in \{0, \dots, 9\}$. We demonstrate this for positive real numbers $x > 0$. To this end, in the above proof we set

$$\varepsilon = 10^{-k_1} = 0, \underbrace{00\dots0}_{k_1-1 \text{ places}} 1$$

and $n = 10^{k_1}$ for some arbitrary $k_1 \geq 1$ with $10^{-k_1} < x$. This yields a rational number of the form

$$r_1 = \frac{m-1}{10^{k_1}} = c_l^1 c_{l-1}^1 \dots c_1^1 . d_1^1 d_2^1 d_3^1 \dots d_{k_1}^1,$$

i.e. a decimal number with at most k_1 decimal places, which is positive since $x - r_1 < 10^{-k_1}$ implies $r_1 > x - 10^{-k_1} > 0$.

Now if $r_1 = x$ or $r_1 + 10^{-k_1} = x$ then we are done, because we have found the decimal fraction representing x (which in this case is finite).

Otherwise, by the construction in the proof, $r_1 < x < r_1 + 10^{-k_1}$, i.e. $0 < x - r_1 < 10^{-k_1}$. Choosing $k_2 \geq k_1 + 1$ with $10^{-k_2} < x - r_1$, we obtain a new finite decimal fraction r_2 with at most k_2 decimal places, for which $r_2 \leq x \leq r_2 + 10^{-k_2}$. If $r_2 = x$, we are again finished; otherwise we have

$$x - r_2 \leq 10^{-k_2} < x - r_1 \quad \Rightarrow \quad r_2 > r_1$$

and

$$r_2 - r_1 = (r_2 - x) + (x - r_1) < 10^{-k_1}.$$

Thus $r_2 - r_1$ has the form

$$r_2 - r_1 = 0, \underbrace{00\dots0}_{k_1 \text{ places}} \tilde{d}_{k_1+1} \tilde{d}_{k_1+2} \dots \tilde{d}_{k_2},$$

and therefore

$$\begin{aligned} r_2 = r_1 + (r_2 - r_1) &= c_l^1 c_{l-1}^1 \dots c_1^1 . d_1^1 d_2^1 d_3^1 \dots d_{k_1}^1 \tilde{d}_{k_1+1} \tilde{d}_{k_1+2} \dots \tilde{d}_{k_2} \\ &=: c_l^2 c_{l-1}^2 \dots c_1^2 . d_1^2 d_2^2 d_3^2 \dots d_{k_1}^2 d_{k_1+1}^2 d_{k_1+2}^2 \dots d_{k_2}^2. \end{aligned}$$

So the integer part and the first k decimal places of both numbers agree, i.e. r_2 arises from r_1 by appending $k_2 - k_1$ further decimal places.

Continuing this procedure produces decimal fractions $r_1, r_2, r_3, \dots$ of the form

$$r_j = c_l^j c_{l-1}^j \dots c_1^j . d_1^j d_2^j d_3^j \dots d_{k_j}^j,$$

so that either $r_j = x$ for some $j \in \mathbb{N}$, in which case x is a finite decimal (and thus a rational number), or we obtain ever longer decimal fractions, where the digits of the shorter fractions agree with those of the longer ones. Since $k_j \geq j$, at least j decimal places are fixed for each r_j in the construction. Defining the infinite decimal fraction

$$x' := c_l^1 c_{l-1}^1 \dots c_1^1 , d_1^1 d_2^2 d_3^3 \dots,$$

the integer part and the first k_j decimal places of x' agree with those of r_j. Hence,

$$|x' - r_j| = x' - r_j = 0, \underbrace{00\dots0}_{k_j \text{ places}} d_{k_j+1}^{k_j+1} d_{k_j+2}^{k_j+2} \dots \leq 10^{-k_j},$$

and therefore for all $j \geq 1$

$$|x' - x| \leq |x' - r_j| + |r_j - x| \leq 2 \cdot 10^{-k_j}.$$

Since $k_j \geq j$, the right-hand side becomes arbitrarily small as j becomes larger and larger, and so the inequality can only hold if $|x' - x| = 0$, i.e. $x = x'$.

[6] In this construction, we assume familiarity with decimal fractions from school. Formally, we will define them in Example 9.2 later in this lecture.

8. Real sequences

Date:
March 5, 2026

8.A Definition and Examples.

A real sequence is an infinite collection of real numbers a_0, a_1, a_2, ..., such that to every natural number $n \in \mathbb{N}$ exactly one number $a_n \in \mathbb{R}$ is assigned.

A sequence is written as $(a_n)_{n\in\mathbb{N}}$ or $(a_0, a_1, a_2, \ldots)$. Sometimes it is useful or necessary to begin the numbering at some $n_0 > 0$; in such cases we will always explicitly mention it.

EXAMPLES
different sequences

8.1. **Examples.**

(a) For $a_n = a \in \mathbb{R}$ for all $n \in \mathbb{N}$ we obtain the *constant sequence* $(a, a, a, a, \ldots)$.
(b) Let $a_n = 1/n$ for all $n \in \mathbb{N}$ with $n \geq 1$. The resulting sequence is $(1, 1/2, 1/3, 1/4, \ldots)$.
(c) For $a_n = (-1)^n$ we obtain $(1, -1, 1, -1, 1, \ldots)$.
(d) For $a_n = n/(n+1)$ we obtain $(0, 1/2, 2/3, 3/4, \ldots)$.
(e) For $a_n = n/2^n$ we obtain $(0, 1/2, 1/2, 3/8, 1/4, 5/32, \ldots)$.
(f) Any nested interval construction in the sense of Definition 5.8 is uniquely defined by the two sequences $(a_0, a_1, a_2, \ldots)$ and $(b_0, b_1, b_2, \ldots)$.
(g) For $a_n = b^n$ and arbitrary $b \in \mathbb{R}$ we obtain $(1, b, b^2, b^3, \ldots)$.
(h) Sequences are often defined recursively by a rule. For example, with $a_0 = 1$, $a_1 = 1$ and the rule $a_n = a_{n-2} + a_{n-1}$ for $n \geq 2$, we obtain the sequence

$$(1, 1, 2, 3, 5, 8, 13, 21, \ldots).$$

This is the so-called Fibonacci sequence.

♣

Many applications of mathematics lead to sequences. The following example introduces some (simple) applications that we will later define more precisely.

EXAMPLES
sequences from applications

8.2. **Examples.**

(a) **(Water tank)** A water tank can be filled or emptied through valves. On the tank there is a scale for the water level, where the mark 0 indicates the desired level. A technician programs a computer to measure the water level every second. If it is too high, the outlet valve is automatically opened; if it is too low, the inlet valve is opened. This produces a sequence of water levels $(a_0, a_1, a_2, \ldots)$.
(b) **(Square root calculation)** A popular algorithm for computing the square root $\sqrt{x}$ for $x \in \mathbb{R}$, $x > 0$, works as follows: we choose any $a_0 > 0$ and compute a_{n+1} from a_n recursively using the formula

$$a_{n+1} = \frac{1}{2}\left(a_n + \frac{x}{a_n}\right). \tag{8.1}$$

This generates a sequence $(a_0, a_1, a_2, \ldots)$.
(c) **(Market equilibrium)** A farmer harvests an amount a_0 kg of potatoes and realizes that the market price per kilogram of potatoes is far too low to cover his costs. He decides to plant less the next year, so that the harvest is a_1 kg with $a_1 < a_0$. At the end of this year he realizes that the market price has risen considerably and decides to increase the amount again the following year to a_2 kg with $a_2 > a_1$. If he adjusts the amount for the following year in this way depending on the market price each year, a sequence $(a_0, a_1, a_2, \ldots)$ is created.

♣

The sequences from the examples will be taken up again later and analyzed in detail.

8.B Convergence.

DEF Convergent sequence

8.3. **Definition.** A real sequence $(a_n)_{n\in\mathbb{N}}$ is called *convergent* if there exists an $a \in \mathbb{R}$ such that:

For every $\varepsilon > 0$ there exists an $N(\varepsilon) \in \mathbb{N}$ such that the inequality $d(a_n, a) < \varepsilon$ holds for all $n \geq N(\varepsilon)$.

In this case we write

$$\lim_{n\to\infty} a_n = a \quad \text{or} \quad a_n \to a \text{ as } n \to \infty$$

(spoken: a_n converges to a).

The value a is then called the *limit* of the sequence $(a_n)_{n\in\mathbb{N}}$. If $a = 0$ we call $(a_n)_{n\in\mathbb{N}}$ a *null sequence*. ♢

A slightly different formulation of this definition is obtained if we consider the interval $(a-\varepsilon, a+\varepsilon)$ for $\varepsilon > 0$. This is called the *ε-neighborhood of a*. From the definition of intervals and absolute value it follows that

$$a_n \in (a-\varepsilon, a+\varepsilon) \;\Leftrightarrow\; a_n > a-\varepsilon \text{ and } a_n < a+\varepsilon \;\Leftrightarrow\; d(a_n, a) < \varepsilon.$$

The convergence condition from Definition 8.3 can therefore also be written as:

For every $\varepsilon > 0$ there exists an $N(\varepsilon) \in \mathbb{N}$ such that all terms a_n with $n \geq N(\varepsilon)$ lie in the ε-neighborhood $(a-\varepsilon, a+\varepsilon)$.

Thus we can illustrate convergence graphically as in Figure 1.

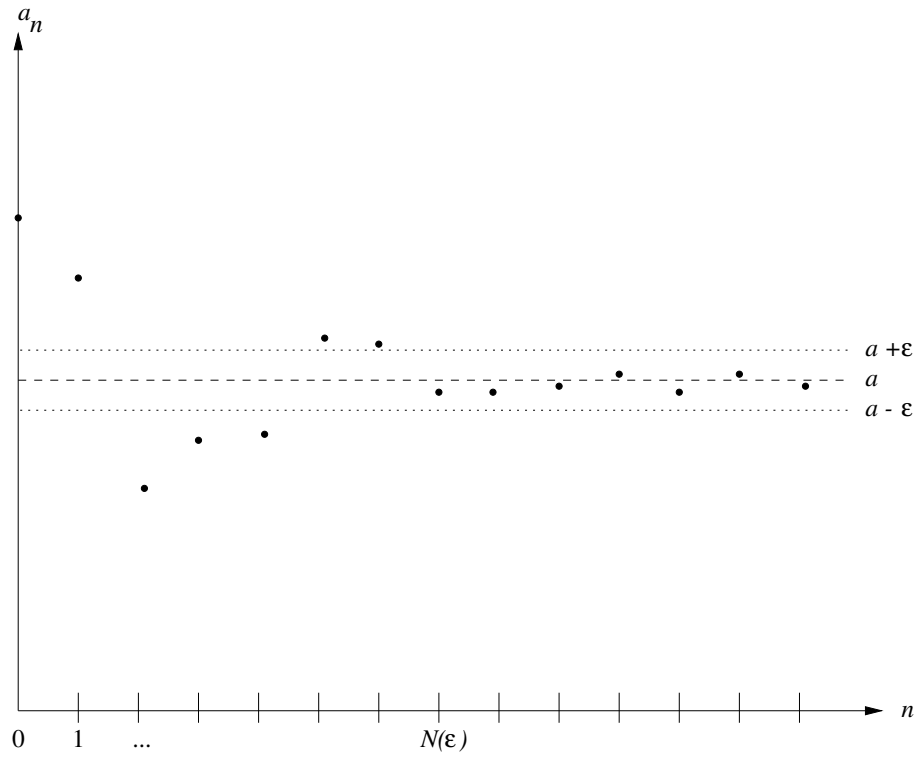


FIGURE 1. Illustration of the convergence of a sequence

If one is not interested in the index $N(\varepsilon)$, the condition can also be stated as:

For every $\varepsilon > 0$ only finitely many a_n lie outside $(a-\varepsilon, a+\varepsilon)$.

Reason: If Definition 8.3 holds with $N(\varepsilon)$, then at most the finitely many terms $(a_0, \ldots, a_{N(\varepsilon)-1})$ lie outside $(a-\varepsilon, a+\varepsilon)$. Conversely, if only finitely many terms lie outside $(a-\varepsilon, a+\varepsilon)$ and N is the maximum index of these terms, then Definition 8.3 holds with $N(\varepsilon) = N+1$.

Another equivalent formulation of convergence is given by the following theorem.

8.4. **Theorem.** *A real sequence $(a_n)_{n\in\mathbb{N}}$ converges to some $a \in \mathbb{R}$ if and only if there exists a null sequence $(b_n)_{n\in\mathbb{N}}$ and an $n_0 \in \mathbb{N}$ such that $d(a_n, a) \le b_n$ for all $n \in \mathbb{N}$ with $n \ge n_0$.*

THM Convergence and null sequences

Proof. We first show that from the convergence $b_n \to 0$ it follows that $a_n \to a$. Let $\varepsilon > 0$ be given. From $b_n \to 0$ it follows that there exists an $N(\varepsilon)$ with $b_n \le |b_n| = d(b_n, 0) < \varepsilon$ for all $n \ge N(\varepsilon)$. Hence

$$d(a_n, a) \le b_n < \varepsilon$$

for all $n \ge \max\{N(\varepsilon), n_0\}$.

Thus, only finitely many terms a_n lie outside $(a-\varepsilon, a+\varepsilon)$, from which the convergence $a_n \to a$ follows.

Conversely, suppose $a_n \to a$, i.e. for every $\varepsilon > 0$ there exists an $N(\varepsilon) \in \mathbb{N}$ with $d(a_n, a) < \varepsilon$ for all $n \ge N(\varepsilon)$. Define $b_n := d(a_n, a)$ and $n_0 := 0$. Then obviously $d(a_n, a) \le b_n$, and moreover

$$d(b_n, 0) = |b_n| = \Big|d(a_n, a)\Big| = d(a_n, a) < \varepsilon,$$

for all $n \ge N(\varepsilon)$. Hence $b_n \to 0$. ❑

The opposite of convergence is divergence.

8.5. **Definition.** A sequence $(a_n)_{n\in\mathbb{N}}$ is called *divergent* if it is not convergent. ♢

DEF Divergent sequence

We now investigate some of the sequences from Example 8.1 with respect to convergence or divergence:

Example 8.1(a): The sequence is convergent with $\lim_{n\to\infty} a_n = a$, since: let $\varepsilon > 0$ be given. Then for all $n \ge 0$ we have the inequality

$$d(a_n, a) = d(a, a) = 0 < \varepsilon.$$

Thus we obtain the required condition with $N(\varepsilon) = 0$. ♣

Example 8.1(b): The sequence is convergent with $\lim_{n\to\infty} a_n = 0$, since: let $\varepsilon > 0$ be given and choose a natural number $N(\varepsilon)$ with $N(\varepsilon) > 1/\varepsilon$. Then $1/N(\varepsilon) < \varepsilon$, and hence for all $n \ge N(\varepsilon)$ we have

$$d(a_n, a) = \left|\frac{1}{n} - 0\right| = \frac{1}{n} \le \frac{1}{N(\varepsilon)} < \varepsilon.$$

Thus the sequence with $a_n = 1/n$ is a null sequence. ♣

Warning. Convergent sequences do not need to *attain* their limit, i.e., there need not be an $\hat{n}$ with $a_{\hat{n}} = \lim_{n\to\infty} a_n$. The values a_n just have to get closer and closer to the limit, but may not reach it in general. Example 8.1(b) is typical for this: the values $\frac{1}{n}$ get closer and closer to 0 as n grows but never actually become 0.

Example 8.1(c): The sequence is divergent, which we show by contradiction. Suppose that there exists an $a \in \mathbb{R}$ such that the sequence converges. Then by definition, for $\varepsilon = 1$ there exists an $N(\varepsilon) \in \mathbb{N}$ such that

$$d(a_n, a) < 1$$

holds for all $n \geq N(\varepsilon)$. For even n, we have $a_n = 1$, hence

$$1 > d(a_n, a) > a_n - a = 1 - a \quad \Rightarrow \quad a > 0.$$

For odd n, we have $a_n = -1$, hence

$$1 > d(a_n, a) > a - a_n = a - (-1) = a + 1 \quad \Rightarrow \quad a < 0.$$

But a cannot be both greater and less than zero, so we obtain a contradiction. ♣

Example 8.1(d): The sequence is convergent with $\lim_{n\to\infty} a_n = 1$, since

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

Since $b_n := \frac{1}{n}$ converges to 0 by part (b), it follows by Theorem 8.4 that a_n converges to $a = 1$. ♣

Example 8.1(e): The sequence is convergent with $\lim_{n\to\infty} a_n = 0$, i.e. it is a null sequence. To see this, one first proves by induction for all $n \geq 4$ the inequality $n^2 \leq 2^n$. Thus we obtain

$$\frac{n^2}{2^n} \leq 1 \quad \Rightarrow \quad \frac{n}{2^n} \leq \frac{1}{n}$$

and hence

$$d(a_n, 0) = \left|\frac{n}{2^n} - 0\right| = \frac{n}{2^n} \leq \frac{1}{n}.$$

Since $\frac{1}{n} \to 0$, the claim now follows from Theorem 8.4. ♣

For the investigation of further sequences from Example 8.1 as well as for the investigation of Example 8.2(b), we need some additional preliminary work, which we will do in the following. For Example 8.2(a) and (c), we cannot yet make any statements, since we have not specified a mathematical rule for the resulting sequences — we will consider this later. Nevertheless, convergence is already a meaningful concept here. For example, the question whether the produced amounts in Example 8.2(c) approach a constant value a (a so-called *market equilibrium*) over the years is mathematically nothing other than the question of the convergence of the sequence.

Similarly, the question whether the technician's strategy in Example 8.2(a) leads the water level to settle around 0 is mathematically nothing other than asking whether the resulting sequence of water levels is a null sequence.

8.C Properties of convergent sequences and rules for computation.

DEF
Bounded sequence

8.6. **Definition.** A real sequence $(a_n)_{n\in\mathbb{N}}$ is called *bounded from above (respectively bounded from below)* if there exists $K \in \mathbb{R}$ such that $a_n \leq K$ (respectively $a_n \geq K$) for all $n \in \mathbb{N}$. The sequence is called *bounded* if it is bounded both from above and from below. ♢

8.7. **Theorem.** *Every convergent sequence is bounded.*

THM Convergent sequences are bounded

Proof. We show boundedness from above; the proof for boundedness from below is analogous.

Assume the sequence converges with limit $a \in \mathbb{R}$. We apply Definition 8.3 with $\varepsilon = 1$. For all $n \geq N(\varepsilon)$ it follows that

$$a_n = a_n - a + a \leq |a_n - a| + a = d(a_n, a) + a < 1 + a.$$

Now choose $K = \max\{a_0, a_1, \ldots, a_{N(\varepsilon)-1}, 1 + a\}$. Then $a_n \leq K$ for all $n \in \mathbb{N}$. Hence the sequence is bounded from above. ❑

The sequence $a_n = (-1)^n$ shows that the converse is not true, since it is bounded from above (by $K = 1$) and below (by $K = -1$), but does not converge.

Formally, we have shown the implication: "the sequence converges $\Rightarrow$ the sequence is bounded." This also implies "the sequence is not bounded $\Rightarrow$ the sequence does not converge," or equivalently, "the sequence is not bounded $\Rightarrow$ the sequence diverges." With this, we can now study two further sequences from Example 8.1.

Example 8.1(h): The Fibonacci sequence diverges, since by induction one shows the inequality $a_n \geq n$ for all $n \in \mathbb{N}$, hence the sequence is unbounded. ♣

Example 8.1(g): The convergence or divergence of the sequence $a_n = b^n$ depends on the value of b:

Case 1: $|b| < 1$. In this case, by Theorem 5.2(b), for every $\varepsilon > 0$ there exists $N(\varepsilon) \in \mathbb{N}$ (there simply called n) such that $|b|^{N(\varepsilon)} < \varepsilon$. Hence, for all $n \geq N(\varepsilon)$,

$$d(a_n, 0) = |b^n| = |b|^n \leq |b|^{N(\varepsilon)} < \varepsilon.$$

Thus, the sequence converges to $a = 0$.

Case 2: $b = 1$. Then $b^n = 1$ for all n, hence the sequence converges to $a = 1$, cf. Example 8.1(a).

Case 3: $b = -1$. Then $b^n = (-1)^n$ for all n, hence the sequence diverges, cf. Example 8.1(c).

Case 4: $|b| > 1$. Then, by Theorem 5.2(a), $|b|^n$ is unbounded. Hence also b^n is unbounded, so the sequence diverges. ♣

We continue with further properties and rules for sequences.

8.8. **Theorem.** *If a sequence $(a_n)_{n\in\mathbb{N}}$ converges both to $a \in \mathbb{R}$ and to $\tilde{a} \in \mathbb{R}$, then $a = \tilde{a}$.*

THM Uniqueness of the Limit

Proof. Assume for contradiction that $a \neq \tilde{a}$ and set $\varepsilon := d(a, \tilde{a})/2$. By Definition 8.3, there exists $n \in \mathbb{N}$ such that $d(a_n, a) < \varepsilon$ and $d(a_n, \tilde{a}) < \varepsilon$. By the triangle inequality for the distance d,

$$d(a, \tilde{a}) \leq d(a, a_n) + d(a_n, \tilde{a}) < \varepsilon + \varepsilon = 2\varepsilon = d(a, \tilde{a}).$$

This implies $d(a, \tilde{a}) < d(a, \tilde{a})$, a contradiction. ❑

8.9. **Theorem.** *Let $(a_n)_{n\in\mathbb{N}}$ and $(b_n)_{n\in\mathbb{N}}$ be two convergent sequences with limits $\lim_{n\to\infty} a_n = a$ and $\lim_{n\to\infty} b_n = b$. Then the sequence $(c_n)_{n\in\mathbb{N}}$ with $c_n := a_n + b_n$ is also convergent with*

$$\lim_{n\to\infty} c_n = a + b.$$

That is,

$$\lim_{n\to\infty} (a_n + b_n) = \lim_{n\to\infty} a_n + \lim_{n\to\infty} b_n.$$

THM Sum Rule

Proof. Let $\varepsilon > 0$ and let $N_a(\varepsilon/2)$ and $N_b(\varepsilon/2)$ be the corresponding indices from Definition 8.3. Set $N(\varepsilon) := \max\{N_a(\varepsilon/2), N_b(\varepsilon/2)\}$. Then for all $n \geq N(\varepsilon)$,

$$d(c_n, a+b) = d(a_n+b_n, a+b) = |a_n - a + b_n - b| \leq |a_n - a| + |b_n - b| < \varepsilon/2 + \varepsilon/2 = \varepsilon.$$

Thus, the convergence of $a_n + b_n$ is established. ❑

Example: Consider the sequence $c_n = (n+1)/n$. Writing it as

$$c_n = \frac{n+1}{n} = 1 + \frac{1}{n} = a_n + b_n,$$

we can use the convergence of $a_n \to 1$ (Example 8.1(a)) and $b_n \to 0$ (Example 8.1(b)) to conclude, without further calculation, that $c_n \to 1 + 0 = 1$.

8.10. **Theorem.** *Let $(a_n)_{n\in\mathbb{N}}$ and $(b_n)_{n\in\mathbb{N}}$ be two convergent sequences. Then the sequence $(a_n b_n)_{n\in\mathbb{N}}$ is also convergent, and*

$$\lim_{n\to\infty} (a_n b_n) = \left(\lim_{n\to\infty} a_n\right)\left(\lim_{n\to\infty} b_n\right).$$

THM Product Rule

Proof. Let $a := \lim_{n\to\infty} a_n$ and $b := \lim_{n\to\infty} b_n$. By Theorem 8.7, $(a_n)_{n\in\mathbb{N}}$ is bounded. Let K be the maximum of the absolute values of the upper and lower bounds, so $|a_n| \leq K$ for all $n \in \mathbb{N}$. By enlarging K if necessary, we can also assume $|b| \leq K$.

Let $\varepsilon > 0$ and choose $N_a(\varepsilon)$ and $N_b(\varepsilon)$ according to Definition 8.3 for a_n and b_n. Then for all $n \geq N(\varepsilon) := \max\{N_a(\varepsilon/(2K)), N_b(\varepsilon/(2K))\}$,

$$\begin{aligned} d(a_n b_n, ab) &= |a_n b_n - ab| = |a_n(b_n - b) + (a_n - a)b| \\ &\leq |a_n|\,|b_n - b| + |a_n - a|\,|b| \\ &= |a_n|\,d(b_n, b) + d(a_n, a)\,|b| < K \cdot \frac{\varepsilon}{2K} + \frac{\varepsilon}{2K} \cdot K = \varepsilon. \end{aligned}$$

❑

8.11. **Corollary.**[7] *Let $(b_n)_{n\in\mathbb{N}}$ be a convergent sequence and $\lambda \in \mathbb{R}$. Then the sequence $(\lambda b_n)_{n\in\mathbb{N}}$ is also convergent with*

$$\lim_{n\to\infty} (\lambda b_n) = \lambda \lim_{n\to\infty} b_n.$$

COR Constant multiple

Proof. This follows immediately from Theorem 8.10 with the constant sequence $a_n = \lambda$ for all $n \in \mathbb{N}$. ❑

[7] A *corollary* denotes a — usually simple — consequence of previous results.

With this, we can now also treat the water tank from **Example 8.2(a)**, once we mathematically formalize the opening and closing of the valves.

Assume that if $a_n < 0$, the inlet valve is opened just enough so that the water level rises in one second by the amount $\alpha|a_n|$ for some $\alpha > 0$. If $a_n > 0$, assume that the outlet valve is opened just enough so that the level decreases by $\beta|a_n|$ for some $\beta > 0$. Thus, the level a_{n+1} after one second satisfies either

$$a_{n+1} = a_n + \alpha|a_n| = a_n - \alpha a_n \quad \text{or} \quad a_{n+1} = a_n - \beta|a_n| = a_n - \beta a_n.$$

(Note that in the first case $a_n < 0$, hence $\alpha|a_n| = -\alpha a_n$.)

If we now assume $\alpha = \beta$[8], then both cases yield the same rule:

$$a_{n+1} = a_n - \alpha a_n = (1-\alpha)a_n.$$

By induction one shows that the resulting sequence is of the form $a_n = a_0(1-\alpha)^n$, which we can also write as

$$a_n = \lambda b^n,$$

with $\lambda = a_0$ and $b = (1-\alpha)$.

If $b^n \to 0$, then by Corollary 8.11, also $a_n = \lambda b^n \to 0$, which is our desired result.

Now consider the case that the initial level $a_0 \neq 0$ (otherwise there would be nothing to do). Then $\lambda \neq 0$. If $\lambda b^n \to 0$, then by Theorem 8.10, also $b^n = \frac{1}{\lambda}(\lambda b^n) \to 0$ (note: here $\lambda \neq 0$ is crucial so that $1/\lambda$ exists).

Thus, for $a_0 \neq 0$:

$$\lim_{n\to\infty} a_n = 0 \quad \Leftrightarrow \quad \lim_{n\to\infty} b^n = 0.$$

By Example 8.1(g), the sequence b^n converges to 0 precisely when $|b| < 1$. Hence, the level stabilizes at 0 exactly when $|1-\alpha| < 1$, which is equivalent to $0 < \alpha < 2$.

The technician should therefore choose α between 0 and 2. In particular, the seemingly intuitive conclusion that the level converges faster to 0 if α is chosen very large (i.e., if a lot of water is added or drained every second) is in fact false! ♣

8.12. **Corollary.** *Let $(a_n)_{n\in\mathbb{N}}$ and $(b_n)_{n\in\mathbb{N}}$ be two convergent sequences. Then the sequence $(a_n - b_n)_{n\in\mathbb{N}}$ is also convergent, and*

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

COR Difference rule

Proof. Since $a_n - b_n = a_n + (-1)b_n$, the claim follows from Theorem 8.9 and Corollary 8.11. ❑

8.13. **Theorem.** *Let $(a_n)_{n\in\mathbb{N}}$ and $(b_n)_{n\in\mathbb{N}}$ be two convergent sequences with $\lim_{n\to\infty} b_n = b \neq 0$. Then there exists n_0 such that $b_n \neq 0$ for all $n \geq n_0$, and the sequence $(a_n/b_n)_{n\geq n_0}$ is convergent with*

$$\lim_{n\to\infty}\left(\frac{a_n}{b_n}\right) = \frac{\lim\limits_{n\to\infty} a_n}{\lim\limits_{n\to\infty} b_n}.$$

THM Quotient Rule

[8] The case $\alpha \neq \beta$ is much more complicated. If desired, we can treat this in the discussion session.

Proof. We first consider the convergence of the sequence $1/b_n$.

Since $b \neq 0$, it follows that $|b|/2 > 0$. Thus, for all $n \geq N_b(|b|/2)$, by the reverse triangle inequality,

$$|b| - |b_n| \leq |b - b_n| = d(b, b_n) < \frac{|b|}{2} \quad \Rightarrow \quad |b_n| > |b| - \frac{|b|}{2} = \frac{|b|}{2} > 0,$$

which in particular implies $b_n \neq 0$. Hence, we may set $n_0 = N_b(|b|/2)$.

Now let $\varepsilon > 0$ and define $N(\varepsilon) := \max\{n_0, N_b(\varepsilon|b|^2/2)\}$. Then for all $n \geq N(\varepsilon)$,

$$d\left(\frac{1}{b_n}, \frac{1}{b}\right) = \left|\frac{1}{b_n} - \frac{1}{b}\right| = \left|\frac{b - b_n}{b_n b}\right| = \frac{1}{|b_n||b|}|b_n - b| = \frac{1}{|b_n||b|}d(b_n, b) < \frac{2}{|b|^2} \cdot \frac{\varepsilon|b|^2}{2} = \varepsilon,$$

where in the last inequality we used $\frac{1}{|b_n|} < \frac{2}{|b|}$, which follows from $|b_n| > \frac{|b|}{2}$. This shows that $1/b_n$ converges with limit $1/b$.

The convergence of the sequence a_n/b_n now follows from

$$\frac{a_n}{b_n} = a_n \cdot \frac{1}{b_n}$$

together with Theorem 8.10. ❑

Thus, we have proved the essential calculation rules for convergent sequences. With these rules, one can compute limits of sequences with complicated terms, although one usually needs the right idea for a suitable approach.

As an **example**, consider the sequence $(a_n)_{n\in\mathbb{N}}$ with

$$a_n = \frac{5n^3 + 7n}{n^3 - 3}.$$

The quotient rule is not directly applicable here, because neither the numerator nor the denominator converge individually. However, for $n \geq 1$ we can divide numerator and denominator by n^3, which gives[9]

$$a_n = \frac{5 + 7/n^2}{1 - 3/n^3}.$$

Now we can apply our theorems to the individual terms: In the numerator, $1/n^2 \to 0$ by Theorem 8.10 (applied to $1/n$ and $1/n$) and Example 8.1(b). Applying Theorem 8.10 and Example 8.1(b) again to $1/n^2$ and $1/n$, we also get $1/n^3 \to 0$. By Corollary 8.11, it follows that $7/n^2 \to 0$ and $3/n^3 \to 0$. Then, by Theorem 8.9, the numerator converges to 5, and by Corollary 8.12, the denominator converges to 1. Since $1 \neq 0$, Theorem 8.13 applies, and the whole fraction converges with limit $5/1 = 5$. ♣

Finally, we consider inequalities for limits.

8.14. **Theorem.** *Let $(a_n)_{n\in\mathbb{N}}$ and $(b_n)_{n\in\mathbb{N}}$ be two convergent sequences such that there exists $n_0 \in \mathbb{N}$ with $a_n \leq b_n$ for all $n \geq n_0$. Then*

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

THM
Inequalities for limits

[9] Excluding the term for $n = 0$ is irrelevant for convergence and the limit, since the definition of convergence only requires the property for sufficiently large n.

Proof. Let $a := \lim_{n\to\infty} a_n$ and $b := \lim_{n\to\infty} b_n$, and assume $a > b$. For all $n \geq \max\{N_a(\varepsilon), N_b(\varepsilon)\}$ with $\varepsilon := (a/2 - b/2) > 0$, we have

$$a_n > a - \varepsilon = \frac{a}{2} + \frac{b}{2} \quad \text{and} \quad b_n < b + \varepsilon = \frac{a}{2} + \frac{b}{2}.$$

Thus,

$$a_n > \frac{a}{2} + \frac{b}{2} > b_n,$$

which contradicts the assumption $a_n \leq b_n$, since n can be chosen arbitrarily large (in particular larger than n_0). ❑

Warning. From the strict inequality $a_n < b_n$ for all n, one does in general **not** obtain the strict inequality

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

As a counterexample, consider the sequences $a_n = 0$ and $b_n = 1/n$. Here $a_n < b_n$ holds, but both sequences have the limit 0.

By choosing either a_n or b_n in Theorem 8.14 to be constant equal to A or B, we immediately obtain the following corollary.

8.15. **Corollary.** *Let $(a_n)_{n\in\mathbb{N}}$ be a convergent sequence and $A, B \in \mathbb{R}$, $n_0 \in \mathbb{N}$ such that $A \leq a_n \leq B$ for all $n \geq n_0$. Then*

COR
Limit inequality

$$A \leq \lim_{n\to\infty} a_n \leq B.$$

8.D Improper Convergence.

Divergent sequences can behave in very different ways. Let us, for example, consider the sequences $(a_n)_{n\in\mathbb{N}}$ with

(i) $a_n = 2^n$
(ii) $a_n = -2^n$
(iii) $a_n = (-2)^n$
(iv) $a_n = (-1)^n$

We observe that the terms of sequence (i) become larger and larger, the terms of sequence (ii) become smaller and smaller (in the sense of becoming "more negative"), and the terms of sequence (iii) alternate between becoming larger and becoming smaller. All of these sequences are unbounded and therefore divergent. Sequence (iv) is bounded, but as shown in Example 8.1(c), it is also divergent.

Thus we have four sequences, all of which are divergent, but in very different ways. The following definition includes two of these cases.

8.16. **Definition.** A sequence of real numbers is said to *converge improperly* to $+\infty$ (resp. to $-\infty$), if for every $K \in \mathbb{R}$ there exists an $M(K) \in \mathbb{N}$ such that

$$a_n > K \ \text{ (resp. } a_n < K) \quad \text{for all } n \geq M(K).$$

In this case we also write[10]

$$\lim_{k\to\infty} a_k = \infty \quad (\text{resp. } \lim_{k\to\infty} a_k = -\infty).$$

Instead of "improperly convergent" one also says *properly divergent.* ♢

DEF
Improper convergence

What matters in this definition is not that each term of the sequence is greater (resp. smaller) than the preceding one, but that every arbitrary bound is eventually exceeded (resp. undershot) for all sufficiently large n.

Sequence (i) is thus improperly convergent to $+\infty$, sequence (ii) improperly convergent to $-\infty$. Sequences (iii) and (iv) are not improperly convergent.

With the concept of improper convergence we can also consider limits of sequences of the form $1/a_n$, when the limit of a_n is either zero or does not exist in the proper sense.

8.17. **Theorem.** *(a) Let $(a_n)_{n\in\mathbb{N}}$ be improperly convergent to ∞ (resp. $-\infty$). Then there exists an $n_0 \in \mathbb{N}$ such that $a_n \neq 0$ for all $n \geq n_0$ and the sequence $1/a_n$, $n \geq n_0$, is convergent with*

$$\lim_{n\to\infty} \frac{1}{a_n} = 0,$$

i.e. it is a null sequence.

(b) Let $(a_n)_{n\in\mathbb{N}}$ be a null sequence such that for some $n_0 \in \mathbb{N}$ the inequality $a_n > 0$ holds for all $n \geq n_0$ (resp. $a_n < 0$ for all $n \geq n_0$). Then the sequence $1/a_n$, $n \geq n_0$, is improperly convergent with

$$\lim_{n\to\infty} \frac{1}{a_n} = \infty \ \ (\textit{resp. } = -\infty).$$

THM
Improper convergence and null sequences

Proof. We prove the case "∞" for both parts; the proof for the case "$-\infty$" is analogous.

(a) Setting $n_0 = M(1)$ for $M(K)$ from Definition 8.16, we obtain from $a_n > 1$ that $a_n \neq 0$ for all $n \geq n_0$. For a given $\varepsilon > 0$, we set $N(\varepsilon) = \max\{n_0, M(K)\}$ with $K = 1/\varepsilon$. Then it follows that

$$d\left(\frac{1}{a_n}, 0\right) = \left|\frac{1}{a_n}\right| = \frac{1}{a_n} < \frac{1}{K} = \varepsilon.$$

(b) Let $K > 0$ be given. Setting $M(K) := \max\{n_0, N(\varepsilon)\}$ with $\varepsilon = 1/K$ it follows for all $n \geq N(K)$ that

$$\frac{1}{a_n} > \frac{1}{\varepsilon} = K.$$ ❑

[10] Note that ∞ and $-\infty$ are merely symbols here, not real numbers. The existence of a real number ∞ would, because of $\infty + 1 = \infty$ and $\infty + 0 = \infty$, imply that $1 = 0$, which is explicitly excluded in the field axioms.

8.E Interlude: Supremum, Infimum, Maximum, and Minimum.

When considering a sequence $(a_n)_{n\in\mathbb{N}}$, the question arises whether it has a largest or smallest element. We will discuss this question for general sets $M \subset \mathbb{R}$, for which sequences $M = \{a_n \,|\, n \in \mathbb{N}\}$ are just a particular example. A basic prerequisite for the existence of a largest or smallest element in M is certainly that M contains neither infinitely large positive numbers nor infinitely small negative numbers. This is precisely the content of the following definition.

DEF Boundedness of sets

8.18. **Definition.** A subset $M \subset \mathbb{R}$ is called *bounded from above (respectively, bounded from below)* if there exists a number $s \in \mathbb{R}$ such that

$$x \le s \qquad \text{(respectively, } x \ge s\text{)}$$

for all $x \in M$. The set M is called *bounded* if it is both bounded above and bounded below. ♢

EXAMPLES (un)bounded sets

8.19. **Examples.** (a) Every interval $[a,b]$, (a,b), $[a,b)$, and $(a,b]$ in the sense of Definition 5.7 is bounded. Here, every $s \le a$ is a lower bound and every $s \ge b$ is an upper bound.

(b) The set $\mathbb{N}$ is bounded from below by any $s \le 0$, but unbounded from above. The set $\mathbb{Z}$ is unbounded both from above and below.

(c) The set $M := \{1/n \mid n \in \mathbb{N} \setminus \{0\}\}$ is bounded; from above by any $s \ge 1$ and from below by any $s \le 0$. ♣

What may seem surprising at first is the fact that a bounded set need not contain a largest or smallest number. This can be easily visualized with open intervals. For example, the interval $I = (0,1)$ consists of all numbers strictly greater than 0 and strictly less than 1. If there were a smallest number $x_{\min} \in I$, then $x_{\min} \le 0$ must hold, because for $x_{\min} > 0$, also $x_{\min}/2 > 0$, and the number $x = \min\{x_{\min}/2, 1/2\}$ lies strictly between 0 and 1 and therefore in I. Since $x \le x_{\min}/2 < x_{\min}$, $x_{\min}$ cannot be greater than zero. On the other hand, no number $x \le 0$ lies in I, so there is no smallest number in I.

Similarly, there is no smallest number in $M := \{1/n \mid n \in \mathbb{N} \setminus \{0\}\}$: For any $x > 0$, we can choose $n > 1/x$, giving $1/n < x$, so the smallest number must be ≤ 0. But M contains no numbers ≤ 0.

This observation is the reason why, instead of largest (or smallest) elements, the following slightly more cumbersome definition is used.

DEF Supremum and infimum

8.20. **Definition.** A number $s \in \mathbb{R}$ is called the *supremum* of a set $M \subset \mathbb{R}$ if s is the smallest upper bound of the set, i.e.,

(i) s is an upper bound of M,

(ii) every number $\tilde{s} < s$ is not an upper bound of M.

If a supremum exists, it is unique[11] and is denoted by

$$s = \sup M.$$

Analogously, the *infimum* $s = \inf M$ is defined as the largest lower bound. ♢

EXAMPLES
Supremum and infimum

8.21. **Examples.** For the sets from Examples 8.19 we obtain:

(a) $\sup[a,b] = \sup(a,b) = \sup[a,b) = \sup(a,b] = b$ and
$\inf[a,b] = \inf(a,b) = \inf[a,b) = \inf(a,b] = a$

(b) $\inf \mathbb{N} = 0$

(c) For $M = \{1/n \mid n \in \mathbb{N} \setminus \{0\}\}$, we obtain $\sup M = 1$ and $\inf M = 0$.

For sets M that are unbounded from above, one writes $\sup M = \infty$, and for sets unbounded from below, $\inf M = -\infty$. With this notation, $\sup \mathbb{N} = \infty$. ♣

The following theorem shows that this is always satisfied for subsets of $\mathbb{R}$ bounded above (or below).

THM
Existence of supremum and infimum

8.22. **Theorem.** *Every non-empty subset $M \subset \mathbb{R}$ that is bounded from above (respectively, from below) has a supremum (respectively, infimum).*

Proof. We show the claim for the supremum; the statement for the infimum follows analogously. The existence of the supremum follows from the completeness axiom if we can construct a nested sequence of intervals that always contains the supremum. To this end, we construct a nested sequence of intervals $I_n = [a_n, b_n]$ with $a_n < b_n$ such that b_n is always an upper bound and a_n is never an upper bound. From the definition of the supremum as the smallest upper bound, it follows that it is contained in each such interval, and by the completeness axiom, it exists in $\mathbb{R}$.

Let M be a non-empty set with an upper bound s. Set $b_0 := s$, choose any element $x \in M$ and set $a_0 := x - 1$. Then certainly $a_0 < b_0$, and a_0 is by construction not an upper bound.

Now let $n \in \mathbb{N}$ be arbitrary with $a_n < b_n$ having the above properties. We proceed similarly to the proof of Theorem 5.9: set $c_n := (a_n + b_n)/2$ (midpoint of the interval) and define

$$a_{n+1} := a_n, \quad b_{n+1} := c_n \quad \text{if } c_n \text{ is an upper bound of } M,$$

and

$$a_{n+1} := c_n, \quad b_{n+1} := b_n \quad \text{otherwise.}$$

By construction, $a_{n+1} < b_{n+1}$, a_{n+1} is not an upper bound, and b_{n+1} is an upper bound. Continuing this construction inductively gives a nested sequence of intervals, which, by the completeness axiom, contains exactly one element $x \in \mathbb{R}$. This element is an upper bound, because if it were not, there would be $a \in M$ with $a > x$. Then each b_n would be larger than a and would have a distance $d(x, b_n) > a - x$ from x. But then x would not lie in intervals I_n with $|I_n| < a - x$. On the other hand, there cannot be a smaller upper bound $\tilde{s} < x$, because then $d(x, a_n) > x - \tilde{s}$ and x would not lie in any I_n with $|I_n| < x - \tilde{s}$. Therefore, x is the smallest upper bound, and hence the supremum of M. ❑

The importance of the completeness of $\mathbb{R}$ (i.e., the validity of the completeness axiom) for the existence of the supremum (or infimum) can be seen if one considers $\mathbb{Q}$ instead of $\mathbb{R}$. For example, the set

$$M = \{x \in \mathbb{Q} \mid x > 0 \text{ and } x^2 < 2\} \subset \mathbb{Q}$$

[11] Suppose s and s' satisfy the two properties, then by (ii) $s' \geq s$ and $s \geq s'$ must hold, from which equality follows.

has no supremum in $\mathbb{Q}$, because: by the definition of M, $s \in \mathbb{Q}$ is an upper bound of M if and only if $s > 0$ and $s^2 \geq 2$. We now prove that there can be no smallest upper bound in $\mathbb{Q}$. Let $s \in \mathbb{Q}$ be any upper bound, i.e., $s > 0$ and $s^2 \geq 2$. But since there is no $s \in \mathbb{Q}$ with $s^2 = 2$, it follows that $s^2 > 2$. Now choose $n \in \mathbb{N}$ with $1/n < (s^2 - 2)/(2s)$ and $1/n < s$, and set $\tilde{s} := s - 1/n$. Then $\tilde{s} \in \mathbb{Q}$, $\tilde{s} > 0$, $\tilde{s} < s$, and

$$\tilde{s}^2 = (s - 1/n)^2 = s^2 - 2s/n + 1/n^2 > s^2 - 2s/n > s^2 - (s^2 - 2) = 2.$$

Hence $\tilde{s}$ is also an upper bound, contradicting the definition of s as the smallest upper bound.

8.23. **Remark.** We derived the existence of a supremum here from the completeness axiom. In fact, the converse is also true, i.e., the completeness axiom can be derived from the existence of the supremum. For a given nested sequence of intervals $I_n = [a_n, b_n]$, define the set $M = \{a_0, a_1, a_2, \dots\}$. From the nesting, we have $a_0 \leq a_1 \leq a_2 \leq \dots$ and $b_0 \geq b_1 \geq b_2 \geq \dots$. From this and $a_n \leq b_n$, it follows for all $m, n \in \mathbb{N}$ that $a_n \leq b_m$, i.e., every b_m is an upper bound of M. Therefore $\sup M \leq b_n$ for all n, and by definition, $\sup M \geq a_n$ for all n. Hence $\sup M$ is contained in all intervals. If $\sup M$ exists in $\mathbb{R}$, the completeness axiom is satisfied.

REMARK
Existence of supremum and completeness axiom

Alternatively, one can replace the completeness axiom in the definition of $\mathbb{R}$ with the condition: "every set bounded above has a supremum in $\mathbb{R}$." ♠

A particularly nice case occurs when the supremum (or infimum) of a set M is contained in the set M itself, i.e., when the set has a largest (or smallest) element.

8.24. **Definition.** If a set M contains a largest (or a smallest) element $s \in M$, i.e., an element that satisfies $x \leq s$ (or $x \geq s$) for all $x \in M$, then s is called the *maximum* (or *minimum*) of the set M, denoted by $s = \max M$ (or $s = \min M$). ♢

DEF
Maximum, Minimum

The relationship between maximum and supremum is as follows: If the maximum $s = \max M \in M$ exists, it coincides with the supremum, since by definition it is an upper bound and for any $\tilde{s} < s$, $\tilde{s} < s \in M$ is not an upper bound. Conversely: if the supremum $\sup M$ exists, it is a maximum if and only if it lies in M. Therefore, a maximum is always a supremum, but a supremum is only a maximum if it lies in M. Consequently, an upper-bounded set $M \subset \mathbb{R}$ has a maximum if and only if $\sup M \in M$. The analogous statement holds for minimum and infimum.

8.25. **Examples.**

EXAMPLES

(a) The open interval (a, b) has no maximum because $\sup(a, b) = b \notin (a, b)$, whereas the closed interval $[a, b]$ does have a maximum since $\sup[a, b] = b \in [a, b]$.

(b) Every non-empty subset $N \subset \mathbb{N}$ of the natural numbers has a minimum, which we prove by contradiction: Suppose no minimum exists in N. We prove by induction over n that the numbers $0, \dots, n$ cannot belong to N: If $n = 0 \in N$, then this would be the minimum, since there is no smaller number in $\mathbb{N}$. Assuming by induction that $0, \dots, n$ are not in N, then $n+1$ cannot be in N, because otherwise it would be the minimum. Hence no $n \in \mathbb{N}$ is in N, which contradicts the fact that N is non-empty.

(c) Every lower-bounded non-empty subset $A \subset \mathbb{Z}$ of the integers has a minimum. If $A \subset \mathbb{N}$, this follows directly from (b). If A contains negative numbers, choose $k \in \mathbb{N}$ with $k > -s$, where s is a lower bound of A. Then define the set

$$\widetilde{A} := \{a + k \mid a \in A\}.$$

From this definition follows

$$\tilde{a} \in \widetilde{A} \Leftrightarrow \tilde{a} - k \in A \qquad \text{and} \qquad a \in A \Leftrightarrow a + k \in \widetilde{A}.$$

For each $\tilde{a} \in \widetilde{A}$, let $a := \tilde{a} - k$. Then $s \leq a$ and hence $\tilde{a} = a + k > a - s \geq 0$. Thus $\widetilde{A}$ is a subset of $\mathbb{N}$ and therefore has a minimum $\min \widetilde{A} \in \widetilde{A}$ by (b). This satisfies $\min \widetilde{A} - k \in A$, and for every $a \in A$, with $\tilde{a} := a + k$, the inequality $\min \widetilde{A} - k \leq \tilde{a} - k = a$ holds. Consequently, $\min \widetilde{A} - k$ is a minimum of A. ♣

The following theorem shows an important consequence of the existence of a maximum.

8.26. **Theorem.** *Let $s \in \mathbb{R}$ be given such that for a given non-empty set M, the inequality $x < s$ holds for all $x \in M$. Then*

THM
Inequalities for suprema and maxima

$$\sup M \leq s.$$

If the maximum $\max M$ exists, then the strict inequality

$$\sup M = \max M < s$$

holds. An analogous statement holds for infimum and minimum with the inequalities $\geq$ and $>$.

Proof. The first statement follows from the fact that s is an upper bound and the supremum is the least upper bound.

The second statement follows because $x := \sup M = \max M \in M$, and thus by assumption $\sup M = x < s$. ❑

The difference between the two statements is that in the first case equality may hold, whereas in the second case a strict inequality is guaranteed. In other words: only if a maximum exists we can ensure that the elementwise strict inequality $x < s$ carries over to the supremum; in general, elementwise "$<$" may become the weaker "$\leq$" when passing to the supremum. This is similar to the behaviour of limits of sequences, cf. the warning after Theorem 8.14.

8.27. **Example.** Consider the open interval (a, b). Since all $x \in (a, b)$ satisfy $x < b$ by definition, $s = b$ is a choice for s in Theorem 8.26. However, the strict inequality does not hold, because $\sup(a, b) = b$. EXAMPLE

For the closed interval $[a, b]$, $s = b$ does not satisfy the assumptions of Theorem 8.26, because for $x = b$, the strict inequality $x < s$ does not hold. ♣

8.F Monotone Sequences and Cauchy Sequences.

With the help of the existence of the supremum we can derive a convergence criterion that plays an important role in analysis, since for a given sequence it is often relatively easy to check. For this we need the following concept.

8.28. **Definition.** A real sequence $(a_n)_{n\in\mathbb{N}}$ is called

DEF Monotone sequence

- *monotonically increasing*, if for all $n \in \mathbb{N}$ the inequality $a_{n+1} \geq a_n$ holds,
- *monotonically decreasing*, if for all $n \in \mathbb{N}$ the inequality $a_{n+1} \leq a_n$ holds,
- *monotone*, if it is either monotonically increasing or monotonically decreasing. ♢

One can verify (sometimes with some calculations) that, for example, the sequences from Example 8.1(a), (b), (d), (f), (h) are monotone. The sequence $(b^n)_{n\in\mathbb{N}}$ from Example 8.1(g) is monotone if and only if $b \geq 0$. The sequences $\big((-1)^n\big)_{n\in\mathbb{N}}$ from Example 8.1(c) and $\left(\frac{n}{2^n}\right)_{n\in\mathbb{N}}$ from Example 8.1(e) are not monotone (although the latter is decreasing for $n \geq 1$).

8.29. **Theorem.** *Every bounded and monotone real sequence $(a_n)_{n\in\mathbb{N}}$ converges.*

THM Monotone and bounded sequences converge

Proof. We prove the theorem for monotonically increasing sequences; the proof for monotonically decreasing sequences is analogous.

Since by assumption, the sequence $(a_n)_{n\in\mathbb{N}}$ is bounded, the set $\{a_n \mid n \in \mathbb{N}\}$ has a supremum $s \in \mathbb{R}$ (Thm. 8.22). We show that $(a_n)_{n\in\mathbb{N}}$ converges to s. Let $\varepsilon > 0$. That s is the supremum means that $a_n \leq s$ for all $n \in \mathbb{N}$ and that there is some $N(\varepsilon) \in \mathbb{N}$ such that $a_{N(\varepsilon)} > s - \varepsilon$ (otherwise $s - \varepsilon$ would be a smaller upper bound). Then for all $n \geq N(\varepsilon)$, we find that

$$|a_n - s| \overset{a_n \leq s}{=} s - a_n \overset{a_{N(\varepsilon)} \leq a_n}{\leq} s - a_{N(\varepsilon)} < \varepsilon$$

as desired. ❑

Example. Coming back to Example 8.1(f), with this theorem we can now prove convergence for the sequences a_n and b_n defined by the boundaries of the nested intervals in Example 8.1(f). Since the intervals are nested, it must hold that $a_{n+1} \geq a_n$ and $b_{n+1} \leq b_n$, i.e. the sequences are monotone. From monotonicity it follows immediately that $K = a_0$ is a lower bound for a_n since $a_n \geq a_0$. Since $a_n < b_n \leq b_0$, $K = b_0$ is an upper bound. Thus a_n is bounded, and analogously one sees that b_n is also bounded. Consequently, both sequences are convergent by Theorem 8.29.

EXAMPLE Interval nestings

In fact, we can prove even more here, namely that both limits $a := \lim_{n\to\infty} a_n$ and $b := \lim_{n\to\infty} b_n$ lie in every interval I_n. To see this, fix any $n \in \mathbb{N}$. Corollary 8.15 gives, since $a_k \geq a_n$ for all $k \geq n$, the inequality $a \geq a_n$; analogously it follows that $b \leq b_n$. Since $a_n < b_n$, Theorem 8.14 also yields $a \leq b$. Consequently,

$$a_n \leq a \leq b \leq b_n \quad \Rightarrow \quad a \in I_n \text{ and } b \in I_n.$$

Since in Section 5.B we already proved that there can be only one element lying in all intervals I_n, it follows in particular that $a = b$. ♣

Example. With the theorems from this section, we can now analyze the method of root calculation from **Example 8.2(b)** and in particular prove that the recursively defined sequence

EXAMPLE
Square root

$$a_0 > 0 \text{ arbitrary}, \quad a_{n+1} = \frac{1}{2}\left(a_n + \frac{x}{a_n}\right)$$

actually converges to some $a \in \mathbb{R}$ with $a^2 = x$. The existence of such an $a \in \mathbb{R}$ is not assumed; rather, it arises as a "byproduct" of the following proof, which is divided into six steps (0–5).

Step 0. As a preliminary step, we first prove that there can be at most one positive solution $y \in \mathbb{R}$ of the equation $y^2 = x$. Suppose that $\tilde{y} > 0$ also satisfies $\tilde{y}^2 = x$. Then

$$0 = x - x = y^2 - \tilde{y}^2 = (y + \tilde{y})(y - \tilde{y}).$$

Since $y + \tilde{y} > 0$, it must follow that $y - \tilde{y} = 0$, hence $y = \tilde{y}$. Therefore there is at most one positive solution of the equation $y^2 = x$.

Step 1. We have $a_n > 0$ for all n. This is easily proven by induction, since with $a_n > 0$ the expression in parentheses is always > 0.

Step 2. It holds that $a_n^2 \geq x$ for all $n \geq 1$, because:

$$\begin{aligned} a_{n+1}^2 - x &= \left(\frac{1}{2}\left(a_n + \frac{x}{a_n}\right)\right)^2 - x \\ &= \frac{1}{4}\left(a_n^2 + 2x + \frac{x^2}{a_n^2}\right) - x \\ &= \frac{1}{4}\left(a_n^2 - 2x + \frac{x^2}{a_n^2}\right) \\ &= \frac{1}{4}\left(a_n - \frac{x}{a_n}\right)^2 \geq 0. \end{aligned}$$

Step 3. It holds that $a_{n+1} \leq a_n$ for all $n \geq 1$, because:

$$a_n - a_{n+1} = a_n - \frac{1}{2}\left(a_n + \frac{x}{a_n}\right) = \frac{1}{2a_n}(a_n^2 - x) \geq 0.$$

Step 4. From Step 3 it follows that the sequence a_n for $n \geq 1$ is monotonically decreasing. Hence it is bounded above by a_1. Since it is also bounded below by 0 according to Step 1, it is monotone and bounded overall, and therefore converges by Theorem 8.29 to some $a \in \mathbb{R}$. From Step 1 we conclude $a_n > 0$, and thus by Theorem 8.14 also $a \geq 0$. By Theorem 8.10, the sequence a_n^2 converges to a^2, and since by Step 2 we have $a_n^2 > x$, it follows from Theorem 8.14 that $a^2 \geq x$. Hence, since $x > 0$, we also have $a > 0$.

Step 5. By the rules for limits (note that Theorem 8.13 applies because $a > 0$, and that the sequence $(a_{n+1})_{n\in\mathbb{N}}$ also converges to a), we obtain

$$a = \lim_{n\to\infty} a_{n+1} = \lim_{n\to\infty} \frac{1}{2}\left(a_n + \frac{x}{a_n}\right) = \frac{1}{2}\left(\lim_{n\to\infty} a_n + \frac{x}{\lim\limits_{n\to\infty} a_n}\right) = \frac{1}{2}\left(a + \frac{x}{a}\right).$$

It follows that

$$a = \frac{1}{2}\left(a + \frac{x}{a}\right) \quad \Rightarrow \quad 2a^2 = a^2 + x \quad \Rightarrow \quad a^2 = x.$$

Thus a is a positive solution of the equation $a^2 = x$. We have therefore simultaneously proven the existence of such a solution, which we henceforth (as usual) denote by $\sqrt{x}$. ♣

Before we move on to Cauchy sequences, we state a useful theorem on limits.

THM Squeeze Theorem

8.30. **Theorem.** *Let $(a_n)_{n\in\mathbb{N}}$, $(b_n)_{n\in\mathbb{N}}$, and $(c_n)_{n\in\mathbb{N}}$ be three sequences and let $l \in \mathbb{R}$ be such that*

(i) $\lim_{n\to\infty} a_n = l = \lim_{n\to\infty} c_n$ *and*
(ii) *there is some $n_0 \in \mathbb{N}$ such that for all $n \geq n_0$, $a_n \leq b_n \leq c_n$.*

Then $(b_n)_{n\in\mathbb{N}}$ also converges to l.

Proof. Let $\varepsilon > 0$ and set $N(\varepsilon) := \max\{n_0, N_a(\varepsilon), N_c(\varepsilon)\}$, where $N_a(\varepsilon)$, $N_c(\varepsilon)$ come from the convergence of $(a_n)_{n\in\mathbb{N}}$ and of $(c_n)_{n\in\mathbb{N}}$. Then for every $n \geq N(\varepsilon)$, we have

$$-\varepsilon < a_n - l \leq b_n - l \leq c_n - l < \varepsilon\,, \qquad \text{so} \qquad |b_n - l| < \varepsilon\,. \qquad ❑$$

A.-L. Cauchy 1789–1857

The previous Definition 8.3 of convergence has the disadvantage that one needs to know the limit a before one can check whether a sequence is convergent. This sometimes makes its use cumbersome or even impossible. With the Cauchy sequences, named after the French mathematician Augustin-Louis Cauchy (1789–1857), one avoids this problem, as Theorem 8.32 below will show.

DEF Cauchy sequence

8.31. **Definition.** A real sequence $(a_n)_{n\in\mathbb{N}}$ is called a *Cauchy sequence*, if for every $\varepsilon > 0$ there exists a $C(\varepsilon) \in \mathbb{N}$ such that

$$d(a_n, a_m) < \varepsilon$$

holds for all $n, m \geq C(\varepsilon)$. ♢

Illustratively, this definition means that for $n, m \geq C(\varepsilon)$ all terms a_m lie within an ε-neighborhood of a_n. In contrast to the definition of convergence in Definition 8.3, this neighborhood is defined without knowledge of a limit a. The following theorem shows that this condition is equivalent to convergence.

THM Cauchy sequences are convergent

8.32. **Theorem.** *A real sequence $(a_n)_{n\in\mathbb{N}}$ is convergent if and only if it is a Cauchy sequence.*

Proof. We prove the implications "Convergence ⇒ Cauchy sequence" and "Cauchy sequence ⇒ Convergence", beginning with the first.

"Convergence ⇒ Cauchy sequence": We use $N(\varepsilon)$ from Definition 8.3 and set $C(\varepsilon) := N(\varepsilon/2)$. Then for all $n, m \geq C(\varepsilon)$, by the triangle inequality,

$$d(a_n, a_m) \leq d(a_n, a) + d(a, a_m) < \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon.$$

Hence $(a_n)_{n\in\mathbb{N}}$ is a Cauchy sequence.

"Cauchy sequence ⇒ Convergence": Suppose $(a_n)_{n\in\mathbb{N}}$ is a Cauchy sequence. Similarly to the proof of Theorem 8.7, one sees that every Cauchy sequence is bounded. We define two sequences $(b_n)_{n\in\mathbb{N}}$ and $(c_n)_{n\in\mathbb{N}}$ as follows:

$$b_n := \inf\{a_k \mid k \geq n\} \quad \text{and} \quad c_n := \sup\{a_k \mid k \geq n\}\,.$$

With this definition the inequality

$$b_n \leq a_k \leq c_n \quad \text{for all } k \geq n. \tag{8.2}$$

holds. We also have for all $n \in \mathbb{N}$ that $b_{n+1} \geq b_n$ since b_n is a lower bound for $\{a_k \mid k \geq n+1\}$; similarly, $c_n \leq c_{n+1}$. So both $(b_n)_{n\in\mathbb{N}}$ and $(c_n)_{n\in\mathbb{N}}$ are monotone and bounded, so they converge to limits $\lim_{n\to\infty} b_n = b$ and $\lim_{n\to\infty} c_n = c$, respectively, by Theorem 8.29.

We now show that $b = c$; then the Squeeze Theorem 8.30 shows that $\lim_{n\to\infty} a_n = b = c$ as well. Since $b_n \leq c_n$ for all n, it suffices to show that for every $\varepsilon > 0$ there is $N(\varepsilon) \in \mathbb{N}$ such that for all $n \geq N(\varepsilon)$, $c_n - b_n < \varepsilon$. We set $N(\varepsilon) := C(\varepsilon/3)$ with $C(\delta)$ from the definition of "Cauchy sequence". Then for all $m, n \geq N(\varepsilon)$, we have $d(a_m, a_n) < \varepsilon/3$. This implies that $b_n \geq a_{N(\varepsilon)} - \varepsilon/3$ and $c_n \leq a_{N(\varepsilon)} + \varepsilon/3$ and thus $c_n - b_n \leq 2\varepsilon/3 < \varepsilon$ for all $n \geq N(\varepsilon)$, as desired. ❑

The sequence of lower endpoints and the sequence of upper endpoints of an interval nesting form Cauchy sequences (this is fairly easy to see), and one can deduce that their limits must be contained in all the intervals. (This is similar to the discussion above using monotonicity of these sequences.) So the completeness axiom follows from Theorem 8.29. This gives us the following circle of implications (given the axioms of an ordered field).

Completeness axiom
⇓
Every non-empty bounded subset of $\mathbb{R}$ has a supremum
⇓
Every bounded monotone sequence converges
⇓
Every Cauchy sequence converges
⇓
Completeness axiom

This we see that the completeness axiom is actually equivalent to the three properties in the middle of the chain. We could just as well have required one of these properties in place of the completeness axiom, as is in fact done in some books.

8.G Subsequences.

8.33. **Definition.** Let $(a_n)_{n\in\mathbb{N}}$ be a real sequence and **DEF** Subsequence

$$n_0 < n_1 < n_2 < \dots$$

an increasing sequence of natural numbers. Then the sequence

$$(a_{n_k})_{k\in\mathbb{N}} := (a_{n_0}, a_{n_1}, a_{n_2}, \dots)$$

is called a *subsequence* of $(a_n)_{n\in\mathbb{N}}$. ♢

Note that from $n_k < n_{k+1}$ it follows immediately that $n_{k+1} \geq n_k + 1$ and thus, by induction, $n_k \geq k$.

8.34. **Example.** (a) For the sequence $a_n = (-1)^n$ we obtain, with $n_k = 2k$, i.e., $n_0 = 0, n_1 = 2, n_2 = 4, \dots$, the subsequence $(1, 1, 1, 1, \dots)$. For $n_k = 2k+1$ we obtain the subsequence $(-1, -1, -1, -1, \dots)$. **EXAMPLE**

(b) Consider the sequence $a_n = (-1)^n + (1/2)^n$, see Fig. 2. Here we obtain, with $n_k = 2k$, i.e., $(0, 2, 4, 6, \dots)$, the subsequence $a_{n_k} = (-1)^{2k} + (1/2)^{2k} = 1 + (1/2)^{2k}$, i.e., $(2, 5/4, 17/16, \dots)$. For $n_k = 2k+1$, i.e., $(1, 3, 5, 7, \dots)$, we obtain $a_{n_k} = -1 + (1/2)^{2k+1}$, i.e., $(-1/2, -7/8, -31/32, \dots)$. ♣

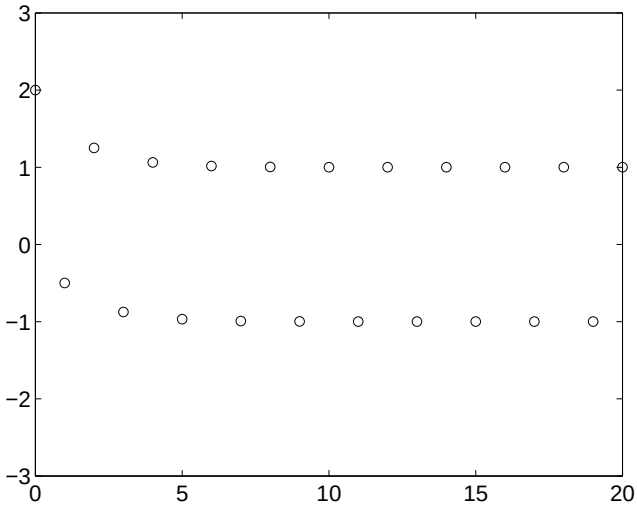


FIGURE 2. Illustration of the sequence $a_n = (-1)^n + (1/2)^n$

If the sequence $(a_n)_{n\in\mathbb{N}}$ converges, then by the definition of convergence, every subsequence also converges to the same limit, because if $|a_n - a| < \varepsilon$ for all $n \geq N(\varepsilon)$, then since $n_k \geq k$, we also have $|a_{n_k} - a| < \varepsilon$ for all $k \geq N(\varepsilon)$.

However, a sequence can contain convergent subsequences even if it does not converge itself. An example is the sequence $a_n = (-1)^n$, since for this sequence the two subsequences given above converge to 1 and -1, respectively, but the sequence itself is well known not to converge.

8.35. **Definition.** A $p \in \mathbb{R}$ is called an *accumulation point* of a sequence $(a_n)_{n\in\mathbb{N}}$ if there exists a subsequence $(a_{n_k})_{k\in\mathbb{N}}$ that converges to p. ♢

DEF Accumulation point

An alternative characterization, which makes the term "accumulation point" more intuitive, is given by the following theorem.

8.36. **Theorem.** *A point $p \in \mathbb{R}$ is an accumulation point of a sequence $(a_n)_{n\in\mathbb{N}}$ if and only if in every interval of the form $(p-\varepsilon, p+\varepsilon)$, $\varepsilon > 0$, there are infinitely many terms a_n.*

THM on accumulation points

Proof. Suppose p is an accumulation point and let a_{n_k} be the subsequence converging to p. For every $\varepsilon > 0$ there exists $K(\varepsilon)$ such that infinitely many terms a_{n_k}, $k \geq K(\varepsilon)$, lie in $(p-\varepsilon, p+\varepsilon)$.

Conversely, suppose p is a point such that for all $\varepsilon > 0$ there are infinitely many terms a_n in $(p-\varepsilon, p+\varepsilon)$. We construct the indices n_k for a subsequence converging to p: set $n_0 := 1$ and consider $\varepsilon = 1/k$ for each $k \geq 1$. Then there are infinitely many terms in $(p-\varepsilon, p+\varepsilon)$, and we choose n_k as one of the corresponding indices such that $n_k \geq n_{k-1}$. Then for the subsequence defined in this way, $a_{n_k} \in (p-\varepsilon, p+\varepsilon)$ and thus $d(a_{n_k}, p) \leq \varepsilon = 1/k$, from which convergence follows because $1/k$ is a null sequence. ❑

From the observation before Definition 8.35 it follows immediately that a convergent sequence has exactly one accumulation point. The following definition also gives us, in the case of non-convergence, a way to determine upper and lower bounds for the accumulation points of a sequence, as we will prove next.

8.37. **Definition.** (a) For a real sequence $(a_n)_{n\in\mathbb{N}}$ we define the sequence

$$b_n := \sup\{a_k \mid k \geq n\}.$$

If these suprema exist for all $n \in \mathbb{N}$ (i.e., are $< \infty$) and the sequence $(b_n)_{n\in\mathbb{N}}$ converges, then we say that the *limit superior* of $(a_n)_{n\in\mathbb{N}}$ exists and define it as

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

(b) For a real sequence $(a_n)_{n\in\mathbb{N}}$ we define the sequence

$$c_n := \inf\{a_k \mid k \geq n\}.$$

If these infima exist for all $n \in \mathbb{N}$ (i.e., are $> -\infty$) and the sequence $(c_n)_{n\in\mathbb{N}}$ converges, then we say that the *limit inferior* of $(a_n)_{n\in\mathbb{N}}$ exists and define it as

$$\liminf_{n\to\infty} a_n := \lim_{n\to\infty} c_n. \qquad \diamondsuit$$

DEF Superior and inferior limit

8.38. **Example.** (a) For the sequence $a_n = (-1)^n$ we have for all $n \in \mathbb{N}$,

$$b_n = \sup\{a_k \mid k \geq n\} = 1 \quad \text{and} \quad c_n = \inf\{a_k \mid k \geq n\} = -1.$$

The sequences b_n and c_n are thus constant and converge to 1 and -1, respectively. Hence,

$$\limsup_{n\to\infty} a_n = 1 \quad \text{and} \quad \liminf_{n\to\infty} a_n = -1.$$

(b) For the sequence $a_n = (-1)^n + (-1/2)^n$ we have $b_n = 1 + (1/2)^n$ if n is even and $b_n = 1 + (1/2)^{n+1}$ if n is odd. The limit is thus 1, so $\limsup_{n\to\infty} a_n = 1$. Similarly, $\liminf_{n\to\infty} a_n = -1$. $\clubsuit$

EXAMPLE

8.39. **Theorem.** *A sequence $(a_n)_{n\in\mathbb{N}}$ is bounded if and only if both the limit superior and the limit inferior exist.*

THM on bounded sequences

Proof. First, we show that boundedness implies existence. We prove this for the limit superior; the proof for the limit inferior is analogous.

From boundedness above, it follows by Theorem 8.22 that the suprema in the definition of b_n exist. From the definition of the sequence b_n we have

$$b_{n+1} = \sup\{a_k \mid k \geq n+1\} \leq \max\{\sup\{a_k \mid k \geq n+1\}, a_n\} = \sup\{a_k \mid k \geq n\} = b_n.$$

Thus, the sequence b_n is monotonically decreasing and therefore bounded above by b_0. Since a_n is bounded below, there exists $K \in \mathbb{R}$ with $a_n \geq K$, which implies by the definition of the supremum that $b_n \geq K$ for all $n \in \mathbb{N}$. Hence, b_n is bounded below by K and thus bounded overall and monotonic. The existence of the limit now follows from Theorem 8.29.

Conversely, if the limit superior exists, then by definition the suprema in the definition of b_n exist. In particular, b_0 is an upper bound for the sequence, so it is bounded above. Similarly, the existence of the limit inferior implies that c_0 is a lower bound for the sequence. Hence, it is bounded. $\square$

To prove the connection between subsequences and the limit superior and inferior, we need the following auxiliary result.

LEMMA

8.40. **Lemma.** *For a sequence $(a_n)_{n\in\mathbb{N}}$ that is bounded above (resp. bounded below), for every b_j (resp. c_j) from Definition 8.37 and every $\varepsilon > 0$, there exists an $n \geq j$ such that*

$$d(b_j, a_n) < \varepsilon \qquad (\textit{resp. } d(c_j, a_n) < \varepsilon).$$

Proof. We prove the statement for b_j.

Since $b_j := \sup\{a_n \mid n \geq j\}$, it follows immediately that $a_n - b_j \leq 0$ for all $n \geq j$. Thus, it suffices to show that for every $\varepsilon > 0$ there exists an a_n, $n \geq j$, with $b_j - a_n < \varepsilon$ in order to prove $d(b_j, a_n) < \varepsilon$.

Assume that for a given $\varepsilon > 0$ there is no such a_n, i.e., for all $n \geq j$ we have $b_j - a_n \geq \varepsilon$. Then $a_n \leq b_j - \varepsilon$, so $b_j - \varepsilon$ is an upper bound of the set $\{a_n \mid n \geq j\}$. But this would be a smaller upper bound than b_j, contradicting the definition of the supremum as the least upper bound. ❑

THM
lim inf and lim sup are accumulation points

8.41. **Theorem.** *Let $(a_n)_{n\in\mathbb{N}}$ be a sequence for which the limit superior (resp. limit inferior) exists. Then*

$$p := \limsup_{n\to\infty} a_n \quad (\textit{resp. } p := \liminf_{n\to\infty} a_n)$$

is an accumulation point, and for all other accumulation points $q \in \mathbb{R}$ we have

$$q \leq \limsup_{n\to\infty} a_n \quad (\textit{resp. } q \geq \liminf_{n\to\infty} a_n).$$

Proof. We prove the statement for the limit superior and construct a subsequence converging to p inductively as follows: First, set $n_0 = 0$. For each $k \geq 1$, set $j = n_{k-1} + 1$ and $\varepsilon = 1/j$, choose the corresponding index n from Lemma 8.40, and set $n_k = n$. Then, by construction, $n_k \geq j = n_{k-1} + 1 > n_{k-1}$ (and thus, since $n_0 = 0$, in particular $n_k \geq k$), and

$$|b_{n_{k-1}+1} - a_{n_k}| = |b_j - a_n| < \varepsilon = 1/j = 1/(n_{k-1} + 1) \leq 1/k.$$

Now let $p := \limsup_{n\to\infty} a_n = \lim_{n\to\infty} b_n$ and let $\varepsilon > 0$ be arbitrary. Choose $N(\varepsilon) \in \mathbb{N}$ such that $|b_k - p| < \varepsilon/2$ and $1/k < \varepsilon/2$ for all $k \geq N(\varepsilon)$. Then

$$|a_{n_k} - p| \leq |a_{n_k} - b_{n_{k-1}+1}| + |b_{n_{k-1}+1} - p| < 1/k + \varepsilon/2 < \varepsilon$$

for all $k \geq N(\varepsilon)$, which proves that a_{n_k} converges to p.

To show the claimed inequality for other accumulation points q, assume there exists a subsequence with

$$q := \lim_{k\to\infty} a_{n_k} > \limsup_{n\to\infty} a_n =: p.$$

Then for $\varepsilon = (q-p)/2$ there exists $N(\varepsilon) \in \mathbb{N}$ such that $a_{n_k} > q - \varepsilon$ for all $k \geq N(\varepsilon)$. It follows that

$$\begin{aligned} b_n &= \sup\{a_k \mid k \geq n\} \geq \sup\{a_{n_k} \mid n_k \geq n\} \geq q - \varepsilon \\ &= q - (q-p)/2 = q/2 + p/2 = p + (q-p)/2, \end{aligned}$$

and thus $\lim_{n\to\infty} b_n \geq p + (q-p)/2 > p$, which contradicts the definition of p. ❑

An important theorem of analysis — named after the mathematicians Bernard Bolzano (1781–1848) and Karl Weierstrass (1815–1897) — now follows directly from the two preceding theorems.

8.42. **Theorem.** *Every bounded sequence $(a_n)_{n\in\mathbb{N}}$ has a convergent subsequence.* **THM** Bolzano–Weierstrass Theorem

Proof. By Theorem 8.39, a bounded sequence has a limit superior, and by Theorem 8.41, this is an accumulation point. Hence, there exists a subsequence converging to the limit superior, which is therefore convergent. ❑

From the Bolzano–Weierstrass Theorem, one can derive the theorem (already proved here by another method) 8.29 on the convergence of bounded and monotonic sequences without using the completeness axiom (which we will not do here). Since the Bolzano–Weierstrass Theorem itself relies on the completeness axiom (as this guarantees the existence of the supremum in the proof of Theorem 8.39), this theorem is also equivalent to the completeness axiom.

We conclude this section with two corollaries of the preceding theorems.

8.43. **Corollary.** *Let $(a_n)_{n\in\mathbb{N}}$ be a real sequence.* **COR**

(a) The sequence is convergent if and only if the limit superior and the limit inferior exist and coincide. In this case,

$$\lim_{n\to\infty} a_n = \limsup_{n\to\infty} a_n = \liminf_{n\to\infty} a_n.$$

(b) If the sequence is bounded and has exactly one accumulation point, then it converges.

Proof. (a) If the sequence is convergent, then by the observation after Definition 8.35 it has exactly one accumulation point p. Since every convergent sequence is bounded by Theorem 8.7, both the limit superior and the limit inferior exist by Theorem 8.39. By Theorem 8.41, these must coincide; otherwise, there would be two distinct accumulation points.

Conversely, if the limit superior and limit inferior exist and coincide, then by Theorem 8.41 every accumulation point equals $p = \limsup_{n\to\infty} a_n$. Suppose the entire sequence does not converge to p. Then there exists $\varepsilon > 0$ such that infinitely many $n \in \mathbb{N}$ satisfy $|a_n - p| \geq \varepsilon$. Arrange all such n in increasing order as $n_0 < n_1 < n_2 < \ldots$, forming a subsequence $(a_{n_k})_{k\in\mathbb{N}}$. Since $(a_n)_{n\in\mathbb{N}}$ is bounded by Theorem 8.39, the subsequence $(a_{n_k})_{k\in\mathbb{N}}$ is also bounded and thus, by Theorem 8.42, has a convergent subsequence $(a_{n_{k_j}})_{j\in\mathbb{N}}$, which is also a subsequence of $(a_n)_{n\in\mathbb{N}}$ and hence converges to p. Therefore, there exists $j \in \mathbb{N}$ with $|a_{n_{k_j}} - p| < \varepsilon$, contradicting the definition of the n_k.

(b) If the sequence is bounded, then by Theorem 8.39 both the limit superior and the limit inferior exist, and by Theorem 8.41 they are accumulation points. If there is only one accumulation point, then the two limits must coincide, and convergence follows from (a). ❑

9. Real Series

Date:
March 5, 2026

9.A Definition and Examples.

DEF
Series

9.1. **Definition.** For a real sequence $(a_n)_{n\in\mathbb{N}}$, the sequence

$$c_n := \sum_{k=0}^{n} a_k, \quad n \in \mathbb{N}$$

is called an *(infinite) series*. If the sequence $(c_n)_{n\in\mathbb{N}}$ converges, we also call the series $\sum_{k=0}^{n} a_k$ *convergent* and denote the limit by

$$\sum_{k=0}^{\infty} a_k := \lim_{n\to\infty} c_n.$$

The finite sum $\sum_{k=0}^{n} a_k$ is also called a *partial sum*. ♢

Warning. Some authors use the expression "$\sum_{k=0}^{\infty} a_k$" as a symbolic notation for the infinite series itself, even if the limit does not exist. In this lecture, we will use this expression exclusively to denote the limit. The phrase "the limit exists" is equivalent to "the series converges".

Often it is convenient to start a series at an index $k_0 \geq 1$ instead of 0, i.e.

$$\sum_{k=k_0}^{n} a_k.$$

All statements in this chapter apply analogously to such series if we consider the restriction $n \geq k_0$.

An example of an infinite series can be obtained from **Example 8.2(c)** by suitably specifying the rules by which a farmer plans his potato production. Let us assume that the price p_n that he can obtain per kilogram in year n depends directly on the quantity he offers, and decreases the more he offers. For instance, we can assume that the price can be expressed as

$$p_n = c - da_n,$$

where c and d are positive real numbers. This means that the price is exactly c if no potatoes are offered at all, and decreases linearly as the offered quantity increases. Of course, this is a very simplified assumption, since a market normally operates according to much more complicated rules. For simplicity, we make this assumption here to obtain a formula we can work with[12].

Furthermore, assume that the farmer produces more potatoes a_{n+1} in the following year the higher the price p_n was. For example, he could use the formula

$$a_{n+1} = bp_n$$

with $b > 0$, adjusting the quantity simply proportionally to the previous year's price.

If the farmer produces the quantity a_n in year n, the price is $p_n = c - da_n$, and thus the quantity in the following year is

$$a_{n+1} = bp_n = bc - bda_n.$$

[12] This procedure is almost always unavoidable when modeling real situations mathematically, i.e., translating them into mathematical formulas. In economic contexts, this is usually more obvious than in technical or natural science applications, but simplifications are almost always necessary there as well. It is important, however, not to forget these simplifying assumptions when interpreting the results.

Starting with an arbitrary quantity a_0, we get

$$a_1 = bc - bda_0,$$

$$a_2 = bc - bda_1 = bc - bd(bc - bda_0) = bc(1 - bd) + b^2d^2a_0,$$

$$a_3 = bc - bda_2 = bc - bd(bc(1 - bd) + b^2d^2a_0) = bc(1 - bd + b^2d^2) - b^3d^3a_0.$$

By induction, one can then prove that for any n the equation

$$a_n = bc\left(\sum_{k=0}^{n-1}(-bd)^k\right) + (-bd)^n a_0$$

holds.

If we now want to know whether the produced quantity approaches a fixed quantity a over the years (a so-called *market equilibrium*), this is mathematically nothing other than the question of the convergence of the sequence a_n. Since the sequence contains the series $\sum_{k=0}^{n-1}(-bd)^k$ as a component, we need to investigate when this series converges. ♣

EXAMPLE
Decimal fraction

9.2. **Example.** Another example of series are the infinite decimal fractions already used in Section 5.B without a precise definition. For simplicity, we restrict ourselves to positive decimal fractions between 0 and 10, which can be written as

$$d_0.d_1d_2d_3\ldots$$

with $d_k \in \{0, \ldots, 9\}$. The value of this infinite decimal fraction is then

$$\sum_{k=0}^{\infty} d_k 10^{-k},$$

i.e., to ensure that this is a well-defined value, we first need to prove that the series $\sum_{k=0}^{n} a_k$ with $a_k = d_k 10^{-k}$ converges at all. ♣

From the sum rule for convergent sequences in Theorem 8.9, it follows immediately, because

$$\sum_{k=0}^{n}(a_k + b_k) = \sum_{k=0}^{n} a_k + \sum_{k=0}^{n} b_k$$

(which follows from the associative and commutative laws), that for two convergent series $\sum_{k=0}^{n} a_k$ and $\sum_{k=0}^{n} b_k$, the series $\sum_{k=0}^{n}(a_k + b_k)$ also converges, and

$$\sum_{k=0}^{\infty}(a_k + b_k) = \sum_{k=0}^{\infty} a_k + \sum_{k=0}^{\infty} b_k$$

holds. Similarly, for any $\lambda \in \mathbb{R}$, because

$$\sum_{k=0}^{n}(\lambda a_k) = \lambda \sum_{k=0}^{n} a_k,$$

(which follows from the distributive law) with Corollary 8.11, for any convergent series $\sum_{k=0}^{n} a_k$, the equation

$$\sum_{k=0}^{\infty}(\lambda a_k) = \lambda \sum_{k=0}^{\infty} a_k$$

holds.

Multiplication of convergent series, however, is considerably more complicated.

The reason is that in general

$$\sum_{k=0}^{n}(a_k \cdot b_k) \neq \left(\sum_{k=0}^{n} a_k\right) \cdot \left(\sum_{k=0}^{n} b_k\right)$$

For example, already for $n = 1$ we have

$$\sum_{k=0}^{1}(a_k \cdot b_k) = a_0b_0 + a_1b_1$$

but

$$\left(\sum_{k=0}^{1} a_k\right) \cdot \left(\sum_{k=0}^{1} b_k\right) = (a_0 + a_1)(b_0 + b_1) = a_0b_0 + a_0b_1 + a_1b_0 + a_1b_1.$$

We will present the correct multiplication formula later in Theorem 9.20.

9.B Convergence Criteria for Infinite Series.

We now want to state conditions under which we can ensure that an infinite series converges. Ideally, one would like a simple criterion of the form "the series converges if and only if the a_k satisfy condition x." Unfortunately, there is no such simple condition x. The conditions we give below are either only sufficient or only necessary, or they are not simple, or they apply only to series with special properties.

We begin with a series that is already known.

THM
Convergence of geometric series

9.3. **Theorem.** *The geometric series*

$$\sum_{k=0}^{n} x^k$$

converges if and only if $|x| < 1$. In this case, we have

$$\sum_{k=0}^{\infty} x^k = \frac{1}{1-x}.$$

Proof. According to Theorem 4.3, for $x \neq 1$

$$\sum_{k=0}^{n} x^k = \frac{1 - x^{n+1}}{1-x} = \frac{1}{1-x} - \frac{1}{1-x}x^{n+1}.$$

In the case $|x| < 1$, using the rules for limits and Example 8.1(g)

$$\sum_{k=0}^{\infty} x^k = \lim_{n\to\infty}\sum_{k=0}^{n} x^k = \frac{1}{1-x} - \frac{1}{1-x}\underbrace{\lim_{n\to\infty} x^{n+1}}_{=0} = \frac{1}{1-x}.$$

In the case $|x| > 1$ or $x = -1$, x^{n+1} does not converge by Example 8.1(g), so the series also does not converge. In the case $x = 1$, we have $\sum_{k=0}^{n} x^k = n + 1$, which also does not converge. ❑

Many series, however, are not of this form. Another a Cauchy sequence. It is indeed an "if and only if" criterion, but not directly for the terms of the series, rather for partial sums.

9.4. **Theorem.** *Let $(a_n)_{n\in\mathbb{N}}$ be a real sequence. Then the limit $\sum_{k=0}^{\infty} a_k$ exists if and only if for all $\varepsilon > 0$ there exists a $C(\varepsilon) \in \mathbb{N}$ such that the inequality*

$$\left| \sum_{k=m}^{n} a_k \right| < \varepsilon$$

holds for all $n \geq m \geq C(\varepsilon)$.

THM
Cauchy's Convergence Criterion for Series

Proof. For the sequence

$$c_n = \sum_{k=0}^{n} a_k$$

we have

$$d(b_n, b_{m-1}) = \left| \sum_{k=m}^{n} a_k \right| .$$

Hence, the statement follows directly from Theorem 8.32, because the sequence c_n is a Cauchy sequence if and only if the stated condition holds. ❑

From Theorem 9.4, the following necessary condition for the a_k immediately follows.

9.5. **Theorem.** *If the limit $\sum_{k=0}^{\infty} a_k$ exists, then $\lim_{n\to\infty} a_n = 0$.*

THM

Proof. We have $a_n = \sum_{k=n}^{n} a_k$. Thus, for any $\varepsilon > 0$ and all $n \geq C(\varepsilon)$ from Theorem 9.4, the inequality $|a_n - 0| = |a_n| \leq \varepsilon$ holds, and therefore a_n converges to 0. ❑

9.6. **Example.** The series $\sum_{k=0}^{n}(-1)^k$ does not converge, because the sequence $(-1)^n$ does not converge to zero. ♣

EXAMPLE
Alternating series

A sufficient criterion for convergence can be derived from Theorems 8.7 and 8.29.

9.7. **Theorem.** *Let $(a_n)_{n\in\mathbb{N}}$ be a real sequence with $a_n \geq 0$ for all $n \in \mathbb{N}$. Then the limit $\sum_{k=0}^{\infty} a_k$ exists if and only if the series is bounded, i.e., if there exists a $K \in \mathbb{R}$ such that*

$$\sum_{k=0}^{n} a_k \leq K \text{ for all } n \in \mathbb{N}.$$

THM
Convergent series are bounded

Proof. Since $a_n \geq 0$, the sequence $b_n = \sum_{k=0}^{n} a_k \geq 0$ is monotonically increasing. It is bounded if and only if it is bounded above. Therefore, the statement follows from Theorems 8.7 and 8.29. ❑

EXAMPLE
Harmonic series

9.8. **Example.** Consider the series

$$\sum_{k=1}^{n} \frac{1}{k} \quad \text{and} \quad \sum_{k=1}^{n} \frac{1}{k^2}.$$

For the first series — the *harmonic series* — consider the partial sums

$$\sum_{k=1}^{2^{n+1}} \frac{1}{k}$$

and split this sum into the terms with indices $\{1, 2\}$ and $\{2^p + 1, \ldots, 2^{p+1}\}$ for $p = 1, \ldots, n$ (i.e., $\{1, 2\}$, $\{3, 4\}$, $\{5, 6, 7, 8\}$, $\{9, 10, \ldots, 16\}$, etc.). Then we have

$$\begin{aligned}
\sum_{k=1}^{2^{n+1}} \frac{1}{k} &= 1 + \frac{1}{2} + \sum_{p=1}^{n} \left(\sum_{k=2^p+1}^{2^{p+1}} \frac{1}{k} \right) \\
&= 1 + \frac{1}{2} + \left(\frac{1}{3} + \frac{1}{4} \right) + \left(\frac{1}{5} + \frac{1}{6} + \frac{1}{7} + \frac{1}{8} \right) + \ldots + \sum_{k=2^n+1}^{2^{n+1}} \frac{1}{k} \\
&\geq 1 + \frac{1}{2} + \left(\frac{1}{4} + \frac{1}{4} \right) + \left(\frac{1}{8} + \frac{1}{8} + \frac{1}{8} + \frac{1}{8} \right) + \ldots + \sum_{k=2^n+1}^{2^{n+1}} \frac{1}{2^{n+1}}.
\end{aligned}$$

Each partial sum contains 2^p terms each $\geq 1/2^{p+1}$, so for each partial sum

$$\sum_{k=2^p+1}^{2^{p+1}} \frac{1}{k} \geq \sum_{k=2^p+1}^{2^{p+1}} \frac{1}{2^{p+1}} = 2^p \frac{1}{2^{p+1}} = \frac{1}{2}.$$

Hence

$$\sum_{k=1}^{2^{n+1}} \frac{1}{k} \geq 1 + \frac{1}{2} + n\frac{1}{2} = \frac{n+3}{2}$$

and the series is unbounded. Consequently, the harmonic series does not converge according to Theorem 9.7, even though one might initially suspect convergence because $1/n \to 0$ as $n \to \infty$.

For the series $\sum_{k=1}^{n} \frac{1}{k^2}$, we now show that it is bounded by $K = 2$ and therefore converges by Theorem 9.7. Let $n \in \mathbb{N}$ be given and let $m \in \mathbb{N}$ be such that $n \leq 2^{m+1} - 1$. Then, using a similar index splitting as above and Theorem 4.3,

$$\begin{aligned}
\sum_{k=1}^{n} \frac{1}{k^2} &\leq \sum_{k=1}^{2^{m+1}-1} \frac{1}{k^2} = \sum_{p=0}^{m} \left(\sum_{k=2^p}^{2^{p+1}-1} \frac{1}{k^2} \right) \leq \sum_{p=0}^{m} \left(2^p \frac{1}{(2^p)^2} \right) \\
&= \sum_{p=0}^{m} \left(\frac{1}{2} \right)^p = \frac{1 - (1/2)^{m+1}}{1 - 1/2} \leq \frac{1}{1 - 1/2} = 2.
\end{aligned}$$

Later, in Example 9.16, we will show that all series of the form

$$\sum_{k=1}^{n} \frac{1}{k^p}$$

converge for $p \geq 2$. Using more advanced analytic methods, one can also prove that $\sum_{k=1}^{\infty} \frac{1}{k^2} = \pi^2/6$. ♣

To conclude this section, we consider a convergence criterion for series whose terms alternate in sign at each step. Series of this type are called *alternating*. For the

following theorem, it is convenient to write the terms in the form $(-1)^n a_n$ with $a_n \geq 0$.

THM Leibniz's Convergence Criterion for Alternating Series

9.9. **Theorem.** *For a monotonically decreasing sequence $(a_n)_{n\in\mathbb{N}}$ with $a_n \geq 0$ for all $n \in \mathbb{N}$ and $\lim_{n\to\infty} a_n = 0$, the limit*

$$\sum_{k=0}^{\infty} (-1)^k a_k$$

exists.

Proof. Consider the sequence $b_n = \sum_{k=0}^{n} (-1)^k a_k$ and the two subsequences

$$b_{2n} = (b_0, b_2, b_4, \ldots) \quad \text{and} \quad b_{2n+1} = (b_1, b_3, b_5, \ldots).$$

Due to the monotonicity of the a_k, we have

$$b_{2n+2} - b_{2n} = (-1)^{2n+2} a_{2n+2} + (-1)^{2n+1} a_{2n+1} = a_{2n+2} - a_{2n+1} \leq 0$$

and

$$b_{2n+3} - b_{2n+1} = (-1)^{2n+3} a_{2n+3} + (-1)^{2n+2} a_{2n+2} = -a_{2n+3} + a_{2n+2} \geq 0.$$

Hence, b_{2n} is monotonically decreasing and b_{2n+1} is monotonically increasing. Moreover, since $b_{2n+1} - b_{2n} = (-1)^{2n+1} a_{2n+1} \leq 0$, we have

$$b_{2n+1} \leq b_{2n} \leq b_0 \quad \text{and} \quad b_{2n} \geq b_{2n+1} \geq b_1$$

for all $n \in \mathbb{N}$. Therefore, both subsequences are monotone and bounded, and thus converge. For the limits, the rules for limits give

$$\lim_{n\to\infty} b_{2n} - \lim_{n\to\infty} b_{2n+1} = \lim_{n\to\infty} (b_{2n} - b_{2n+1}) = \lim_{n\to\infty} -(-1)^{2n+1} a_{2n+1} = \lim_{n\to\infty} a_{2n+1} = 0,$$

so the two limits coincide. Denote this common limit by b. Then, for every $\varepsilon > 0$, there exist indices $N_1(\varepsilon), N_2(\varepsilon) \in \mathbb{N}$ such that

$$|b_{2n} - b| < \varepsilon \quad \text{for all } n \geq N_1(\varepsilon) \quad \text{and} \quad |b_{2n+1} - b| < \varepsilon \quad \text{for all } n \geq N_2(\varepsilon).$$

For all $n \geq N(\varepsilon) := \max\{2N_1(\varepsilon), 2N_2(\varepsilon) + 1\}$, it follows that

$$|b_n - b| < \varepsilon,$$

which proves the convergence. ❑

EXAMPLE Alternating harmonic series

9.10. **Example.** In Example 9.8, we saw that the harmonic series $\sum_{k=1}^{n} \frac{1}{k}$ does not converge. In contrast, the *alternating harmonic series*

$$\sum_{k=1}^{n} \frac{(-1)^{k-1}}{k} = 1 - \frac{1}{2} + \frac{1}{3} - \frac{1}{4} \pm \ldots \pm \frac{1}{n}$$

satisfies the conditions of Theorem 9.9 and therefore converges. The exact value of the limit requires methods that will be introduced later. ♣

The alternating harmonic series is a nice example showing that the convergence behavior of a series can change if one rearranges the terms. We can, for instance,

reorder the terms as follows:

$$\begin{aligned}
&1 - \frac{1}{2} + \frac{1}{3} - \frac{1}{4} \\
&+ \left(\frac{1}{5} + \frac{1}{7}\right) - \frac{1}{6} \\
&+ \left(\frac{1}{9} + \frac{1}{11} + \frac{1}{13} + \frac{1}{15}\right) - \frac{1}{8} \\
&+ \left(\frac{1}{17} + \frac{1}{19} + \frac{1}{21} + \frac{1}{23} + \frac{1}{25} + \frac{1}{27} + \frac{1}{29} + \frac{1}{31}\right) - \frac{1}{10} \\
&+ \dots
\end{aligned}$$

That is, starting from $1/5$, for $p = 1, 2, 3, \dots$ we always take the next 2^p unused positive terms and then the next unused negative term. One can easily see that every term of the alternating harmonic series eventually appears exactly once. Nevertheless, the new series has a completely different convergence behavior: in each line, the sum of the fractions in parentheses is greater than $1/4$, from which at most $1/6$ is subtracted. Therefore, each line contributes more than $1/12$, so the partial sum after n lines exceeds $n/12$ and grows without bound. By rearrangement, the convergent alternating series becomes unbounded and thus divergent.

In the following section, we will consider a stronger notion of convergence for series, under which both the convergence property and the limit are preserved under any rearrangement.

9.C Absolute Convergence.

9.11. **Definition.** A series $\sum_{k=0}^{n} a_k$ is called *absolutely convergent* if the series **DEF** Absolute convergence

$$\sum_{k=0}^{n} |a_k|$$

converges. ♢

That absolute convergence is indeed a stronger concept than "ordinary" convergence[13] is shown by the following theorem and the subsequent discussion.

9.12. **Theorem.** *Every absolutely convergent series is also convergent in the usual sense.* **THM**

Proof. If the series converges absolutely, then according to the Cauchy convergence criterion from Theorem 9.4, for every $\varepsilon > 0$ there exists a $C(\varepsilon) \in \mathbb{N}$ such that

$$\sum_{k=m}^{n} |a_k| = \left| \sum_{k=m}^{n} |a_k| \right| < \varepsilon$$

for all $n \geq m \geq C(\varepsilon)$. Using the triangle inequality $(n - m)$ times, we obtain

$$\left| \sum_{k=m}^{n} a_k \right| \leq \sum_{k=m}^{n} |a_k| < \varepsilon,$$

and hence the series converges according to Theorem 9.4. ❑

[13] One says that a property A is *stronger* than a property B if the implication $A \Rightarrow B$ holds, but the implication $B \Rightarrow A$ does not.

Warning. Note that the statement of the theorem does not imply that the two limits $\sum_{k=0}^{\infty} |a_k|$ and $\sum_{k=0}^{\infty} a_k$ are equal. It only states that the second limit exists if the first one exists. The converse does not hold, as shown by the alternating harmonic series, which converges but does not converge absolutely. This shows that convergence does not imply absolute convergence; therefore, absolute convergence is indeed a stronger property. The series $\sum_{k=1}^{n} 1/k^2$, on the other hand, converges absolutely, because for $a_k = 1/k^2$, it is obvious that $|a_k| = a_k$ since $1/k^2 > 0$.

The following theorem shows that in the case of absolute convergence, any rearrangement does not change the convergence behavior, and the limit even remains the same. We first define a rearrangement formally.

9.13. **Definition.** A series $\sum_{k=0}^{n} a_{\tau(k)}$ with a map $\tau : \mathbb{N} \to \mathbb{N}$ is called a *rearrangement* of a series $\sum_{k=0}^{n} a_k$ if τ is bijective. ♢

DEF rearrangement

9.14. **Theorem.** *Let $\sum_{k=0}^{n} a_k$ be an absolutely convergent series. Then every rearrangement $\sum_{k=0}^{n} a_{\tau(k)}$ of the series also converges, and we have*

$$\sum_{k=0}^{\infty} a_{\tau(k)} = \sum_{k=0}^{\infty} a_k.$$

THM Convergent rearrangements

Proof. Let

$$b := \sum_{k=0}^{\infty} a_k.$$

We need to prove that $\sum_{k=0}^{n} a_{\tau(k)}$ converges to b. Since the original series converges absolutely, by Theorem 9.4 for every $\varepsilon > 0$ there exists $k_0 = C(\varepsilon/2) \in \mathbb{N}$ such that

$$c_m := \sum_{k=k_0}^{m} |a_k| < \frac{\varepsilon}{2}$$

for all $m \geq k_0$. The sequence c_m is thus monotone and bounded, and therefore converges to a limit $\sum_{k=k_0}^{\infty} |a_k|$, for which according to Theorem 8.14

$$\sum_{k=k_0}^{\infty} |a_k| \leq \frac{\varepsilon}{2}.$$

It follows[14]

$$\left| b - \sum_{k=0}^{k_0-1} a_k \right| = \left| \sum_{k=k_0}^{\infty} a_k \right| \leq \sum_{k=k_0}^{\infty} |a_k| \leq \frac{\varepsilon}{2}.$$

Since every k appears in the sequence $(\tau(j))_{j\in\mathbb{N}}$, for each k there exists a j_k such that $k = \tau(j_k)$. Now set $N(\varepsilon) := \max\{j_0, j_1, \ldots, j_{k_0-1}\}$, so that every $k = 0, \ldots, k_0 - 1$ appears in the set $\{\tau(0), \tau(1), \ldots, \tau(N(\varepsilon))\}$. Then for all $m \geq N(\varepsilon)$ we have

$$\left| \sum_{k=0}^{m} a_{\tau(k)} - \sum_{k=0}^{k_0-1} a_k \right| = \left| \sum_{\substack{k=0 \\ k\neq j_0,\ldots,j_{k_0-1}}}^{m} a_{\tau(k)} \right| \leq \sum_{\substack{k=0 \\ k\neq j_0,\ldots,j_{k_0-1}}}^{m} |a_{\tau(k)}| \leq \sum_{k=k_0}^{\max\{\tau(0),\ldots,\tau(m)\}} |a_k| < \frac{\varepsilon}{2},$$

[14] If the upper summation limit is finite, the inequality used here follows from the triangle inequality. By Theorem 8.14 it carries over to the limits.

where in the penultimate step we used that each index k occurs at most once in the sequence. Thus, for all $m \geq N(\varepsilon)$,

$$\left|\sum_{k=0}^{m} a_{\tau(k)} - b\right| \leq \left|\sum_{k=0}^{m} a_{\tau(k)} - \sum_{k=0}^{k_0-1} a_k\right| + \left|\sum_{k=0}^{k_0-1} a_k - b\right| < \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon,$$

and hence the claimed convergence. ❑

Since absolute convergence evidently is a useful property (we provide further evidence in the last theorem of this section), it is reasonable to find criteria with which we can verify it for a given series. The following two theorems provide two such criteria.

THM
Majorant Criterion or Comparison Test

9.15. **Theorem.** *Let $\sum_{k=0}^{n} c_k$ be a convergent series with $c_k \geq 0$ for all $k \in \mathbb{N}$. Let $k_0 \in \mathbb{N}$ and $(a_k)_{k\in\mathbb{N}}$ be a sequence with $|a_k| \leq c_k$ for all $k \in \mathbb{N}$ with $k \geq k_0$. Then the series*

$$\sum_{k=0}^{n} a_k$$

converges absolutely. The series $\sum_{k=0}^{n} c_k$ is then called a majorant *of the series $\sum_{k=0}^{n} a_k$.*

Proof. In the case $k_0 = 0$, the statement follows from

$$\left|\sum_{k=m}^{n} |a_k|\right| = \sum_{k=m}^{n} |a_k| \leq \sum_{k=m}^{n} c_k = \left|\sum_{k=m}^{n} c_k\right|$$

analogously to the proof of Theorem 9.12 from Theorem 9.4.

In the case $k_0 > 0$, define $\tilde{a}_k := c_k$ for $k = 0, \ldots, k_0 - 1$ and $\tilde{a}_k := a_k$ for $k \geq k_0$. Then the series $\sum_{k=0}^{n} \tilde{a}_k$ satisfies the condition of the theorem for $k_0 = 0$ and converges by the first part of the proof. From the definition of $\tilde{a}_k$ it follows for $n \geq k_0$ that

$$\sum_{k=0}^{n} a_k = \sum_{k=0}^{n} \tilde{a}_k - \sum_{k=0}^{k_0-1} c_k + \sum_{k=0}^{k_0-1} a_k,$$

and since the series $\sum_{k=0}^{n} \tilde{a}_k$ converges as $n \to \infty$ and the other two terms on the right-hand side are independent of n, the series $\sum_{k=0}^{n} a_k$ also converges. ❑

EXAMPLE
Convergence of $\frac{1}{k^p}$

9.16. **Example.** Consider the series

$$\sum_{k=0}^{n} \frac{1}{k^p}$$

for $p \geq 2$. For $p = 2$ we have already proved convergence in Example 9.8. Since $k \geq 1$, we have $k^q \geq 1$ and therefore $1/k^q \leq 1$ for all $q \geq 1$. For $p \geq 3$ we obtain

$$\left|\frac{1}{k^p}\right| = \frac{1}{k^p} = \frac{1}{k^{p-2}} \cdot \frac{1}{k^2} \leq \frac{1}{k^2}.$$

Hence, absolute convergence (and thus usual convergence) follows from the majorant criterion with $k_0 = 0$. ♣

9.17. **Example.** For the decimal fractions from Example 9.2, we have

EXAMPLE Decimal fractions

$$|d_k 10^{-k}| \le 9 \cdot 10^{-k}.$$

Since the series

$$\sum_{k=0}^{n} 9 \cdot 10^{-k} = 9 \sum_{k=0}^{n} \left(\frac{1}{10}\right)^k$$

converges to $9\frac{1}{1-1/10} = 10$ by Theorem 9.3 and obviously has positive terms, every decimal fraction series therefore converges absolutely by the majorant criterion. In particular, every infinite decimal fraction thus has a well-defined value. ♣

9.18. **Theorem.** *Let $\sum_{k=0}^{n} a_k$ be a series with $a_k \neq 0$ for all $k \ge k_0$. Let $\theta \in \mathbb{R}$ with $0 < \theta < 1$ such that*

THM Quotient Criterion or Ratio Test

$$\left|\frac{a_{k+1}}{a_k}\right| \le \theta \quad \text{for all } k \ge k_0. \tag{9.1}$$

Then the series $\sum_{k=0}^{n} a_k$ converges absolutely.

Proof. By complete induction, the assumption implies the inequality

$$|a_k| \le \theta^{k-k_0} |a_{k_0}|$$

for all $k \in \mathbb{N}$ with $k \ge k_0$. The series

$$\sum_{k=0}^{n} |a_{k_0}| \theta^{k-k_0}$$

is therefore a majorant for $\sum_{k=0}^{n} a_k$ for $k \ge k_0$. Since

$$\sum_{k=0}^{n} |a_{k_0}| \theta^{k-k_0} = |a_{k_0}| \theta^{-k_0} \sum_{k=0}^{n} \theta^k$$

converges by Theorem 9.3 because $|\theta| < 1$, convergence follows by Theorem 9.15. ❑

9.19. **Examples.**

EXAMPLES

(a) The series $\sum_{k=0}^{n} a_k$ with $a_k = \frac{k^2}{2^k}$ converges, since

$$\left|\frac{a_{k+1}}{a_k}\right| = \frac{(k+1)^2 2^k}{2^{k+1} k^2} = \frac{1}{2}\frac{(k+1)^2}{k^2} = \frac{1}{2}\left(1 + \frac{1}{k}\right)^2,$$

and this expression is smaller than $8/9 < 1$ for all $k \ge k_0 = 3$.

(b) The example of the harmonic series ($a_k = 1/k$) shows that the formulation of the inequality in (9.1) as "$\left|\frac{a_{k+1}}{a_k}\right| \le \theta < 1$" is important, and the weaker inequality "$\left|\frac{a_{k+1}}{a_k}\right| < 1$" is not sufficient. This weaker inequality is indeed satisfied for the harmonic series because

$$\left|\frac{a_{k+1}}{a_k}\right| = \frac{k}{k+1} < 1.$$

However, since the fraction $k/(k+1)$ approaches 1 for large k, no $\theta < 1$ exists such that (9.1) holds for arbitrarily large k. This must be the case, because the harmonic series is divergent.

(c) The example $a_k = 1/k^2$ shows that the ratio test is sufficient but not necessary for absolute convergence. Similarly to the harmonic series, one sees that no

$\theta < 1$ can be found such that (9.1) holds. Nevertheless, the series is absolutely convergent.

(d) From the definition of lim sup, it follows that the condition of Theorem 9.18 is satisfied if and only if

$$\limsup_{k\to\infty} \left| \frac{a_{k+1}}{a_k} \right| < 1$$

(Exercise). ♣

To conclude this chapter, we return once more to the multiplication of convergent series mentioned at the beginning. As already noted there, the series $\sum_{k=0}^{n} a_k \cdot b_k$ is not the correct expression to represent the product of two series as a single series. The following theorem shows the correct expression. Note that it only holds for absolutely convergent series, since its proof requires an estimate of partial sums in terms of absolute values.

THM Cauchy Product of Series

9.20. **Theorem.** *Let $\sum_{k=0}^{n} a_k$ and $\sum_{k=0}^{n} b_k$ be absolutely convergent series. For each $k \in \mathbb{N}$, define*

$$c_k := \sum_{j=0}^{k} a_{k-j} b_j.$$

Then the series $\sum_{k=0}^{n} c_k$ is also absolutely convergent, and

$$\sum_{k=0}^{\infty} c_k = \left(\sum_{k=0}^{\infty} a_k \right) \left(\sum_{k=0}^{\infty} b_k \right).$$

Proof. Let $a := \sum_{k=0}^{\infty} a_k$, $b := \sum_{k=0}^{\infty} b_k$, and $d_n := \sum_{k=0}^{n} c_k$. We first show that d_n converges in the usual sense and that

$$\lim_{n\to\infty} d_n = ab.$$

Absolute convergence will be proved in the subsequent step.

Let

$$d_n^* := \left(\sum_{k=0}^{n} a_k \right) \left(\sum_{k=0}^{n} b_k \right).$$

By Theorem 8.10, $d_n^* \to ab$ as $n \to \infty$. To show that $d_n \to ab$ as well, we prove that the sequence $(d_n^* - d_n)$ converges with

$$\lim_{n\to\infty} (d_n^* - d_n) = 0. \tag{9.2}$$

Then the claim follows from

$$d_n = \underbrace{d_n^*}_{\to ab} - \underbrace{(d_n^* - d_n)}_{\to 0}$$

by Corollary 8.12.

To prove (9.2), we rewrite d_n^* using the distributive law:

$$d_n^* = \left(\sum_{i=0}^{n} a_i \right) \left(\sum_{j=0}^{n} b_j \right) = \sum_{i=0}^{n} \sum_{j=0}^{n} a_i b_j =: \sum_{i,j=0}^{n} a_i b_j.$$

We rewrite the expression for d_n by introducing $i = k - j$ as an "artificial" new summation index with the condition $i + j = k$:

$$d_n = \sum_{k=0}^{n} \sum_{j=0}^{k} a_{k-j} b_j = \sum_{k=0}^{n} \sum_{j=0}^{k} \sum_{\substack{i=0 \\ i+j=k}}^{k} a_i b_j = \sum_{k=0}^{n} \sum_{\substack{i,j=0 \\ i+j=k}}^{k} a_i b_j = \sum_{\substack{i,j=0 \\ i+j\le n}}^{n} a_i b_j.$$

It follows that

$$d_n^* - d_n = \sum_{\substack{i,j=0 \\ i+j>n}}^{n} a_i b_j. \tag{9.3}$$

Now, let

$$p_n := \left(\sum_{k=0}^{n} |a_k|\right)\left(\sum_{k=0}^{n} |b_k|\right) = \sum_{i,j=0}^{n} |a_i b_j|.$$

By Theorem 8.10, p_n converges. Given $\varepsilon > 0$, there exists $n_0 := C(\varepsilon) \in \mathbb{N}$ such that

$$|p_n - p_{n_0}| < \varepsilon \quad \text{for all } n \geq n_0.$$

The difference $p_n - p_{n_0}$ can also be written as

$$p_n - p_{n_0} = \sum_{i,j=0}^{n} |a_i b_j| - \sum_{i,j=0}^{n_0} |a_i b_j| = \sum_{\substack{i,j=0 \\ i\geq n_0+1 \text{ or } j\geq n_0+1}}^{n} |a_i b_j|. \tag{9.4}$$

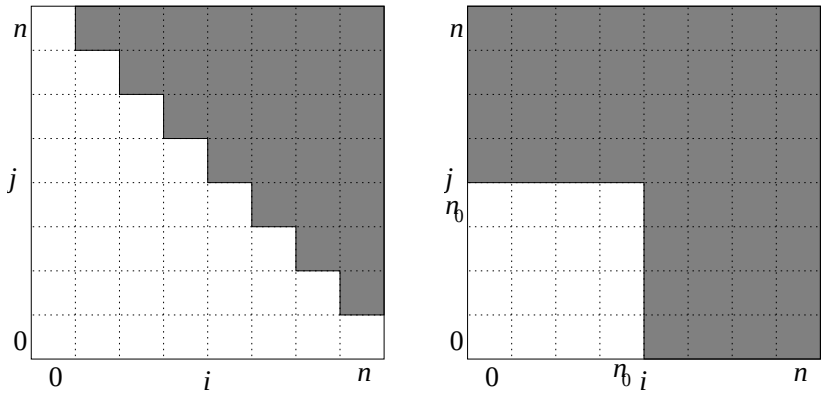


FIGURE 3. Index pairs (i, j) in (9.3) (left) and (9.4) (right), shaded in gray. Illustration for $n = 7$ and $n_0 = 3$.

Comparing the summation indices in the last sum of (9.3) with those in the last sum of (9.4) (shown graphically in Figure 3), we see that for $n \geq 2n_0$, every index pair in (9.3) also appears in (9.4), since if $i + j > n \geq 2n_0$, then $i \geq n_0 + 1$ or $j \geq n_0 + 1$; otherwise $i + j \leq 2n_0$. Hence,

$$|d_n^* - d_n| = \left|\sum_{\substack{i,j=0 \\ i+j>n}}^{n} a_i b_j\right| \leq \sum_{\substack{i,j=0 \\ i+j>n}}^{n} |a_i b_j| \leq \sum_{\substack{i,j=0 \\ i\geq n_0+1 \text{ or } j\geq n_0+1}}^{n} |a_i b_j| = p_n - p_{n_0} < \varepsilon.$$

This proves the convergence (9.2) with $N(\varepsilon) = 2n_0 = 2C(\varepsilon)$.

It remains to show absolute convergence. Applying the first part of the proof to the series $\sum_{k=0}^{n} |a_k|$ and $\sum_{k=0}^{n} |b_k|$, we obtain, due to the absolute convergence of the original series, that the series $\sum_{k=0}^{n} c'_k$ with

$$c'_n := \sum_{k=0}^{n} |a_{n-k}|\,|b_k|$$

converges. Since

$$|c_n| = \left|\sum_{k=0}^{n} a_{n-k} b_k\right| \leq \sum_{k=0}^{n} |a_{n-k}|\,|b_k| = c'_n,$$

the series $\sum_{k=0}^{n} c'_k$ is a convergent majorant, which implies that $\sum_{k=0}^{n} c_k$ converges absolutely. ❑

9.D Complex sequences and series.

Just as in $\mathbb{R}$, one can also define sequences $(z_n)_{n\in\mathbb{N}} = (a_n + ib_n)_{n\in\mathbb{N}}$ in $\mathbb{C}$. Convergence is defined exactly as in $\mathbb{R}$ via the distance $d(z_1, z_2) := |z_1 - z_2|$.

DEF Convergence of a complex sequence

9.21. **Definition.** A complex sequence $(z_n)_{n\in\mathbb{N}}$ is called *convergent*, if there exists a $z \in \mathbb{C}$ such that:

For every $\varepsilon > 0$ there exists an $N(\varepsilon) \in \mathbb{N}$ such that the inequality $d(z_n, z) < \varepsilon$ holds for all $n \geq N(\varepsilon)$. ♢

The notation $\lim_{n\to\infty} z_n := z$ is then used just as in $\mathbb{R}$.

The following theorem is the reason why all theorems for real convergent sequences can be transferred to complex sequences.

THM Convergence of complex and real sequences

9.22. **Theorem.** *A complex sequence $(z_n)_{n\in\mathbb{N}} = (a_n + ib_n)_{n\in\mathbb{N}}$ is convergent if and only if the real sequences $(a_n)_{n\in\mathbb{N}}$ and $(b_n)_{n\in\mathbb{N}}$ are convergent. In this case one has*

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

Proof. Suppose z_n converges with limit $z = a + ib$. Then for all $\varepsilon > 0$ there exists an $N(\varepsilon) \in \mathbb{N}$ with

$$|z_n - z| < \varepsilon \qquad \text{for all } n \geq N(\varepsilon).$$

Now we have

$$|z_n - z| = \sqrt{(a_n - a)^2 + (b_n - b)^2} \geq \sqrt{(a_n - a)^2} = |a_n - a|;$$

analogously one shows $|z_n - z| \geq |b_n - b|$. It follows that

$$|a_n - a| < \varepsilon \quad \text{and} \quad |b_n - b| < \varepsilon \qquad \text{for all } n \geq N(\varepsilon),$$

hence a_n and b_n converge to a and b.

Conversely, assume a_n and b_n converge with limits a and b. Let $N_a(\varepsilon)$ and $N_b(\varepsilon)$ be the corresponding indices from Definition 8.3. Then for all $\varepsilon > 0$ and $n \geq N(\varepsilon) := \max\{N_a(\varepsilon/2), N_b(\varepsilon/2)\}$ and $z = a + ib$ we have the inequality

$$|z_n - z|^2 = (a_n - a)^2 + (b_n - b)^2 < \frac{\varepsilon^2}{4} + \frac{\varepsilon^2}{4} = \frac{\varepsilon^2}{2},$$

and therefore

$$|z_n - z| < \sqrt{\frac{\varepsilon^2}{2}} = \frac{\varepsilon}{\sqrt{2}} < \varepsilon.$$

Thus z_n converges to z. ❑

In a similar way one proves that z_n is a Cauchy sequence if and only if a_n and b_n are Cauchy sequences.

Theorem 9.22 forms the basis for transferring all computational rules for real limits to complex limits. For complex conjugate sequences, the theorem also immediately implies that z_n converges if and only if $\bar{z}_n$ converges, and in this case

$$\lim_{n\to\infty} \bar{z}_n = \overline{\lim_{n\to\infty} z_n}. \tag{9.5}$$

For complex series $\sum_{k=0}^{n} z_k$, all definitions from $\mathbb{R}$ can be carried over word by word, in particular convergence and absolute convergence. The comparison test, the ratio test, and the theorem on the Cauchy product of series all hold in $\mathbb{C}$ just as in $\mathbb{R}$, with the majorant in the comparison test still being a real series.

9.E Power series.

In this section we consider a particular class of series, the so called power series. These are defined by

$$\sum_{k=0}^{n} c_k(x-x_0)^k.$$

Here c_0, c_1, ... and x_0 are fixed real or complex numbers and x denotes a real or complex variable. From now on we consider the complex case but the statements for the real case are completely analogous. For each $n \in \mathbb{N}$ we can understand

$$f_n(x) := \sum_{k=0}^{n} c_k(x-x_0)^k \tag{9.6}$$

as a map[15] $f_n\colon \mathbb{C} \to \mathbb{C}$. The question is now, for which x does the limit

$$\lim_{n\to\infty} f_n(x) = \sum_{k=0}^{\infty} c_k(x-x_0)^k$$

exist, i.e., for which x does this sequence converge, possibly even absolutely? This is formalized by defining the *convergence radius* of the series, which is the largest number $r \geq 0$ such that (9.6) converges for all $x \in \mathbb{C}$ with $d(x,x_0) < r$. The following theorem shows that the knowledge of a single point $x \neq x_0$ for which the sequence converges already suffices to give an estimate for r and even ensure absolute convergence.

THM Convergence radius of power series

9.23. **Theorem.** *Consider a power series (9.6), which converges for some $x = x_1 \in \mathbb{C}$ with $x_1 \neq x_0$.*

Then the power series converges absolutely for all $x \in D := \{x \in \mathbb{C} \,|\, d(x,x_0) < d(x_1,x_0)\}$. In particular, the convergence radius r of the series satisfies the inequality $r \geq d(x_1,x_0)$.

Proof. Define $g_k(x) := c_k(x-x_0)^k$. By assumption, the series $\sum_{k=0}^{n} g_k(x_1)$ converges, hence by Theorem 9.5 we have $|g_k(x_1)| \to 0$ as $k \to \infty$. From Theorem 8.7 it follows[16] that there exists $M \in \mathbb{R}$ with $|g_k(x_1)| \leq M$ for all $k \in \mathbb{N}$.

Thus, for all $x \in D$:

$$|g_k(x)| = |c_k(x-x_0)^k| = |c_k(x_1-x_0)^k| \cdot \left|\frac{x-x_0}{x_1-x_0}\right|^k \leq M\theta^k$$

with

$$\theta = \left|\frac{x-x_0}{x_1-x_0}\right| < \frac{\rho}{|x_1-x_0|} = 1.$$

Since the sequence

$$\sum_{k=0}^{n} M\theta^k = M\sum_{k=0}^{n} \theta^k$$

converges by Theorem 9.3 and Corollary 8.11, it is a convergent majorant for the sequence $\sum_{k=0}^{n} g_k(x)$. Thus, by Theorem 9.15 this sequence converges absolutely. ❑

[15] If x_0 and the c_k are real numbers, we can also understand f_n as a real map $f_n : \mathbb{R} \to \mathbb{R}$.

[16] Here we use the fact, justified in the previous section, that all statements on sequences and series made so far also hold in the complex case.

10. Vector spaces: definition and examples

Date:
March 5, 2026

In this part of the course we introduce the basic concepts of Linear Algebra. What is "Linear Algebra"? Linear Algebra studies "linear structures", which we will define below as *vector spaces* (or *linear spaces*), and the *linear maps* between them that are compatible with the linear structure. These are rather abstract notions, but this is actually the main strength of Linear Algebra: linear spaces and maps occur very frequently throughout mathematics in many different contexts. Precisely because one ignores ("abstracts from") their concrete individual features and focuses on their essential common properties, it is possible to apply the results of Linear Algebra in all these different situations. Historically, it was a long and arduous process to arrive at this level of abstraction, but at the end of this development, we have a very powerful, widely applicable, and successful theory. This had the effect that *linear* problems are considered to be *easy*, whereas *nonlinear* problems are frequently particularly *difficult*. Here are two examples coming from physics.

J.-B.-J. Fourier (1768–1830)

- The Heat Equation, which describes the temporal development of the temperature distribution in a solid, is a *linear* partial differential equation. The relevant theory for its solution was already developed by Jean-Baptiste-Joseph Fourier in 1822 ("Théorie analytique de la chaleur").
- By contrast, the *Navier-Stokes-Equations*, which describe the movement of fluids, are *nonlinear* partial differential equations, and the question whether there are always smooth solutions for all times given reasonable initial conditions is an open problem (there is recent progress suggesting that this may not be the case, though). It is one of the seven Millennium Problems of the Clay Foundation; a solution would make you a million US-dollars richer.

What is the meaning of the word "linear"? Here are three linear equations or systems of equations:

EXAMPLES
linear equations

10.1. **Examples.**

(1) We want to find $w, x, y, z \in \mathbb{R}$ such that

$$w + x + y + z = 0 \qquad \text{and} \qquad x + 2y + 3z = 0\,.$$

You probably have learned in school how to solve such systems of equations (and we will discuss this in some depth in these lectures as well). The solutions are

$$(w, x, y, z) = (a, -2a + b, a - 2b, b) \qquad \text{with } a, b \in \mathbb{R}.$$

(2) We look for sequences $(a_n)_{n\in\mathbb{N}}$ of real numbers that satisfy

$$a_{n+2} = a_{n+1} + a_n \qquad \text{for all } n \in \mathbb{N}.$$

The well-known sequence $(0, 1, 1, 2, 3, 5, 8, \ldots)$ of Fibonacci numbers is one solution, but there are more. All solutions can be written in the form

$$a_n = a\Big(\frac{1+\sqrt{5}}{2}\Big)^n + b\Big(\frac{1-\sqrt{5}}{2}\Big)^n \qquad \text{with } a, b \in \mathbb{R}.$$

(3) We want (twice differentiable) functions $f\colon \mathbb{R} \to \mathbb{R}$ such that

$$f''(x) + f(x) = 0 \qquad \text{for all } x \in \mathbb{R}.$$

In this case the solutions are given by

$$f(x) = a\cos x + b\sin x \qquad \text{with } a, b \in \mathbb{R}. \qquad \clubsuit$$

Even though the objects considered in these examples are of quite different nature (quadruples of real numbers, sequences of real numbers, twice differentiable real functions), the structure of the solution set is in all three cases very similar. That this is necessarily the case is a general result for linear equations. More concretely, the linearity manifests itself in the form that the *sum* of two solutions is again a solution and that every *multiple* of a solution is again a solution. These two operations, addition and taking multiples, i.e., multiplication by a "scalar" (which is a real number in all the examples above), make up the linear structure, which we formalize in the following definition.

10.2. **Definition.** Let K be a field. A *K-vector space* or *vector space over K* or *linear space over K* is a quintuple $(V, +, \mathbf{0}, -, \cdot)$, where V is a set, $+ \colon V \times V \to V$ is a binary operation, $\mathbf{0} \in V$ is an element, and $- \colon V \to V$ and $\cdot \colon K \times V \to V$ (scalar multiplication) are maps, such that $(V, +, \mathbf{0}, -)$ is a commutative group and the following additional axioms are satisfied. **DEF** vector space

(1) $\forall v \in V \colon 1 \cdot v = v$ (here $1 \in K$ is the unit element of the field K).

(2) (associativity of scalar multiplication)
$\forall \lambda, \mu \in K \; \forall v \in V \colon \lambda \cdot (\mu \cdot v) = (\lambda\mu) \cdot v$.

(3) (distributive laws)
$\forall \lambda, \mu \in K \; \forall v \in V \colon (\lambda + \mu) \cdot v = \lambda \cdot v + \mu \cdot v$ and
$\forall \lambda \in K \; \forall v, w \in V \colon \lambda \cdot (v + w) = \lambda \cdot v + \lambda \cdot w$.

We frequently abbreviate $\lambda \cdot v$ to λv. The elements of a vector space are also called *vectors*. $\mathbf{0} \in V$ is the *zero vector* of V.

A vector space over $\mathbb{R}$ is also called a *real vector space*; similarly, a vector space over $\mathbb{C}$ is a *complex vector space*. ♢

Note that the symbol "+" in these axioms has two different meanings: it can denote the addition in the field K, but also the addition in V!

For completeness and as a reminder, here are the four axioms for a commutative group $(V, +, \mathbf{0}, -)$.

(1) (associativity of addition) $\forall v_1, v_2, v_3 \in V \colon (v_1 + v_2) + v_3 = v_1 + (v_2 + v_3)$.

(2) (commutativity of addition) $\forall v, w \in V \colon v + w = w + v$.

(3) (zero element) $\forall v \in V \colon v + \mathbf{0} = v$.

(4) (negatives) $\forall v \in V \colon v + (-v) = \mathbf{0}$.

As usual, $v - w$ is an abbreviation for $v + (-w)$.

In the usual way, it is enough to specify the addition and scalar multiplication; the zero element and negation are uniquely determined (if they exist). When the operations are clear from the context, one refers to the "K-vector space V", or even just the "vector space V" when K is also clear from the context.

We show some simple properties.

10.3. **Lemma.** *Let $(V, +, \mathbf{0}, -, \cdot)$ be a K-vector space. Then:* **LEMMA** rules in a vector space

(1) $\forall v \in V \colon 0 \cdot v = \mathbf{0}$.

(2) $\forall \lambda \in K \colon \lambda \cdot \mathbf{0} = \mathbf{0}$.

(3) $\forall v \in V \colon (-1) \cdot v = -v$.

(4) $\forall \lambda \in K\ \forall v \in V\colon \lambda \cdot v = \mathbf{0} \iff \lambda = 0 \quad or \quad v = \mathbf{0}.$

Proof.

(1) Using one of the distributive laws, we obtain that

$$0 \cdot v + 0 \cdot v = (0+0) \cdot v = 0 \cdot v\,;$$

then adding $-(0 \cdot v)$ on both sides gives

$$0 \cdot v = 0 \cdot v + 0 \cdot v - 0 \cdot v = 0 \cdot v - 0 \cdot v = \mathbf{0}\,.$$

(2) This can be done in a similar way using the other distributive law:

$$\mathbf{0} = \lambda \cdot \mathbf{0} - \lambda \cdot \mathbf{0} = \lambda \cdot (\mathbf{0} + \mathbf{0}) - \lambda \cdot \mathbf{0} = \lambda \cdot \mathbf{0} + \lambda \cdot \mathbf{0} - \lambda \cdot \mathbf{0} = \lambda \cdot \mathbf{0}\,.$$

(3) We have

$$v + (-1) \cdot v = 1 \cdot v + (-1) \cdot v = (1 + (-1)) \cdot v = 0 \cdot v = \mathbf{0}\,,$$

so $(-1) \cdot v$ must be the uniquely determined negative $-v$ of v.

(4) Let $\lambda \in K$ and $v \in V$. The implication "$\Leftarrow$" was already proved in the first two parts of this lemma. To show "$\Rightarrow$", we assume $\lambda \cdot v = \mathbf{0}$. If $\lambda = 0$, then the conclusion holds. Otherwise there exists the inverse $\lambda^{-1} \in K$, and (using part (2) and the associativity of scalar multiplication) we obtain

$$\mathbf{0} = \lambda^{-1} \cdot \mathbf{0} = \lambda^{-1} \cdot (\lambda \cdot v) = (\lambda^{-1}\lambda) \cdot v = 1 \cdot v = v\,. \qquad \square$$

We now give some examples of vector spaces.

10.4. **Examples.** Let K be a field.

EXAMPLES
vector spaces

(1) The smallest possible K-vector space has the zero vector as its only element: $V = \{\mathbf{0}\}$, and we have $\mathbf{0}+\mathbf{0} = \mathbf{0}$ and $\lambda \cdot \mathbf{0} = \mathbf{0}$ for all $\lambda \in K$. This vector space is the *zero (vector) space*. As vector spaces go, it is not terribly interesting, but it plays a similar role in Linear Algebra as the empty set does in set theory.

(2) The next example is the field K itself with its addition and multiplication. The axioms of a vector space are then just some of the field axioms.

(3) The following class of examples is very important; they provide the standard examples of K-vector spaces. As the underlying set we take K^n, the set of n-tuples of elements of K. The operations are defined "component-wise":

$$(x_1, x_2, \ldots, x_n) + (y_1, y_2, \ldots, y_n) = (x_1 + y_1, x_2 + y_2, \ldots, x_n + y_n) \qquad \text{and}$$
$$\lambda \cdot (x_1, x_2, \ldots, x_n) = (\lambda x_1, \lambda x_2, \ldots, \lambda x_n)\,.$$

Then it is easy to verify the axioms. We carry this out for one of the distributive laws to give a flavor of the argument.

$$\begin{aligned}
\lambda \cdot \big((x_1, x_2, \ldots, x_n) + (y_1, y_2, \ldots, y_n)\big) &= \lambda \cdot (x_1 + y_1, x_2 + y_2, \ldots, x_n + y_n) \\
&= \big(\lambda(x_1 + y_1), \lambda(x_2 + y_2), \ldots, \lambda(x_n + y_n)\big) \\
&= (\lambda x_1 + \lambda y_1, \lambda x_2 + \lambda y_2, \ldots, \lambda x_n + \lambda y_n) \\
&= (\lambda x_1, \lambda x_2, \ldots, \lambda x_n) + (\lambda y_1, \lambda y_2, \ldots, \lambda y_n) \\
&= \lambda \cdot (x_1, x_2, \ldots, x_n) + \lambda \cdot (y_1, y_2, \ldots, y_n)\,.
\end{aligned}$$

We see that this follows directly from the distributive law $\lambda(x+y) = \lambda x + \lambda y$ of K. The other axioms can be dealt with in an analogous way.

This example contains the two preceding ones as limiting cases. Taking $n = 0$, the set K^0 has only one element (the zero tuple) and is therefore a zero space. Taking $n = 1$, we have $K^1 = K$, and we obtain K as a vector space over K.

For $K = \mathbb{R}$ and $K = \mathbb{C}$ we obtain the real vector space $\mathbb{R}^n$ and the complex vector space $\mathbb{C}^n$ for every $n \in \mathbb{N}$.

(4) The preceding example can be generalized further. We can consider K^n as the special case $I = \{1, 2, \ldots, n\}$ of the set K^I of families of elements of K indexed by I. (Recall that families $(x_i)_{i\in I}$ with $x_i \in K$ are just a different notation for maps $I \to K$.) We turn K^I into a K-vector space by defining addition and scalar multiplication again component-wise. In "family language", this looks as follows.

$$(x_i)_{i\in I} + (y_i)_{i\in I} = (x_i + y_i)_{i\in I} \qquad \text{and}$$
$$\lambda \cdot (x_i)_{i\in I} = (\lambda x_i)_{i\in I}\,.$$

We can also write this in "map language" (then one also says "point-wise" instead of "component-wise") as

$$f + g\colon I \longrightarrow K\,, \quad i \longmapsto f(i) + g(i)\,, \qquad \text{i.e.,} \quad (f+g)(i) = f(i) + g(i) \qquad \text{and}$$
$$\lambda \cdot f\colon I \longrightarrow K\,, \quad i \longmapsto \lambda f(i)\,, \qquad \text{i.e.,} \quad (\lambda \cdot f)(i) = \lambda f(i).$$

The verification of the axioms works in essentially the same way as for K^n. As an example, we deal with the other distributive law in "map language": Let $\lambda, \mu \in K$ and let $f\colon I \to K$ be a map. Then for $i \in I$ we have

$$\begin{aligned}\big((\lambda+\mu)\cdot f\big)(i) &= (\lambda+\mu)f(i) = \lambda f(i) + \mu f(i)\\ &= (\lambda\cdot f)(i) + (\mu \cdot f)(i) = (\lambda \cdot f + \mu \cdot f)(i)\,,\end{aligned}$$

which shows that $(\lambda+\mu)\cdot f = \lambda \cdot f + \mu \cdot f$. For example, we can consider the real vector space $\mathbb{R}^{\mathbb{R}} = \mathrm{Map}(\mathbb{R}, \mathbb{R})$ of all real functions or the vector space $\mathbb{R}^{\mathbb{N}}$ of all sequences of real numbers.

(5) The field $\mathbb{C}$ of complex numbers is a real vector space. The addition is the usual addition of $\mathbb{C}$, and the scalar multiplication is the multiplication of $\mathbb{C}$, but restricted to $\mathbb{R} \times \mathbb{C}$. (The *restriction* of a map $f\colon X \to Y$ to a subset $T \subset X$ is the map $f|_T\colon T \to Y$, $x \mapsto f(x)$; we shrink (or restrict) the domain, but leave the mapping rule as is.) If we consider $\mathbb{C}$ as $\mathbb{R}^2$, then this is the same real vector space as in (3) above with $K = \mathbb{R}$ and $n = 2$. ♣

DEF
restriction

We obtain further examples of vector spaces as *linear subspaces* of other vector spaces; we will discuss this in the next section.

In the examples 10.1 of linear equations from the beginning of this section, we are looking for solutions in certain real vector spaces. These are $\mathbb{R}^4$ in the first example, $\mathbb{R}^{\mathbb{N}}$ in the second, and a linear subspace of $\mathrm{Map}(\mathbb{R}, \mathbb{R})$ in the third.

11. Linear subspaces and linear hulls

Date:
March 5, 2026

Given a vector space V, we frequently want to work with a suitable subset instead of with the full vector space. This raises the question when such a subset is again a vector space (with the addition and scalar multiplication of V, restricted to the subset). To make this question meaningful, it is necessary that addition and scalar multiplication on the subset are *well-defined*, which here means that sums and multiples of elements of the subset are again in the subset. And in any case, we need the zero vector. This leads to the following definition.

DEF
linear subspace

11.1. **Definition.** Let K be a field, V a K-vector space and $U \subset V$ a subset of V. Then U is a *linear subspace* of V, if the following conditions are satisfied.

(1) $\mathbf{0} \in U$,

(2) $\forall u_1, u_2 \in U : u_1 + u_2 \in U$
("U is closed under addition"),

(3) $\forall \lambda \in K \; \forall u \in U : \lambda \cdot u \in U$
("U is closed under scalar multiplication"). ♢

We show that this definition makes sense.

LEMMA
lin. subspace is vector space

11.2. **Lemma.** *Let K be a field, V a K-vector space and $U \subset V$ a linear subspace. Then for all $u \in U$, the negative $-u$ is also in U.*

We (temporarily) write $+_U$ for the restricted addition $+_U : U \times U \to U$, $(u_1, u_2) \mapsto u_1 + u_2$, $-_U$ for the restricted negation $-_U : U \to U$, $u \mapsto -u$, and $\cdot_U$ for the restricted scalar multiplication $\cdot_U : K \times U \to U$, $(\lambda, u) \mapsto \lambda \cdot u$. Then $(U, +_U, \mathbf{0}, -_U, \cdot_U)$ is a K-vector space.

Proof. The first claim says $\forall u \in U : -u \in U$. This follows from the definition of "linear subspace" since $-u = (-1) \cdot u$; see Lemma 10.3. By the definition again and using the first claim, we can then define $+_U$, $-_U$ and $\cdot_U$ as stated (note that the images always are in U). We then need to verify the vector space axioms for U. These all are universally quantified statements that are supposed to hold for all elements $u_1, u_2, \ldots$ of U. Since V is a vector space, these axioms are even valid for all elements of V, so they certainly also hold for all elements of the subset U. ❑

In the literature you frequently find a definition of "vector space" (and similarly for groups, rings, fields, ...) that starts from the triple $(V, +, \cdot)$ and then requires the *existence* of a zero element and of inverses with respect to addition. In contrast to that, we have included the zero element and the negation map in the "data" of the vector space. This has the advantage that the axioms then all are universally quantified statements that are easier to check, like in the proof above. On the other hand, one has to figure out first what the zero element is and what the negation map looks like. In the proof of the lemma above, we do that by showing that U is also closed under the negation map, which allows us to define the negation map $-_U$. If one uses the other way of stating the axioms, then one has to carry out this step when showing the existence of negatives in U. In the end, one has to do the same, but in a different order.

But note that it is not so obvious how to state the axioms precisely in the "existential" form—the property of being the negative depends on what the zero element is, so the existence of the negative must be included in the scope of the existential quantifier for the zero element:

$$\exists \mathbf{0} \in V : \Big(\forall v \in V : v + \mathbf{0} = v \wedge \forall v \in V \; \exists v' \in V : v + v' = \mathbf{0} \Big)$$

This makes working correctly with this kind of definition a little bit unpleasant.

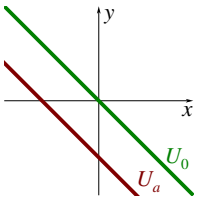
The notation "$+_U$" and so on was just used in the lemma above to clearly distinguish between the addition etc. of V and the addition etc. of U. We usually simply write "$+$" and so on.

EXAMPLES trivial subspaces

11.3. **Examples.** Every vector space has the linear subspaces $U = \{\mathbf{0}\} \subset V$ (which is a zero space) and $U = V$. ♣

EXAMPLE subspaces of $\mathbb{R}^2$

11.4. **Example.** Let $a \in \mathbb{R}$. We consider the real vector space $V = \mathbb{R}^2$ and set $U_a = \{(x, y) \in \mathbb{R}^2 \mid x + y = a\}$. For which a is U_a a linear subspace of $\mathbb{R}^2$?



We have to check the conditions in the definition. The first condition says that the zero vector $\mathbf{0} = (0, 0)$ must be an element of U_a. This means that $0+0 = a$, so this is only possible when $a = 0$. We check the other two conditions in this case.

- U_0 is closed under addition: for elements $u_1 = (x_1, y_1)$ and $u_2 = (x_2, y_2)$ of U_0 we have $u_1 + u_2 = (x_1 + x_2, y_1 + y_2)$ and
$$(x_1 + x_2) + (y_1 + y_2) = (x_1 + y_1) + (x_2 + y_2) = 0 + 0 = 0\,,$$
hence $u_1 + u_2 \in U_0$.
- U_0 is closed under scalar multiplication: for $u = (x, y) \in U_0$ and $\lambda \in \mathbb{R}$ we have $\lambda \cdot u = (\lambda x, \lambda y)$, and
$$\lambda x + \lambda y = \lambda(x + y) = \lambda \cdot 0 = 0\,,$$
hence $\lambda \cdot u \in U_0$. ♣

Further interesting examples are given by "sequence spaces" and "function spaces" that are linear subspaces of the vector space $\mathbb{R}^{\mathbb{N}}$ of sequences of real numbers or the vector space $\mathrm{Map}(\mathbb{R}, \mathbb{R})$ of real functions.

EXAMPLES sequence spaces

11.5. **Examples.** Let $V = \mathbb{R}^{\mathbb{N}}$ be the real vector space whose elements are all sequences of real numbers.

(1) Let $U_b = \{(a_n)_{n\in\mathbb{N}} \in \mathbb{R}^{\mathbb{N}} \mid (a_n)_{n\in\mathbb{N}} \text{ is bounded}\}$. Then U_b is a linear subspace of $\mathbb{R}^{\mathbb{N}}$.

Proof. We check the conditions. The constant zero sequence (with $a_n = 0$ for all $n \in \mathbb{N}$) is bounded, so $\mathbf{0} \in U_b$. Now assume that $(a_n)_{n\in\mathbb{N}}$ and $(b_n)_{n\in\mathbb{N}}$ are two bounded sequences. Then there are $A, B \in \mathbb{R}$ such that $|a_n| \le A$ and $|b_n| \le B$ for all $n \in \mathbb{N}$ (this is the definition of "bounded"). We deduce that $|a_n + b_n| \le |a_n| + |b_n| \le A + B$, so the the sum $(a_n)_{n\in\mathbb{N}} + (b_n)_{n\in\mathbb{N}} = (a_n + b_n)_{n\in\mathbb{N}}$ of the two sequences is also bounded. Now let $\lambda \in \mathbb{R}$; then $|\lambda a_n| \le |\lambda| A$, which shows that $\lambda \cdot (a_n)_{n\in\mathbb{N}} = (\lambda a_n)_{n\in\mathbb{N}}$ is bounded as well. ❑

(2) Let $U_z = \{(a_n)_{n\in\mathbb{N}} \in \mathbb{R}^{\mathbb{N}} \mid (a_n)_{n\in\mathbb{N}} \text{ tends to zero as } n \to \infty\}$. Then U_z is a linear subspace of $\mathbb{R}^{\mathbb{N}}$ (and also of U_b) (exercise).

(3) Let $U_c = \{(a_n)_{n\in\mathbb{N}} \in \mathbb{R}^{\mathbb{N}} \mid (a_n)_{n\in\mathbb{N}} \text{ converges}\}$. Then U_c is a linear subspace of $\mathbb{R}^{\mathbb{N}}$ (and also of U_b).

Proof. The zero sequence converges (to zero), so it is in U_c. We have learned that the sum of two convergent sequences as again convergent and that every (constant) multiple of a convergent sequence is convergent. This shows that the three conditions are satisfied. ❑

We have the inclusions $U_z \subset U_c \subset U_b$ among these linear subspaces (a sequence that tends to zero converges, and a convergent sequence is bounded). ♣

EXAMPLES
function spaces

11.6. **Examples.** Let $V = \mathrm{Map}(\mathbb{R}, \mathbb{R})$ be the real vector space whose elements are all functions $\mathbb{R} \to \mathbb{R}$.

(1) Let $\mathcal{C}(\mathbb{R}) = \{f \in \mathrm{Map}(\mathbb{R}, \mathbb{R}) \mid f \text{ is continuous}\}$. Then $\mathcal{C}(\mathbb{R})$ is a linear subspace of V.

Proof. The zero function $x \mapsto 0$ is continuous. We will soon learn that sums and constant multiples of continuous functions are again continuous. ❑

(2) Let $n \in \mathbb{N}$ and set
$\mathcal{C}^n(\mathbb{R}) = \{f \in \mathrm{Map}(\mathbb{R}, \mathbb{R}) \mid f \text{ is } n\text{-times differentiable and } f^{(n)} \text{ is continuous}\}$;
this is the space of n-times continuously differentiable functions. We will soon see that this is a linear subspace of V.

(3) Let $a > 0$ and $\mathcal{P}(a) = \{f \in \mathrm{Map}(\mathbb{R}, \mathbb{R}) \mid \forall x \in \mathbb{R}\colon f(x + a) = f(x)\}$ the set of all periodic functions with period a (for example, sin and cos are elements of $\mathcal{P}(2\pi)$). Then $\mathcal{P}(a)$ is a linear subspace of V.

Proof. The zero function is periodic (with any period), hence an element of $\mathcal{P}(a)$. Let $f, g \in \mathcal{P}(a)$ and $\lambda \in \mathbb{R}$. We show that $f + g, \lambda f \in \mathcal{P}(a)$. For all $x \in \mathbb{R}$ we have

$$(f+g)(x+a) = f(x+a) + g(x+a) \overset{f,g\in\mathcal{P}(a)}{=} f(x) + g(x) = (f+g)(x) \qquad \text{and}$$

$$(\lambda f)(x+a) = \lambda f(x+a) \overset{f\in\mathcal{P}(a)}{=} \lambda f(x) = (\lambda f)(x)\,.$$

This shows that the three conditions are satisfied. ❑

♣

The notion of linear subspace is central in *Coding Theory*.

EXAMPLE
linear codes

11.7. **Example.** Let F be a finite field (for example, $F = \mathbb{F}_2$) and $n \in \mathbb{N}$. Then a linear subspace of F^n is a *linear code* of length n over F. The *Hamming Code* is a concrete example. It is a linear code of length 7 over $\mathbb{F}_2$, which can be defined as

$$H = \{(x_1, x_2, x_3, x_4, x_1{+}x_2{+}x_4, x_1{+}x_3{+}x_4, x_2{+}x_3{+}x_4) \in \mathbb{F}_2^7 \mid x_1, x_2, x_3, x_4 \in \mathbb{F}_2\}\,.$$

Coding Theory studies the "size" (more precisely, the *dimension*, which we will introduce soon) of a code and also the maximal number of errors it can correct. To be able to correct errors, every pair of distinct code words (elements of the code) must differ at as many positions as possible. Using the linear structure, we can take differences, which allows us to reduce to the case that one of the code words is the zero word. Then the question becomes, how many non-zero entries do the non-zero code words minimally have? This *minimal distance* is 3 for the Hamming code H, which means that it "can correct one error". (If a code word is changed in one position, then it can be reconstructed since all other code words differ from the modified word in at least two places.) ♣

Let us come back to the examples of function spaces. We had seen that the space $\mathcal{C}(\mathbb{R})$ of continuous real functions and the space $\mathcal{P}(a)$ of a-periodic real functions are both linear subspaces of the vector space of all real functions. What

about continuous a-periodic functions? Is $\mathcal{C}(\mathbb{R}) \cap \mathcal{P}(a)$ necessarily also a linear subspace?

11.8. **Lemma.** *Let V be a K-vector space with two linear subspaces U_1 and U_2. Then the intersection $U_1 \cap U_2$ is again a linear subspace of V.*

LEMMA
intersection of subspaces

Proof. As usual, we have to check the three conditions for a linear subspace.

(1) Since U_1 and U_2 are linear subspaces, we have $\mathbf{0} \in U_1$ and $\mathbf{0} \in U_2$, so $\mathbf{0} \in U_1 \cap U_2$.

(2) Let $u, u' \in U_1 \cap U_2$. Then we have $u, u' \in U_1$ and $u, u' \in U_2$. Since U_1 and U_2 are linear subspaces, it follows that $u + u' \in U_1$ and $u + u' \in U_2$, hence $u + u' \in U_1 \cap U_2$.

(3) Let $\lambda \in K$ and $u \in U_1 \cap U_2$. Then we have $u \in U_1$ and $u \in U_2$. Since U_1 and U_2 are linear subspaces, it follows that $\lambda u \in U_1$ and $\lambda u \in U_2$, hence $\lambda u \in U_1 \cap U_2$. ❑

A straight-forward induction proof generalizes this to any (finite) number of linear subspaces:

$$U_1, U_2, \ldots, U_n \subset V \text{ linear subspaces} \\ \implies U_1 \cap U_2 \cap \ldots \cap U_n \subset V \text{ linear subspace.}$$

11.9. **Example.** The space

$$\mathcal{C}(\mathbb{R}) \cap \mathcal{P}(a) = \{f \in \mathrm{Map}(\mathbb{R}, \mathbb{R}) \mid f \text{ is continuous and } a\text{-periodic}\}$$

is a linear subspace of $\mathrm{Map}(\mathbb{R}, \mathbb{R})$. ♣

EXAMPLE
continuous periodic functions

What about *unions* of linear subspaces? Usually this does *not* give a linear subspace. For example, the union of two linear subspaces U_1 and U_2 is a linear subspace only when one of the two is contained in the other (exercise).

We now want to describe the *smallest* (in the sense of inclusion) linear subspace of a given K-vector space V that contains some family $(v_i)_{i \in I}$ of elements of V. To this end, we consider which element of V must necessarily be present in such a linear subspace. For simplicity, we first restrict to finite families (or tuples) $(v_1, \ldots, v_n)$. Assume that $U \subset V$ is a linear subspace that contains $v_1, \ldots, v_n$. Then by the closure properties of linear subspaces, we will have $\lambda_1 v_1, \ldots, \lambda_n v_n \in U$ for all choices of $\lambda_1, \ldots, \lambda_n \in K$, and then also the sum $\lambda_1 v_1 + \ldots + \lambda_n v_n$ must be an element of U. We give these sums a name and generalize to arbitrary families.

11.10. **Definition.** Let K be a field and V a K-vector space.

DEF
linear combination

(1) If $v_1, v_2, \ldots, v_n \in V$ and $\lambda_1, \lambda_2, \ldots, \lambda_n \in K$, then

$$\sum_{i=1}^{n} \lambda_i v_i = \lambda_1 v_1 + \lambda_2 v_2 + \ldots + \lambda_n v_n$$

is a *(K-)linear combination of* $(v_1, v_2, \ldots, v_n)$.

We call λ_j the *coefficient* of v_j in the linear combination.

(2) If $(v_i)_{i \in I}$ is a family of elements of V, then a *(K-)linear combination of the family* $(v_i)_{i \in I}$ is a linear combination of a subfamily $(v_j)_{j \in J}$ with $J \subset I$ finite.

(3) We extend the sum notation and write

$$\sum_{i\in I} \lambda_i v_i$$

for a linear combination of $(v_i)_{i\in I}$, where we set $\lambda_i = 0$ for every $i \in I \setminus J$ (with J as in (2)). We write $K^{(I)}$ for the set of all families $(\lambda_i)_{i\in I} \in K^I$ such that $\lambda_i \neq 0$ only for finitely many $i \in I$. These are exactly the families of coefficients for which the sum notation above makes sense.

(4) A linear combination $\sum_{i\in I} \lambda_i v_i$ is *nontrivial* if there is some $i \in I$ such that $\lambda_i \neq 0$. It is *trivial* otherwise, i.e., if $\lambda_i = 0$ for all $i \in I$. ♢

It is not hard to see that when I is finite, the second definition gives the same linear combinations as the first: when indices are missing in J, we can add them with a zero coefficient.

Warning. In any linear combination there are only **finitely many** vectors (that occur with a nonzero coefficient)! In Linear Algebra there are no truly infinite sums!

Now we show that the set of all such linear combinations is already a linear subspace.

11.11. **Lemma.** *Let V be a K-vector space and let $(v_i)_{i\in I}$ be a family of elements of V. Then the set of all linear combinations of $(v_i)_{i\in I}$ is a linear subspace of V.*

LEMMA
linear combinations form subspace

Proof. For simplicity, we give the proof only for finite families $(v_1, \ldots, v_n)$. Let U be the set of all linear combinations of $(v_1, \ldots, v_n)$. We have to check the three conditions for a linear subspace.

- $\mathbf{0} \in U$ (take $\lambda_i = 0$ for all i).
- U is closed under addition since for $\lambda_1, \ldots, \lambda_n, \mu_1, \ldots, \mu_n \in K$,

$$\begin{aligned}(\lambda_1 v_1 + \lambda_2 v_2 + \ldots + \lambda_n v_n) + (\mu_1 v_1 + \mu_2 v_2 + \ldots + \mu_n v_n)\\ = (\lambda_1 + \mu_1) v_1 + (\lambda_2 + \mu_2) v_2 + \ldots + (\lambda_n + \mu_n) v_n\end{aligned}$$

  is again a linear combination of $(v_1, \ldots, v_n)$.
- U is closed under scalar multiplication since for $\lambda, \lambda_1, \ldots, \lambda_n \in K$,

$$\lambda(\lambda_1 v_1 + \lambda_2 v_2 + \ldots + \lambda_n v_n) = (\lambda\lambda_1) v_1 + (\lambda\lambda_2) v_2 + \ldots + (\lambda\lambda_n) v_n$$

  is again a linear combination of $(v_1, \ldots, v_n)$.

For general families, one obtains $\mathbf{0}$ from any finite linear combination with all zero coefficients (for example, the empty one). Closure under scalar multiplication follows in the same way as above. To show that U is closed under addition, we enlarge the two finite index subsets to a common finite superset in I (e.g., their union) and set the additional coefficients to zero. This does not change the value of the linear combinations. Then we can add them in the same way as above. ❑

It is a good exercise to figure out which vector space axioms are used in which steps of this proof.

11.12. **Definition.** Let V be a K-vector space and let $(v_i)_{i\in I}$ be a family of elements of V. The *linear hull* of the family $(v_i)_{i\in I}$ or the linear subspace *spanned* by the family $(v_i)_{i\in I}$ is the linear subspace of V consisting of all linear combinations of $(v_i)_{i\in I}$. **DEF** linear hull

We denote it $\langle (v_i)_{i\in I}\rangle$ or $\langle v_i \mid i \in I\rangle$. Instead of $\langle (v_i)_{i\in\{1,\dots,n\}}\rangle$, we also write $\langle v_1, v_2, \dots, v_n\rangle$. If we want to indicate the field K of scalars, we write $\langle \dots\rangle_K$. ♢

Note that the field of scalars does make a difference. Considering $\mathbb{C}$ as a real vector space, we have for example $\langle 1\rangle = \langle 1\rangle_{\mathbb{R}} = \mathbb{R} \subset \mathbb{C}$. If we instead consider $\mathbb{C}$ as a complex vector space, then we have $\langle 1\rangle = \langle 1\rangle_{\mathbb{C}} = \mathbb{C}$.

The linear hull of a family is indeed the smallest linear subspace containing its terms: any such linear subspace must contain all linear combinations of the family and thus must contain the linear hull.

11.13. **Example.** We can take the empty family in Definition 11.12. What is the linear subspace spanned by the empty family? **EXAMPLE** $\langle\rangle = \{\mathbf{0}\}$

The only linear combination of the empty family () is the empty sum (of element of V), which by definition is the zero vector, so $\langle\rangle = \{\mathbf{0}\}$. ♣

11.14. **Definition.** Let V be a K-vector space and $(v_i)_{i\in I}$ a family of elements of V. Then $(v_i)_{i\in I}$ is a *(K-)spanning family* or *(K-)generating family* of V, if $V = \langle (v_i)_{i\in I}\rangle$. ♢ **DEF** spanning family

For example, the empty family is a spanning family of the zero space.

11.15. **Example.** Let K be a field and $n \in \mathbb{N}$. The standard vector space K^n has elements **EXAMPLE** spanning family of K^n

$$\mathbf{e}_1 = (1, 0, 0, \dots, 0), \quad \mathbf{e}_2 = (0, 1, 0, \dots, 0), \quad \dots, \quad \mathbf{e}_n = (0, 0, 0, \dots, 0, 1)\,.$$

All the components of $\mathbf{e}_j$ are zero except the jth, which is 1. Then $(\mathbf{e}_1, \mathbf{e}_2, \dots, \mathbf{e}_n)$ is a spanning family of K^n since every element of K^n is a linear combination of these elements:

$$(x_1, x_2, \dots, x_n) = x_1\mathbf{e}_1 + x_2\mathbf{e}_2 + \dots + x_n\mathbf{e}_n\,. \qquad ♣$$

11.16. **Example.** In general, a vector space has many spanning families. For example, **EXAMPLE** many spanning families

$$(\mathbf{e}_1, \mathbf{e}_2),\ \big((1,1),(1,-1)\big),\ \big((1,2),(2,3),(3,4)\big),\ \big((a,b)\big)_{a,b\in\mathbb{Z}} \text{ and } \big((x,y)\big)_{x,y\in\mathbb{R}}$$

all are spanning families of the real vector space $V = \mathbb{R}^2$. ♣

EXAMPLE vector space of polynomial functions

11.17. **Example.** In the real vector space $V = \mathrm{Map}(\mathbb{R}, \mathbb{R})$ we consider the *power functions*

$$f_0 \colon x \longmapsto 1\,, \quad f_1 \colon x \longmapsto x\,, \quad f_2 \colon x \longmapsto x^2\,, \quad \ldots\,, \quad f_n \colon x \longmapsto x^n\,, \quad \ldots\,.$$

What does the linear subspace P of V spanned by $(f_0, f_1, f_2, \ldots) = (f_n)_{n\in\mathbb{N}}$ look like?

The elements of P are precisely the linear combinations of finitely many power functions. By adding power functions with coefficient 0 (this does not change the value of the linear combination) and combining like terms, we see that all the linear combinations can be written in the form

$$f = a_0 f_0 + a_1 f_1 + \ldots + a_n f_n$$

with $n \in \mathbb{N}$ and $a_0, a_1, \ldots, a_n \in \mathbb{R}$. Then we have (for $x \in \mathbb{R}$)

$$f(x) = a_0 f_0(x) + a_1 f_1(x) + \ldots + a_n f_n(x) = a_0 + a_1 x + a_2 x^2 + \ldots + a_n x^n\,.$$

So the elements of P are precisely the *polynomial functions*. ♣

We note some further simple properties of the linear hull.

LEMMA properties of linear hulls

11.18. **Lemma.** *Let V be a K-vector space.*

(1) *If $I \subset J$ are sets and $(v_j)_{j\in J}$ is a family of elements of V, then we have $\langle v_i \mid i \in I\rangle \subset \langle v_j \mid j \in J\rangle$.*

(2) *Let $(s_i)_{i\in I}$ be a spanning family of V. A family $(v_j)_{j\in J}$ of elements of V is a spanning family of V if and only if $\forall i \in I \colon s_i \in \langle v_j \mid j \in J\rangle$.*

Proof. Exercise. ❑

12. Linear independence, basis, and dimension

Date: March 5, 2026

We have seen that a K-vector space V can have very many spanning families; one of them, for example, is the family $(v)_{v\in V}$ of all elements of V. Using this as a spanning family seems a bit wasteful, so one can ask the question whether there are minimal spanning families in a suitable sense and how they can be characterized. To approach this, consider some spanning family $(v_i)_{i\in I}$ of V that is not minimal in the sense that there is an element $i_0 \in I$ such that $(v_i)_{i\in I\setminus\{i_0\}}$ is still a spanning family of V. Then we can write v_{i_0} as a linear combination of elements of this subfamily,

$$v_{i_0} = \lambda_1 v_{i_1} + \lambda_2 v_{i_2} + \ldots + \lambda_n v_{i_n}$$

with $i_1, i_2, \ldots, i_n \in I \setminus \{i_0\}$ pairwise distinct and $\lambda_1, \lambda_2, \ldots, \lambda_n \in K$. Setting $\lambda_0 = -1$, we can write this more symmetrically as

$$\lambda_0 v_{i_0} + \lambda_1 v_{i_1} + \lambda_2 v_{i_2} + \ldots + \lambda_n v_{i_n} = \mathbf{0}\,.$$

So there is a *nontrivial* linear combination (note that $\lambda_0 = -1 \neq 0$) of the family that gives the zero vector.

Conversely, if there exists such a nontrivial linear combination of $(v_i)_{i\in I}$ that gives the zero vector, say

$$\lambda_0 v_{i_0} + \lambda_1 v_{i_1} + \lambda_2 v_{i_2} + \ldots + \lambda_n v_{i_n} = \mathbf{0}\,.$$

with $i_0, i_1, i_2, \ldots, i_n \in I$ pairwise distinct, then we have $\lambda_j \neq 0$ for at least one $j \in \{0, 1, 2, \ldots, n\}$. If necessary, we can re-number the i_js and λ_js so that $\lambda_0 \neq 0$. Then the equality above is equivalent to

$$v_{i_0} = (-\lambda_0^{-1}\lambda_1)v_{i_1} + (-\lambda_0^{-1}\lambda_2)v_{i_2} + \ldots + (-\lambda_0^{-1}\lambda_n)v_{i_n}\,.$$

So for some index i_0 we can write v_{i_0} as a linear combination of elements v_i of the subfamily with index set $I \setminus \{i_0\}$. This implies that $(v_i)_{i\in I\setminus\{i_0\}}$ is still a spanning family of V by Lemma 11.18—for all $i \in I$, v_i is a linear combination of $(v_i)_{i\in I\setminus\{i_0\}}$. This is clear for $i \neq i_0$ (write $v_i = 1 \cdot v_i$), and v_{i_0} is given by the linear combination above.

We conclude that $(v_i)_{i\in I}$ is a minimal spanning family if and only if the zero vector *cannot* be written as a nontrivial linear combination of this family. This property is very important and deserves its own name.

DEF linearly independent

12.1. **Definition.** Let V be a K-vector space and let I be a set. A family $(v_i)_{i\in I}$ of elements of V is *(K-)linearly independent*, if the only linear combination of $(v_i)_{i\in I}$ that gives the zero vector is the trivial linear combination:

$$\forall (\lambda_i)_{i\in I} \in K^{(I)}\colon \Big(\sum_{i\in I} \lambda_i v_i = \mathbf{0} \implies \forall i \in I\colon \lambda_i = 0\Big)$$

Otherwise, the family $(v_i)_{i\in I}$ is *(K-)linearly dependent*. ♢

In other words, a family of vectors is linearly dependent if and only if the zero vector can be written as a nontrivial linear combination of vectors taken from the family.

Note that a family of vectors is linearly independent if and only if every *finite* subfamily is linearly independent.

From our initial considerations in this section, we see that a spanning family is minimal if and only if it is linearly independent. We also obtained the following result.

12.2. **Lemma.** *Let V be a vector space and $v_1, \dots, v_n \in V$. Then $(v_1, v_2, \dots, v_n)$ is linearly dependent if and only if one of these vectors can be written as a linear combination of the others.*

LEMMA linear dependence

Important: Linear independence is a key concept in Linear Algebra. It is very important that you understand it!

12.3. **Example.** Let us consider an extreme case: is the empty family linearly independent or linearly dependent?

EXAMPLE the empty family is linearly independent

There is only one linear combination of the empty family; it is the empty sum, whose value is $\mathbf{0}$. Is this linear combination trivial or not? Recall that "nontrivial" means that there is a nonzero coefficient. However, the empty linear combination does not have any coefficients, so in particular, there is no nonzero coefficient. So the empty linear combination is trivial. This shows that the empty family is linearly independent.

This fits with the observation that a spanning family is minimal if and only if it is linearly independent; clearly, the empty family must be minimal! ♣

12.4. **Example.** When is a singleton family (v) linearly independent?

EXAMPLE when is (v) linearly independent?

The linear combinations have the form λv with a scalar λ. Lemma 10.3 shows that $\lambda v = \mathbf{0}$ implies $\lambda = 0$ or $v = \mathbf{0}$. This shows that (v) is linearly independent when v is not the zero vector. On the other hand, $1 \cdot \mathbf{0} = \mathbf{0}$ is a nontrivial linear combination giving the zero vector, so $(\mathbf{0})$ is linearly dependent. ♣

12.5. **Example.** According to our earlier considerations, a pair (v_1, v_2) is linearly dependent if and only if one of the two vectors is a multiple of the other, $v_2 = \lambda v_1$ or $v_1 = \lambda v_2$ for some $\lambda \in K$. (If $v_1 = \mathbf{0}$, $v_2 \neq \mathbf{0}$, then v_1 is a multiple of v_2, but not conversely.) ♣

EXAMPLE linear independence of two vectors

12.6. **Example.** Here is a very concrete (and typical) example. Is the triple of vectors (v_1, v_2, v_3) with $v_1 = (1, 1, 1, 1)$, $v_2 = (1, 2, 3, 4)$ and $v_3 = (1, 3, 5, 7)$ in $V = \mathbb{R}^4$ linearly dependent or independent?

EXAMPLE 3 vectors in $\mathbb{R}^4$

We have to check the condition in the definition. So let $\lambda_1, \lambda_2, \lambda_3 \in \mathbb{R}$ be such that

$$\lambda_1 v_1 + \lambda_2 v_2 + \lambda_3 v_3 = \mathbf{0} = (0, 0, 0, 0)\,.$$

Now the question is whether this necessarily implies $\lambda_1 = \lambda_2 = \lambda_3 = 0$. We plug in the concrete values of v_1, v_2, v_3 and obtain

$$(\lambda_1 + \lambda_2 + \lambda_3, \lambda_1 + 2\lambda_2 + 3\lambda_3, \lambda_1 + 3\lambda_2 + 5\lambda_3, \lambda_1 + 4\lambda_2 + 7\lambda_3) = (0, 0, 0, 0)\,,$$

which is equivalent to the four equations

$$\begin{aligned} \lambda_1 + \ \lambda_2 + \ \lambda_3 &= 0 \\ \lambda_1 + 2\lambda_2 + 3\lambda_3 &= 0 \\ \lambda_1 + 3\lambda_2 + 5\lambda_3 &= 0 \\ \lambda_1 + 4\lambda_2 + 7\lambda_3 &= 0 \end{aligned}$$

This system of equations has the nontrivial solution $(\lambda_1, \lambda_2, \lambda_3) = (1, -2, 1)$. This means that the triple is linearly dependent. ♣

12.7. **Example.** The spanning family $(\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_n)$ of K^n is linearly independent:

EXAMPLE $(\mathbf{e}_1, \ldots, \mathbf{e}_n)$ is linearly independent

$$\lambda_1\mathbf{e}_1 + \lambda_2\mathbf{e}_2 + \ldots + \lambda_n\mathbf{e}_n = (\lambda_1, \lambda_2, \ldots, \lambda_n)$$

is the zero vector if and only if all coefficients are zero. ♣

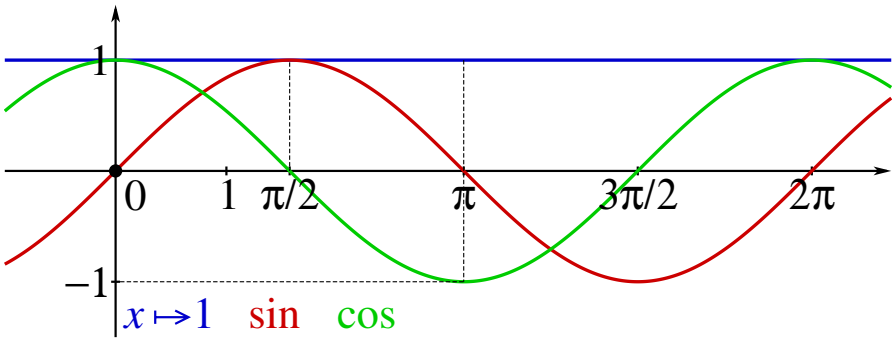


FIGURE 4. Illustration for Example 12.8

12.8. **Example.** The quintuple $(x \mapsto 1, \sin, \cos, \sin^2, \cos^2)$ of functions in the space $\mathcal{C}(\mathbb{R})$ of continuous real functions is linearly dependent since we have

EXAMPLE linear independence in $\mathcal{C}(\mathbb{R})$

$$\forall x \in \mathbb{R}\colon\ \sin^2(x) + \cos^2(x) - 1 = 0\,.$$

This gives a nontrivial linear combination that represents the zero function,

$$(-1)\cdot(x \mapsto 1) + 0\cdot\sin + 0\cdot\cos + 1\cdot\sin^2 + 1\cdot\cos^2 = \mathbf{0}\,.$$

On the other hand, $(x \mapsto 1, \sin, \cos)$ is linearly independent:
assuming that $\lambda_1 + \lambda_2\sin(x) + \lambda_3\cos(x) = 0$ for all $x \in \mathbb{R}$, we obtain by setting $x = 0, \pi, \pi/2$ the three equations

$$\lambda_1 + \lambda_3 = \lambda_1 - \lambda_3 = \lambda_1 + \lambda_2 = 0$$

and from these $\lambda_1 = \lambda_2 = \lambda_3 = 0$. ♣

12.9. **Example.** The family of power functions $(f_n)_{n\in\mathbb{N}}$ with $f_n\colon x \mapsto x^n$ is linearly independent. This means

EXAMPLE power functions

$$\forall n \in \mathbb{N}\ \forall a_0, a_1, \ldots, a_n \in \mathbb{R}\colon$$
$$(\forall x \in \mathbb{R}\colon\ a_0 + a_1x + \ldots + a_nx^n = 0) \Rightarrow a_0 = a_1 = \ldots = a_n = 0\,.$$

We can prove this by induction. In the base case $n = 0$, we have the trivial statement $a_0 = 0 \Rightarrow a_0 = 0$. For the inductive step we assume that the claim holds for n and show it for $n+1$. So let $a_0, a_1, \ldots, a_{n+1} \in \mathbb{R}$ be such that

$$\forall x \in \mathbb{R}\colon\ a_0 + a_1x + a_2x^2 + \ldots + a_{n+1}x^{n+1} = 0\,.$$

Evaluating this at $x = 0$ gives $a_0 = 0$, so we get that

$$\forall x \in \mathbb{R}\colon\ x(a_1 + a_2x + \ldots + a_{n+1}x^n) = 0\,,$$

hence

$$\forall x \in \mathbb{R}\setminus\{0\}\colon\ a_1 + a_2x + \ldots + a_{n+1}x^n = 0\,.$$

Polynomial functions are continuous (we will discuss that soon), so by taking the limit as x tends to zero, we see that the relation remains valid for $x = 0$. So we have that

$$\forall x \in \mathbb{R}\colon\ a_1 + a_2x + \ldots + a_{n+1}x^n = 0\,.$$

We can now apply the inductive hypothesis, which gives us that $a_1 = a_2 = \ldots = a_{n+1} = 0$ as desired.

An alternative approach is to use the (known from school?) fact that a polynomial function of degree n (this means that $a_n \neq 0$ above) can have at most n distinct zeros. In particular, it cannot be the zero function, which has infinitely many zeros (all real numbers). So the only way to obtain the zero function is when all coefficients a_i are zero. ♣

We record a simple but useful observation that relates back to our considerations from the beginning of this section.

12.10. **Lemma.** *Let V be a vector space.* **LEMMA** linear hull of linearly independent vectors

(1) *Let $(v_1, v_2, \ldots, v_n)$ be a linearly independent tuple of vectors in V. Then for all $v \in V$,*

$$v \in \langle v_1, v_2, \ldots, v_n \rangle \iff (v, v_1, v_2, \ldots, v_n) \text{ linearly dependent.}$$

(2) *Let $(v_i)_{i\in I}$ be a linearly independent family of vectors in V. Let $v \in V$ be arbitrary. Assume further that $i_0 \notin I$ and set $I' = I \cup \{i_0\}$ and $v_{i_0} = v$ (so we extend our family by the vector v). Then we have*

$$v \in \langle v_i \mid i \in I \rangle \iff (v_i)_{i\in I'} \text{ linearly dependent.}$$

Proof.

(1) "⇒": $v \in \langle v_1, v_2, \ldots, v_n \rangle$ means that $v = \lambda_1 v_1 + \lambda_2 v_2 + \ldots + \lambda_n v_n$ is a linear combination of the vectors v_j. Then

$$(-1)v + \lambda_1 v_1 + \lambda_2 v_2 + \ldots + \lambda_n v_n = \mathbf{0}$$

and this is a nontrivial linear combination, so $(v, v_1, \ldots, v_n)$ is linearly dependent.

"⇐": Since $(v, v_1, \ldots, v_n)$ is linearly dependent, there is a nontrivial linear combination

$$\lambda v + \lambda_1 v_1 + \lambda_2 v_2 + \ldots + \lambda_n v_n = \mathbf{0}\,.$$

We claim that $\lambda \neq 0$: otherwise, we would obtain a nontrivial linear combination of $(v_1, v_2, \ldots, v_n)$ that gives the zero vector, which contradicts the linear independence of these vectors. Knowing that $\lambda \neq 0$, we can solve the equation for v,

$$v = -\lambda^{-1}\lambda_1 v_1 - \lambda^{-1}\lambda_2 v_2 - \ldots - \lambda^{-1}\lambda_n v_n\,,$$

which shows that $v \in \langle v_1, v_2, \ldots, v_n \rangle$.

(2) This follows from the first part since each linear combination contains only finitely many vectors v_i. ❑

Linearly independent *spanning families* play a fundamental role in Linear Algebra.

12.11. **Definition.** Let V be a K-vector space. A family $(v_i)_{i\in I}$ of elements of V is a *(K-)basis* of V, if it is a linearly independent spanning family of V. ♢ **DEF** basis

12.12. **Examples.**

EXAMPLES
bases

(1) If V is a vector space and $(v_i)_{i\in I}$ is a linearly independent family of elements of V, then $(v_i)_{i\in I}$ is a basis of its linear hull $\langle v_i \mid i \in I\rangle$.

(2) The empty family is a basis of the zero space $\{\mathbf{0}\}$.

(3) The tuple $(\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_n)$ is a K-basis of K^n, the *standard basis* of K^n.

DEF
standard basis of K^n

(4) The sequence $(f_n)_{n\in\mathbb{N}}$ of power functions is a basis of the vector space P of polynomial functions. ♣

At the beginning of this section, we had seen that a spanning family is minimal if and only if it is linearly independent (and therefore a basis). We record this here and add a similar statement about linearly independent families.

12.13. **Lemma.** *Let V be a vector space and $B = (b_i)_{i\in I}$ a family of vectors in V. Then the following statements are equivalent.*

LEMMA
characterizing bases

(1) *B is a basis of V.*

(2) *B is a minimal spanning family of V.*

(3) *B is a maximal linearly independent family in V.*

"Maximal" means in this context that for every $v \in V$ the family ceases to be linearly independent if we extend it by v.

Proof. According to Definition 12.11, a basis is a linearly independent spanning family. We show the equivalences "(1)⇔(2)" and "(1)⇔(3)".

"(1)⇔(2)": We have already shown this equivalence at the beginning of this section.

"(1)⇒(3)": We need to show that every strictly larger family is linearly dependent. This follows directly from Lemma 12.10, (2), as we know that $v \in \langle B\rangle$ for all $v \in V$.

"(3)⇒(1)": We know that B is linearly independent; we have to show that B is a spanning family. So let $v \in V$. Then by assumption, the family B' obtained by extending B by v is linearly dependent. From Lemma 12.10. (2) we obtain $v \in \langle B\rangle$. Since $v \in V$ was arbitrary, we see that $\langle B\rangle = V$, so B is a generating family of V. ❑

We can express the properties of being a spanning family, linearly independent, or a basis also in terms of the number of linear combinations that represent a given vector $v \in V$. We give here a version for finite families (n-tuples).

12.14. **Lemma.** *Let V be a K-vector space and let $v_1, v_2, \ldots, v_n \in V$. We define the associated linear combination map*

LEMMA
properties via number of lin. comb.

$$\phi_{(v_1,v_2,\ldots,v_n)}\colon K^n \longrightarrow V, \quad (\lambda_1, \lambda_2, \ldots, \lambda_n) \longmapsto \lambda_1 v_1 + \lambda_2 v_2 + \ldots + \lambda_n v_n\,.$$

Then:

DEF
linear combination map

(1) *$(v_1, v_2, \ldots, v_n)$ is a **spanning family** of V, if every vector $v \in V$ can be represented in **at least one** way as a linear combination of $(v_1, v_2, \ldots, v_n)$, equivalently, if and only if $\phi_{(v_1,v_2,\ldots,v_n)}$ is **surjective**.*

(2) *$(v_1, v_2, \ldots, v_n)$ is **linearly independent** of V, if every vector $v \in V$ can be represented in **at most one** way as a linear combination of $(v_1, v_2, \ldots, v_n)$, equivalently, if and only if $\phi_{(v_1,v_2,\ldots,v_n)}$ is **injective**.*

(3) $(v_1, v_2, \ldots, v_n)$ *is a* ***basis*** *of V, if every vector $v \in V$ can be represented in* ***exactly one*** *way as a linear combination of $(v_1, v_2, \ldots, v_n)$, equivalently, if and only if $\phi_{(v_1,v_2,\ldots,v_n)}$ is* ***bijective***.

Proof. The equivalence of the last two statements in each part is just the definition of "surjective", "injective", and "bijective".

Part (1) follows directly from Definition 11.14.

We prove part (2).

"$\Rightarrow$": We assume that $(v_1, v_2, \ldots, v_n)$ is linearly independent. Let $v \in V$. If there are two linear combinations representing v,

$$v = \lambda_1 v_1 + \lambda_2 v_2 + \ldots + \lambda_n v_n = \mu_1 v_1 + \mu_2 v_2 + \ldots + \mu_n v_n$$

with $\lambda_1, \lambda_2, \ldots, \lambda_n, \mu_1, \mu_2, \ldots, \mu_n \in K$, then we form the difference,

$$(\lambda_1 - \mu_1)v_1 + (\lambda_2 - \mu_2)v_2 + \ldots + (\lambda_n - \mu_n)v_n = \mathbf{0}\,.$$

Since $(v_1, v_2, \ldots, v_n)$ is linearly independent, this must be the trivial linear combination, and we obtain $\lambda_1 = \mu_1$, $\lambda_2 = \mu_2$, $\ldots$, $\lambda_n = \mu_n$, showing that the original two linear combinations were in fact the same.

"$\Leftarrow$": We assume that every $v \in V$ can be written in at most one way as a linear combination of $(v_1, v_2, \ldots, v_n)$. This must then hold in particular for $v = \mathbf{0}$. The trivial linear combination gives $\mathbf{0}$, so it must be the only one that represents $\mathbf{0}$. This shows that $(v_1, v_2, \ldots, v_n)$ is linearly independent.

Part (3) then follows from (1) and (2). ❑

The preceding lemma extends to arbitrary families $(v_i)_{i\in I}$ of vectors. We can define the linear combination map quite generally as

$$\phi_{(v_i)_{i\in I}} \colon K^{(I)} \longrightarrow V, \quad (\lambda_i)_{i\in I} \longmapsto \sum_{i\in I} \lambda_i v_i\,.$$

Then we have again the following.

(1) $(v_i)_{i\in I}$ spanning family $\iff$ $\phi_{(v_i)_{i\in I}}$ surjective.

(2) $(v_i)_{i\in I}$ linearly independent $\iff$ $\phi_{(v_i)_{i\in I}}$ injective.

(3) $(v_i)_{i\in I}$ basis $\iff$ $\phi_{(v_i)_{i\in I}}$ bijective.

$K^{(I)}$ is precisely the K-linear subspace of K^I that is spanned by the families $\mathbf{e}_i = (\delta_{ij})_{j\in I}$ for $i \in I$. Here $\delta_{ij} = 1$ for $i = j$ and $\delta_{ij} = 0$ for $i \neq j$ (the so-called *Kronecker delta*); the family $\mathbf{e}_i$ has a one as its ith components, and all other components are zero. This generalizes the standard basis of K^n to the vector space $K^{(I)}$.

So a basis $(v_1, v_2, \ldots, v_n)$ of V provides us with a bijective map $K^n \to V$. This allows us to specify the elements of V by the coefficient n-tuple $(\lambda_1, \lambda_2, \ldots, \lambda_n) \in K^n$ (also, addition and scalar multiplication of V correspond to those of K^n). This is certainly very nice! This raises the question whether every vector space has a basis. We will answer this positively for finitely generated vector spaces. (A vector space is *finitely generated*, if it has a finite generating family.) We actually prove a stronger statement that will have many useful applications.

DEF finitely generated vector space

12.15. **Theorem.** *Let V be a vector space and let $v_1, v_2, \dots, v_n$ and $w_1, w_2, \dots, w_m$ be elements of V such that*

THM
Basis
Extension
Theorem

(1) $(v_1, v_2, \dots, v_n)$ is linearly independent and

(2) $(v_1, v_2, \dots, v_n, w_1, w_2, \dots, w_m)$ is a spanning family of V.

Then we can extend $(v_1, v_2, \dots, v_n)$ to a basis of V by adding suitable vectors w_j.

More precisely, there exist $k \in \mathbb{N}$ and indices $j_1, j_2, \dots, j_k \in \{1, 2, \dots, m\}$ such that

$$(v_1, v_2, \dots, v_n, w_{j_1}, w_{j_2}, \dots, w_{j_k})$$

is a basis of V.

The natural numbers n and m in the assumptions can be zero (then there are no vectors v_i or w_j, respectively), and the number k in the conclusion can also be zero (then $(v_1, \dots, v_n)$ is already a basis, which is the case, for example, when $m = 0$).

In the case $n = 0$ the theorem says that every finite spanning family contains a basis. This is plausible since we can always remove elements from the spanning family as long as it is not (yet) minimal. At some point (we can remove an element at most m times) we must arrive at a minimal spanning family, which is then a basis. The proof will use the same idea.

Proof. The proof is by induction on m. The idea behind it is that we start with $(v_1, \dots, v_n, w_1, \dots, w_m)$ and successively remove elements w_j until we obtain a linearly independent family. We fix the n-tuple $(v_1, \dots, v_n)$ and assume according to assumption (1) that it is linearly independent. The statement $A(m)$ that we prove by induction is then that the claim in the theorem holds for this fixed tuple $(v_1, \dots, v_n)$ and *every* m-tuple $(w_1, \dots, w_m)$ satisfying (2).

The base case is $m = 0$. Then $(v_1, \dots, v_n)$ is a spanning family by assumption (2) and linearly independent by assumption (1) and hence a basis. This shows that the claim holds with $k = 0$.

For the inductive step we assume that the claim holds for all m-tuples satisfying (2); we then want to show it for an arbitrary $(m+1)$-tuple $(w_1, \dots, w_{m+1})$ such that $(v_1, \dots, v_n, w_1, \dots, w_{m+1})$ is a spanning family of V. There are two possibilities.

- Either $(v_1, \dots, v_n, w_1, \dots, w_{m+1})$ is linearly independent. Then this is a basis, so the claim holds with $k = m + 1$ and $j_1 = 1, \dots, j_{m+1} = m + 1$.
- Or else $(v_1, \dots, v_n, w_1, \dots, w_{m+1})$ is linearly dependent. Then there is a nontrivial linear combination

  $$\lambda_1 v_1 + \dots + \lambda_n v_n + \mu_1 w_1 + \dots + \mu_m w_m + \mu_{m+1} w_{m+1} = \mathbf{0}\,.$$

  If all μ_j were zero, then we would obtain a nontrivial linear combination of the v_i that represents the zero vector, which contradicts their linear independence. So there must be some $j_0 \in \{1, 2, \dots, m+1\}$ such that $\mu_{j_0} \neq 0$. We can then solve the equation above for w_{j_0}. This shows

  $$w_{j_0} \in \langle v_1, \dots, v_n, w_1, \dots, w_{j_0-1}, w_{j_0+1}, \dots w_{m+1} \rangle\,,$$

  which implies by Lemma 11.18 (2) that $(v_1, \dots, v_n, w_1, \dots, w_{j_0-1}, w_{j_0+1}, \dots, w_{m+1}$ is a spanning family of V. This means that assumption (2) is satisfied for the m-tuple $(w_1, \dots, w_{j_0-1}, w_{j_0+1}, \dots, w_{m+1})$, so that we can apply the inductive hypothesis, which gives the desired conclusion.

To be very precise, we set $w'_i = w_i$ for $i \in \{1, 2, \dots, j_0 - 1\}$ and $w'_i = w_{i+1}$ for $i \in \{j_0, \dots, m\}$; then we apply the inductive hypothesis for $(w'_1, \dots, w'_m)$. This gives $k \leq m$ and $j'_1, \dots, j'_k$ such that $(v_1, \dots, v_n, w'_{j'_1}, \dots, w'_{j'_k})$ is a basis. We then set $j_i = j'_i$ if $j'_i < j_0$ and $j_i = j'_i + 1'$, if $j'_i > j_0$; then $(v_1, \dots, v_n, w_{j_1}, \dots, w_{j_k})$ is a basis that has the form in the claim. ❑

12.16. **Corollary.** *Every vector space that has a finite spanning family has a basis.* **COR** existence of a basis

Proof. This is a consequence of Theorem 12.15 in the case $n = 0$. More precisely, we see that one can find a basis that consists of elements taken form any given finite spanning family. ❑

What about the case when there is no finite spanning family? There is still a Basis Extension Theorem, which we state here for sets (instead of families). A subset $S \subset V$ is a spanning set, linearly independent, or a basis set of V if the "identical family" $(v)_{v \in S}$ is a spanning family, linearly independent, or a basis of V.

Theorem. *Let V be a vector space and let A and S be subsets of V such that A is linearly independent and $A \cup S$ is a spanning set of V. Then there exists a subset $B \subset S$ such that $A \cup B$ is a basis set of V.* **THM** Basis Extension Theorem

This can no longer be proved by induction. One needs another tool, for example *Zorn's Lemma*. It says the following.

Theorem. *Let X be a set and $\mathcal{S} \subset \mathcal{P}(X)$ a set of subsets of X. A **chain** in $\mathcal{S}$ is a subset $\mathcal{C} \subset \mathcal{S}$ such that any two elements of $\mathcal{C}$ are comparable:* **THM** Zorn's Lemma

$$\forall T_1, T_2 \in \mathcal{C} \colon T_1 \subset T_2 \quad \text{or} \quad T_2 \subset T_1 \,.$$

*If every such chain $\mathcal{C}$ has an **upper bound** in $\mathcal{S}$, meaning that given $\mathcal{C}$, there is an element $U \in \mathcal{S}$ such that*

$$\forall T \in \mathcal{C} \colon T \subset U \,,$$

*then $\mathcal{S}$ has **maximal** elements: there is (at least one) $S \in \mathcal{S}$ such that*

$$\forall T \in \mathcal{S} \colon S \subset T \Rightarrow S = T$$

(i.e., there is no set in $\mathcal{S}$ that strictly contains S).

One can show that Zorn's Lemma is equivalent to the Axiom of Choice (assuming the "harmless" axioms of set theory as given); compare the discussion in small print on page 21.

We can then prove the Basis Extension Theorem as follows. The set X in Zorn's Lemma is S, and $\mathcal{S} = \{B \subset S \mid A \cup B \text{ linearly independent}\}$. We need to verify the assumption in Zorn's Lemma. So let $\mathcal{C} \subset \mathcal{S}$ be a chain. We set $U = \bigcup \mathcal{C}$ (this is the union of all subsets of S that are elements of the chain $\mathcal{C}$). Clearly, $T \subset U$ for all $T \in \mathcal{C}$. So it remains to show that $U \in \mathcal{S}$, which means that $A \cup U$ is linearly independent. Assume that this is not the case. Then we would have a nontrivial linear combination of elements of $A \cup U$ that represents the zero vector. This linear combination involves only finitely many elements $v_1, v_2, \dots, v_n$ of U. Since $U = \bigcup \mathcal{C}$, there is a $C_j \in \mathcal{C}$ with $v_j \in C_j$ for every $j = 1, \dots, n$. After renumbering, we can assume that $C_1 \subset C_2 \subset \dots \subset C_n$ (this is where we use that $\mathcal{C}$ is a chain). But then we have $v_1, v_2, \dots, v_n \in C_n$, so that $A \cup C_n$ would be linearly dependent, contradicting $C_n \in \mathcal{S}$. This shows that $A \cup U$ is linearly independent. (It is decisive here that linear combinations are finite!) So U is an upper bound of $\mathcal{C}$ in $\mathcal{S}$, and the assumption in Zorn's Lemma is satisfied. We conclude that $\mathcal{S}$ has a maximal element B. Since $B \in \mathcal{S}$, we know that $A \cup B$ is linearly independent. If $A \cup B$ were not a spanning set, then there would be some $v \in S$ such that $v \notin \langle A \cup B \rangle$.

This would imply that $A \cup (B \cup \{v\})$ is also linearly independent, so we would have $B \cup \{v\} \in \mathcal{S}$. But this cannot be the case since B is maximal (v is not in B because of $v \notin \langle A \cup B \rangle$). So $A \cup B$ is also a spanning set and therefore a basis.

This immediately implies (setting $A = \emptyset$ and $S = V$)

Corollary. *Every vector space has a basis.*

COR existence of basis

So the Axiom of Choice implies for example that $\mathbb{R}$ as a $\mathbb{Q}$-vector space (the structure is given by the addition of $\mathbb{R}$ and the multiplication of $\mathbb{R}$ restricted to $\mathbb{Q} \times \mathbb{R}$, in the same way as we can consider $\mathbb{C}$ as an $\mathbb{R}$-vector space) must have a basis. However, nobody has ever seen such a basis! As already mentioned, the Axiom of Choice and therefore also Zorn's Lemmas are inherently nonconstructive; the proof above does not give any indication what the desired subset B might look like. This is in contrast to the finite case, where the proof does give an algorithm for how to pick the suitable vectors.

Another important consequence is that we can extend any tuple of linearly independent vectors to obtain a basis (if the vector space is finitely generated).

12.17. **Corollary.** *Let V be a vector space with a finite spanning family and let $(v_1, v_2, \dots, v_n) \in V^n$ be linearly independent. Then there are $k \in \mathbb{N}$ and vectors $v_{n+1}, v_{n+2}, \dots, v_{n+k} \in V$ such that $(v_1, v_2, \dots, v_{n+k})$ is a basis of V.*

COR extension to basis

Proof. Let $(w_1, w_2, \dots, w_m)$ be a finite spanning family of V. Then $v_1, \dots, v_n$ and $w_1, \dots, w_m$ satisfy the assumptions of Theorem 12.15. The statement of the theorem then gives the claim upon setting $v_{n+1} = w_{j_1}, \dots, v_{n+k} = w_{j_k}$. ❑

12.18. **Example.** We find a basis of the linear subspace

EXAMPLE basis

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

To do this, we find as many linearly independent vectors as possible and then check whether they form a spanning family. For example, $(1, 0, 1)$ and $(0, 1, 1)$ are linearly independent elements of U; we have

$$\lambda(1, 0, 1) + \mu(0, 1, 1) = \mathbf{0} \iff (\lambda, \mu, \lambda + \mu) = (0, 0, 0) \iff \lambda = \mu = 0 .$$

These two vectors also form a generating family since for $(x, y, z) \in U$ we have $z = x + y$, and so

$$(x, y, z) = (x, y, x + y) = x(1, 0, 1) + y(0, 1, 1) \in \langle (1, 0, 1), (0, 1, 1) \rangle .$$

This shows that $\big((1, 0, 1), (0, 1, 1)\big)$ is a basis of U. ♣

We now deduce the Exchange Lemma.

12.19. **Lemma.** *Let V be a vector space and let $(v_1, v_2, \dots, v_n)$ and $(w_1, \dots, w_m)$ be two bases of V. For each $i \in \{1, 2, \dots, n\}$ there is a $j \in \{1, 2, \dots, m\}$ such that $(v_1, \dots, v_{i-1}, w_j, v_{i+1}, \dots, v_n)$ is also a basis of V.*

LEMMA Exchange Lemma

So we exchange the element v_i of the first basis with an element of the second basis.

Proof. Without loss of generality, we can assume $i = n$ (we can renumber if necessary). We apply the Basis Extension Theorem 12.15 to $v_1, v_2, \dots, v_{n-1}$ and $w_1, w_2, \dots, w_m$. The assumptions are satisfied since subfamilies of linearly independent families are linearly independent and since the w_j are already a spanning family by themselves. So there are $k \in \mathbb{N}$ and indices $j_1, \dots, j_k \in \{1, 2, \dots, m\}$ such that $(v_1, \dots, v_{n-1}, w_{j_1}, w_{j_2}, \dots, w_{j_k})$ is a basis of V. The claim we want to prove is that one can take $k = 1$ (then $j = j_1$). Clearly, we must have $k > 0$; $(v_1, v_2, \dots, v_{n-1})$ is no longer a spanning family (we have removed an element from a minimal spanning family). We show that $(v_1, v_2, \dots, v_{n-1}, w_{j_1})$ is a spanning family; this implies the claim.

We have $w_{j_1} \in V = \langle v_1, v_2, \dots, v_n \rangle$. By Lemma 12.10 we conclude that the $(n+1)$-tuple $(v_1, v_2, \dots, v_{n-1}, v_n, w_{j_1})$ is linearly dependent. Since $(v_1, v_2, \dots, v_{n-1}, w_{j_1})$ is linearly independent as part of the basis $(v_1, \dots, v_{n-1}, w_{j_1}, w_{j_2}, \dots, w_{j_k})$, it follows by Lemma 12.10 again that $v_n \in \langle v_1, v_2, \dots, v_{n-1}, w_{j_1} \rangle$. Clearly $v_1, v_2, \dots, v_{n-1}$ are also elements of this linear hull, which therefore contains a spanning family of V. This implies $\langle v_1, v_2, \dots, v_{n-1}, w_{j_1} \rangle = V$ as claimed. ❑

12.20. **Corollary.** *Let V be a vector space and let $(v_1, \dots, v_n)$ and $(w_1, \dots, w_m)$ be two (finite) bases of V. Then they have the same size, i.e., $n = m$.* **COR** size of bases

Proof. We assume that $n > m$ and deduce a contradiction (the case $n < m$ can be dealt with in the same way). Applying Lemma 12.19 n times (successively with $i = 1, 2, \dots, n$), we obtain indices $j_1, j_2, \dots, j_n \in \{1, 2, \dots, m\}$ such that $(w_{j_1}, w_{j_2}, \dots, w_{j_n})$ is a basis of V. Since m is strictly less than n, there must be repetitions in this tuple. But then $(w_{j_1}, w_{j_2}, \dots, w_{j_n})$ is not linearly independent, so it is not a basis. This gives the desired contradiction. ❑

We introduce notation for the number of elements of a set.

12.21. **Definition.** Let M be a set. We write $\#M$ for the number of elements of M. If M is infinite, we set $\#M = \infty$. ♢ **DEF** $\#M$

In the literature frequently the notation $|M|$ is used. We prefer $\#M$ to reduce the danger of confusion (with, e.g., absolute values).

We assume an intuitive understanding of the meaning of "number of elements" of a set. It is actually not that simple to define this in a formally rigorous way. For example, we can define for a set M and $n \in \mathbb{N}$,

$$\#M = n \iff \exists f \colon M \to \mathbb{N}_{<n} \quad \text{with } f \text{ bijective}.$$

(where $\mathbb{N}_{<n} = \{m \in \mathbb{N} \mid m < n\} = \{0, 1, 2, \dots, n-1\}$ has n elements.) This formalizes the idea that one can number the elements from 0 to $n-1$ (or similarly from 1 to n). One then has to show that such a bijection cannot exist for two sets $\mathbb{N}_{<n}$ with different n and that a set is infinite if and only if there is no such bijection for any n.

The first can be shown by induction. Let us assume there is a bijection $f \colon \mathbb{N}_{<n+m} \to \mathbb{N}_{<n}$ with $m > 0$. This is impossible when $n = 0$ (there is no map from the nonempty set $\mathbb{N}_{<m}$ into the empty set $\mathbb{N}_{<0}$). For $n > 0$ we construct from f a new bijection $f' \colon \mathbb{N}_{<n-1+m} \to \mathbb{N}_{<n-1}$, but this cannot exist by the inductive hypothesis, giving the needed contradiction. To do this, we set $f'(k) = f(k)$, except when $f(k) = n-1$; in this case we set $f'(k) = f(n-1+m)$.

For the second, one needs a definition what an "infinite set" is. (One could *define* this notion so that the statement above holds.) Two possibilities are

$$M \text{ is infinite} \iff \exists f\colon \mathbb{N} \to M \quad \text{with } f \text{ injective}$$

and

$$M \text{ is infinite} \iff \exists f\colon M \to M \quad \text{with } f \text{ injective, but not surjective}.$$

It is an interesting exercise to show the equivalence of these two definitions.

Using the second definition, one shows by induction that the sets $\mathbb{N}_{<n}$ are finite (i.e., not infinite)—an injective map $\mathbb{N}_{<n} \to \mathbb{N}_{<n}$ must in fact be bijective. Then one has to show that a finite set M is bijective with one of the sets $\mathbb{N}_{<n}$. We construct an injective map f from an initial segment of $\mathbb{N}$ into M as follows. If M is empty, then the map is empty; the initial segment is $\mathbb{N}_{<0} = \emptyset$ and $\#M = 0$. Otherwise, there exists some $m_0 \in M$; we set $f(0) = m_0$ and $M_1 = M \setminus \{m_0\}$. Assuming that f is already defined on $\mathbb{N}_{<n}$ and we have constructed the subset M_n of M at the same time, then either M_n is empty, in which case we have found a bijection $f\colon \mathbb{N}_{<n} \to M$ and $\#M = n$. Or else there exists some $m_n \in M_n$, so that we can set $f(n) = m_n$ and $M_{n+1} = M_n \setminus \{m_n\}$. This extends f to $\mathbb{N}_{<n+1}$. If this construction does not terminate, then this gives us an injective map $f\colon \mathbb{N} \to M$, and so M must be infinite (according to the first definition).

We have already seen that infinite sets come in different sizes: sets like $\mathbb{N}$, $\mathbb{Z}$, or $\mathbb{Q}$ are countably infinite, whereas $\mathbb{R}$ (and also $\mathbb{C}$) are uncountable. In this sense, one can show that for each set M, its power set $\mathcal{P}(M)$ is strictly larger (in fact, $\mathbb{R}$ has the same size as $\mathcal{P}(\mathbb{N})$, but $\mathcal{P}(\mathbb{R})$ is even larger than $\mathbb{R}$).

We are now in a position to define the dimension of a vector space.

12.22. **Definition.** Let V be a vector space. If V has a finite basis $(v_1, \dots, v_n)$, then we say that V has *dimension* n or is *n-dimensional* and write $\dim V = n$. If V does not have a finite basis, then we say that V is *infinite-dimensional* and write $\dim V = \infty$. If V has dimension n for some $n \in \mathbb{N}$, then we correspondingly say that V is *finite-dimensional* and write $\dim V < \infty$.

When we want to emphasize that we consider the dimension of V as a linear space over K, we use the more precise notation $\dim_K V$. ♢

DEF
dimension

For example, we have $\dim_{\mathbb{C}} \mathbb{C} = 1$ ((1) is a $\mathbb{C}$-basis), but $\dim_{\mathbb{R}} \mathbb{C} = 2$ ($(1, \boldsymbol{i})$ is an $\mathbb{R}$-basis).

Corollary 12.20 tells us that this definition makes sense, because all finite bases of V (if they exist) have the same number of elements.

12.23. **Examples.**

EXAMPLES
dimension

(1) The empty family is a basis of the zero space, so $\dim\{\mathbf{0}\} = 0$. Conversely, if V is a vector space with $\dim V = 0$, then it has an empty basis, and so $V = \{\mathbf{0}\}$.

(2) For $n \in \mathbb{N}$ we have $\dim K^n = n$ since K^n has the standard basis $(\mathbf{e}_1, \mathbf{e}_2, \dots, \mathbf{e}_n)$ of size n.

(3) For the vector space P of polynomial functions we have $\dim P = \infty$ since it has an infinite basis and therefore cannot be finite-dimensional (see Corollary 12.25 below). ♣

The following statements show why the dimension of a vector space is important.

12.24. **Theorem.** *Let $m, n \in \mathbb{N}$, let V be an n-dimensional vector space, and let $v_1, v_2, \ldots, v_m \in V$.*

THM
properties
of dimension

(1) *If $(v_1, v_2, \ldots, v_m)$ is linearly independent, then $m \leq n$. If in addition $m = n$, then $(v_1, v_2, \ldots, v_m)$ is a basis of V.*

(2) *If $(v_1, v_2, \ldots, v_m)$ is a spanning family of V, then $m \geq n$. If in addition $m = n$, then $(v_1, v_2, \ldots, v_m)$ is a basis of V.*

Proof.

(1) According to Corollary 12.17, we can extend $(v_1, v_2, \ldots, v_m)$ to obtain a basis of V (be adding suitable vectors). The basis thus obtained has n elements; this implies $m \leq n$. If $m = n$, then no elements had to be added, so we already had a basis to begin with.

(2) According to the Basis Extension Theorem 12.15 (taking $n = 0$ there), there is a basis that is obtained by removing suitable vectors v_j from $(v_1, v_2, \ldots, v_m)$. This basis has n elements; this implies $m \geq n$. If $m = n$, then no vectors are removed, so we already had a basis to begin with. ❑

Because this theorem is so important, we give an alternative formulation.

The first part of the statements in the theorem can be stated as follows.

(1) *In an **n-dimensional** vector space, **more than n** vectors are **always linearly dependent**.*

(2) *The **linear hull** of **m** Vectors has **dimension at most m**,*

$$\dim\langle v_1, v_2, \ldots, v_m\rangle \leq m\,.$$

The first of these is a strong *existential statement*. It says the following. *For any $v_1, v_2, \ldots, v_m \in V$ such that $m > \dim V$, there is a nontrivial linear combination*

$$\lambda_1 v_1 + \lambda_2 v_2 + \ldots + \lambda_m v_m = \mathbf{0}\,.$$

The second part of the two statements in the theorem above can be stated as follows.

(1) *In an n-dimensional vector space, n linearly independent vectors are already a basis.*

(2) *In an n-dimensional vector space, a spanning family of n vectors is already a basis.*

So linearly independent families give lower bounds and spanning families give upper bounds for the dimension (of a finite-dimensional vector space):

- The dimension of V is the **maximal** number of **linearly independent** vectors in V.
- The dimension of V is the **minimal** number of vectors **spanning** V.

This makes the following characterization of infinite-dimensional vector spaces plausible.

12.25. **Corollary.** *Let V be a vector space. The following statements are equivalent.* **COR** dim = ∞

(1) *There exists an (infinite) sequence $(v_n)_{n\in\mathbb{N}}$ of linearly independent vectors in V.*

(2) $\dim V = \infty$.

Proof. "(1) ⇒ (2)": We assume that $(v_n)_{n\in\mathbb{N}}$ is linearly independent. If we had $\dim V = m < \infty$, then by Theorem 12.24, the $m+1$ vectors $v_0, v_1, \ldots, v_m$ would be linearly dependent, contradicting the assumption. So V must be infinite-dimensional.

"(2) ⇒ (1)": We assume that V is infinite-dimensional. This means that V has no finite basis, in particular, no finite linearly independent subset of V can be a spanning set. We recursively construct a linearly independent sequence $(v_n)_{n\in\mathbb{N}}$ in V. Assuming that $(v_0, v_1, \ldots, v_{n-1})$ has already been constructed and is linearly independent (this is trivially true for $n = 0$, where we have the empty tuple), we find another vector v_n such that $(v_0, v_1, \ldots, v_{n-1}, v_n)$ is linearly independent, as follows. Since $(v_0, v_1, \ldots, v_{n-1})$ is not a spanning family, there issome vector $v_n \in V \setminus \langle v_0, v_1, \ldots, v_{n-1}\rangle$. Then $(v_0, \ldots, v_{n-1}, v_n)$ is linearly independent by Lemma 12.10. Since all finite initial segments of the sequence $(v_n)_{n\in\mathbb{N}}$ are linearly independent, the sequence itself must also be linearly independent. ❑

We give an application of the fact that $n+1$ vectors in an n-dimensional vector space must be linearly dependent.

12.26. **Definition.** We say that a polynomial function $f \in P$ has *degree* $< n$ (and write $\deg(f) < n$), if it can be written in the form **DEF** degree of a polynomial function

$$f(x) = a_0 + a_1 x + \ldots + a_{n-1}x^{n-1}$$

(with $a_0, a_1, \ldots, a_{n-1} \in \mathbb{R}$). We say that f has *degree* n ($\deg(f) = n$), if $\deg(f) < n+1$, but not $\deg(f) < n$. Then f has the form

$$f(x) = a_0 + a_1 x + \ldots + a_n x^n$$

with $a_n \neq 0$. ♢

You know from school that a polynomial function of degree n has at most n real zeros. This can also be stated as follows.

12.27. **Lemma.** *If f is a polynomial function such that $\deg(f) < n$ and f has at least n real zeros, then f must be the zero function.* **LEMMA** polynomial = 0

12.28. **Example.** *Let $x_1, \ldots, x_n \in \mathbb{R}$ be pairwise distinct and $y_1, \ldots, y_n \in \mathbb{R}$. Then there exists a polynomial function f with $\deg(f) < n$ such that $f(x_j) = y_j$ for all $j \in \{1, 2, \ldots, n\}$.* **EXAMPLE** interpolation

Proof. We consider the following $n+1$ vectors in $\mathbb{R}^n$.

$$\begin{aligned} v_0 &= (1, 1, 1, \ldots, 1) \\ v_1 &= (x_1, x_2, x_3, \ldots, x_n) \\ v_2 &= (x_1^2, x_2^2, x_3^2, \ldots, x_n^2) \\ &\vdots \qquad\qquad \vdots \\ v_{n-1} &= (x_1^{n-1}, x_2^{n-1}, x_3^{n-1}, \ldots, x_n^{n-1}) \\ v_n &= (y_1, y_2, y_3, \ldots, y_n) \end{aligned}$$

We know that $v_0, v_1, \ldots, v_n$ must be linearly dependent since $\dim \mathbb{R}^n = n < n+1$. This means that there are $\lambda_0, \lambda_1, \ldots, \lambda_n \in \mathbb{R}$, not all zero, such that

$$\lambda_0 + \lambda_1 x_j + \lambda_2 x_j^2 + \ldots + \lambda_{n-1} x_j^{n-1} + \lambda_n y_j = 0$$

for all $j \in \{1, 2, \ldots, n\}$. Now, λ_n must be nonzero. Otherwise, the polynomial function

$$x \longmapsto \lambda_0 + \lambda_1 x + \ldots + \lambda_{n-1} x^{n-1}$$

of degree $< n$ would have at least the n zeros $x_1, x_2, \ldots, x_n$, so it would have to be the zero function by Lemma 12.27. But this would mean that $\lambda_0 = \lambda_1 = \ldots = \lambda_{n-1} = 0$, and so the linear combination above would be trivial, a contradiction. So $\lambda_n \neq 0$. We set

$$a_0 = -\frac{\lambda_0}{\lambda_n}, \quad a_1 = -\frac{\lambda_1}{\lambda_n}, \quad \ldots, \quad a_{n-1} = -\frac{\lambda_{n-1}}{\lambda_n}$$

and

$$f(x) = a_0 + a_1 x + \ldots + a_{n-1} x^{n-1};$$

then we have

$$f(x_j) = a_0 + a_1 x_j + \ldots + a_{n-1} x_j^{n-1} = y_j$$

as desired. ❑

These considerations also show that the vectors $v_j = (x_1^j, \ldots, x_n^j)$ for $j \in \{0, 1, \ldots, n-1\}$ are linearly independent. ♣

The dimension measures the "size" of a vector space. This becomes apparent when we consider the dimension of a linear subspace.

12.29. **Theorem.** *Let V be a vector space and $U \subset V$ a linear subspace. Then $\dim U \leq \dim V$. If V is finite-dimensional and $\dim U = \dim V$, then $U = V$.*

THM dimension of subspaces

We declare that $n \leq \infty$ holds for all $n \in \mathbb{N}$, also $\infty \leq \infty$.

Proof. If $\dim V = \infty$, the statement is trivially true. So we can now assume that $\dim V = n \in \mathbb{N}$. If we had $\dim U = \infty$, then Corollary 12.25 says that there are infinitely many linearly independent elements in U and therefore also in V, a contradiction. So U is finite-dimensional with $\dim U = m \in \mathbb{N}$. A basis of U consists of m linearly independent vectors of V, so Theorem 12.24 shows that $\dim U = m \leq n = \dim V$. If we have $m = n$, then the basis of U is already a basis of V, which implies $U = V$. ❑

12.30. **Example.** An infinite-dimensional vector space does have proper linear subspaces that are themselves infinite-dimensional. For example, in the vector space P of polynomial functions we can consider the subspace P_e of even polynomial functions,

$$P_e = \{f \in P \mid \forall x \in \mathbb{R}\colon f(-x) = f(x)\}\,.$$

(You should check that P_e is indeed a linear subspace of P!) Since the function $x \mapsto x$ is an element of P, but is not in P_e, we see that $P_e \neq P$. On the other hand, the even power functions $x \mapsto x^{2n}$ for $n \in \mathbb{N}$ are linearly independent, so $\dim P_e = \infty$. ♣

EXAMPLE
$\dim U = \dim V = \infty$
$U \subsetneq V$

13. Linear maps

Date: March 5, 2026

Let V be a K-vector space and let $v_1, v_2, \ldots, v_n \in V$. We consider the associated linear combination map $\phi = \phi_{(v_1,\ldots,v_n)}$ given by

$$\phi\colon K^n \longrightarrow V, \qquad (x_1, x_2, \ldots, x_n) \longmapsto x_1v_1 + x_2v_2 + \ldots + x_nv_n\,.$$

Then, if $\boldsymbol{x} = (x_1, x_2, \ldots, x_n) \in K^n$, $\boldsymbol{y} = (y_1, y_2, \ldots, y_n) \in K^n$ and $\lambda \in K$, we can compute

$$\begin{aligned}
\phi(\boldsymbol{x} + \boldsymbol{y}) &= \phi\big((x_1, x_2, \ldots, x_n) + (y_1, y_2, \ldots, y_n)\big)\\
&= \phi(x_1 + y_1, x_2 + y_2, \ldots, x_n + y_n)\\
&= (x_1 + y_1)v_1 + (x_2 + y_2)v_2 + \ldots + (x_n + y_n)v_n\\
&= (x_1v_1 + x_2v_2 + \ldots + x_nv_n) + (y_1v_1 + y_2v_2 + \ldots + y_nv_n)\\
&= \phi(x_1, x_2, \ldots, x_n) + \phi(y_1, y_2, \ldots, y_n)\\
&= \phi(\boldsymbol{x}) + \phi(\boldsymbol{y})
\end{aligned}$$

and

$$\begin{aligned}
\phi(\lambda\boldsymbol{x}) &= \phi\big(\lambda(x_1, x_2, \ldots, x_n)\big)\\
&= \phi(\lambda x_1, \lambda x_2, \ldots, \lambda x_n)\\
&= (\lambda x_1)v_1 + (\lambda x_2)v_2 + \ldots + (\lambda x_n)v_n\\
&= \lambda(x_1v_1 + x_2v_2 + \ldots + x_nv_n)\\
&= \lambda\phi(x_1, x_2, \ldots, x_n)\\
&= \lambda\phi(\boldsymbol{x})\,.
\end{aligned}$$

(Note that addition and scalar multiplication occurs in K^n and in V here. Which is which?)

The map ϕ is therefore compatible with addition and scalar multiplication: the image of a sum is the sum of the images, and the image of a scalar multiple is the corresponding multiple of the image. Such maps are called *linear maps*.

13.1. **Definition.** Let K be a field and let V_1 and V_2 be two K-vector spaces. A map $\phi\colon V_1 \to V_2$ is *(K-)linear* or a *homomorphism (of K-vector spaces)*, if it satisfies the following two conditions.

DEF
linear map
homomorphism

(1) $\forall v, w \in V_1\colon \phi(v + w) = \phi(v) + \phi(w)$.

(2) $\forall \lambda \in K\ \forall v \in V_1\colon \phi(\lambda v) = \lambda\phi(v)$.

A linear map is a *monomorphism*, if it is injective, an *epimorphism*, if it is surjective, and an *isomorphism*, if it is bijective. A linear map $\phi\colon V \to V$ is an *endomorphism* of V; ϕ is an *automorphism* of V, if ϕ is bijective in addition. Two vector spaces V_1 and V_2 are *isomorphic*, $V_1 \cong V_2$, if there exists an isomorphism $\phi\colon V_1 \to V_2$. ♢

mono-, epi-, iso-, endo-, automorphism
isomorphic

13.2. **Examples.**

EXAMPLES
linear maps

(1) For any two K-vector spaces V_1 and V_2 the *zero map* $V_1 \to V_2$, $v \mapsto \mathbf{0}$ is a linear map.

(2) For any K-vector space V the identity map $\mathrm{id}_V\colon V \to V$ is an automorphism of V.

(3) If V is a vector space and $U \subset V$ is a linear subspace, then the inclusion map $U \to V$ is linear. ♣

13.3. **Example.** Let V, $v_1, v_2, \dots, v_n$, and ϕ as above in the discussion opening this section. Then ϕ is a homomorphism. By Lemma 12.14, we also have the following. **EXAMPLE** linear comb. map is linear

- $(v_1, v_2, \dots, v_n)$ is linearly independent $\iff$ ϕ is a monomorphism.
- $(v_1, v_2, \dots, v_n)$ is a spanning family of V $\iff$ ϕ is an epimorphism.
- $(v_1, v_2, \dots, v_n)$ is a basis of V $\iff$ ϕ is an isomorphism.

The last point also implies that

$$\dim V < \infty \implies V \cong K^{\dim V}. \qquad \clubsuit$$

We now check that a linear map is compatible with all the structure of a linear space and that linear maps are well-behaved with respect to composition and inversion of maps.

13.4. **Lemma.** *Let V_1, V_2 and V_3 be K-vector spaces.* **LEMMA** properties of linear maps

(1) *Let $\phi\colon V_1 \to V_2$ be a linear map. Then*

$$\phi(\mathbf{0}) = \mathbf{0} \qquad \text{and} \qquad \forall v \in V_1\colon \phi(-v) = -\phi(v)\,.$$

(2) *Let $\phi_1\colon V_1 \to V_2$ and $\phi_2\colon V_2 \to V_3$ be linear maps. Then $\phi_2 \circ \phi_1\colon V_1 \to V_3$ is also linear.*

(3) *Let $\phi\colon V_1 \to V_2$ be an isomorphism. Then the inverse map $\phi^{-1}\colon V_2 \to V_1$ is again an isomorphism.*

We see from part (3) that in the definition of "isomorphic" it does not matter whether we require an isomorphism $V_1 \to V_2$ or an isomorphism $V_2 \to V_1$.

Proof.

(1) We have $\phi(\mathbf{0}) + \phi(\mathbf{0}) = \phi(\mathbf{0} + \mathbf{0}) = \phi(\mathbf{0})$. Adding $-\phi(\mathbf{0})$ on both sides gives $\phi(\mathbf{0}) = \mathbf{0}$.
Now let $v \in V_1$; then $\phi(-v) = \phi((-1)v) = (-1)\phi(v) = -\phi(v)$.

(2) We need to verify the two properties in Definition 13.1 for $\phi_2 \circ \phi_1$. So let $v, w \in V_1$ and $\lambda \in K$. Then we have

$$\begin{aligned}(\phi_2 \circ \phi_1)(v + w) &= \phi_2\big(\phi_1(v + w)\big) = \phi_2\big(\phi_1(v) + \phi_1(w)\big)\\ &= \phi_2\big(\phi_1(v)\big) + \phi_2\big(\phi_1(w)\big) = (\phi_2 \circ \phi_1)(v) + (\phi_2 \circ \phi_1)(w)\end{aligned}$$

and

$$(\phi_2 \circ \phi_1)(\lambda v) = \phi_2\big(\phi_1(\lambda v)\big) = \phi_2\big(\lambda\phi_1(v)\big) = \lambda\phi_2\big(\phi_1(v)\big) = \lambda(\phi_2 \circ \phi_1)(v)\,.$$

(3) We verify the conditions from Definition 13.1 for ϕ^{-1}. So let $v, w \in V_2$ and $\lambda \in K$. We set $v' = \phi^{-1}(v)$ and $w' = \phi^{-1}(w)$, so that $v = \phi(v')$ and $w = \phi(w')$. Then we have

$$\phi^{-1}(v + w) = \phi^{-1}\big(\phi(v') + \phi(w')\big) = \phi^{-1}\big(\phi(v' + w')\big) = v' + w' = \phi^{-1}(v) + \phi^{-1}(w)$$

and

$$\phi^{-1}(\lambda v) = \phi^{-1}\big(\lambda\phi(v')\big) = \phi^{-1}\big(\phi(\lambda v')\big) = \lambda v' = \lambda\phi^{-1}(v)\,. \qquad \square$$

Before we study further properties, we introduce some more notation.

13.5. **Definition.** Let $f\colon X \to Y$ be a map between arbitrary sets X and Y. For a subset T of X we write

$$f(T) = \{f(x) \mid x \in T\} \subset Y$$

for the set of images under f of the elements of T; we call $f(T)$ the *image of T under f*. When $T = X$ we also write $\operatorname{im}(f)$ for $f(X)$; $\operatorname{im}(f)$ is the *image* or *range* of f. For a subset U of Y we write

$$f^{-1}(U) = \{x \in X \mid f(x) \in U\} \subset X$$

for the set of all preimages of the elements of U; we call $f^{-1}(U)$ the *preimage of U under f*. ♢

DEF
images and preimages of subsets

$T \subset T' \subset X$ implies $f(T) \subset f(T')$, and $U \subset U' \subset Y$ implies $f^{-1}(U) \subset f^{-1}(U')$.

In the literature you can frequently find the notation $f^{-1}(y) = \{x \in X \mid f(x) = y\}$ for the set of all preimages of the *element* $y \in Y$. This can lead to confusion since when f is bijective, then $f^{-1}(y)$ also means the (unique) preimage of y and not the preimage *set* of y. We will carefully distinguish the "data types" (elements vs. subsets) and always write $f^{-1}(\{y\})$ for this set.

When f is bijective, then $f^{-1}(U)$ has two possible meanings: it can be the set preimage of U under f or the set image of U under the inverse map f^{-1}. Fortunately, both versions give the same set.

A **warning:** The notation introduced here may tempt you to think that necessarily $f^{-1}(f(T)) = T$ and $f(f^{-1}(U)) = U$. But this is *wrong* in general! What is always true is that we have $f^{-1}(f(T)) \supset T$ and $f(f^{-1}(U)) \subset U$; in both cases, the inclusions can be strict, however. Find examples!

As the zero vector plays a special role in a vector space, the set of its preimages under a linear map is an important quantity.

13.6. **Definition.** Let $\phi\colon V_1 \to V_2$ be a linear map. The *kernel* of ϕ is the set of all preimages of $\mathbf{0} \in V_2$ under ϕ,

$$\ker(\phi) = \phi^{-1}(\{\mathbf{0}\}) = \{v \in V_1 \mid \phi(v) = \mathbf{0}\} \subset V_1\,. \qquad ♢$$

DEF
kernel

Lemma 13.4 shows that $\mathbf{0} \in \ker(\phi)$.

13.7. **Example.** Let $V \subset \mathbb{R}^{\mathbb{N}}$ be the vector space of convergent sequences. Then

$$\lim\colon V \longrightarrow \mathbb{R}, \qquad (a_n)_{n\in\mathbb{N}} \longmapsto \lim_{n\to\infty} a_n$$

is an $\mathbb{R}$-linear map. This follows from the properties of limits.

The kernel $\ker(\lim)$ is precisely the set of sequences converging to zero. ♣

EXAMPLE
limit is linear

The kernel has the following important property.

13.8. **Lemma.** *Let $\phi\colon V_1 \to V_2$ be a linear map. Then we have that*

$$\phi \text{ is injective} \iff \ker(\phi) = \{\mathbf{0}\}\,.$$

LEMMA
injective $\iff$ ker $= \{\mathbf{0}\}$

Proof. "⇒": Let ϕ be injective and $v \in \ker(\phi)$. Then we have $\phi(v) = \mathbf{0} = \phi(\mathbf{0})$, implying that $v = \mathbf{0}$.

"⇐": We assume that $\ker(\phi) = \{\mathbf{0}\}$. Let further $v, w \in V_1$ with $\phi(v) = \phi(w)$. Then $\mathbf{0} = \phi(v) - \phi(w) = \phi(v - w)$, so $v - w \in \ker(\phi) = \{\mathbf{0}\}$, which implies $v - w = \mathbf{0}$ and hence $v = w$. ❑

When $\ker(\phi) = \{\mathbf{0}\}$, we say that the kernel is *trivial*.

DEF trivial kernel

As is to be expected, linear maps behave well with respect to linear subspaces.

13.9. **Theorem.** *Let $\phi\colon V_1 \to V_2$ be a K-linear map.*

THM linear maps and subspaces

(1) *If $U_1 \subset V_1$ is a linear subspace, then $\phi(U_1) \subset V_2$ is again a linear subspace. In particular, the image $\operatorname{im}(\phi) = \phi(V_1) \subset V_2$ is a linear subspace of V_2. Furthermore, the restricted map $\phi|_{U_1}\colon U_1 \to V_2$ is also linear.*

(2) *If $U_2 \subset V_2$ is a linear subspace, then $\phi^{-1}(U_2) \subset V_1$ is again a linear subspace. In particular, the kernel $\ker(\phi) = \phi^{-1}(\{\mathbf{0}\}) \subset V_1$ is a linear subspace of V_1.*

(3) *Define*

$$M_1 = \{U_1 \mid U_1 \subset V_1 \text{ linear subspace with } \ker(\phi) \subset U_1\} \qquad \text{and}$$
$$M_2 = \{U_2 \mid U_2 \subset V_2 \text{ linear subspace with } U_2 \subset \operatorname{im}(\phi)\}\,.$$

The maps $M_1 \to M_2$, $U_1 \mapsto \phi(U_1)$, and $M_2 \to M_1$, $U_2 \mapsto \phi^{-1}(U_2)$, are inverse bijections.

The fact that the kernel of a linear map is a linear subspace is frequently useful when we want to show that a certain subset is a linear subspace. In many cases, a linear subspace can be written as a kernel in a natural way.

Proof.

(1) We need to check the subspace conditions for $\phi(U_1)$.
- $\mathbf{0} = \phi(\mathbf{0}) \in \phi(U_1)$ since $\mathbf{0} \in U_1$.
- Let $v, w \in \phi(U_1)$. Then there are $v', w' \in U_1$ such that $\phi(v') = v$, $\phi(w') = w$. Since $v'+w' \in U_1$, we obtain that $v+w = \phi(v')+\phi(w') = \phi(v'+w') \in \phi(U_1)$.
- Let $v \in \phi(U_1)$ and $\lambda \in K$. Then there is $v' \in U_1$ such that $\phi(v') = v$. Since $\lambda v' \in U_1$, we obtain that $\lambda v = \lambda\phi(v') = \phi(\lambda v') \in \phi(U_1)$.

Since V_1 itself is a linear subspace of V_1, we see from this that $\operatorname{im}(\phi)$ is a linear subspace of V_2.
That $\phi|_{U_1}$ is linear follows from the fact that we can write $\phi|_{U_1}$ as the composition of the (linear) inclusion map $U_1 \to V_1$ and ϕ.

(2) We check the subspace conditions for $\phi^{-1}(U_2)$.
- $\mathbf{0} \in \phi^{-1}(U_2)$ since $\phi(\mathbf{0}) = \mathbf{0} \in U_2$.
- Let $v, w \in \phi^{-1}(U_2)$. Then $\phi(v), \phi(w) \in U_2$. This implies $\phi(v+w) = \phi(v)+\phi(w) \in U_2$, and so $v+w \in \phi^{-1}(U_2)$.
- Let $v \in \phi^{-1}(U_2)$ and $\lambda \in K$. Then $\phi(v) \in U_2$, which gives $\phi(\lambda v) = \lambda\phi(v) \in U_2$, and so $\lambda v \in \phi^{-1}(U_2)$.

Since $\{\mathbf{0}\}$ is a linear subspace of V_2, we see from this that $\ker(\phi)$ is a linear subspace of V_1.

(3) We first show that the two maps make sense (are "well-defined"). If $U_1 \subset V_1$, then $\phi(U_1) \subset \phi(V_1) = \operatorname{im}(\phi)$, and if $U_2 \subset V_2$, then $\phi^{-1}(U_2) \supset \phi^{-1}(\{\mathbf{0}\}) = \ker(\phi)$. By parts (1) and (2), linear subspaces are mapped to linear subspaces. So we indeed obtain maps between the two sets.

We now show that the two maps are inverses. This then also implies that they

are bijective. So let $U_1 \subset V_1$ be a linear subspace with $\ker(\phi) \subset U_1$. Then

$$\begin{aligned} v \in \phi^{-1}\big(\phi(U_1)\big) &\iff \phi(v) \in \phi(U_1) \\ &\iff \exists v' \in U_1 \colon \phi(v) = \phi(v') \\ &\iff \exists v' \in U_1 \colon \phi(v - v') = \mathbf{0} \\ &\iff \exists v' \in U_1 \colon v - v' \in \ker(\phi) \\ &\iff v \in U_1 \,. \end{aligned}$$

(Here is a proof of the last equivalence. "$\Leftarrow$": choose $v' = v$. "$\Rightarrow$": let $v'' = v - v' \in \ker(\phi) \subset U_1$; then $v = v' + v'' \in U_1$.) This shows $\phi^{-1}\big(\phi(U_1)\big) = U_1$.

Now let $U_2 \subset \operatorname{im}(\phi)$ be a linear subspace of V_2. Then

$$\begin{aligned} v \in \phi\big(\phi^{-1}(U_2)\big) &\iff \exists v' \in \phi^{-1}(U_2) \colon \phi(v') = v \\ &\iff \exists v' \in V_1 \colon \phi(v') \in U_2 \text{ and } \phi(v') = v \\ &\iff v \in U_2 \quad \text{and} \quad v \in \operatorname{im}(\phi) \\ &\iff v \in U_2 \cap \operatorname{im}(\phi) \\ &\iff v \in U_2 \,. \end{aligned}$$

This shows $\phi\big(\phi^{-1}(U_2)\big) = U_2$. ❑

One can visualize the relation described in the last part of this theorem in a schematic way as in Figure 5.

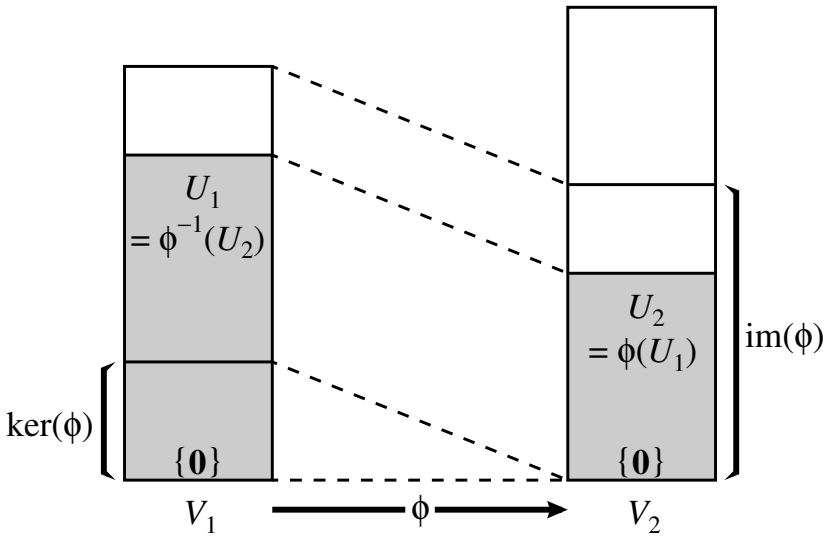


FIGURE 5. Visualization of Theorem 13.9

13.10. **Example.** Let K be a field, X a set and V a linear subspace of $K^X = \operatorname{Map}(X, K)$ (for example, we can take $X = K = \mathbb{R}$ and take the space $\mathcal{C}(\mathbb{R})$ of continuous real functions for V). Let further $x \in X$. Then the *evaluation map*

EXAMPLE evaluation **DEF** evaluation map

$$\operatorname{ev}_x \colon V \longrightarrow K, \quad f \longmapsto f(x)$$

is linear. This is a direct consequence of the definition of addition and scalar multiplication of functions:

$$\begin{aligned} \operatorname{ev}_x(f + g) &= (f + g)(x) = f(x) + g(x) = \operatorname{ev}_x(f) + \operatorname{ev}_x(g) \quad \text{and} \\ \operatorname{ev}_x(\lambda f) &= (\lambda f)(x) = \lambda f(x) = \lambda \operatorname{ev}_x(f) \,. \end{aligned}$$

(Indeed, addition and scalar multiplication on K^X are defined in this way precisely so that the evaluation maps become linear!)

If T is a subset of X, then

$$\{f \in V \mid \forall x \in T\colon f(x) = 0\} = \bigcap_{x \in T} \ker(\mathrm{ev}_x)$$

is a linear subspace of V.

In the special case $X = \{1, 2, \ldots, n\}$ we have $K^X = K^n$; then the maps ev_j (for $j \in \{1, 2, \ldots, n\}$) are the *projections* and are denoted pr_j,

DEF projection

$$\mathrm{pr}_j\colon K^n \longrightarrow K, \quad (a_1, a_2, \ldots, a_n) \longmapsto a_j\,.$$

They are therefore also linear. (One could say that the linear structure on K^n is defined *in such a way that* the pr_j are linear.) ♣

We now show that a linear map is determined by the images of a basis.

13.11. **Theorem.** *Let V be a K-vector space with basis $(b_1, b_2, \ldots, b_n)$ and let W be another K-vector space. Let further $w_1, w_2, \ldots, w_n \in W$. Then there exists exactly one K-linear map $\phi\colon V \to W$ such that $\phi(b_j) = w_j$ for all $j \in \{1, 2, \ldots, n\}$.*

THM bases and linear maps

Proof. We first show uniqueness (i.e., there is at most one such map). Assume that $\phi_1, \phi_2\colon V \to W$ are linear maps such that $\phi_1(b_j) = w_j = \phi_2(b_j)$ for all $j \in \{1, 2, \ldots, n\}$. Let $v \in V$ be an arbitrary vector. Then v is a linear combination of the basis vectors,

$$v = \lambda_1 b_1 + \lambda_2 b_2 + \ldots + \lambda_n b_n\,.$$

Using this, we obtain

$$\begin{aligned}
\phi_1(v) &= \phi_1(\lambda_1 b_1 + \lambda_2 b_2 + \ldots + \lambda_n b_n) \\
&= \lambda_1 \phi_1(b_1) + \lambda_2 \phi_1(b_2) + \ldots + \lambda_n \phi_1(b_n) \\
&= \lambda_1 w_1 + \lambda_2 w_2 + \ldots + \lambda_n w_n \\
&= \lambda_1 \phi_2(b_1) + \lambda_2 \phi_2(b_2) + \ldots + \lambda_n \phi_2(b_n) \\
&= \phi_2(\lambda_1 b_1 + \lambda_2 b_2 + \ldots + \lambda_n b_n) \\
&= \phi_2(v)\,,
\end{aligned}$$

which shows that $\phi_1 = \phi_2$.

This uniqueness proof tells us how to prove existence (i.e., there is at least one such map): If such a linear map $\phi\colon V \to W$ with $\phi(b_j) = w_j$ for all $j \in \{1, 2, \ldots, n\}$ exists, then for $v \in V$ as above we must have

$$\phi(v) = \lambda_1 w_1 + \lambda_2 w_2 + \ldots + \lambda_n w_n\,.$$

We now must check

(1) that ϕ is well-defined (makes sense as a map), which here means that $\phi(v)$ does not depend on the way we write v as a linear combination of the b_j, and

(2) that the map ϕ thus defined is linear.

The first point follows from the fact that v can be written in exactly one way as a linear combination of the basis vectors (Lemma 12.14). The linearity of ϕ can be checked by a straight-forward computation. A more elegant way is to notice that $\phi = \phi_{(w_1, w_2, \ldots, w_n)} \circ \phi^{-1}_{(b_1, b_2, \ldots, b_n)}$ (using the linear combination maps associated to $(b_1, b_2, \ldots, b_n)$ and to $(w_1, w_2, \ldots, w_n)$—note that $\phi_{(b_1, b_2, \ldots, b_n)}$ is an isomorphism);

then the claim follows from the linearity of the linear combination maps (Example 13.3) and from Lemma 13.4. ❑

The analogous statement holds for (not necessarily finite) basis sets.

If V and W are K-vector spaces and $B \subset V$ is a basis, then given any map $f\colon B \to W$, there exists exactly one linear map $\phi\colon V \to W$ such that $\phi(b) = f(b)$ for all $b \in B$ (more concisely, $\phi|_B = f$).

The proof proceeds in essentially the same way, using the general linear combination maps $K^{(B)} \to V$ and $K^{(B)} \to W$.

The statement of Theorem 13.11 can be summarized as follows, which gives an indication how we can use it in practice. We fix a K-vector space V with basis $(b_1, b_2, \dots, b_n)$ and another K-vector space W.

- We can define a linear map $V \to W$ by specifying the images of the basis vectors b_j in any desired way.
- Two linear maps $V \to W$ are already the same, if the images of the basis vectors b_j under both maps agree.

Since a linear map is uniquely determined by the image of a basis, it should be possible to detect whether it is injective or surjective in terms of this image.

THM inj./surj. linear map

13.12. **Theorem.** *Let V and W be K-vector spaces and let $\phi\colon V \to W$ be linear. Let further $(b_1, b_2, \dots, b_n)$ be a basis of V.*

(1) *ϕ is injective if and only if $\big(\phi(b_1), \phi(b_2), \dots, \phi(b_n)\big)$ is linearly independent.*

(2) *ϕ is surjective if and only if $\big(\phi(b_1), \phi(b_2), \dots, \phi(b_n)\big)$ is a spanning family of W.*

(3) *ϕ is an isomorphism if and only if $(\phi(b_1), \phi(b_2), \dots, \phi(b_n))$ is a basis of W.*

Proof. Set $w_1 = \phi(b_1)$, $w_2 = \phi(b_2)$, $\dots$, $w_n = \phi(b_n)$. As in the proof of Theorem 13.11 we then have $\phi = \phi_{(w_1,w_2,\dots,w_n)} \circ \phi^{-1}_{(b_1,b_2,\dots,b_n)}$. Since $\phi_{(b_1,b_2,\dots,b_n)}$ is bijective, it follows that ϕ is injective (respectively, surjective) if and only if $\phi_{(w_1,w_2,\dots,w_n)}$ has that property (note that $\phi \circ \phi_{(b_1,b_2,\dots,b_n)} = \phi_{(w_1,w_2,\dots,w_n)}$). The claims then follow immediately from the statements in Lemma 12.14 (see also Example 13.3). ❑

We can draw two important conclusions.

COR finite-dim. spaces of same dim. are isomorphic

13.13. **Corollary.** *If V and W are two K-vector spaces that have the same finite dimension n, then V and W are isomorphic.*

Proof. Let $(b_1, b_2, \dots, b_n)$ be a basis of V and let $(b'_1, b'_2, \dots, b'_n)$ be a basis of W. Theorem 13.11 gives us a linear map $\phi\colon V \to W$ such that $\phi(b_j) = b'_j$ for all $j \in \{1, 2, \dots, n\}$. By Theorem 13.12, ϕ is an isomorphism. ❑

13.14. **Corollary.** *Let V and W be two K-vector spaces that have the same finite dimension n and let $\phi\colon V \to W$ be a linear map. Then the following statements are equivalent.*

COR lin. maps when same dimension

(1) *ϕ is an isomorphism.*

(2) *ϕ is injective.*

(3) *ϕ is surjective.*

Proof. Clearly, (1) implies (2) and (3). Let $(b_1, \dots, b_n)$ be a basis of V. By Theorem 13.12, ϕ is injective if and only if $\big(\phi(b_1), \dots, \phi(b_n)\big)$ is linearly independent. However, n linearly independent vectors form a basis (because $\dim W = n$; see Theorem 12.24); by Theorem 13.12, ϕ is then an isomorphism. Similarly, ϕ is surjective if and only if $\big(\phi(b_1), \dots, \phi(b_n)\big)$ spans W. A spanning family with n elements is again a basis, hence ϕ is an isomorphism. ❑

To help remember this important fact, think of maps between finite sets, where the following is true.

Let X and Y be two finite sets with $\#X = \#Y = n$ and let $f\colon X \to Y$ be a map. Then the following statements are equivalent.

(1) *f is bijective.*

(2) *f is injective.*

(3) *f is surjective.*

The product notation

$$\prod_{i \in I} a_i$$

used in the next example is defined in an analogous way as the sum notation with the summation sign $\sum$, but multiplying the elements instead of adding them. For $I = \emptyset$, the "empty product" has by definition the value 1.

13.15. **Example.** The vector space $P_{<n}$ of polynomial functions of degree $< n$ is spanned by the n power functions $x \mapsto x^j$ for $j \in \{0, 1, \dots, n-1\}$. These functions are linearly independent, hence $P_{<n}$ has dimension n.

EXAMPLE interpolation

Now let $x_1, x_2, \dots, x_n \in \mathbb{R}$ be pairwise distinct. For $j \in \{1, 2, \dots, n\}$ we define the polynomial function $p_j \in P_{<n}$ via

$$p_j(x) = \prod_{i \in \{1,\dots,n\} \setminus \{j\}} \frac{x - x_i}{x_j - x_i} = \frac{x - x_1}{x_j - x_1} \cdots \frac{x - x_{j-1}}{x_j - x_{j-1}} \cdot \frac{x - x_{j+1}}{x_j - x_{j+1}} \cdots \frac{x - x_n}{x_j - x_n}.$$

For example, when $n = 3$ and $(x_1, x_2, x_3) = (-1, 0, 1)$, this gives the following.

$$p_1(x) = \frac{(x - x_2)(x - x_3)}{(x_1 - x_2)(x_1 - x_3)} = \frac{x(x-1)}{(-1)(-2)} = \tfrac{1}{2}x^2 - \tfrac{1}{2}x$$

$$p_2(x) = \frac{(x - x_1)(x - x_3)}{(x_2 - x_1)(x_2 - x_3)} = \frac{(x+1)(x-1)}{1 \cdot (-1)} = -x^2 + 1$$

$$p_3(x) = \frac{(x - x_1)(x - x_2)}{(x_3 - x_1)(x_3 - x_2)} = \frac{(x+1)x}{2 \cdot 1} = \tfrac{1}{2}x^2 + \tfrac{1}{2}x$$

The relevant property of these functions is that for $i, j \in \{1, 2, \dots, n\}$ we have

$$p_j(x_i) = \begin{cases} 1 & \text{if } i = j; \\ 0 & \text{if } i \neq j. \end{cases}$$

We define the linear map

$$\phi\colon P_{<n} \longrightarrow \mathbb{R}^n, \quad f \longmapsto (f(x_1), f(x_2), \ldots, f(x_n))$$

(linearity follows from the fact that ϕ is constructed from evaluation maps) and another linear map $\psi\colon \mathbb{R}^n \to P_{<n}$ by specifying its images on the standard basis,

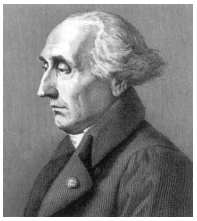
$$\psi(\mathbf{e}_j) = p_j \qquad \text{for all } j \in \{1, 2, \ldots, n\}.$$

(note that ψ is the linear combination map $\phi_{(p_1,\ldots,p_n)}$.) Then we have $\phi \circ \psi = \mathrm{id}_{\mathbb{R}^n}$:

$$\phi\big(\psi(\mathbf{e}_j)\big) = \phi(p_j) = \big(p_j(x_1), \ldots, p_j(x_n)\big) = \mathbf{e}_j\,,$$

so $\phi \circ \psi$ and the identity map agree on a basis, hence they are equal. This implies that ψ is injective and ϕ is surjective. Since $\dim \mathbb{R}^n = n = \dim P_{<n}$, Corollary 13.14 then implies that both maps are (inverse) isomorphisms. This implies the following.

Let $x_1, x_2, \ldots, x_n \in \mathbb{R}$ be pairwise distinct and let $y_1, y_2, \ldots, y_n \in \mathbb{R}$ be arbitrary. Then there exists exactly one *polynomial function $f \in P_{<n}$ such that $f(x_j) = y_j$ for all $j \in \{1, 2, \ldots, n\}$, namely, $f = y_1 p_1 + y_2 p_2 + \ldots + y_n p_n$.*



J.-L. Lagrange
1736–1813

To see this, note that the condition for f means $f \in P_{<n}$ and $\phi(f) = (y_1, y_2, \ldots, y_n)$. But the latter is equivalent to

$$f = \psi(y_1, y_2, \ldots, y_n) = y_1 p_1 + y_2 p_2 + \ldots + y_n p_n\,.$$

This formula for the interpolation polynomial is known as Lagrange's interpolation formula.

(That ϕ is bijective can also be seen as follows. From Example 12.28 we already know that ϕ is surjective. Since $\dim P_{<n} = n = \dim \mathbb{R}^n$, we obtain from Corollary 13.14 that ϕ must be bijective. However, this does not tell us what the inverse map looks like.) ♣

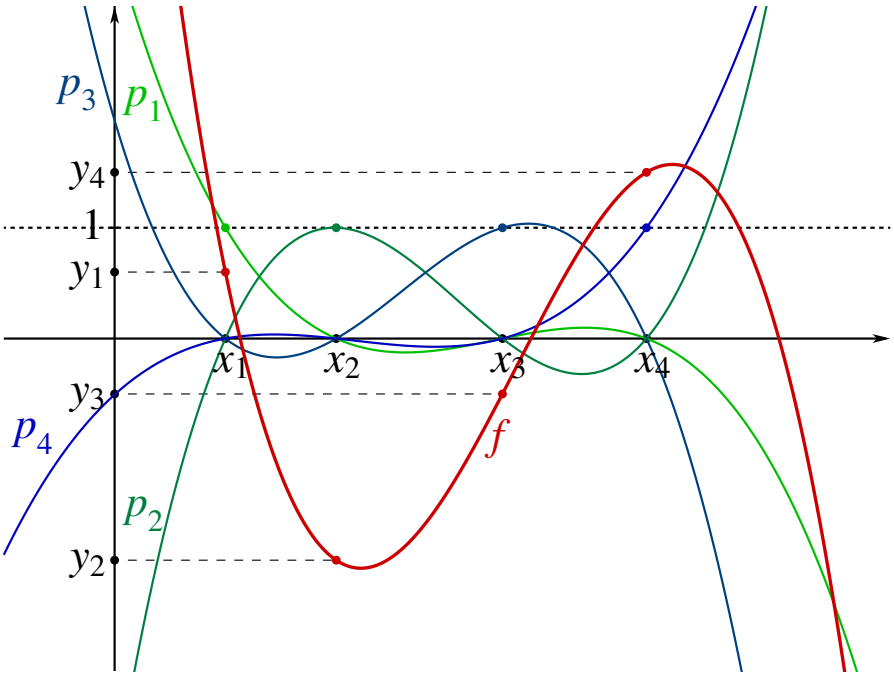


FIGURE 6. Interpolation polynomial (see Example 13.15); we show a case with four data points $(x_1, y_1), \ldots, (x_4.y_4)$.

We stay with polynomial functions and give some more examples of linear maps.

Examples. EX lin. maps on polynomials

(1) We saw in Example 13.10 that for $a \in \mathbb{R}$ the evaluation map

$$\mathrm{ev}_a \colon P \longrightarrow \mathbb{R}, \quad f \longmapsto f(a)$$

is linear.

(2) *Differentiation* of polynomial functions is linear,

$$D \colon P \longrightarrow P, \quad f \longmapsto f' .$$

If $f(x) = a_0 + a_1 x + \ldots + a_n x^n$, then $f'(x) = a_1 + 2a_2 x + \ldots + n a_n x^{n-1}$. So one could define D as the linear map that sends $f_n \colon x \mapsto x^n$ to $n f_{n-1}$ when $n > 0$ and sends f_0 to the zero map.

D is surjective (an epimorphism), and the kernel of D consists exactly of the constant functions. (This example shows that Theorem 13.12 does not necessarily hold for infinite-dimensional vector spaces.)

(3) The computation of the *definite integral* from a to b is linear,

$$I_{a,b} \colon P \longrightarrow \mathbb{R}, \quad f \longmapsto \int_a^b f(x)\,dx .$$

For the power functions f_n we have $I_{a,b}(f_n) = \dfrac{b^{n+1} - a^{n+1}}{n+1}$.

(4) The *indefinite integral* with base-point $a \in \mathbb{R}$ is linear,

$$I_a \colon P \longrightarrow P, \quad f \longmapsto \left(x \mapsto \int_a^x P(t)\,dt\right).$$

This is the linear map such that $I_a(f_n) = \left(x \mapsto \frac{1}{n+1}(x^{n+1} - a^{n+1})\right)$. This map I_a is injective, but not surjective—its image is precisely the kernel of ev_a (the integral functions all vanish at a).

(5) The *translation* by $a \in \mathbb{R}$ is linear,

$$T_a \colon P \longrightarrow P, \quad f \longmapsto (x \mapsto f(x-a)) .$$

T_a is an automorphism of P; its inverse is T_{-a}.

There are many relations between these maps, for example:

$$\mathrm{ev}_b \circ I_a = I_{a,b}, \qquad D \circ I_a = \mathrm{id}_P, \qquad (I_a \circ D)(f) = f - \mathrm{ev}_a(f) f_0,$$
$$T_a \circ D = D \circ T_a, \qquad T_a \circ T_b = T_{a+b}, \qquad I_a \circ T_b = T_b \circ I_{a-b},$$
$$I_{a,b} \circ T_c = I_{a-c,b-c}, \qquad \mathrm{ev}_a \circ T_b = \mathrm{ev}_{a-b} .$$

One easily checks these by applying both sides to power functions (this is sufficient, because the power functions form a basis of P).

We will learn soon that differentiation and integration are linear maps in general. ♣

The kernel and image of a linear map are important quantities. Their dimensions have their own names.

13.16. **Definition.** If $\phi \colon V \to W$ is a linear map, then DEF rank nullity

$$\mathrm{rk}(\phi) = \dim \mathrm{im}(\phi)$$

is the *rank* of ϕ and $\dim \ker(\phi)$ is the *nullity* of ϕ. ♢

There is a simple relation between the rank and the nullity. See Figure 5 for a visualization.

THM
Rank–Nullity Theorem

13.17. **Theorem.** *Let $\phi\colon V \to W$ be a linear map. Then*

$$\dim\ker(\phi) + \mathrm{rk}(\phi) = \dim V\,.$$

Here we set $n + \infty = \infty + n = \infty + \infty = \infty$ for $n \in \mathbb{N}$.

Proof. First assume that $\dim\ker(\phi) = \infty$. Then we must also have $\dim V = \infty$ since $\ker(\phi)$ is a linear subspace of V (Theorem 12.29). So both sides of the equation are ∞.

Now assume that $\mathrm{rk}(\phi) = \infty$. Then we can find infinitely many linearly independent vectors $w_j \in \mathrm{im}(\phi)$ (with $j \in \mathbb{N}$). For each j, pick a preimage $v_j \in V$ of w_j; then the v_j are also linearly independent, as we now show. Assume that $\lambda_0 v_0 + \lambda_1 v_1 + \ldots + \lambda_n v_n = \mathbf{0}$. Applying ϕ on both sides gives

$$\begin{aligned}\lambda_0 w_0 + \lambda_1 w_1 + \ldots + \lambda_n w_n &= \lambda_0\phi(v_0) + \lambda_1\phi(v_1) + \ldots + \lambda_n\phi(v_n)\\ &= \phi(\lambda_0 v_0 + \lambda_1 v_1 + \ldots + \lambda_n v_n) = \mathbf{0}\,.\end{aligned}$$

Since $w_0, w_1, \ldots, w_n$ are linearly independent, all coefficients λ_j must be zero. So $v_0, v_1, \ldots, v_n$ are linearly independent. So we have infinitely many linearly independent vectors in V; then $\dim V = \infty$, and again both sides of the equation are ∞.

So we can now assume that $\dim\ker(\phi) = k$ and $\mathrm{rk}(\phi) = r$ are both finite. We can choose a basis $(b_1, \ldots, b_k)$ of $\ker(\phi)$ and a basis $(b'_1, \ldots, b'_r)$ of $\mathrm{im}(\phi)$. We can pick vectors $b_{k+1}, \ldots, b_{k+r}$ such that $\phi(b_{k+1}) = b'_1$, $\phi(b_{k+2}) = b'_2$, $\ldots$, $\phi(b_{k+r}) = b'_r$. Now we claim that $(b_1, \ldots, b_k, b_{k+1}, \ldots, b_{k+r})$ is a basis of V. This implies $k+r = \dim V$, which is what we want to prove.

- Spanning family:
  Let $v \in V$. Since $(b'_1, \ldots, b'_r)$ is a basis of $\mathrm{im}(\phi)$, there are scalars $\mu_1, \ldots, \mu_r$ such that $\phi(v) = \mu_1 b'_1 + \ldots + \mu_r b'_r$. Then

  $$\begin{aligned}\phi\big(v - (\mu_1 b_{k+1} + \ldots + \mu_r b_{k+r})\big) &= \phi(v) - \big(\mu_1\phi(b_{k+1}) + \ldots + \mu_r\phi(b_{k+r})\big)\\ &= \phi(v) - (\mu_1 b'_1 + \ldots + \mu_r b'_r) = \mathbf{0}\,,\end{aligned}$$

  so $v - (\mu_1 b_{k+1} + \ldots + \mu_r b_{k+r}) \in \ker(\phi)$. There are then scalars $\lambda_1, \ldots, \lambda_k$ such that

  $$v - (\mu_1 b_{k+1} + \ldots + \mu_r b_{k+r}) = \lambda_1 b_1 + \ldots + \lambda_k b_k\,.$$

  This shows that

  $$v = \lambda_1 b_1 + \ldots + \lambda_k b_k + \mu_1 b_{k+1} + \ldots + \mu_r b_{k+r}$$

  is a linear combination of $(b_1, \ldots, b_{k+r})$.
- Linearly independent:
  Let $\lambda_1, \ldots, \lambda_{k+r}$ be scalars such that

  $$\lambda_1 b_1 + \ldots + \lambda_k b_k + \lambda_{k+1} b_{k+1} + \ldots + \lambda_{k+r} b_{k+r} = \mathbf{0}\,.$$

  Then

  $$\begin{aligned}\mathbf{0} &= \phi(\lambda_1 b_1 + \ldots + \lambda_k b_k + \lambda_{k+1} b_{k+1} + \ldots + \lambda_{k+r} b_{k+r})\\ &= \lambda_1\phi(b_1) + \ldots + \lambda_k\phi(b_k) + \lambda_{k+1}\phi(b_{k+1}) + \ldots + \lambda_{k+r}\phi(b_{k+r})\\ &= \lambda_{k+1}\phi(b_{k+1}) + \ldots + \lambda_{k+r}\phi(b_{k+r})\\ &= \lambda_{k+1} b'_1 + \ldots + \lambda_{k+r} b'_r\end{aligned}$$

since $\phi(b_1) = \ldots = \phi(b_k) = \mathbf{0}$. Since $(b'_1, \ldots, b'_r)$ is linearly independent, it follows that $\lambda_{k+1} = \ldots = \lambda_{k+r} = 0$. We plug this back into the original relation and obtain

$$\lambda_1 b_1 + \ldots + \lambda_k b_k = \mathbf{0}\,.$$

Since $(b_1, \ldots, b_k)$ is also linearly independent, we obtain $\lambda_1 = \ldots = \lambda_k = 0$, so the original linear combination is trivial. ❑

EXAMPLE
dimension of a kernel

13.18. **Example.** Let $V = K^n$ with $n \geq 1$ and

$$\phi\colon V \longrightarrow K, \quad (x_1, x_2, \ldots, x_n) \longmapsto x_1 + x_2 + \ldots + x_n\,.$$

Then ϕ is linear (easy computation). So

$$U = \{(x_1, x_2, \ldots, x_n) \in V \mid x_1 + x_2 + \ldots + x_n = 0\} = \ker(\phi) \subset V$$

is a linear subspace of V. We have $\operatorname{im}(\phi) = K$:

$$\text{for } \lambda \in K \text{ we have } \phi\big((\lambda, 0, \ldots, 0)\big) = \lambda.$$

This shows that $\operatorname{rk}(\phi) = \dim_K K = 1$. This implies by Theorem 13.17 that

$$\dim U = \dim \ker(\phi) = \dim V - \operatorname{rk}(\phi) = n - 1\,.$$ ♣

Contrary to this example, it is often easier to determine the kernel of a linear map and its dimension directly than the rank. We then obtain the rank from the Rank-Nullity Theorem 13.17.

We can generalize the construction of the vector space $K^X = \operatorname{Map}(X, K)$ further.

DEF
vector space V^X

13.19. **Definition.** Let K be a field and V a K-vector space. Let further X be a set. Then we can define a linear space structure on $V^X = \operatorname{Map}(X, V)$ by

$$f + g\colon x \mapsto f(x) + g(x) \qquad \text{and} \qquad \lambda f\colon x \mapsto \lambda f(x)\,.$$

The proof is analogous to the proof for K^X.

When $X = \{1, 2, \ldots, n\}$, we have $V^X = V^n$. ♢

Since we can add arbitrary maps $V \to W$ and multiply them by a scalar, we can certainly do this with *linear* maps. Not surprisingly, the maps we obtain are again linear.

THM
vector space of lin. maps

13.20. **Theorem.** *Let V and W be two K-vector spaces. The set of all linear maps $V \to W$ is a K-linear subspace of* $\operatorname{Map}(V, W)$.

Proof. We need to check the subspace conditions.

- The zero map is linear.
- Let $\phi, \psi\colon V \to W$ be linear. We show that $\phi + \psi$ is also linear. So let $v, v' \in V$, $\lambda \in K$. Then

$$\begin{aligned}(\phi + \psi)(v + v') &= \phi(v + v') + \psi(v + v') = \phi(v) + \phi(v') + \psi(v) + \psi(v') \\ &= \phi(v) + \psi(v) + \phi(v') + \psi(v') = (\phi + \psi)(v) + (\phi + \psi)(v')\end{aligned}$$

and

$$\begin{aligned}(\phi + \psi)(\lambda v) &= \phi(\lambda v) + \psi(\lambda v) = \lambda\phi(v) + \lambda\psi(v) \\ &= \lambda(\phi(v) + \psi(v)) = \lambda(\phi + \psi)(v)\,.\end{aligned}$$

- Let $\phi\colon V \to W$ be linear and $\lambda \in K$. We show that $\lambda\phi$ is also linear. So let $v, v' \in V$, $\mu \in K$. Then

$$\begin{aligned}(\lambda\phi)(v+v') &= \lambda\phi(v+v') = \lambda(\phi(v)+\phi(v')) \\ &= \lambda\phi(v) + \lambda\phi(v') = (\lambda\phi)(v) + (\lambda\phi)(v')\end{aligned}$$

and

$$\begin{aligned}(\lambda\phi)(\mu v) &= \lambda\phi(\mu v) = \lambda \cdot \mu\phi(v) \\ &= \mu \cdot \lambda\phi(v) = \mu(\lambda\phi)(v)\,.\end{aligned}$$

❑

13.21. **Definition.** We write $\operatorname{Hom}(V,W)$ (or $\operatorname{Hom}_K(V,W)$) for the vector space of linear maps $V \to W$. When $V = W$, we also write $\operatorname{End}(V) = \operatorname{Hom}(V,V)$ (or $\operatorname{End}_K(V)$) for the vector space of all endomorphisms of V. ♢

DEF
$\operatorname{Hom}(V,W)$
$\operatorname{End}(V)$

13.22. **Theorem.** *Let V and W be two K-vector spaces with $\dim V = n < \infty$. Let further $(b_1, b_2, \ldots, b_n)$ be a basis of V. Then*

$$\Phi\colon \operatorname{Hom}(V,W) \longrightarrow W^n, \quad \phi \longmapsto \big(\phi(b_1), \phi(b_2), \ldots, \phi(b_n)\big)$$

is an isomorphism. In particular, when $\dim W = m < \infty$, we have

$$\dim \operatorname{Hom}(V,W) = \dim W^n = n \dim W = mn = (\dim V)(\dim W)\,.$$

THM
$\operatorname{Hom}(V,W)$
$\cong W^{\dim V}$

Proof. It is clear that Φ is linear (it is composed from evaluation maps; the evaluation maps $\operatorname{Map}(V,W) \to W$, $\phi \mapsto \phi(v)$, are also linear in this more general context; the proof is the same as before). According to Theorem 13.11, for each choice of the images of $b_1, \ldots, b_n$ in W there is exactly one linear map; this means that Φ is bijective. Isomorphic vector spaces have the same dimension. The proof of $\dim W^n = n \dim W$ is an exercise. ❑

13.23. **Corollary.** *Let V and W be two K-vector spaces; let further $(b_1, b_2, \ldots, b_n)$ be a basis of V and $(b'_1, b'_2, \ldots, b'_m)$ a basis of W.*
For $(i,j) \in \{1,2,\ldots,m\} \times \{1,2,\ldots,n\}$ let $\phi_{ij}\colon V \to W$ be the linear map such that $\phi_{ij}(b_k) = \mathbf{0}$ for $k \neq j$ and $\phi_{ij}(b_j) = b'_i$. Then $(\phi_{ij})_{(i,j)\in\{1,2,\ldots,m\}\times\{1,2,\ldots,n\}}$ is a basis of $\operatorname{Hom}(V,W)$.

COR
basis of
$\operatorname{Hom}(V,W)$

Proof. By Theorem 13.11, unique ϕ_{ij} as stated exist. We show that the $\phi_{ij} \in \operatorname{Hom}(V,W)$ are linearly independent. So let λ_{ij} be scalars such that

$$\sum_{i=1}^{m}\sum_{j=1}^{n} \lambda_{ij}\phi_{ij} = \mathbf{0}\,.$$

Let $k \in \{1,2,\ldots,n\}$. Evaluating in b_k, we obtain

$$\mathbf{0} = \Big(\sum_{i=1}^{m}\sum_{j=1}^{n} \lambda_{ij}\phi_{ij}\Big)(b_k) = \sum_{i=1}^{m}\sum_{j=1}^{n} \lambda_{ij}\phi_{ij}(b_k) = \sum_{i=1}^{m} \lambda_{ik} b'_i$$

In the last step we have used that $\phi_{ij}(b_k) = \mathbf{0}$ for $k \neq j$; the inner sum is therefore reduced to $\lambda_{ik}\phi_{ik}(b_k) = \lambda_{ik}b'_i$. Since the b'_i are linearly independent, it follows that $\lambda_{ik} = 0$ for all i. Since k was arbitrary, we see that all $\lambda_{ij} = 0$; this was to be shown. By Theorem 13.22, we have that $\dim \operatorname{Hom}(V,W) = nm$ is the number of linearly independent elements $\phi_{ij} \in \operatorname{Hom}(V,W)$, and by Theorem 12.24, the ϕ_{ij} then are already a basis of $\operatorname{Hom}(V,W)$. ❑

When $V = K^n$, $W = K^m$ and we use the standard bases, then ϕ_{ij} does the following. Take the jth component of $(x_1, x_2, \dots, x_n) \in K^n$ and put it into the ith component of the result; the other components are set to zero.

EXAMPLE basis of $\mathrm{Hom}(\mathbb{R}^3, \mathbb{R}^2)$

13.24. **Example.** As a simple example, we consider $V = \mathbb{R}^3$ and $W = \mathbb{R}^2$ with their standard bases $(\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3)$ and $(\mathbf{e}'_1, \mathbf{e}'_2)$. The basis of $\mathrm{Hom}(\mathbb{R}^3, \mathbb{R}^2)$ in Corollary 13.23 then looks as follows.

$$\begin{aligned}
\phi_{11} &\colon (x, y, z) \longmapsto (x, 0) \\
\phi_{12} &\colon (x, y, z) \longmapsto (y, 0) \\
\phi_{13} &\colon (x, y, z) \longmapsto (z, 0) \\
\phi_{21} &\colon (x, y, z) \longmapsto (0, x) \\
\phi_{22} &\colon (x, y, z) \longmapsto (0, y) \\
\phi_{23} &\colon (x, y, z) \longmapsto (0, z)
\end{aligned}$$

Every linear map $\phi\colon \mathbb{R}^3 \to \mathbb{R}^2$ can be written as a linear combination of these six maps: there are $a, b, c, d, e, f \in \mathbb{R}$ such that

$$\phi = a\phi_{11} + b\phi_{12} + c\phi_{13} + d\phi_{21} + e\phi_{22} + f\phi_{23},$$

so

$$\phi(x, y, z) = (ax + by + cz, dx + ey + fz)\,. \qquad \clubsuit$$

DEF endomorphism ring

The endomorphisms of a vector space V even form a ring, the *endomorphism ring* of V.

THM $\mathrm{End}(V)$ is a ring

13.25. **Theorem.** *Let V be a K-vector space. Then $\mathrm{End}(V)$ is a ring with the addition of the K-vector space $\mathrm{End}(V) = \mathrm{Hom}(V, V)$ and the composition of maps as multiplication; the identity map id_V is the unit element.*

Proof. The vector space axioms that are valid in $\mathrm{End}(V)$ give us the ring axioms for addition. It remains to show that the multiplication is associative and has neutral element id_V, and that the two distributive laws of a ring hold. So let $f, g, h \in \mathrm{End}(V)$. The associative law $(f \circ g) \circ h = f \circ (g \circ h)$ holds for maps in complete generality, and the same is true for $\mathrm{id}_V \circ f = f = f \circ \mathrm{id}_V$. To verify the distributive laws, we apply both sides to an arbitrary $v \in V$ and compute

$$\begin{aligned}
\big((f + g) \circ h\big)(v) &= (f + g)\big(h(v)\big) = f\big(h(v)\big) + g\big(h(v)\big) \\
&= (f \circ h)(v) + (g \circ h)(v) = (f \circ h + g \circ h)(v)\,;
\end{aligned}$$

this shows $(f + g) \circ h = f \circ h + g \circ h$, and

$$\begin{aligned}
(f \circ (g + h))(v) &= f\big((g + h)(v)\big) = f\big(g(v) + h(v)\big) \\
&= f\big(g(v)\big) + f\big(h(v)\big) = (f \circ g)(v) + (f \circ h)(v) \\
&= (f \circ g + f \circ h)(v)\,;
\end{aligned}$$

this shows $f \circ (g + h) = f \circ g + f \circ h$. (We have used that f is linear—where?). ❑

The endomorphism ring is not commutative when $\dim V \geq 2$ (exercise). When $\dim V = 1$, we have $\mathrm{End}(V) = K$ since all endomorphisms are given by multiplication by a scalar. When $\dim V = 0$, $\mathrm{End}(V)$ is the zero ring.

The automorphisms of V form a group, the *automorphism group* $\mathrm{Aut}(V)$ of V.

14. Continuity

Date:
March 5, 2026

14.A Real functions.

14.1. **Definition.** A *real function* is a map f that assigns to each number x from a subset $D \subset \mathbb{R}$ a number $f(x) \in A \subset \mathbb{R}$. The set D is called the *domain*, the set A is called the *codomain* and the set $f(D) := \{f(x) \,|\, x \in D\}$ is called the *range* or *image* of f. Notation: $f : D \to A$. ♢

DEF
Real function

Real functions are represented graphically in a coordinate system by plotting, for each $x \in D$ on the horizontal axis, the value $f(x)$ on the vertical axis. The resulting figure is called the *graph* of the function.

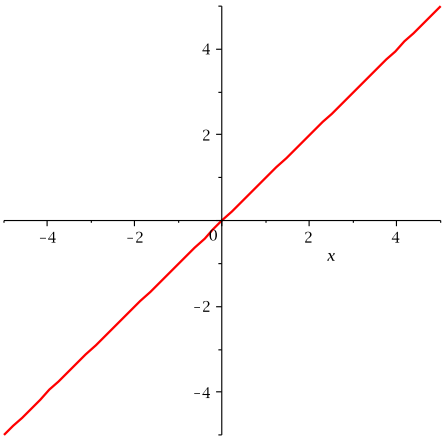
Functions are often defined by a formula depending on x. We then write $f : x \mapsto \ldots$, where the corresponding formula is inserted in place of the dots. Often, one simply writes $f(x) = \ldots$.

Here are some examples.

14.2. **Examples.**

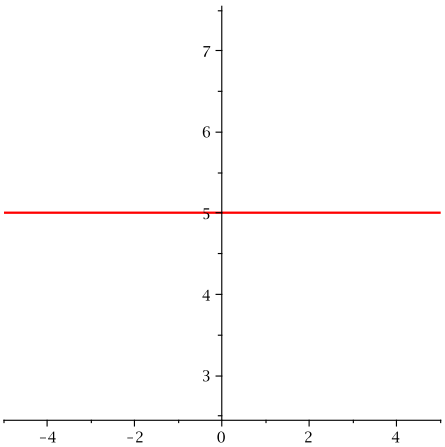
EXAMPLES
Real functions

(a) **(Identity function)** $f : \mathbb{R} \to \mathbb{R}$ with $f : x \mapsto x$



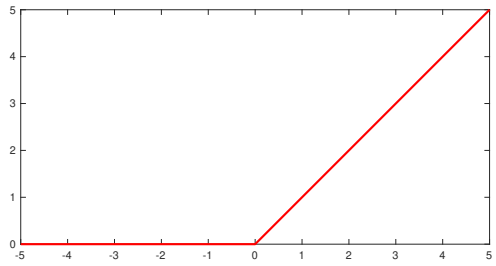
Graph of the identity function $f(x) = x$

(b) **(Constant function)** $f : \mathbb{R} \to \mathbb{R}$ with $f : x \mapsto c$ for a fixed $c \in \mathbb{R}$.



Graph of the constant function $f(x) = c$, here for $c = 5$

(c) **(Rectified Linear Unit, ReLU)** $f : \mathbb{R} \to \mathbb{R}$ with $f : x \mapsto (x)_+$, where $(x)_+ := \max\{0, x\}$

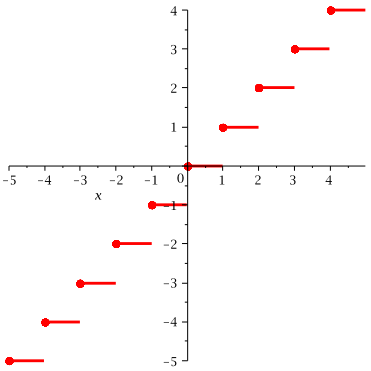


Graph of the ReLU function $f(x) = (x)_+$

(d) **(Floor function)** We define the so-called *floor* $[x]$ of a real number x as the greatest integer $k \leq x$, or formally:

$$[x] := \max\{k \in \mathbb{Z} \,|\, k \leq x\}.$$

The corresponding function is $f : \mathbb{R} \to \mathbb{R}$ with $f : x \mapsto [x]$.



Graph of the floor function $f(x) = [x]$

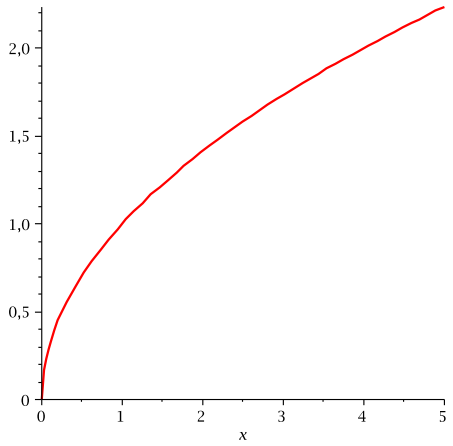
The points in the graph illustrate where the value lies at the position where the graph "jumps".

Note: The domain D is a set that we can choose ourselves. For instance, we can consider the function $f : x \mapsto |x|$ only for x between 0 and 1 by choosing the domain as $D = [0, 1]$. Of course, we could also choose $D = \mathbb{R}$. However, if the expression defining $f(x)$ is not defined for some $x \in \mathbb{R}$, we must exclude these x from D in order to obtain a well-defined function. This happens, for example, in the following case (e). The set of all x for which the expression $f(x)$ is defined is called the *maximal domain*. We cannot choose D larger than the maximal domain (unless we provide a new expression for $f(x)$ for those x in which the original one is not defined). Yet, we can always choose it smaller if we wish.

(e) **(Square root)** All previous examples could be defined on $D = \mathbb{R}$. The square root, however, is only defined in $\mathbb{R}$ for numbers ≥ 0, so we must restrict the domain of the square root function. To do this, we extend Definition 5.7 of intervals by the following unbounded intervals for $a, b \in \mathbb{R}$:

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

where we call $[a, \infty)$ and $(-\infty, b]$ closed,[17] and (a, ∞) and $(-\infty, b)$ open. The maximal domain of the square root function is, in this notation, $D = [0, \infty)$, so we can write $f : [0, \infty) \to \mathbb{R}$ with $f : x \mapsto \sqrt{x}$.



Graph of the square root function $f(x) = \sqrt{x}$

The ranges $f(D) := \{f(x) \,|\, x \in D\}$ of these functions can easily be determined from the given formulas. We have:

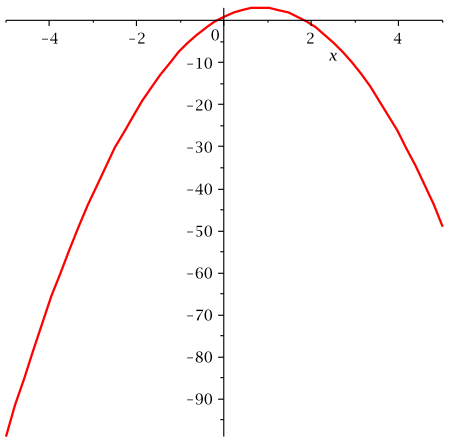
(a) $f(D) = \mathbb{R}$
(b) $f(D) = \{c\}$
(c) $f(D) = [0, \infty)$
(d) $f(D) = \mathbb{Z}$
(e) $f(D) = [0, \infty)$

In particular, we see that the range is not always the entire $\mathbb{R}$.

Further examples of functions are

(f) **(Polynomial functions)** $f : \mathbb{R} \to \mathbb{R}$, $f : x \mapsto a_n x^n + a_{n-1} x^{n-1} + \ldots + a_1 x + a_0$, where $n \in \mathbb{N}$ and $a_0, \ldots, a_n \in \mathbb{R}$ are fixed values, the so-called *coefficients* of the polynomial. As a concrete example, for $n = 2$, $a_0 = 1$, $a_1 = 5$, and $a_2 = -3$ we obtain the function

$$f : x \mapsto -3x^2 + 5x + 1.$$



Graph of the polynomial function $f(x) = -3x^2 + 5x + 1$

[17] Note that we also called the compact intervals $[a, b]$ in Definition 5.7 closed. An interval is therefore compact if and only if it is bounded and closed.

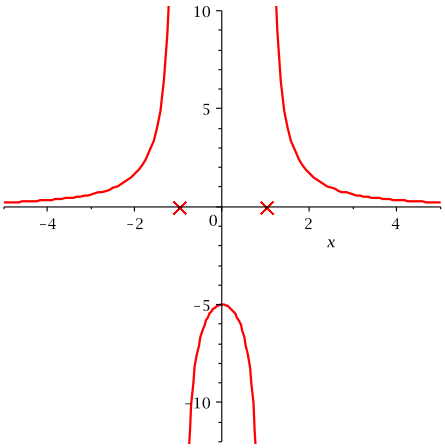
(g) **(Rational functions)** For two polynomial functions g and h we define the set $D := \{x \in \mathbb{R} \,|\, h(x) \neq 0\}$. Then the *rational function* is defined as

$$f : D \to \mathbb{R}, \quad f : x \mapsto \frac{g(x)}{h(x)}.$$

Concrete examples of rational functions are

$$f(x) = \frac{5}{x^2 - 1} \quad \text{and} \quad f(x) = \frac{x^2 - 1}{x - 1}$$

with maximal domains $D = \mathbb{R} \setminus \{-1, 1\}$ and $D = \mathbb{R} \setminus \{1\}$, respectively.

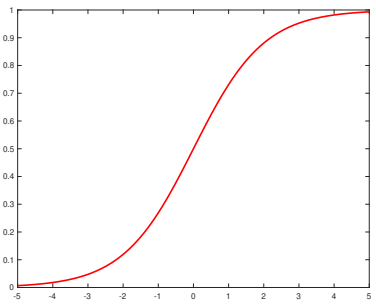


Graph of the rational function $f(x) = \frac{5}{x^2-1}$. The crosses mark the $x \notin D$.

The ranges of polynomial and rational functions cannot be determined as easily as in the previous examples. Although one can get an idea from the graphs of what they look like, the actual computation is generally rather involved.

(h) **(Sigmoid function)** This is actually a class of functions which, like the ReLU function, play an important role in neural networks in machine learning. One representative is the logistic function[18]

$$f : x \mapsto \frac{1}{1 + \exp(-x)}.$$



Graph of the sigmoid function $f(x) = \frac{1}{1+\exp(-x)}$

(i) For every sequence $(a_n)_{n\in\mathbb{N}}$ one can define a function $f : \mathbb{N} \to \mathbb{R}$ by $f : n \mapsto a_n$. Conversely, for every function $f : \mathbb{N} \to \mathbb{R}$, one can define the sequence $(a_n)_{n\in\mathbb{N}}$ by $a_n := f(n)$. Hence, sequences and functions with domain $D = \mathbb{N}$ are simply two different notations for the same mathematical objects.

[18] The exponential function exp will be defined in the next chapter.

♣

Functions are ubiquitous in all applications of mathematics. Whether in physics one considers electrical conductivity as a function of temperature, in electrical engineering the current as a function of voltage, or in economics the income tax rate as a function of salary: in all cases these relationships are expressed mathematically by functions. In AI and machine learning they define, e.g., how the data is processed in a neural network.

From given functions, new functions can be composed in various ways.

DEF Combinations of real functions

14.3. **Definition.** For given functions $f, g : D \to \mathbb{R}$ and $\lambda \in \mathbb{R}$, we define the functions

$$\begin{aligned} f+g : D \to \mathbb{R}, \quad & f+g : \; x \mapsto f(x)+g(x) \\ \lambda f : D \to \mathbb{R}, \quad & \lambda f : \; x \mapsto \lambda f(x) \\ fg : D \to \mathbb{R}, \quad & fg : \; x \mapsto f(x)g(x) \\ \frac{f}{g} : D' \to \mathbb{R}, \quad & \frac{f}{g} : \; x \mapsto \frac{f(x)}{g(x)}, \end{aligned}$$

where the domain D' in the last case must be restricted to

$$D' := \{x \in D \,|\, g(x) \neq 0\}.$$

For two functions $f : D \to \mathbb{R}$ and $g : E \to \mathbb{R}$ with $f(D) \subset E$, we define the *composition* (also called *chaining* or *successive application*) of f and g as

$$g \circ f : D \to \mathbb{R}, \quad g \circ f : x \mapsto g(f(x)).$$

♢

For $f(x) = 2x$ and $g(x) = x^2$, for example, we have $(f+g)(x) = x^2 + 2x$ and $(g \circ f)(x) = (2x)^2 = 4x^2$. For $f(x) = x^2$ and $g(x) = \sqrt{x}$ we obtain $(g \circ f)(x) = \sqrt{x^2} = |x|$.

14.B Limits of Functions.

When examining the graphs of the various functions from the previous section, certain differences become apparent. One graph that stands out is (g), because it is not defined everywhere. This is due to the gaps in the maximal domain of definition, which result directly from the formula for $f(x)$.

The graph (d) also stands out because it has "jumps." In the next section, we want to define mathematically and formally the property "a graph has no jumps." To do this, we must first generalize the concept of limits, already known from sequences, to functions.

DEF Limit of a function

14.4. **Definition.** Let $f : D \to \mathbb{R}$, and let $a \in \mathbb{R}$ be a point for which there exists a convergent real sequence $(x_n)_{n\in\mathbb{N}}$ with $x_n \in D$ for all $n \in \mathbb{N}$ and $\lim_{n\to\infty} x_n = a$.[19]

If there exists a $c \in \mathbb{R}$ such that for every sequence $(x_n)_{n\in\mathbb{N}}$ with $x_n \in D$ for all $n \in \mathbb{N}$ and $\lim_{n\to\infty} x_n = a$, the sequence $f(x_n)$ converges with

$$\lim_{n\to\infty} f(x_n) = c,$$

[19] This is always the case for $a \in D$, since one can simply choose $x_n = a$ for all $n \in \mathbb{N}$. However, it can also hold for $a \notin D$. For example, in the case $D = \mathbb{R} \setminus \{1\}$, we can choose the sequence $x_n = 1 + 1/n$, $n \geq 1$, which lies entirely in D but converges to $1 \notin D$.

then we say that the limit of f at a exists, call c the *limit* of f at a, and define

$$\lim_{x\to a} f(x) := c.$$

With the notation of improper convergence from Definition 8.16, we can in the same way define the limits

$$\lim_{x\to\infty} f(x) \quad \text{and} \quad \lim_{x\to-\infty} f(x).$$

♢

EXAMPLES
Limit of a function

14.5. **Examples.** (a) $\lim_{x\to 0} x^2 = 0$, since for every sequence $x_n \to 0$ we have

$$\lim_{n\to\infty} x_n^2 = \lim_{n\to\infty} x_n \cdot \lim_{n\to\infty} x_n = 0.$$

(b) For every $k \in \mathbb{Z}$, the limit $\lim_{x\to k}[x]$ does not exist, since for the sequences $x_n := k + 1/n$ and $x'_n := k - 1/n$ we have for all $n \geq 2$

$$[x_n] = k \quad \text{and} \quad [x'_n] = k - 1,$$

and thus

$$\lim_{n\to\infty}[x_n] = k \quad \text{and} \quad \lim_{n\to\infty}[x'_n] = k - 1.$$

Hence, the condition that all sequences of the form $[x_n]$ converge to one and the same limit c is violated.

(c) For the rational function

$$f(x) = \frac{x^2 - 1}{x - 1}$$

with maximal domain $D = \mathbb{R}\backslash\{1\}$, the limit $\lim_{x\to 1} f(x)$ exists even though $1 \notin D$. Indeed, let x_n be any sequence with $x_n \to 1$ and $x_n \in D$, i.e. $x_n \neq 1$ for all $n \in \mathbb{N}$. Then

$$f(x_n) = \frac{x_n^2 - 1}{x_n - 1} = \frac{(x_n - 1)(x_n + 1)}{x_n - 1} = x_n + 1 \to 2.$$

Thus, we always obtain the same limit $c = 2$.

(d) An example of improper limits is the following: For a polynomial of the form $f(x) = x^k + a_{k-1}x^{k-1} + \ldots + a_0$, we have

$$\lim_{x\to\infty} f(x) = \infty \quad \text{and} \quad \lim_{x\to-\infty} f(x) = \begin{cases} \infty, & \text{if } k \text{ is even,} \\ -\infty, & \text{if } k \text{ is odd.} \end{cases}$$

To prove the first statement, we must show that for every sequence $x_n \to \infty$ and every $K > 0$, there exists an $M(K) \in \mathbb{N}$ such that $f(x_n) > K$ for all $n \geq M(K)$. For $x \neq 0$, we can write $f(x) = x^k g(x)$ with

$$g(x) = 1 + \frac{a_{k-1}}{x} + \ldots + \frac{a_0}{x^k}.$$

Since $x_n \to \infty$ implies $x_n^k \to \infty$ for all $k \geq 1$, it follows from Theorem 8.17 that each of the fractions in g tends to zero, and hence

$$\lim_{n\to\infty} g(x_n) = 1.$$

Thus, for sufficiently large n, we have $g(x_n) \geq \frac{1}{2}$ and therefore $f(x_n) \geq x_n^k/2$. Since $x_n \to \infty$, it follows that $f(x_n) > K$ for all sufficiently large n, which proves the existence of $M(K)$.

The second statement follows, for even k, from

$$f(-x) = x^k - a_{k-1}x^{k-1} + a_{k-2}x^{k-2} - \ldots - a_1 x + a_0 =: h(x),$$

and for odd k, from

$$f(-x) = -\Big[\, x^k - a_{k-1}x^{k-1} + a_{k-2}x^{k-2} - \ldots + a_1 x - a_0 \,\Big] = -h(x).$$

Note that the functions $h(x)$ in the two cases are not identical, but in both cases satisfy the assumptions of the first statement, hence $\lim_{x\to\infty} h(x) = \infty$. For $x_n \to -\infty$, we have $(-x_n) \to \infty$, and therefore $h(-x_n) \to \infty$. Thus, for even k (with $x = -x_n$),

$$f(x_n) = f(-(-x_n)) = h(-x_n) \to \infty,$$

and for odd k,

$$f(x_n) = f(-(-x_n)) = -h(-x_n) \to -\infty.$$

♣

14.C Continuity.

Continuity is precisely the property of a function that its graph has no jumps within its domain of definition. From the examples in the previous section, it is clear that only the floor function fails to have this property—specifically at the points $k \in \mathbb{Z}$. The observation made in Example 14.5(b), that the limit of this function does not exist at $k \in \mathbb{Z}$, gives us a way to express the intuitive criterion "no jumps" as a mathematically rigorous condition.

DEF
Continuity at a point x

14.6. **Definition.** A function $f : D \to \mathbb{R}$ is called *continuous at a point* $a \in D$ if the limit for $x \to a$ exists and the equation

$$\lim_{x\to a} f(x) = f(a)$$

holds. The function is called *continuous* if it is continuous at every point $a \in D$.

♢

EXAMPLES
Continuity at a point x

14.7. **Examples.** (a) The function $f : x \mapsto x^2$ is continuous at $x = 0$, since $\lim_{x\to 0} x^2 = 0$ (see Example 14.5(a)) and $f(0) = 0^2 = 0$. In fact, the function is continuous for all $x \in D = \mathbb{R}$, as follows from Corollary 14.9 below.

(b) The floor function $f : x \mapsto [x]$ is not continuous at points $x = k$ for $k \in \mathbb{Z}$, because, as shown in Example 14.5(b), the limit does not exist there.

(c) The constant function $f : x \mapsto c$ and the identity function $f : x \mapsto x$ are continuous. For the constant function $f(x) = c$, we have $f(x_n) = c$ for any sequence x_n. Hence, for all $a \in \mathbb{R}$ and every sequence $x_n \to a$, it follows that $\lim_{n\to\infty} f(x_n) = c = f(a)$.

For the identity, we have $f(x_n) = x_n$. If a sequence x_n converges to a, then so does $f(x_n)$, and thus

$$\lim_{n\to\infty} f(x_n) = a = f(a).$$

(d) The absolute value function $f : x \mapsto |x|$ is continuous, since for every convergent sequence $x_n \to a$ we have, by the reverse triangle inequality,

$$|f(x_n) - f(a)| = \big|\, |x_n| - |a| \,\big| \le |x_n - a|.$$

If $|x_n - a| \to 0$, then $\lim_{n\to\infty} f(x_n) = f(a)$ follows from Theorem 8.4. ♣

Verifying continuity using the limit definition from 14.4 is generally tedious, and should therefore be avoided when possible. Instead, one usually attempts to deduce the continuity of a given function from the continuity of known functions. The following theorem shows how this can be done.

14.8. **Theorem.** *(a) Let $f, g : D \to \mathbb{R}$ be functions that are continuous at some point $a \in D$. Then, for any $\lambda \in \mathbb{R}$, the functions*

$$f + g, \quad \lambda f, \quad \text{and} \quad fg$$

are also continuous at a. If, in addition, $a \in D' := \{x \in D \mid g(x) \neq 0\}$, then the function

$$\frac{f}{g}$$

is continuous at a as well.

(b) For two functions $f : D \to \mathbb{R}$ and $g : E \to \mathbb{R}$ with $f(D) \subset E$, if f is continuous at $a \in D$ and g is continuous at $f(a) \in E$, then the composition $g \circ f$ is continuous at a.

THM Continuity of combined functions

Proof. (a) Let $(x_n)_{n\in\mathbb{N}}$ be a sequence with $x_n \in D$ (or $x_n \in D'$ in the last case) and $x_n \to a$. We must show that the function values $f(x_n) + g(x_n)$, etc., converge with $\lim_{n\to\infty}(f(x_n) + g(x_n)) = f(a) + g(a)$, and so on. This follows directly from the assumed convergences $f(x_n) \to f(a)$ and $g(x_n) \to g(a)$ as $n \to \infty$, together with the corresponding theorems for sums, products, and quotients of limits.

(b) Let $(x_n)_{n\in\mathbb{N}}$ be a sequence with $x_n \in D$. From the continuity of f, we have $\lim_{n\to\infty} f(x_n) = f(a)$. Thus, the sequence $(f(x_n))_{n\in\mathbb{N}}$ converges to $f(a)$, and by the continuity of g it follows that

$$\lim_{n\to\infty} g \circ f(x_n) = \lim_{n\to\infty} g(f(x_n)) = g(f(a)) = g \circ f(a).$$

Hence, $g \circ f$ is continuous at a. ❑

14.9. **Corollary.** *Every rational function is continuous.*

COR Continuity of rational functions

Proof. Every rational function has the form

$$f(x) = \frac{g(x)}{h(x)}$$

with $g(x) = a_n x^n + a_{n-1}x^{n-1} + \ldots + a_1 x + a_0$ and $h(x) = b_m x^m + b_{m-1}x^{m-1} + \ldots + b_1 x + b_0$. Thus, it can be written as

$$f(x) = \frac{a_n p(x)^n + a_{n-1}p(x)^{n-1} + \ldots + a_1 p(x) + q(x)}{b_m p(x)^m + b_{m-1}p(x)^{m-1} + \ldots + b_1 p(x) + r(x)}$$

with $p(x) = x$, $q(x) = a_0$, and $r(x) = b_0$. Therefore, the continuity of f at every $a \in D$ follows by repeated application of Theorem 14.8(a) to the individual components of f, using the continuity of the identity and constant functions established in Example 14.7(d). ❑

An example of the application of Theorem 14.8 is the function $f : x \mapsto |x^3|$, which can be written as $g \circ h$ with $g(x) = |x|$ and $h(x) = x^3$. Since g and h are both continuous and have $D = \mathbb{R}$, f is also continuous.

14.D Theorems on Continuous Functions.

Continuous functions possess many properties that are extremely useful in analysis. In this chapter, we will formulate and prove some of the most important of these. We make use of the fact—already mentioned before Example 14.2(e)—that we can restrict the maximal domain of definition of a function arbitrarily. Here, we will mainly consider functions of the form $f : [a, b] \to \mathbb{R}$ on bounded, closed intervals $[a, b]$.

This does not mean that $D = [a, b]$ is necessarily the maximal possible domain of definition of f, but rather that the maximal domain contains $[a, b]$, and that in what follows, we only consider the values $f(x)$ for $x \in [a, b]$. What the function does outside $[a, b]$ is irrelevant for our purposes. Thus, when we speak of *continuous functions* $f : [a, b] \to \mathbb{R}$, this means that f is continuous at every $x \in [a, b]$, though not necessarily at every point in its maximal domain.

14.10. **Theorem.** *Let $a, b \in \mathbb{R}$ with $a < b$, and let $f : [a, b] \to \mathbb{R}$ be a continuous function satisfying $f(a) \le 0$ and $f(b) \ge 0$ (or $f(a) \ge 0$ and $f(b) \le 0$). Then there exists an $x \in [a, b]$ such that $f(x) = 0$. This x is called a* zero *or* root *of f.*

THM Zero Theorem

Proof. We consider the case $f(a) \le 0$ and $f(b) \ge 0$. The case with opposite inequalities follows by applying the proof to $-f$ instead of f.

To prove the statement, we construct an interval nesting $I_n = [a_n, b_n]$ such that for all n, $f(a_n) \le 0$ and $f(b_n) \ge 0$. We proceed analogously to the proof of Theorem 5.9 as follows:

(1) Set $a_0 := a$, $b_0 := b$, and $n := 0$.
(2) Define $c_n := (a_n + b_n)/2$ (the midpoint of I_n).
(3) (i) If $f(c_n) \ge 0$, set $a_{n+1} := a_n$ and $b_{n+1} := c_n$;
    (ii) otherwise, set $a_{n+1} := c_n$ and $b_{n+1} := b_n$.
(4) Set $n := n + 1$ and return to (2).

From this construction, we immediately obtain $f(a_n) \le 0$ and $f(b_n) \ge 0$. The convergence $|I_n| \to 0$ follows from Theorem 8.4, since

$$|I_n| \le \frac{b - a}{2^n}.$$

By Example 8.1(f), the sequences (a_n) and (b_n) converge to the same limit x, and from the continuity of f it follows that $\lim_{n\to\infty} f(a_n) = f(x)$ and $\lim_{n\to\infty} f(b_n) = f(x)$. From Corollary 8.15, we have

$$0 \ge \lim_{n\to\infty} f(a_n) = f(x) = \lim_{n\to\infty} f(b_n) \ge 0,$$

and hence $f(x) = 0$. ❑

A direct consequence of Theorem 14.10 is the following corollary.

14.11. **Corollary.** *Let $a, b, c \in \mathbb{R}$ with $a < b$, and let $f : [a, b] \to \mathbb{R}$ be continuous such that $f(a) \le c$ and $f(b) \ge c$ (or $f(a) \ge c$ and $f(b) \le c$). Then there exists an $x \in [a, b]$ with $f(x) = c$.*

COR Intermediate Value Theorem

Proof. The function $g : x \mapsto f(x) - c$ satisfies all the assumptions of Theorem 14.10. Hence there exists an $x \in [a, b]$ with $g(x) = 0$, and thus

$$f(x) - c = g(x) = 0 \quad \Rightarrow \quad f(x) = c.$$

❑

The following examples illustrate several applications of this theorem and corollary.

EXAMPLES Applications of intermediate value theorem

14.12. **Examples.** (a) For every $y > 0$ and $k \in \mathbb{N}$, there exists an $x > 0$ such that $x^k = y$. We also write this as $x = \sqrt[k]{y}$.

This follows by applying the Intermediate Value Theorem to the continuous function $f(x) = x^k$. From Example 14.5(d), we know that $f(x) \to \infty$ as $x \to \infty$, so there exists some b with $f(b) > y$. On the other hand, for $a = 0$ we have $f(a) = 0^k = 0 < y$. Hence, by Corollary 14.11, there exists an x such that $y = f(x) = x^k$.

Note that the theorem does not apply for $y < 0$ and $k = 2$, since in this case $f(x) = x^k - y \geq 0 - y = -y > 0$ for all $x \in \mathbb{R}$, so no zero exists.

(b) Every polynomial of the form $f(x) = x^k + a_{k-1}x^{k-1} + \ldots + a_0$ with odd k has (at least) one zero. This follows because f is continuous and, from Example 14.5(d), we have $f(x) \to \infty$ as $x \to \infty$ and $f(x) \to -\infty$ as $x \to -\infty$. For sufficiently large b, $f(b) > 0$, and for sufficiently small a, $f(a) < 0$, so Theorem 14.10 applies.

(c) The example $f(x) = [x]$ (the floor function) shows that continuity is essential for the validity of Corollary 14.11. For this function, $f(1) = 1 > 1/2$ and $f(0) = 0 < 1/2$, yet there is no $x \in [0, 1]$ with $f(x) = 1/2$, because the function jumps discontinuously from 0 to 1 without taking intermediate values. ♣

DEF Bounded function

14.13. **Definition.** A function $f : D \to \mathbb{R}$ is called *bounded* if the set $f(D)$ is bounded, that is, if there exists an $M \in \mathbb{R}$ such that

$$|f(x)| \leq M \quad \text{for all } x \in D.$$

♢

The following theorem expresses a statement that is important both for many theoretical foundations and for numerous applications of mathematics.

THM Boundedness of continuous functions

14.14. **Theorem.** *Every continuous function $f : [a, b] \to \mathbb{R}$ on a compact interval $[a, b]$ is bounded. Moreover, the maximum and minimum*

$$\max\{f(x) \mid x \in [a, b]\} \quad \textit{and} \quad \min\{f(x) \mid x \in [a, b]\}$$

exist; that is, there exist $p, q \in [a, b]$ such that

$$f(p) = \sup\{f(x) \mid x \in [a, b]\} \quad \textit{and} \quad f(q) = \inf\{f(x) \mid x \in [a, b]\}.$$

The arguments p and q are then called maximiser and minimiser, respectively.

Proof. We prove the theorem for the maximum; the proof for the minimum is analogous, and boundedness then follows from the existence of both maximum and minimum with $M = \max\{|f(p)|, |f(q)|\}$.

Let

$$s := \sup\{f(x) \mid x \in [a, b]\},$$

where we write $s = \infty$ if the set is unbounded above. If s is finite, then for each $\varepsilon = 1/n$, $n \geq 1$, there exists $x_n \in [a, b]$ with $f(x_n) > s - \varepsilon$; and if $s = \infty$, then for each $n \in \mathbb{N}$ there exists $x_n \in [a, b]$ with $f(x_n) > n$. In both cases, we obtain a sequence $(x_n)_{n\in\mathbb{N}}$ with

$$\lim_{n\to\infty} f(x_n) = s.$$

Since the sequence (x_n) lies in the bounded interval $[a, b]$, it possesses a convergent subsequence (x_{n_k}) by the Bolzano–Weierstrass Theorem 8.42. Because

$[a,b]$ is closed, its limit satisfies $\lim_{k\to\infty} x_{n_k} =: p \in [a,b]$. By Corollary 8.15, the subsequence $(f(x_{n_k}))_{k\in\mathbb{N}}$ converges to the same limit as $(f(x_n))_{n\in\mathbb{N}}$, hence $\lim_{k\to\infty} f(x_{n_k}) = s$. Since f is continuous on $[a,b]$, we obtain

$$f(p) = \lim_{k\to\infty} f(x_{n_k}) = s.$$

Thus $s = f(p) \in \mathbb{R}$, i.e. the supremum is finite, and since $f(p) = s$ and $f(p) \in \{f(x) \mid x \in [a,b]\}$, $f(p)$ is the desired maximum. ❑

We will encounter various applications of this theorem throughout this course. However, its significance extends beyond analysis itself. Many applications of mathematics lead to so-called optimization problems, in which one seeks to maximize (e.g. profit, production yield, driving comfort, etc.) or minimize (e.g. loss, pollutant emissions, deviation from a target value, etc.) a quantity. Although Theorem 14.14 does not tell us *how* to achieve this—further mathematical tools will be needed for that—it guarantees that it *makes sense* to search for a maximum or a minimum.

The following examples show that none of the assumptions (continuity, closed interval, bounded interval) can be omitted.

EXAMPLES
Unbounded continuous functions

14.15. **Examples.** (a) Consider the discontinuous function $f(x) = x - [x]$ on the interval $[0,1]$. For any $\varepsilon > 0$ with $\varepsilon < 1$, we have

$$f(1-\varepsilon) = 1 - \varepsilon - [1-\varepsilon] = 1 - \varepsilon - 0 = 1 - \varepsilon,$$

hence

$$\sup\{f(x) \mid x \in [0,1]\} = \sup\{1-\varepsilon \mid \varepsilon \in (0,1)\} = 1.$$

However, there is no $p \in [0,1]$ with $f(p) = 1$, since for $0 \le p < 1$ we have $f(p) = p - [p] = p < 1$, and for $p = 1$, $f(p) = 1 - [1] = 0$.

(b) Consider the function $f(x) = 1/x$. This is a rational function and therefore continuous on its domain $\mathbb{R} \setminus \{0\}$, and in particular on the (left-open) interval $(0,1]$. However, it is unbounded there, since for $x = 1/n \in (0,1]$, $n \ge 1$, the function takes arbitrarily large values $f(x) = 1/(1/n) = n$.

(c) Again consider $f(x) = 1/x$, but now on the unbounded interval $[1,\infty)$. This function is bounded by $M = 1$, but it does not attain a minimum, since for $x = n$, $n \ge 1$, we have $f(x) = 1/n$, which becomes arbitrarily close to zero. Thus,

$$\inf\{f(x) \mid x \in [1,\infty)\} = 0,$$

but there is no $q \in [1,\infty)$ with $f(q) = 1/q = 0$. ♣

14.E The ε-δ Criterion for Continuity.

The criterion for continuity in Definition 14.6 is also known as *sequential continuity*. Sequential continuity can be rather cumbersome to verify, since one must check all possible convergent sequences. It is also impractical when one wishes to know how much the values $f(x_n)$ and $f(a)$ differ—a property that will be important in Definition 14.19 and the following examples.

We therefore introduce, at the end of this chapter, an equivalent definition of continuity that does not rely on sequences and limits.

THM The ε–δ Criterion of Continuity

14.16. **Theorem.** *A real function $f : D \to \mathbb{R}$ is continuous at $a \in D$ if and only if for every $\varepsilon > 0$ there exists a $\delta > 0$ such that for all $x \in D$,*

$$d(x, a) < \delta \;\Rightarrow\; d(f(x), f(a)) < \varepsilon.$$

Intuitively: if x differs only slightly from a, then the function values $f(x)$ and $f(a)$ also differ only slightly.

Proof. We first show that the ε–δ criterion implies the sequential definition of continuity (Definition 14.6). Let $a \in D$ and let $(x_n)_{n\in\mathbb{N}}$ be any sequence with $x_n \to a$ and $x_n \in D$. We must show that $f(x_n) \to f(a)$.

Let $\varepsilon > 0$ be given, and let $\delta > 0$ be the corresponding value from the criterion. Since $x_n \to a$, there exists $N_x(\delta) \in \mathbb{N}$ such that $|x_n - a| < \delta$ for all $n \geq N_x(\delta)$, and hence

$$d(f(x_n), f(a)) < \varepsilon.$$

Thus the definition of convergence is satisfied with $N(\varepsilon) = N_x(\delta)$.

Conversely, assume that f is continuous at a in the sequential sense, but the ε–δ criterion fails. Then there exists some $\varepsilon > 0$ such that for every $\delta > 0$ there exists an $x_\delta \in D$ with $d(x_\delta, a) < \delta$ and $d(f(x_\delta), f(a)) \geq \varepsilon$.

Set $\delta = 1/n$ and define $x_n := x_\delta$. Then $d(x_n, a) < 1/n$, and by Theorem 8.4, $x_n \to a$. However,

$$d(f(x_n), f(a)) \geq \varepsilon,$$

so again by Theorem 8.4, $f(x_n) \not\to f(a)$. This contradicts the continuity of f at a. Hence, the ε–δ criterion must hold. ❑

We now apply this criterion to prove the continuity of the square root function.

EXAMPLE ε-δ continuity for square root function

14.17. **Example.** Consider the function $f : x \mapsto \sqrt{x}$ with the domain $D = [0, \infty)$. For $a > 0$ and $x \geq 0$, we have

$$f(x) - f(a) = \sqrt{x} - \sqrt{a} = (\sqrt{x} - \sqrt{a})\frac{\sqrt{x} + \sqrt{a}}{\sqrt{x} + \sqrt{a}} = \frac{x - a}{\sqrt{x} + \sqrt{a}}.$$

We now consider the two cases $a > 0$ and $a = 0$ separately:

In the case $a > 0$, for all $x \geq 0$ with $|x - a| < \delta$, the inequality

$$d(f(x), f(a)) = \left|\frac{x - a}{\sqrt{x} + \sqrt{a}}\right| < \frac{\delta}{\sqrt{x} + \sqrt{a}} \leq \frac{\delta}{\sqrt{a}}$$

holds. The desired inequality $|f(x) - f(a)| < \varepsilon$ therefore follows if we set $\delta = \varepsilon\sqrt{a}$.

In the case $a = 0$, the inequality $|f(x) - f(a)| < \varepsilon$ is obviously satisfied for $x = a = 0$ and all $\varepsilon > 0$. For all $x > 0$ with $|x - a| < \delta$, we have $x < \delta$ and hence

$$d(f(x), f(a)) = |\sqrt{x} - \sqrt{0}| = |\sqrt{x}| = \sqrt{x} < \sqrt{\delta}.$$

In this case, the desired inequality $d(f(x), f(a)) < \varepsilon$ follows with $\delta = \varepsilon^2$.

This example also illustrates how an appropriate $\delta > 0$ is typically found: we first seek an upper bound for $d(f(x), f(a))$ that depends on δ (and possibly a), and from this we determine a suitable δ depending on the given ε and possibly a. ♣

The ε-δ criterion is useful, for example, in proving the following theorem[20].

14.18. **Theorem.** *Let $f : D \to \mathbb{R}$ be a function that is continuous at a point $a \in D$ and satisfies $f(a) \neq c$. Then there exists a $\delta > 0$ such that $f(x) \neq c$ for all $x \in D \cap (a - \delta, a + \delta)$.*

THM Continuous functions do not change abruptly

Proof. Choose $\varepsilon = d(f(a), c)$ and $\delta > 0$ according to the ε-δ criterion. Then for all $x \in D \cap (a - \delta, a + \delta)$ we have $d(x, a) < \delta$ and thus, by the reverse triangle inequality,

$$d(f(x), c) \geq d(f(a), c) - \underbrace{d(f(a), f(x))}_{<\varepsilon} > d(f(a), c) - \varepsilon = 0,$$

hence $d(f(x), c) > 0$ and thus $f(x) \neq c$. ❑

The δ derived for the square root function in Example 14.17 depends not only on ε but also on a. In fact, the value $\delta = \varepsilon\sqrt{a}$ becomes smaller and smaller as a approaches zero. Comparing this with the graph of the square root function from Example 14.2(e), one can see that this is evidently related to the slope of the graph: the steeper the graph, the smaller δ must be chosen for a given ε. This becomes clear when one visualizes the relationship between ε and δ graphically.

A particularly nice case of continuity occurs when δ can be chosen independently of a. This form of continuity is formally defined as follows.

14.19. **Definition.** A function $f : D \to \mathbb{R}$ is called *uniformly continuous* if the following holds: For every $\varepsilon > 0$ there exists a $\delta > 0$ such that for all $x, x' \in D$,

$$d(x, x') < \delta \;\Rightarrow\; d(f(x), f(x')) < \varepsilon.$$

DEF Uniform continuity

♢

Note the small but essential difference compared to the condition in Theorem 14.16: in uniform continuity, the same δ must guarantee the inequality $d(f(x), f(x')) < \varepsilon$ for all $x' \in D$, whereas in Theorem 14.16, δ may depend on a.

Since the δ in Example 14.17 for the square root function indeed depends on a and becomes smaller for smaller a, one might conclude that the square root function is not uniformly continuous. However, this conclusion would be premature, since we do not know whether the δ calculated in that example is the best possible one. Could it be that a more clever derivation yields a δ independent of a?

Indeed, this is possible—but instead of explicitly calculating such a δ, we proceed more elegantly and prove the following theorem.

14.20. **Theorem.** *Every continuous function $f : [a, b] \to \mathbb{R}$ defined on a compact interval is uniformly continuous.*

THM Uniform continuity on compact intervals

Proof. Suppose f is not uniformly continuous. Analogous to the second part of the proof of Theorem 14.16, this means: There exists an $\varepsilon > 0$ such that for every $\delta > 0$ there are points $x_\delta, x'_\delta \in D$ with $d(x_\delta, x'_\delta) < \delta$ and $d(f(x_\delta), f(x'_\delta)) \geq \varepsilon$.

As in the proof of Theorem 14.16, for each $n \geq 1$ we set $\delta = 1/n$ and obtain two sequence elements $x_n = x_\delta$ and $x'_n = x'_\delta$ with

$$d(x_n, x'_n) \leq 1/n \quad \text{and} \quad d(f(x_n), f(x'_n)) \geq \varepsilon.$$

[20] Of course, this could also be proved using the sequential definition of continuity, but that would be much more complicated and tedious.

This defines two bounded sequences $x_n, x'_n \in [a,b]$. By the Bolzano–Weierstrass theorem, (x_n) has a convergent subsequence $(x_{n_k})_{k\in\mathbb{N}}$ with limit

$$p := \lim_{k\to\infty} x_{n_k} \in [a,b].$$

Since

$$d(x'_{n_k}, p) \le d(x'_{n_k}, x_{n_k}) + d(x_{n_k}, p) < \frac{1}{n_k} + d(x_{n_k}, p) \to 0,$$

the subsequence $(x'_{n_k})_{k\in\mathbb{N}}$ also converges to p. As f is continuous on $[a,b]$, it follows that

$$\lim_{k\to\infty} \big(f(x_{n_k}) - f(x'_{n_k})\big) = \lim_{k\to\infty} f(x_{n_k}) - \lim_{k\to\infty} f(x'_{n_k}) = f(p) - f(p) = 0.$$

Hence, $\lim_{k\to\infty} d(f(x_{n_k}), f(x'_{n_k})) = 0$, contradicting the inequality $d(f(x_n), f(x'_n)) \ge \varepsilon$ valid for all n. Therefore, f must be uniformly continuous. ❑

The theorem shows that uniform continuity is not a "special case", but rather a property that every continuous function on a bounded, closed interval possesses. It may fail only on (half-)open or unbounded intervals.

This also shows that the square root function is indeed uniformly continuous on every interval of the form $[0,b]$, since it is continuous there as shown in Example 14.17. The issue of decreasing δ values for a near zero simply arises because the computed δ in that example was not optimal.

Even though, by the preceding theorem, we no longer need to explicitly calculate the "better" (i.e., a-independent) δ to prove its existence, it is still interesting to see where the estimate in Example 14.17 can be improved. In fact, in the case $a > 0$ we can proceed as follows:

For $a > x$, we have

$$d(f(x), f(a)) = \left|\frac{x-a}{\sqrt{x}+\sqrt{a}}\right| \le \frac{a-x}{\sqrt{a}} \le \frac{a-x}{\sqrt{a-x}} = \sqrt{a-x} < \sqrt{\delta},$$

and for $a < x$,

$$d(f(x), f(a)) = \left|\frac{x-a}{\sqrt{x}+\sqrt{a}}\right| \le \frac{x-a}{\sqrt{x}} \le \frac{x-a}{\sqrt{x-a}} = \sqrt{x-a} < \sqrt{\delta}.$$

We now see that even in the case $a > 0$, one can use $\delta = \varepsilon^2$ to obtain the inequality $d(f(x), f(a)) < \varepsilon$. Since this δ is independent of a, it can be applied to any x and x' (in place of x and a) and thus satisfies the definition of uniform continuity. This works even on the interval $[0,\infty)$, showing that the square root function is actually uniformly continuous on its entire domain $[0,\infty)$.

To illustrate that continuous functions on unbounded intervals are generally not uniformly continuous, consider one final example.

EXAMPLE
Non uniformly continuous function

14.21. **Example.** Consider the function $f : [0,\infty) \to \mathbb{R}$ given by $f(x) = x^2$. Suppose the function were uniformly continuous on $[0,\infty)$. Then, for a given $\varepsilon > 0$, there would exist a $\delta > 0$ such that for all $x, x' \in [0,\infty)$ with $d(x,x') < \delta$, the inequality $d(f(x'), f(x)) < \varepsilon$ holds. In particular, this inequality must hold for $x' = x + \delta/2$, yielding

$$\varepsilon > d(f(x+\delta/2), f(x)) = (x+\tfrac{\delta}{2})^2 - x^2 = x^2 + \delta x + \frac{\delta^2}{4} - x^2 = \delta x + \frac{\delta^2}{4} \ge \delta x.$$

Hence, $\delta < \varepsilon/x$ must hold for all $x \in (0,\infty)$. But this is impossible for any $\delta > 0$, leading to a contradiction. ♣

Date:
March 5, 2026

15.A Definition of the exponential function.

All examples of functions we have considered so far were defined by simple rational formulas or other elementary operations (such as the absolute value or the floor function). In this section, we now consider a function that is important in many areas of analysis and its applications, which is defined by an infinite series.

THM
Absolute convergence of exponential series

15.1. **Theorem.** *For every $x \in \mathbb{R}$, the* exponential series

$$\sum_{k=0}^{n} \frac{x^k}{k!}$$

(with the conventions[21] $0^0 = 1$ and $0! = 1$) is absolutely convergent.

Proof. The claim is immediately clear for $x = 0$, since all terms for $k \geq 1$ are equal to zero. Because of the stated conventions, the sum for $x = 0$ has the value 1. For $x \neq 0$, the claim follows from the ratio test in Theorem 9.18 with $\theta = 1/2$, since

$$\left| \frac{\frac{x^{k+1}}{(k+1)!}}{\frac{x^k}{k!}} \right| = \frac{|x|}{k+1} \leq \frac{1}{2}$$

for all $k \geq 2|x|$. ❑

Thus, the existence of the limit $\sum_{k=0}^{\infty} \frac{x^k}{k!}$ is ensured, and we can formulate the following definition.

DEF
Exponential function

15.2. **Definition.** We define the exponential function[22] $\exp : \mathbb{R} \to \mathbb{R}$ as

$$\exp(x) := \sum_{n=0}^{\infty} \frac{x^n}{n!}.$$

♢

The special significance of the exponential function for analysis will become apparent in many places in this and the following chapters. The function is also important in many applications; we will give some examples. For this, we need the following alternative representation of the exponential function.

THM

15.3. **Theorem.** *For the exponential function, the following holds for all $x \in \mathbb{R}$:*

$$\exp(x) = \lim_{n \to \infty} \left(1 + \frac{x}{n}\right)^n.$$

The proof of this theorem can be carried out with the techniques known so far, but it is rather technical and not very illustrative, hence it is omitted at this point. A short and simple proof will be given in the next chapter after Example 16.11.

With Theorem 15.3, we can now present the already announced applications.

EXAMPLES
Examples for exponential functions in applications

15.4. **Examples.** (a) **(Compound interest)** If an investment of a euros earns

[21] These arise from $x^n = \prod_{k=1}^{n} x$ and $n! = \prod_{k=1}^{n} k$ together with the definition, valid for all $y \in \mathbb{R}$, of the empty product $\prod_{n}^{n-1} y = 1$.

[22] We will later see how to compute this value approximately. However, every scientific calculator, higher programming language, and mathematical software package has the exponential function built in. It is often denoted by "e^x"; the reason for this will become clear shortly.

$p\%$ interest per year, then with annual compounding (at the end of the year) the value, including interest, is $a(1+r)$ euros, where $r = p/100$. (Example: $a = 100$ euros at $p = 5\%$ interest $\rightsquigarrow r = 0.05 \rightsquigarrow$ amount at year?s end $100(1+0.05)$ euros $= 100(1.05)$ euros $= 105$ euros.)

With semiannual compounding (after half a year and after the full year), taking into account compound interest, the value is

$$a\left(1+\frac{r}{2}\right)\left(1+\frac{r}{2}\right) \text{ euros} = a\left(1+\frac{r}{2}\right)^2 \text{ euros}$$

(in the numerical example: $100(1.025)(1.025)$ euros $= 105.0625$ euros).

For $n = 12$ compounding periods (i.e., monthly compounding), we similarly obtain

$$a\left(1+\frac{r}{12}\right)^{12} \text{ euros}$$

(in the example: $a \approx 105.1162$ euros).

If this process is continued for larger n (weekly, daily, hourly, ... compounding), then according to Theorem 15.3 we obtain the amount

$$\lim_{n\to\infty} a\left(1+\frac{r}{n}\right)^n = a\exp(r) \text{ euros};$$

this type of compounding is called *continuous compounding*. In the numerical example, we get $a \approx 105.1272$ euros, which shows that the deviation from monthly compounding is already very small—only about 0.01%. Since the exponential function is much easier to handle than the terms $(1+r/n)^n$, continuous compounding is usually assumed in financial mathematics.

(b) Many functions used as components of neural networks in machine learning contain the exponential function as a part. We have already seen this in Example 14.2 with the logistic function

$$f(x) = \frac{1}{1+\exp(-x)}.$$

Another example is the so-called softplus function, which we will define in Example 16.2(h). ♣

How is the value $\exp(x)$ of the exponential function computed in practice, i.e., how does a calculator or a computer do it? The obvious idea is to approximate the value by

$$\exp(x) \approx \sum_{n=0}^{N} \frac{x^n}{n!}$$

for a sufficiently large $N \in \mathbb{N}$. One can verify that the error made in this approximation is less than $2\frac{|x|^{N+1}}{(N+1)!}$ for $|x| \le 1 + \frac{N}{2}$.

If we want to approximate, for example, the value $\exp(2)$ with an error of at most 10^{-10}, we must choose N large enough so that the inequalities

$$2\frac{2^{N+1}}{(N+1)!} \le 10^{-10} \quad \text{and} \quad 2 \le 1+\frac{N}{2}$$

are satisfied. The second inequality holds for all $N \ge 2$ and the first holds for all $N \ge 17$, as can be checked by simple trial with a calculator. Thus, the first 18 terms of the exponential series must be computed and summed.

15.B Rules and Properties of the Exponential Function.

In this section, we formulate the most important rules for the exponential function.

15.5. **Theorem.** *For all $x, y \in \mathbb{R}$ it holds that*

$$\exp(x+y) = \exp(x)\exp(y).$$

THM Functional equation of the exponential

Proof. We use the Cauchy product of series according to theorem 9.20. This is applicable because of the absolute convergence of the exponential series proven in theorem 15.1.

According to theorem 9.20 it holds that

$$\sum_{k=0}^{\infty} c_k = \left(\sum_{k=0}^{\infty} a_k\right)\left(\sum_{k=0}^{\infty} b_k\right)$$

with

$$c_k := \sum_{j=0}^{k} a_{k-j} b_j.$$

Applied to the exponential series with

$$a_k := \frac{x^k}{k!} \quad \text{and} \quad b_k := \frac{y^k}{k!}$$

we have

$$c_k = \sum_{j=0}^{k} \frac{x^{k-j}}{(k-j)!} \cdot \frac{y^j}{j!} = \frac{1}{k!} \sum_{j=0}^{k} \binom{k}{j} x^{k-j} y^j,$$

with the binomial coefficient

$$\binom{k}{j} = \frac{k!}{(k-j)!j!}.$$

By the binomial theorem 4.4 it follows that

$$c_k = \frac{1}{k!}(x+y)^k$$

and thus

$$\exp(x+y) = \sum_{k=0}^{\infty} \frac{(x+y)^k}{k!} = \sum_{k=0}^{\infty} c_k = \left(\sum_{k=0}^{\infty} a_k\right)\left(\sum_{k=0}^{\infty} b_k\right) = \exp(x)\exp(y).$$

❑

From this theorem, the following properties of the exponential function immediately follow.

15.6. **Theorem.** *(a) For all $x \in \mathbb{R}$ it holds that $\exp(-x) = \frac{1}{\exp(x)} = \exp(x)^{-1}$.*

(b) For all $x \in \mathbb{R}$ it holds that $\exp(x) > 0$.

(c) For all $k \in \mathbb{Z}$ it holds that $\exp(k) = e^k$, where e is the Euler number

$$e := \exp(1) = \sum_{n=0}^{\infty} \frac{1}{n!} = 2.718\,281\,828\,459\ldots$$

THM Properties of the exponential function

Proof. (a) By theorem 15.5 it holds that

$$\exp(x)\exp(-x) = \exp(0) = \sum_{n=0}^{\infty} \frac{0^n}{n!} = 1$$

since we adopted the convention $0^0 = 1$. Therefore $\exp(x) \neq 0$ and

$$\exp(-x) = \frac{1}{\exp(x)}.$$

(b) For $x \geq 0$ the statement is clear because $\exp(x) = \sum_{n=0}^{\infty} \frac{x^n}{n!} \geq x^0 = 1$. For $x < 0$ it follows that $\exp(-x) > 0$ and thus also $\exp(x) = 1/\exp(-x) > 0$.

(c) We first show the statement by induction for all $n \in \mathbb{N}$. For $n = 0$ it holds that $\exp(0) = 1 = e^0$.

For $n \to n+1$ and $n \geq 0$ it holds that

$$\exp(n+1) = \exp(n)\exp(1) = \exp(n)e = e^n e = e^{n+1}.$$

Thus the statement is shown for $k = n \geq 0$.

For $k \in \mathbb{Z}$ with $k < 0$ it holds with $n = -k$

$$\exp(k) = \exp(-n) = \frac{1}{\exp(n)} = \frac{1}{e^n} = e^{-n} = e^k.$$

This proves the statement for $k < 0$. ❑

Statement (c) already gives an indication of why one often writes e^x instead of $\exp(x)$. There is, however, a deeper reason, which we will learn later in this chapter.

With the just-proven rules and the above error estimate we can now prove the following theorem.

15.7. **Theorem.** *The exponential function is continuous.*

THM
Continuity of the exponential function

Proof. For any $a \in \mathbb{R}$ and any sequence with $\lim_{n\to\infty} x_n \to a$ we must prove that $\lim_{n\to\infty} \exp(x_n) = \exp(a)$ holds. Let x_n be such a sequence.

We first consider $a = 0$. From the above error estimate it follows with $N = 0$ for all $|x| \leq 1$

$$|\exp(x) - 1| \leq 2|x|.$$

For $x_n \to 0$ there exists an $n_0 \in \mathbb{N}$ with $|x_n| \leq 1$ for all $n \geq n_0$. Thus it follows that

$$|\exp(x_n) - 1| \leq 2|x_n| \to 0$$

for $n \to \infty$ and thus $\exp(x_n) \to 1 = \exp(0)$.

For $a \neq 0$ it holds that $x_n - a \to 0$ and thus

$$\exp(x_n) = \exp(a)\exp(x_n - a) \to \exp(a)\exp(0) = \exp(a).$$

❑

15.C Monotone Functions.

In the following section, we want to introduce the logarithm as the inverse function of the exponential function. For this, we recall the definition of the inverse function from Definition 2.20. Moreover, we observe that every injective function f on $D \subset \mathbb{R}$ is also bijective if we restrict its codomain A to $f(D)$. Thus, for injective functions, the inverse function $f^{-1} : D' \to \mathbb{R}$ with $D' = f(D)$ exists.

Warning. The inverse function f^{-1} is easily confused with the function $x \mapsto f(x)^{-1} = \frac{1}{f(x)}$.

15.8. **Example.** The function $f(x) = x^2$ is not injective for $D = \mathbb{R}$, because for $x > 0$ we have $-x \neq x$ but $f(x) = x^2 = (-x)^2 = f(-x)$. For $D = [0, \infty)$ it is injective, because for $x, x' \geq 0$ with $x \neq x'$ either $x' > x$ and thus $f(x') > f(x)$ or $x' < x$ and thus $f(x') < f(x)$, in both cases $f(x) \neq f(x')$. Its inverse function can be computed from the relation $?f^{-1}(y)$ is the unique number $x \in [0, \infty)$ with $x^2 = y? \Leftrightarrow f^{-1}(y) = \sqrt{y}$. ♣

EXAMPLE Inverse of $f(x) = x^2$

Note that for inverse functions, it is often not possible to give simple formulas.

For real functions, an intuitive sufficient condition for injectivity can be given.

15.9. **Definition.** A function $f : D \to \mathbb{R}$ is called monotonically increasing (respectively strictly monotonically increasing, monotonically decreasing, or strictly monotonically decreasing) if for all $x, x' \in D$ with $x < x'$ the inequality

$$f(x) \leq f(x') \quad (\text{respectively } f(x) < f(x'),\ f(x) \geq f(x') \text{ or } f(x) > f(x'))$$

holds. ♢

DEF Monotonicity

15.10. **Theorem.** *Every strictly monotonically increasing (respectively decreasing) function $f : D \to \mathbb{R}$ is injective. It therefore has an inverse function f^{-1} with domain $D' = f(D)$.*

THM Strictly monotone functions are injective

Proof. We show the claim for strictly monotonically increasing functions.

Injectivity: Let $x, x' \in D$ with $x \neq x'$ be given. Then either $x' > x$ and thus $f(x') > f(x)$ or $x' < x$ and thus $f(x') < f(x)$, in both cases $f(x) \neq f(x')$. ❑

Moreover, one can prove that f^{-1} in this case is also strictly monotonically increasing (respectively decreasing). If f is also continuous and $D = [a, b]$ is a compact interval, then $D' = [f(a), f(b)]$ (respectively $D' = [f(b), f(a)]$) and f^{-1} is also continuous.

15.D The Logarithm Function.

15.11. **Theorem.** *The exponential function* $\exp : \mathbb{R} \to \mathbb{R}$ *is strictly monotonically increasing and satisfies* $\exp(\mathbb{R}) = (0, \infty)$. *It is therefore injective and has an inverse function with domain* $D' = (0, \infty)$.

THM Inverse function of exponential function

Proof. We first show that exp is strictly monotonically increasing. For $x' > x \geq 0$ we have

$$\exp(x') = \sum_{n=0}^{\infty} \frac{(x')^n}{n!} = 1 + x' + \sum_{n=2}^{\infty} \frac{(x')^n}{n!} \geq 1 + x' + \sum_{n=2}^{\infty} \frac{x^n}{n!} > 1 + x + \sum_{n=2}^{\infty} \frac{x^n}{n!} = \exp(x).$$

For $x' > 0$ and $x \leq 0$ the statement follows because

$$\exp(x') > \exp(0) = 1 \quad \text{and} \quad \exp(x) = \frac{1}{\exp(-x)} \leq \frac{1}{\exp(0)} = 1,$$

and for $0 \geq x' > x$ we have $-x > -x' \geq 0$, so from the first case it follows that $\exp(-x) > \exp(-x')$ and therefore

$$\exp(x') = \frac{1}{\exp(-x')} > \frac{1}{\exp(-x)} = \exp(x).$$

It remains to show that $\exp(\mathbb{R}) = (0, \infty)$. Since $\exp(x) > 0$ for all $x \in \mathbb{R}$, we have $\exp(\mathbb{R}) \subset (0, \infty)$. Because $\exp(x) \geq x$ for $x > 0$, $\exp(x)$ takes arbitrarily large values, and since $\exp(-x) = 1/\exp(x)$ it also takes values arbitrarily close to 0. That all intermediate values are attained follows from continuity by the Intermediate Value Theorem. ❑

15.12. **Definition.** The inverse function of the exponential function is called the (natural) logarithm and is denoted by

$$\ln : (0, \infty) \to \mathbb{R},$$

sometimes also written as log. ♢

DEF logarithm

15.13. **Theorem.** *The natural logarithm is continuous and strictly monotonically increasing.*

THM Continuity and monotonicity of logarithm

Proof. If we restrict the exponential function to a closed interval $[a, b]$, both properties follow directly from Theorem 15.10. Continuity for arbitrary $y \in D' = (0, \infty)$ and monotonicity for arbitrary $y' > y \in D'$ then follow by choosing $a = y/2$ and $b = 2y$ or $b = 2y'$, respectively. ❑

15.14. **Theorem.** *For all $x, y \in (0, \infty)$ it holds that*

$$\ln(xy) = \ln(x) + \ln(y).$$

THM Functional Equation of the Logarithm

Proof. By the functional equation of the exponential function and the inverse function property of the logarithm, we have

$$xy = \exp(\ln(x)) \exp(\ln(y)) = \exp(\ln(x) + \ln(y)).$$

Applying the logarithm to both sides gives

$$\ln(xy) = \ln(\exp(\ln(x) + \ln(y))) = \ln(x) + \ln(y).$$

❑

From this equation it follows that

$$\ln 1 = \ln(1 \cdot 1) = \ln 1 + \ln 1 \;\Rightarrow\; \ln 1 = 0,$$

and for $x > 0$

$$\ln(x) + \ln(1/x) = \ln(x/x) = \ln 1 = 0 \;\Rightarrow\; \ln(1/x) = -\ln(x).$$

Moreover, from the functional equation and the inverse function property it follows for all real $a > 0$ that

$$a^2 = aa = \exp(\ln(a) + \ln(a)) = \exp(2\ln(a)),$$

and by induction for all $n \geq 1$ also

$$a^n = \exp(n \ln(a)).$$

Similarly, for all $n \in \mathbb{N}$ we have

$$a^{-n} = \frac{1}{a^n} = \frac{1}{\exp(n \ln(a))} = \exp(-n \ln(a)).$$

This motivates the following definition.

15.15. **Definition.** For $a, x \in \mathbb{R}$ with $a > 0$ we define the *power* a^x as **DEF** power

$$a^x := \exp(x \ln(a)).$$

This is also called the *exponential function to the base* a and

$$\exp_a(x) := \exp(x \ln a)$$

is written. The inverse function of $\exp_a$ is denoted by $\log_a$ and called the logarithm to base a. ♢

The following calculation rules can be derived from the already known properties of the exponential function and the logarithm (exercise):

$$a^x a^y = a^{(x+y)}, \qquad (a^x)^y = a^{xy}, \qquad a^x b^x = (ab)^x, \qquad a^{\frac{p}{q}} = \sqrt[q]{a^p}.$$

For $a = e = \exp(1)$ it follows because $\ln(e) = 1$ that

$$e^x = \exp_e(x) = \exp(x \ln(e)) = \exp(x),$$

i.e., the usual exponential function is precisely the exponential function to base e. This is the real reason why one writes e^x instead of $\exp(x)$.

From $\exp = \exp_e$ it then immediately follows that $\ln = \log_e$.

A consequence of the equation $\sqrt[n]{a} = \exp_a(1/n)$ is the following corollary.

15.16. **Corollary.** *For all $a > 0$ it holds that $\lim_{n\to\infty} \sqrt[n]{a} = 1$.* **COR**

Proof. Since $\sqrt[n]{a} = \exp_a(1/n)$ and $\lim_{n\to\infty} 1/n = 0$, it follows from the continuity of $\exp_a$ that

$$\lim_{n\to\infty} \sqrt[n]{a} = \lim_{n\to\infty} \exp_a(1/n) = \exp(0) = 1.$$

❑

One might now ask whether a general power a^x could be defined in some other way. However, one can verify (which we omit here for reasons of time) that this is the only possible definition if the equations $a^1 = a$ and $a^{x+y} = a^x a^y$ are to hold. A proof of this can be found, for example, in §12, Satz 6 in the Analysis book by Forster.

15.E Limit Behavior of exp and ln.

We conclude this chapter with some statements about the behavior of exp and ln as $x \to \infty$ and $x \to 0$.

The first statement says that the exponential function $\exp(x)$ grows faster than any power x^k as $x \to \infty$. Formally, this ?faster growth? is expressed by considering the quotient $\exp(x)/x^k$. That exp grows faster is then expressed as follows.

15.17. **Theorem.** *For all $k \in \mathbb{N}$ it holds that* $\lim\limits_{x\to\infty} \dfrac{e^x}{x^k} = \infty$.

THM
Exp grows faster than any polynomial

Proof. We must show that for every $K > 0$ there exists an $x_K > 0$ such that

$$\frac{e^x}{x^k} > K \quad \text{for all } x > x_K.$$

Choose $x_K = K(k+1)!$, then for all $x > x_K$ it follows that

$$e^x = \sum_{n=0}^{\infty} \frac{x^n}{n!} \geq \frac{x^{k+1}}{(k+1)!},$$

and thus

$$\frac{e^x}{x^k} \geq \frac{x^{k+1}}{x^k(k+1)!} = \frac{x}{(k+1)!} > \frac{K(k+1)!}{(k+1)!} = K.$$

❑

From this theorem, many further statements about the limit behavior of exp and ln can be derived, which we state here without proof:

- For all $k \in \mathbb{N}$ it holds that

$$\lim_{x\to\infty} x^k e^{-x} = 0 \quad \text{and} \quad \lim_{\substack{x\to 0\\ x>0}} x^k e^{1/x} = \infty.$$

- For all $a \in \mathbb{R}$ with $|a| < 1$ it holds that

$$\lim_{k\to\infty} k a^k = 0.$$

- It holds that

(15.1)
$$\lim_{\substack{x\to 0\\ x\neq 0}} \frac{e^x - 1}{x} = 1.$$

- It holds that

$$\lim_{x\to\infty} \ln x = \infty \quad \text{and} \quad \lim_{\substack{x\to 0\\ x>0}} \ln x = -\infty.$$

- For every real number $y > 0$ it holds that

$$\lim_{\substack{x\to 0\\ x>0}} x^y = 0 \quad \text{and} \quad \lim_{\substack{x\to 0\\ x>0}} x^{-y} = \infty.$$

- For all $y > 0$ it holds that

$$\lim_{\substack{x\to\infty\\ x>0}} \frac{\ln x}{x^y} = 0 \quad \text{and} \quad \lim_{\substack{x\to 0\\ x>0}} x^y \ln x = 0.$$

15.F The Complex Exponential Function.

The definition of the exponential function can be directly extended to complex arguments.

15.18. **Definition.** We define the exponential function $\exp : \mathbb{C} \to \mathbb{C}$ as

$$\exp(z) := \sum_{n=0}^{\infty} \frac{z^n}{n!}.$$

DEF
Complex exponential function

♢

Analogous to the proofs in the real case, one shows that

$$\left| \exp(z) - \sum_{n=0}^{N} \frac{z^n}{n!} \right| \leq 2 \frac{|z|^{N+1}}{(N+1)!} \quad \text{for } |z| \leq 1 + \frac{N}{2}$$

and the functional equation

$$\exp(z_1 + z_2) = \exp(z_1) \exp(z_2).$$

From this it follows in particular that

$$\exp(z) \exp(-z) = \exp(z - z) = \exp(0) = 1,$$

and thus $\exp(z) \neq 0$ for all $z \in \mathbb{C}$. Moreover, it follows that

$$\exp(z) = \exp(a + ib) = \exp(a) \exp(ib).$$

We can therefore compute the complex exponential function by computing the exponential function for real and purely imaginary numbers and then multiplying them.

As in $\mathbb{R}$, one can also prove that the exponential function is continuous. In the following, as in the real case, we will again use the notation e^z as an alternative to $\exp(z)$ for $z \in \mathbb{C}$.

An important property of the complex exponential function, for which there is no real counterpart, is shown in the following theorem.

15.19. **Theorem.** *For all $z \in \mathbb{C}$ it holds that $e^{\bar{z}} = \overline{e^z}$.*

THM
conjugate complex exponential

Proof. First note that for all $z_1, z_2 \in \mathbb{C}$ the equation $\bar{z}_1 \bar{z}_2 = \overline{z_1 z_2}$ holds. By induction it follows that $\bar{z}^k = \overline{z^k}$ for all $z \in \mathbb{C}$ and all $k \geq 0$. Writing

$$s_n(z) := \sum_{k=0}^{n} \frac{z^k}{k!},$$

we have

$$s_n(\bar{z}) = \sum_{k=0}^{n} \frac{\bar{z}^k}{k!} = \sum_{k=0}^{n} \overline{\left(\frac{z^k}{k!} \right)} = \overline{\sum_{k=0}^{n} \frac{z^k}{k!}} = \overline{s_n(z)}.$$

The claim then follows from

$$e^{\bar{z}} = \lim_{n \to \infty} s_n(\bar{z}) = \lim_{n \to \infty} \overline{s_n(z)} = \overline{\lim_{n \to \infty} s_n(z)} = \overline{e^z},$$

where in the penultimate equality we used the fact that complex conjugation commutes with limits. ❑

15.20. **Corollary.** *For every purely imaginary number $z = ib$ with $b \in \mathbb{R}$ it holds that $|e^z| = 1$.* **COR** $|e^{ib}| = 1$

Proof. We have

$$|e^z|^2 = e^z\overline{e^z} = e^z e^{\bar{z}} = e^z e^{-z} = 1.$$

The claim follows because $\sqrt{1} = \pm 1$ and the modulus of a complex number is always nonnegative. ❑

In the complex plane, all complex numbers of the form e^{ib} lie on the circle of radius 1. An example is shown in Fig. 7.

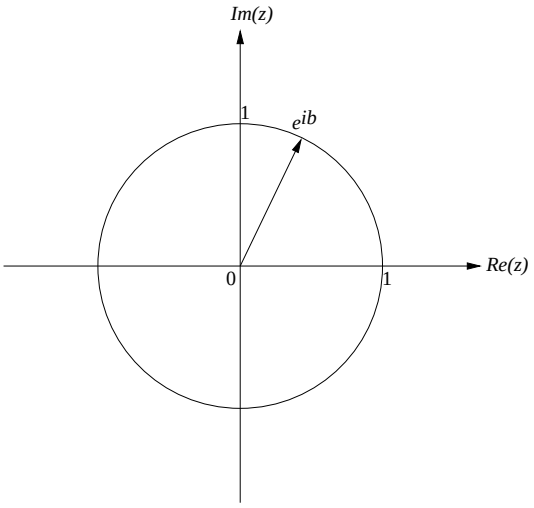


FIGURE 7. Visualization of e^{ib}

An important question here is how the position of e^{ib} on the circle depends on the number $b \in \mathbb{R}$. The number b is precisely a measure of the angle between the real axis and the vector e^{ib}, measured counterclockwise. In the next chapter we will define the number π and see that $e^{i\pi/2} = i$, $e^{i\pi} = -1$, $e^{i3\pi/2} = -i$ and $e^{i2\pi} = e^{i0} = 1$. For these values, b is exactly the length of the arc from the point $1 = 1 + i0$ to the vector e^{ib}, again measured counterclockwise. Consequently, the number b for 0, $\pi/2$, π, etc., is nothing other than the angle measured in radians, which we will use for angle measurement from now on. In other words, it is precisely the *argument* α introduced in the discussion after Definition 6.2. In the degree measure familiar from school, the correspondences are $\pi/2 = 90°$, $\pi = 180°$, $3/2\pi = 270°$ and $2\pi = 360°$. In fact, the relationship between b and the arc length holds not only for these values but for all $b \in [0, 2\pi)$, although we will not prove this in this lecture.

From this visualization it follows that every complex number $a+ib$ can be written as $re^{i\varphi}$ with $r \geq 0$ and $\varphi \in [0, 2\pi)$. Consequently, every pair of real numbers (a, b) can be written as $(r\cos\varphi, r\sin\varphi)$. This representation of real number pairs is called *polar coordinates*.

15.G Sine and Cosine.

From school, it is known that the sine is the ratio of the opposite side to the hypotenuse in a right triangle, and the cosine is the ratio of the adjacent side to the hypotenuse. In the special case where the hypotenuse has length 1, the sine is just the length of the opposite side and the cosine the length of the adjacent side.

Now consider Figure 7 again. We can draw a right triangle whose hypotenuse is the vector e^{ib}. This is done in Fig. 8, where we now denote the real number b by x. Since e^{ix} rotates on the circle in a unique direction (counterclockwise) as x increases, we will use x as a measure of the angle between e^{ix} and the real axis.

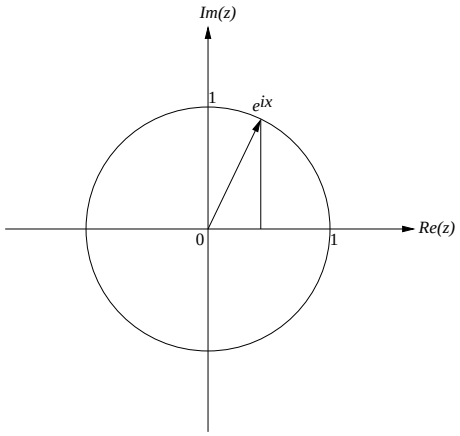


FIGURE 8. Right triangle with hypotenuse e^{ix}

Obviously, the length of the hypotenuse here is exactly $\exp(ix) = 1$, so sine and cosine are given by the lengths of the corresponding legs, see Fig. 9.

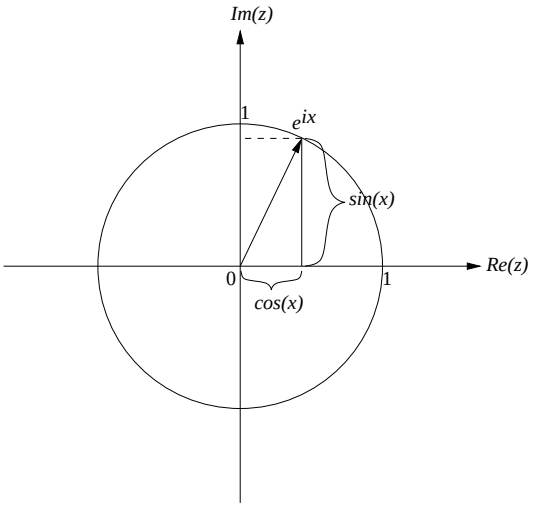


FIGURE 9. Sine and cosine in the right triangle from Fig. 8

The values for sine and cosine are therefore nothing other than the coordinates of the vector $\exp(ix)$. These, in turn, are just the real part and the imaginary part of this function. This observation leads to the following definition.

15.21. **Definition.** For every $x \in \mathbb{R}$ we define

$$\sin x := \operatorname{Im}(e^{ix}) \quad \text{and} \quad \cos x := \operatorname{Re}(e^{ix}).$$

DEF
Sine and Cosine

♢

An immediate consequence of this definition is *Euler's formula*

$$e^{ix} = \cos x + i \sin x.$$

This is named after the Swiss mathematician Leonhard Euler (1707–1783), who

L. Euler
1707–1783

systematically studied the relationship between the exponential function and sine and cosine.

From this definition and the properties of the exponential function, various properties of sine and cosine can be derived:

- $\cos x = \frac{1}{2}(e^{ix} + e^{-ix}), \quad \sin x = \frac{1}{2i}(e^{ix} - e^{-ix})$
- $\cos(-x) = \cos x, \quad \sin(-x) = -\sin x$
- $\cos^2 x + \sin^2 x = 1$
- $\cos : \mathbb{R} \to \mathbb{R}$ and $\sin : \mathbb{R} \to \mathbb{R}$ are continuous
- For all $x, y \in \mathbb{R}$ the addition theorems hold:

$$\cos(x+y) = \cos x \cos y - \sin x \sin y \tag{15.2}$$

$$\sin(x+y) = \sin x \cos y + \cos x \sin y. \tag{15.3}$$

- For all $x, x' \in \mathbb{R}$ it holds that

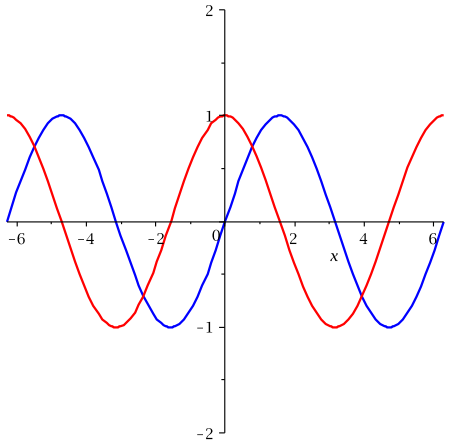
$$\cos x' - \cos x = -2 \sin\left(\frac{x'+x}{2}\right) \sin\left(\frac{x'-x}{2}\right) \tag{15.4}$$

- For all $x \in \mathbb{R}$ it holds that

$$\cos x = \sum_{k=0}^{\infty} (-1)^k \frac{x^{2k}}{(2k)!} = 1 - \frac{x^2}{2!} + \frac{x^4}{4!} - + \dots \tag{15.5}$$

$$\sin x = \sum_{k=0}^{\infty} (-1)^k \frac{x^{2k+1}}{(2k+1)!} = x - \frac{x^3}{3!} + \frac{x^5}{5!} - + \dots \tag{15.6}$$

  These series converge absolutely.
- Using these series, one can approximate the values of $\sin x$ and $\cos x$ and thus, in particular, plot the graphs of these functions.



Graphs of sin (blue) and cos (red)

- It holds that

$$\lim_{\substack{x \to 0 \\ x \neq 0}} \frac{\sin x}{x} = 1. \tag{15.7}$$

One can also use the series representation to prove that the function cos has exactly one zero x^* in the interval $[0, 2]$. We use this to define the number

$$\pi := x^*.$$

If x^* is determined, for example, by interval bisection as in the proof of the Intermediate Value Theorem, one obtains the well-known value $\pi \approx 3.1415926\ldots$.

We can now compute various values of the complex exponential function exactly.

THM values of complex exponential

15.22. **Theorem.** *It holds that*

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

Proof. Since $\sin^2 x + \cos^2 x = 1$ and $\cos\frac{\pi}{2} = 0$, it follows that

$$\sin^2\frac{\pi}{2} = 1 - \cos^2\frac{\pi}{2} = 1.$$

Thus, $\sin\frac{\pi}{2}$ can take the value ± 1, and using the series representation one checks that

$$\sin\frac{\pi}{2} = 1$$

is the correct solution. Therefore,

$$e^{\frac{1}{2}i\pi} = \cos\frac{\pi}{2} + i\sin\frac{\pi}{2} = i.$$

The remaining values then follow for $n = 2, 3, 4$ from the equation

$$e^{n\frac{1}{2}i\pi} = \left(e^{\frac{1}{2}i\pi}\right)^n = i^n.$$

❑

From school, it is known that the circumference of the unit circle is exactly 2π. If we plot e^{ix} for $x = \pi/2$, π, $3\pi/2$, and 2π on the unit circle in the complex plane, we see that the argument x of the exponential function is precisely the length of the arc obtained when moving from the point $1 + i0$ counterclockwise to e^{ix}. This is no coincidence but exactly the reason why we defined π as we did. In fact, this holds not only for the above values of x but for all $x \in [0, 2\pi)$, which can be verified by evaluating e^{ix} for sample values, though we will not prove this here.

From the values of the exponential function, we can construct the following value table for sin and cos:

x	0	$\frac{\pi}{2}$	π	$\frac{3\pi}{2}$	2π
$\sin x$	0	1	0	-1	0
$\cos x$	1	0	-1	0	1

From the addition theorems (15.2) and (15.3) it follows that

$$\cos(x + 2\pi) = \cos x \underbrace{\cos 2\pi}_{=1} - \sin x \underbrace{\sin 2\pi}_{=0} = \cos x \tag{15.8}$$

and similarly

$$\sin(x + 2\pi) = \sin x \tag{15.9}$$

$$\cos(x + \pi) = -\cos x, \qquad \sin(x + \pi) = -\sin x \tag{15.10}$$

as well as

$$\cos x = \sin\left(\frac{\pi}{2} - x\right), \qquad \sin x = \cos\left(\frac{\pi}{2} - x\right). \tag{15.11}$$

Equations (15.8) and (15.9) show that the values of sin and cos repeat when x increases by 2π. We say that sine and cosine are *periodic* with *period* 2π. Note that from these equations, by induction,

$$\cos(x + 2k\pi) = \cos x \quad \text{and} \quad \sin(x + 2k\pi) = \sin x \quad \text{for all } k \in \mathbb{Z} \tag{15.12}$$

follows. Due to periodicity, the sine function has the following zeros:

$$\{k\pi \mid k \in \mathbb{Z}\} = \{\ldots, -3\pi, -2\pi, -\pi, 0, \pi, 2\pi, \ldots\}$$

and the zeros of the cosine are given by the set

$$\{\pi/2 + k\pi \mid k \in \mathbb{Z}\} = \left\{\ldots, -\frac{5\pi}{2}, -\frac{3\pi}{2}, -\frac{\pi}{2}, \frac{\pi}{2}, \frac{3\pi}{2}, \frac{5\pi}{2}, \ldots\right\}.$$

It follows immediately that $e^{ix} = 1$ holds if and only if $x = 2k\pi$ for some $k \in \mathbb{Z}$.

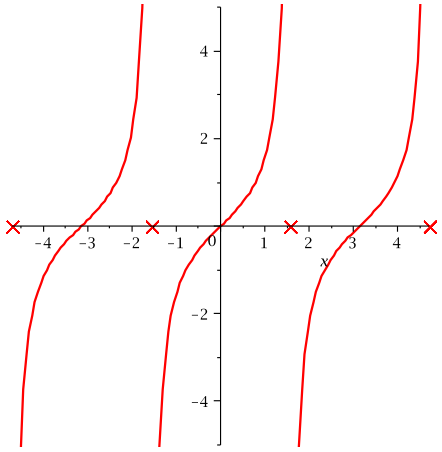
15.23. **Definition.** For $x \in \mathbb{R} \setminus \left\{\frac{\pi}{2} + k\pi\right\}$ we define the *tangent* as

$$\tan x := \frac{\sin x}{\cos x}.$$

DEF Tangent

♢

The tangent is defined for all $x \in \mathbb{R}$ with $\cos x \neq 0$.



Graph of $\tan x$. The discontinuities are marked with crosses.

In certain intervals, sin, cos, and tan are strictly monotonic:

- cos is strictly decreasing on $[0, \pi]$ with $\cos([0, \pi]) = [-1, 1]$.
- sin is strictly increasing on $[-\pi/2, \pi/2]$ with $\sin([-\pi/2, \pi/2]) = [-1, 1]$.
- tan is strictly increasing on $(-\pi/2, \pi/2)$ with $\tan((-\pi/2, \pi/2)) = \mathbb{R}$.

Consequently, the functions are bijective on these restricted domains and codomains, and the inverse functions exist:

$\arccos : [-1, 1] \to [0, \pi]$, $\quad \arcsin : [-1, 1] \to [-\pi/2, \pi/2]$, $\quad \arctan : \mathbb{R} \to (-\pi/2, \pi/2)$,

(pronounced: arc cosine, arc sine, and arc tangent).

16. Differentiation

Date:
March 5, 2026

In this chapter, we introduce the concept of the derivative of a function and discuss the most important calculation rules.

16.A Definition and Examples.

Intuitively, the derivative of a function $f : D \to \mathbb{R}$ at a point $x \in D$ is a real number (denoted by $f'(x) \in \mathbb{R}$) that represents the slope of the function graph at the point $(x, f(x))$. One can derive the definition of the derivative directly from this intuition: Suppose we want to determine the slope of the function graph at a point. We can approximate it by choosing a small real number $h \neq 0$ and considering the slope of the line through the points $(x, f(x))$ and $(x+h, f(x+h))$ (this line is called the *secant*). The smaller $|h|$ is, the more accurately the slope of the secant matches the slope of the function at the point x, see Figure 10.

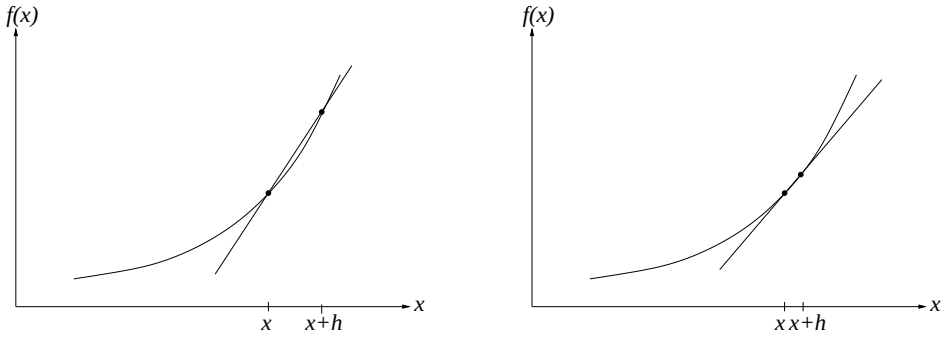


Figure 10. Visualization of the derivative via the slope

Letting $h \to 0$, the slope of the secant converges to the slope of the function graph. Since the slope of the secant is given by the quotient

$$\frac{f(x+h) - f(x)}{h},$$

the so-called *difference quotient*, this leads to the following definition.

16.1. **Definition.** A function $f : D \to \mathbb{R}$ with $D \subset \mathbb{R}$ is called *differentiable* at a point $x \in D$ if the limit **DEF** differentiable

$$f'(x) := \lim_{\substack{h \to 0 \\ h \neq 0, x+h \in D}} \frac{f(x+h) - f(x)}{h}$$

exists. In particular, we assume that at least one sequence $h_n \to 0$ with $h_n \neq 0$ and $x + h_n \in D$ exists.

The value $f'(x) \in \mathbb{R}$ is then called the *derivative* (or also the *differential*) of f at x. ♢

Alternatively, the derivative can also be defined via the limit

$$\lim_{\substack{y \to x \\ y \neq x, y \in D}} \frac{f(y) - f(x)}{y - x},$$

since the two definitions can easily be rewritten into each other using $y = x + h$ or $h = y - x$.

EXAMPLES
Differentiable functions and their derivatives

16.2. **Examples.** In the following examples, we omit the conditions $h \neq 0, x + h \in D$ under the lim to simplify notation. They are, however, always required.

(a) Constant function $f(x) = c$:

$$f'(x) = \lim_{h\to 0} \frac{f(x+h) - f(x)}{h} = \lim_{h\to 0} \frac{c - c}{h} = \lim_{h\to 0} 0 = 0$$

for all $x \in \mathbb{R}$.

(b) Affine linear function $f(x) = a + bx$:

$$f'(x) = \lim_{h\to 0} \frac{f(x+h) - f(x)}{h} = \lim_{h\to 0} \frac{a + b(x+h) - a - bx}{h} = \lim_{h\to 0} \frac{bh}{h} = \lim_{h\to 0} b = b$$

for all $x \in \mathbb{R}$. In particular, for the identity function $f(x) = x$ (i.e., $a = 0$ and $b = 1$) we have $f'(x) = 1$.

(c) Parabola $f(x) = x^2$:

$$\begin{aligned} f'(x) &= \lim_{h\to 0} \frac{f(x+h) - f(x)}{h} = \lim_{h\to 0} \frac{(x+h)^2 - x^2}{h} \\ &= \lim_{h\to 0} \frac{2hx + h^2}{h} = \lim_{h\to 0} (2x + h) = 2x \end{aligned}$$

for all $x \in \mathbb{R}$.

(d) Hyperbola $f(x) = 1/x$ with $D = \mathbb{R} \setminus \{0\}$:

$$\begin{aligned} f'(x) &= \lim_{h\to 0} \frac{f(x+h) - f(x)}{h} = \lim_{h\to 0} \frac{\frac{1}{x+h} - \frac{1}{x}}{h} \\ &= \lim_{h\to 0} \frac{\frac{x-x-h}{(x+h)x}}{h} = \lim_{h\to 0} \frac{-h}{hx^2 + h^2 x} = \lim_{h\to 0} \frac{-1}{x^2 + hx} = -\frac{1}{x^2} \end{aligned}$$

for all $x \in D$.

(e) Exponential function $f(x) = e^x$:

$$f'(x) = \lim_{h\to 0} \frac{e^{x+h} - e^x}{h} = \lim_{h\to 0} \frac{e^x e^h - e^x}{h} = \lim_{h\to 0} e^x \frac{e^h - 1}{h} = e^x \lim_{h\to 0} \frac{e^h - 1}{h} = e^x$$

for all $x \in \mathbb{R}$, where in the last step we used (15.1). The exponential function is therefore equal to its own derivative.

(f) Cosine $f(x) = \cos x$:
Using (15.4), i.e.,

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

with $x' = x + h$, we get

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

By continuity of sine, it follows that

$$\lim_{h\to 0} \sin\left(x + \frac{h}{2}\right) = \sin x$$

and from (15.7) we have

$$\lim_{h\to 0} \frac{\sin\frac{h}{2}}{\frac{h}{2}} = 1.$$

Thus, by the product rule for limits,

$$f'(x) = \lim_{h\to 0} \frac{\cos(x+h) - \cos x}{h} = -\lim_{h\to 0} \sin\left(x + \frac{h}{2}\right) \lim_{h\to 0} \frac{\sin\frac{h}{2}}{\frac{h}{2}} = -\sin x.$$

(g) Sine $f(x) = \sin x$:
With a similar calculation as for cosine, it follows that

$$f'(x) = \cos x.$$

(h) ReLU $f(x) = (x)_+ = \max\{0, x\}$:
Claim: this function is not differentiable at $x = 0$.

Proof. For sequences $h_n \to 0$ with $h_n > 0$ we have

$$\lim_{n\to\infty} \frac{(0+h_n)_+ - (0)_+}{h_n} = \lim_{n\to\infty} \frac{h_n}{h_n} = 1.$$

For sequences $h_n \to 0$ with $h_n < 0$ we have

$$\lim_{n\to\infty} \frac{(0+h_n)_+ - (0)_+}{h_n} = \lim_{n\to\infty} \frac{0}{h_n} = 0.$$

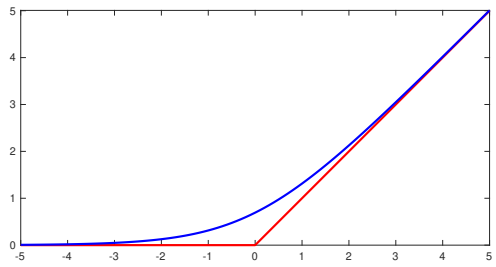
Thus, the limit

$$\lim_{h\to 0} \frac{(0+h)_+ - (0)_+}{h}$$

does not exist. ❑

Looking at the graph of the ReLU function, one sees that it has a kink at $x = 0$. Intuitively, differentiability at a point x is nothing other than the property that the graph has no kink at x. Instead of "differentiable", the term "smooth" is often used. Due to the fact that the non-differentiability of the ReLU function may be undesirable, it is sometimes approximated by the softplus function

$$f(x) = \ln(1 + e^x)$$



Graphs of the ReLU function $f(x) = (x)_+$ (red)
and the softplus function $f(x) = \ln(1 + e^x)$ (blue)

♣

EXAMPLES
Derivatives in applications

16.3. **Examples.** The concept of the derivative has many interpretations in different application areas of mathematics. We give two examples here.

(a) In *optimization*, one wants to find the smallest possible value of a function f. If we are at a point x, it is important for programming an algorithm to know in which direction to search for a smaller value than $f(x)$. If the derivative $f'(x)$ exists and $f'(x) > 0$, then $f(x+h) > f(x)$ and $f(x-h) < f(x)$ for small $h > 0$. So it makes sense to search ?to the left? of x for arguments with smaller values of f. In the case $f'(x) < 0$, it is exactly the opposite: smaller values lie to the right of x. The derivative therefore gives us information about the direction in which an algorithm should search for x with smaller values of $f(x)$.

(b) In *radioactive decay* the decrease in the number of radioactive particles in a small time interval is approximately proportional to the number of particles currently present, and this relationship becomes more accurate as the time interval considered becomes smaller. Formally, this means the following: Let $f(x)$ denote the mass of the particles at time x, and let $\lambda > 0$ denote the proportionality constant of the decrease. Then

$$\frac{f(x+h) - f(x)}{h} = -\lambda f(x) + r(h), \tag{16.1}$$

where the "error term" $r(h)$ converges to zero as $h \to 0$. For $h \to 0$ we then obtain

$$f'(x) = -\lambda f(x). \tag{16.2}$$

We will later see in Example 16.14(b) which function satisfies this equation.

♣

An equivalent description of differentiability is given by the following theorem.

THM
Alternative characterization of differentiability

16.4. **Theorem.** *Let $D \subset \mathbb{R}$ and $x \in D$ be a point such that there exists at least one sequence $h_n \to 0$, $h_n \neq 0$ with $x + h_n \in D$. Then a function $f : D \to \mathbb{R}$ is differentiable at x if and only if there exists an $a \in \mathbb{R}$ and a function $\varphi : \mathbb{R} \to \mathbb{R}$ with*

$$\lim_{h \to 0} \frac{\varphi(h)}{h} = 0$$

such that

$$f(y) = f(x) + a(y - x) + \varphi(y - x)$$

holds for all $y \in D$. In this case, $f'(x) = a$.

Proof. Suppose f is differentiable at x. Set $a := f'(x)$ and define $\varphi(h) := f(x + h) - f(x) - ha$ for all $h \in \mathbb{R}$ with $x + h \in D$ and $\varphi(h) = 0$ otherwise. Then, with $h = y - x$ for all $y \in D$, we have the desired equation

$$f(x) + a(y - x) + \varphi(y - x) = f(x) + ha + \varphi(h) = f(x + h) = f(y).$$

Moreover,

$$\lim_{h \to 0} \frac{\varphi(h)}{h} = \lim_{h \to 0} \frac{f(x+h) - f(x) - ha}{h} = \lim_{h \to 0} \frac{f(x+h) - f(x)}{h} - a = f'(x) - f'(x) = 0,$$

which is the desired property.

Conversely, suppose the equation $f(y) = f(x) + a(y-x) + \varphi(y-x)$ holds with φ as stated in the theorem. Then, with $y = x+h$, i.e., $h = y-x$,

$$\lim_{h\to 0} \frac{f(x+h)-f(x)}{h} = \lim_{h\to 0} \frac{ah+\varphi(h)}{h} = a + \lim_{h\to 0} \frac{\varphi(h)}{h} = a.$$

Thus, f is differentiable at x with $f'(x) = a$. ❑

REMARK
Landau symbol

16.5. **Remark.** (i) The condition

$$\lim_{h\to 0} \frac{\varphi(h)}{h} = 0$$

is often written as

$$\varphi(h) = o(h).$$

The term $o(h)$ is called the *Landau symbol* . Intuitively, this condition means that $\varphi(h)$ converges to zero so much faster than h itself as $h \to 0$ that the quotient still converges to zero. Examples of such functions are, for example, $\varphi(h) = h^2$ (since $\varphi(h)/h = h \to 0$ as $h \to 0$) or $\varphi(h) = h\sqrt{h}$ (since $\varphi(h)/h = \sqrt{h} \to 0$ as $h \to 0$).

(ii) In informal notation, the statement of Theorem 16.4 can be written as

$$f(y) \approx f(x) + f'(x)(y-x),$$

where the symbol ?≈? (read: approximately equal) means that the missing terms are much smaller in absolute value than $|x-y|$ when $|x-y|$ is sufficiently small. Since the missing terms are precisely $\varphi(y-x)$, this follows from the convergence property of φ.

(iii) The approximating function $g(y) := f(x) + f'(x)(y-x)$ from (ii) is geometrically nothing other than the tangent to the graph of f at the point $(x, f(x))$. The function φ thus describes the distance between the tangent and the function itself. The condition $\varphi(h)/h \to 0$ therefore says geometrically that the tangent "gently" touches the graph, i.e., the angle between the tangent and the graph is zero. ♠

EXAMPLES

16.6. **Examples.** (a) For the parabola $f(x) = x^2$ we have $a = f'(x) = 2x$ and thus

$$\varphi(h) = f(x+h) - f(x) - ha = (x+h)^2 - x^2 - 2hx = x^2 + 2hx + h^2 - x^2 - 2hx = h^2.$$

This function satisfies, by Remark 16.5(i), $\lim_{h\to 0} \varphi(h)/h = 0$.

The tangent at the point $(x, f(x))$ is then given by $g(y) = x^2 + 2x(y-x) = -x^2 + 2xy$.

(b) Let f be a function that satisfies the radioactive decay law $f'(x) = -\lambda f(x)$ for some $x \in \mathbb{R}$ (and which is then, of course, differentiable at x). Then, with $y = x+h$, we have

$$f(x+h) = f(x) - \lambda f(x)h + \varphi(h)$$

or, after dividing by h,

$$\frac{f(x+h)-f(x)}{h} = -\lambda f(x) + \frac{\varphi(h)}{h}.$$

This means that the function satisfies equation (16.2) with $r(h) = \varphi(h)/h$, for which $r(h) \to 0$ as $h \to 0$.

Thus, not only can we deduce equation (16.2) from equation (16.1) as in Example 16.3(b), but with Theorem 16.4 we can also deduce equation (16.1) from equation (16.2). ♣

16.B Rules and Properties of Differentiation.

THM Differentiability implies continuity

16.7. **Theorem.** *If a function $f : D \to \mathbb{R}$ is differentiable at a point $x \in D$, then f is also continuous at x.*

Proof. We need to show that $\lim_{y\to x} f(y) = f(x)$. Using the notation from Theorem 16.4, we have $\lim_{y\to x} a(y-x) = 0$ and

$$\lim_{y\to x} \varphi(y-x) = \lim_{y\to x} \frac{\varphi(y-x)}{y-x}(y-x) = \lim_{y\to x} \frac{\varphi(y-x)}{y-x} \lim_{y\to x}(y-x) = 0 \cdot 0 = 0.$$

Thus, the claim follows because

$$\lim_{y\to x} f(y) = \lim_{y\to x} f(x)+a(y-x)+\varphi(y-x) = f(x)+\lim_{y\to x} a(y-x)+\lim_{y\to x} \varphi(y-x) = f(x).$$

❑

THM Differentiability of combined functions

16.8. **Theorem.** *Let $f, g : D \to \mathbb{R}$ be differentiable at $x \in D$ and $\lambda \in \mathbb{R}$. Then the functions $f+g$, λf, and $fg : D \to \mathbb{R}$ are also differentiable at x, and we have*

$$\begin{aligned} (f+g)'(x) &= f'(x)+g'(x) \\ (\lambda f)'(x) &= \lambda f'(x) \\ (fg)'(x) &= f'(x)g(x)+f(x)g'(x) \quad \textit{(Product Rule).} \end{aligned}$$

If $g(y) \neq 0$ for all $y \in D$, then $\frac{f}{g} : D \to \mathbb{R}$ is also differentiable at x, and we have

$$\left(\frac{f}{g}\right)'(x) = \frac{f'(x)g(x)-f(x)g'(x)}{g(x)^2} \quad \textit{(Quotient Rule).}$$

Proof. The first two statements follow from the equalities

$$\frac{(f+g)(x+h)-(f+g)(x)}{h} = \frac{f(x+h)-f(x)}{h} + \frac{g(x+h)-g(x)}{h}$$

and

$$\frac{(\lambda f)(x+h)-(\lambda f)(x)}{h} = \lambda \frac{f(x+h)-f(x)}{h}$$

and the rules for limits.

The product rule follows from

$$\begin{aligned} (fg)'(x) &= \lim_{h\to 0} \frac{f(x+h)g(x+h)-f(x)g(x)}{h} \\ &= \lim_{h\to 0} \frac{1}{h}\Big(f(x+h)(g(x+h)-g(x))+(f(x+h)-f(x))g(x)\Big) \\ &= \lim_{h\to 0} f(x+h)\frac{g(x+h)-g(x)}{h} + \lim_{h\to 0} \frac{f(x+h)-f(x)}{h} g(x) \\ &= f(x)g'(x)+f'(x)g(x), \end{aligned}$$

where in the last step we used the continuity of f at x.

The quotient rule is first proven for the constant function $f(x) := 1$ for all $x \in D$. Then

$$\begin{aligned} \left(\frac{1}{g}\right)'(x) &= \lim_{h\to 0} \frac{1}{h}\left(\frac{1}{g(x+h)} - \frac{1}{g(x)}\right) \\ &= \lim_{h\to 0} \frac{1}{g(x+h)g(x)}\left(\frac{g(x)-g(x+h)}{h}\right) = \frac{-g'(x)}{g(x)^2}, \end{aligned}$$

where we used the continuity of g at x and $g(x+h) \neq 0$ (note that we always assume $x+h \in D$ when forming the limit, even if not written explicitly).

The general case now follows from the product rule with

$$\left(\frac{f}{g}\right)'(x) = \left(f \cdot \frac{1}{g}\right)'(x) = f'(x)\frac{1}{g(x)} + f(x)\left(\frac{1}{g}\right)'(x)$$
$$= \frac{f'(x)}{g(x)} - \frac{f(x)g'(x)}{g(x)^2} = \frac{f'(x)g(x) - f(x)g'(x)}{g(x)^2}.$$

❑

EXAMPLES
Derivatives

16.9. **Examples.**

(a) $f : \mathbb{R} \to \mathbb{R}$, $f(x) = x^n$, $n \in \mathbb{N}$. Claim: $f'(x) = nx^{n-1}$

Proof. By induction on n: For $n = 0$, the claim follows from Example 16.2(a). For $n \to n+1$, for $f(x) = x^{n+1} = x^n x$, the product rule, the induction hypothesis, and Example 16.2(b) (with $a = 0$ and $b = 1$) give

$$f'(x) = (x^n x)' = \underbrace{(x^n)'}_{=nx^{n-1}} x + x^n \underbrace{(x)'}_{=1} = nx^{n-1}x + x^n = (n+1)x^n.$$

❑

(b) $f : \mathbb{R} \setminus \{0\} \to \mathbb{R}$, $f(x) = x^{-n} = 1/x^n$, $n \in \mathbb{N}$.
Set $g(x) := 1/f(x) = x^n$. Then, by the quotient rule,

$$f'(x) = \left(\frac{1}{g}\right)'(x) = \frac{-g'(x)}{g(x)^2} = \frac{-nx^{n-1}}{x^{2n}} = -n\frac{1}{x^{n+1}} = -nx^{-n-1}.$$

Combining (a) and (b), we obtain for all $k \in \mathbb{Z}$ for $f(x) = x^k$ (with $x \neq 0$ if $k < 0$) the equation $f'(x) = kx^{k-1}$. For $k \geq 0$, this follows immediately from (a) with $n = k$, and for $k < 0$ we can apply (b) with $n = -k$ and obtain $f'(x) = -nx^{-n-1} = kx^{k-1}$.

(c) For $f(x) = \tan x = \sin x / \cos x$, the quotient rule gives

$$f'(x) = \frac{\sin'(x)\cos(x) - \sin(x)\cos'(x)}{\cos^2(x)} = \frac{\cos^2(x) + \sin^2(x)}{\cos^2(x)} = \frac{1}{\cos^2(x)}.$$

(d) For $f(x) = a\exp(x)$, for all $a \in \mathbb{R}$, we have $f'(x) = a\exp'(x) = a\exp(x)$ by Example 16.2(e) and Theorem 16.8 with $\lambda = a$. Thus, for the function $f(x) = a\exp(x)$, the derivative equals the function itself.

♣

THM
Derivative of the inverse function

16.10. **Theorem.** *Let $D \subset \mathbb{R}$ be a closed interval, $f : D \to \mathbb{R}$ a continuous and strictly monotonic function, and $f^{-1} : D' \to \mathbb{R}$ its inverse defined on $D' = f(D)$ (which exists and is continuous by Theorem 15.10). Then:*

If f is differentiable at a point $y \in D$ with $f'(y) \neq 0$, then f^{-1} is differentiable at the point $x = f(y)$, and

$$\left(f^{-1}\right)'(x) = \frac{1}{f'(y)} = \frac{1}{f'(f^{-1}(x))}.$$

Proof. Let h_n be any sequence with $h_n \to 0$ such that for $x_n := x + h_n$ the conditions $x_n \in D'$ and $x_n \neq x$ hold. Set $y_n := f^{-1}(x_n)$. Then, by the continuity of f^{-1}, we have $y_n = f^{-1}(x_n) \to f^{-1}(x) = y$. Since $f^{-1}(D') = D$, it follows that $y_n \in D$, and because f^{-1} is strictly monotonic and thus injective, $x_n \neq x$ implies $y_n \neq y$. Thus, we can write

$$\lim_{n\to\infty} \frac{f^{-1}(x+h_n) - f^{-1}(x)}{h_n} = \lim_{n\to\infty} \frac{f^{-1}(x_n) - f^{-1}(x)}{x_n - x} = \lim_{n\to\infty} \frac{y_n - y}{f(y_n) - f(y)} = \frac{1}{f'(y)},$$

where in the last step we used the quotient rule from Theorem 8.13, which we may apply because $f'(y) \neq 0$. ❑

16.11. **Example.** For $\ln(x) = \exp^{-1}(x)$ we have

EXAMPLE
Derivative of logarithm

$$\ln'(x) = \frac{1}{\exp'(\ln x)} = \frac{1}{\exp(\ln x)} = \frac{1}{x}.$$

♣

We can now give the **proof of Theorem 15.3**, i.e., the proof of the equation

$$\exp(x) = \lim_{n\to\infty} \left(1 + \frac{x}{n}\right)^n$$

for all $x \in \mathbb{R}$.

Proof. For $x = 0$, the claim follows immediately since $\exp(0) = 1$. For $x \neq 0$, since $\ln(1) = 0$, we have

$$n \ln\left(1 + \frac{x}{n}\right) = x \frac{\ln\left(1 + \frac{x}{n}\right) - \ln(1)}{\frac{x}{n}}$$

and thus, since $\ln'(1) = 1/1 = 1$,

$$\lim_{n\to\infty} n \ln\left(1 + \frac{x}{n}\right) = x \lim_{n\to\infty} \frac{\ln\left(1 + \frac{x}{n}\right) - \ln(1)}{\frac{x}{n}} = x \ln'(1) = x.$$

Since $a^b = \exp(b \ln(a))$, it follows that

$$\left(1 + \frac{x}{n}\right)^n = \exp\left(n \ln\left(1 + \frac{x}{n}\right)\right),$$

and therefore, by the continuity of exp,

$$\lim_{n\to\infty} \left(1 + \frac{x}{n}\right)^n = \exp\left(\lim_{n\to\infty} n \ln\left(1 + \frac{x}{n}\right)\right) = \exp(x).$$

❑

Further examples of applying Theorem 16.10 are the following functions.

16.12. **Example.** $\arcsin : [-1, 1] \to \mathbb{R}$. For $x \in (-1, 1)$, $y = \arcsin x \in (-\pi/2, \pi/2)$ and thus $\sin' y = \cos y \neq 0$. Therefore, by Theorem 16.10, for $x \in (-1, 1)$,

EXAMPLE
derivative of arcsin and arctan

$$\arcsin'(x) = \frac{1}{\sin'(\arcsin x)} = \frac{1}{\cos(\arcsin x)}.$$

Using the identity $\cos(\arcsin x) = \sqrt{1 - \sin^2(\arcsin x)} = \sqrt{1 - x^2}$, we can simplify the expression to

$$\arcsin'(x) = \frac{1}{\sqrt{1 - x^2}}.$$

Similarly, we compute

$$\arctan'(x) = \cos^2(\arctan x) = \frac{1}{1+x^2}.$$

♣

16.13. **Theorem.** *Let $f : D \to \mathbb{R}$ and $g : E \to \mathbb{R}$ be functions with $f(D) \subset E$. Suppose f is differentiable at $x \in D$ and g is differentiable at $y := f(x) \in E$. Then the composition $g \circ f$ is differentiable at x, and*

$$(g \circ f)'(x) = g'(f(x))f'(x).$$

THM Chain rule

Proof. By Theorem 16.4 (applied with $y = x + h$ and $x = x$ or $y = y + \tilde{h}$ and $x = y$), we have

$$f(x+h) = f(x) + f'(x)h + \varphi_f(h)$$

and

$$g(y+\tilde{h}) = g(y) + g'(y)\tilde{h} + \varphi_g(\tilde{h})$$

with $\varphi_f(h)/h \to 0$ and $\varphi_g(h)/h \to 0$ as $h \to 0$. From the second equation, $\varphi_g(\tilde{h}) = 0$ for $\tilde{h} = 0$, and from the convergence $\varphi_f(h)/h \to 0$ it follows that $\varphi_f(h) = h(\varphi_f(h)/h) \to 0$ as $h \to 0$.

From these equations (with $\tilde{h} = f'(x)h + \varphi_f(h)$), we get

$$\begin{aligned} g \circ f(x+h) &= g(f(x+h)) = g(f(x) + f'(x)h + \varphi_f(h)) \\ &= g \circ f(x) + g'(y)\,(f'(x)h + \varphi_f(h)) + \varphi_g(f'(x)h + \varphi_f(h)) \\ &= g(y) + g'(y)f'(x)h + \underbrace{g'(y)\varphi_f(h) + \varphi_g(\tilde{h})}_{=:\varphi(h)}. \end{aligned}$$

One checks that φ satisfies the condition $\varphi(h)/h \to 0$ from Theorem 16.4. Thus,

$$g \circ f(x+h) = g(y) + g'(y)f'(x)h + \varphi(h) = g \circ f(x) + g'(f(x))f'(x)h + \varphi(h),$$

with $\varphi(h)/h \to 0$, from which the claim follows again by Theorem 16.4 (applied with $y = x + h$). ❑

16.14. **Examples.**

EXAMPLES Examples for chain rule

(a) $h(x) = x^\alpha$ for $\alpha \in \mathbb{R}$. Since $x^\alpha = \exp(\alpha \ln x)$, $h(x)$ can be written as $g \circ f$ with $f(x) = \alpha \ln x$ and $g(x) = \exp(x)$. Then, by the chain rule,

$$h'(x) = g'(f(x))f'(x) = \exp(\alpha \ln x)\alpha\frac{1}{x} = x^\alpha \alpha \frac{1}{x} = \alpha x^{\alpha-1}.$$

The formula already derived in Example 16.9(a) and (b) thus holds not only for integer but also for real exponents.

(b) $h(x) = a\exp(bx)$ with $a, b \in \mathbb{R}$. By the chain rule with $f(x) = bx$ and $g(x) = a\exp(x)$ and Example 16.9(d), we have

$$h'(x) = g'(f(x))f'(x) = a\exp(bx)b = ba\exp(bx),$$

which we can also write as $h'(x) = bh(x)$. In particular, the function $h(x) = a\exp(-\lambda x)$ satisfies the radioactive decay law from Example 16.3(b).

♣

16.15. **Remark.** When we consider a differentiable function $f : D \to \mathbb{R}$ whose domain is a closed interval $D = [a, b]$ or $D = (-\infty, b]$ or $D = [a, \infty)$, then using the definition of the derivative 16.1, for $x = a$ we always obtain the right-hand limit and for $x = b$ we always obtain the left-hand limit. The reason is that for $x = a$ and $h \neq 0$ in D, only values of the form $x + h$ with $h > 0$ are included, and for $x = b$ only values of the form $x + h$ with $h < 0$. The same applies to half-open intervals at the endpoints that belong to the interval.

REMARK Differentiation on closed intervals

A consequence of this fact is that a function that is not differentiable at a point $a \in D$ can become differentiable if we restrict the domain.

As an **example**, consider the ReLU function $f(x) = (x)_+$. As a function on $D = \mathbb{R}$, this is, as we have seen, not differentiable at $x = 0$. However, if we consider the function on the domain $D = [0, \infty)$, then $f(x) = (x)_+ = x$. The function is therefore differentiable for all $x \in D$ and $f'(x) = 1$.

This fact can lead to the false conclusion that the absolute value function is also differentiable at $x = 0$ on $D = \mathbb{R}$, which it is not. The reason is that by restricting to $[0, \infty)$, we have only computed the so-called *right-hand derivative* at $x = 0$.

For open domains of the form $D = (a, b)$ (or also $D = (a, \infty)$, $D = (-\infty, b)$ or $D = (-\infty, \infty) = \mathbb{R}$), this problem does not occur, since D then has no boundary points, i.e., for every $x \in D$ there exist sequences $h_n \to 0$ with $h_n > 0$ and $x + h_n \in D$ as well as sequences $h_n \to 0$ with $h_n < 0$ and $x + h_n \in D$. For this reason, open intervals are often preferred as domains when computing derivatives.

As an **example**, consider again the ReLU function $f(x) = (x)_+$. On the domain $D = (0, \infty)$, as above, $f'(x) = 1$ for all $x \in D$. Since the interval is open, this is now indeed the correct derivative for all $x \in D$. ♠

16.C The Derivative as a Function.

So far, we have considered derivatives only at a fixed point $x \in D$. However, the notation $f'(x)$ for the derivative suggests that the derivative itself can again be regarded as a function. This is indeed the case if we use the following definition.

16.16. **Definition.** A function $f : D \to \mathbb{R}$ is called differentiable (on D) if it is differentiable for all $x \in D$. The derivative can then be regarded as a function

$$f' : D \to \mathbb{R}, \quad f' : x \mapsto f'(x).$$

DEF differentiability on D

♢

If the derivative $f' : D \to \mathbb{R}$ is itself again a function, then it can, of course, also be differentiated again, provided it is differentiable. We can thus define $f^{(2)}(x) := (f')'(x)$ and, if possible, also $f^{(3)}(x) := ((f')')'(x)$, and so on.

This is precisely the content of the following inductive definition.

16.17. **Definition.** (i) For a function $f : D \to \mathbb{R}$ we define the *higher (or higher-order) derivatives* $f^{(k)} : D \to \mathbb{R}$

DEF higher derivatives

- for $k = 0$ as $f^{(0)}(x) := f(x)$ for all $x \in D$,
- for $k = 1, 2, 3, \ldots$ inductively as

$$f^{(k)}(x) := (f^{(k-1)})'(x) \quad \text{for all } x \in D,$$

provided $f^{(k-1)} : D \to \mathbb{R}$ is differentiable.

Note that $f^{(k)} : D \to \mathbb{R}$ is thus defined precisely for those $k \geq 1$ for which $f^{(l)} : D \to \mathbb{R}$ is defined and differentiable for all $l \in \{0, \ldots, k-1\}$.

(ii) If the condition "$f^{(k-1)} : D \to \mathbb{R}$ is differentiable" holds for all $k = 1, \ldots, n$ and some $n \in \mathbb{N}$ (and the definition from (i) is thus applicable for $k = 1, \ldots, n$), then f is called *n times differentiable*. If the condition holds for all $k \in \mathbb{N}$, then f is called *infinitely differentiable*.

(iii) If the function f is n times differentiable and $f^{(n)} : D \to \mathbb{R}$ is continuous, then f is called *n times continuously differentiable*. The set of all n times continuously differentiable functions is denoted by $C^n(D, \mathbb{R})$ (or simply by C^n, if D is clear from the context). ♢

The "zeroth" derivative $f^{(0)}$ according to this definition is nothing other than the function itself, and $f^{(1)}$ is just f'. Instead of $f^{(2)}(x)$ one also writes $f''(x)$, and instead of $f^{(3)}(x)$ sometimes $f'''(x)$. Alternatively, $f^{(n)}(x)$ is also written as

$$\frac{d^n}{dx^n} f(x) \quad \text{or} \quad \frac{d^n f}{dx^n}(x).$$

EXAMPLES
Higher derivatives

16.18. **Examples.**

(a) For $f(x) = \sin x$ we have

$$f^{(1)}(x) = \sin' x = \cos x, \quad f^{(2)}(x) = \cos' x = -\sin x,$$
$$f^{(3)}(x) = -\sin' x = -\cos x, \quad f^{(4)}(x) = -\cos' x = \sin x, \ldots$$

Since $f^{(4)} = f^{(0)}$, this sequence repeats indefinitely, so $\sin x$ is infinitely differentiable.

(b) For $f(x) = x|x|$ with $D = \mathbb{R}$, for $x \in [0, \infty)$ we have $f(x) = x^2$, so $f'(x) = 2x$ for $x > 0$ and $f'_+(0) = 0$. For $x \in (-\infty, 0]$ we have $f(x) = -x^2$ and thus $f'(x) = -2x$ for $x < 0$ and $f'_-(0) = 0$. Therefore, the function is differentiable for all $x \in \mathbb{R}$, since for $x \neq 0$ we have explicitly computed the derivative, and at $x = 0$ the left-hand and right-hand limits agree. Overall,

$$f'(x) = 2|x|.$$

This function is continuous but no longer differentiable at $x = 0$, and thus not differentiable on $D = \mathbb{R}$. The function $f(x) = x|x|$ is therefore once continuously differentiable but not twice differentiable on $D = \mathbb{R}$.

♣

REMARK
Geometric interpretation of the second derivative

16.19. **Remark.** Geometrically, the value of the first derivative $f'(x)$ of a function at a point $x \in D$ is the slope of the function. The second derivative $f''(x)$ describes accordingly the change of the slope. Geometrically, this is the *curvature* of the graph of the function f at the point x. ♠

17. Applications of Differentiability

Date:
March 5, 2026

In this chapter, we discuss various applications of differential calculus.

17.A Extrema and the Mean Value Theorem.

Extrema are the minima and maxima of a function. In particular, we consider here *local* minima and maxima according to the following definition.

DEF local minimum and maximum

17.1. **Definition.** Let $f : D \to \mathbb{R}$ be a function. We say that f has a *local minimum* at some $x \in D$ if there exists an $\varepsilon > 0$ such that

$$f(x) \leq f(y) \quad \text{for all } y \in (x-\varepsilon, x+\varepsilon) \cap D.$$

The minimum is called *strict* if the above inequality is strict ($<$) for $y \neq x$. The point x is called a *(strict) local minimizer* of f. Analogously, the (strict) local maximum is defined using $\geq$ or $>$. ♢

In Section 14 we considered the infimum and supremum of the set

$$\{f(y) \mid y \in D\}.$$

In particular, in Theorem 14.14 we proved that this set has a minimum $f(q)$ and a maximum $f(p)$ if $D = [a, b]$ (i.e., a closed interval) and f is continuous. Thus,

$$f(q) \leq f(y) \quad \text{and} \quad f(p) \geq f(y) \quad \text{for all } y \in D.$$

Hence, $f(q)$ and $f(p)$ are called the *global* minimum and maximum, respectively. It is easy to see that every global minimum (or maximum) is also a local minimum (or maximum). The converse, however, does not necessarily hold.

For differentiable functions, one can identify local minima and maxima using the derivative, provided that the domain is an open interval.

THM Necessary condition for extrema

17.2. **Theorem.** *Let D be an (not necessarily bounded) open interval. Then for every $x \in D$ at which f is differentiable, the following implication holds:*

$$f \text{ has a local minimum or maximum at } x \Rightarrow f'(x) = 0.$$

Proof. We prove the case of a minimum. Choose $\varepsilon > 0$ small enough such that the inequality from the definition of a minimum holds on $(x-\varepsilon, x+\varepsilon)$ and $(x-\varepsilon, x+\varepsilon) \subset D$ (note that this second condition holds automatically for all $\varepsilon > 0$ satisfying $\varepsilon \leq x - a$ or $\varepsilon \leq b - x$ in the case of one-sided or two-sided bounded intervals (a, b), (a, ∞), or $(-\infty, b)$).

Now consider sequences $h_n \to 0$ with $|h_n| < \varepsilon$ for all n. If $h_n \geq 0$ for all n, we obtain

$$a_n := \frac{f(x+h_n) - f(x)}{h_n} \geq 0$$

and if $h_n < 0$ for all n we obtain

$$b_n := \frac{f(x+h_n) - f(x)}{h_n} \leq 0.$$

Since f is differentiable at x, we know that the limit of the difference quotient for $h \to 0$ exists, hence the limits $a := \lim_{n\to\infty} a_n$ and $b := \lim_{n\to\infty} b_n$ must exist and coincide with $f'(x)$. Since $a \geq 0$ and $b \leq 0$, this is only possible if $a = 0$ and $b = 0$, which implies $f'(x) = 0$. ❑

EXAMPLES
Local minima and maxima

17.3. **Examples.** (i) For $f(x) = x^2$ with $D = \mathbb{R}$, the function clearly has a (strict) local minimum at $x = 0$, since $0^2 = 0$ and $x^2 > 0$ for all $x \in \mathbb{R} \setminus \{0\}$. As expected, since $f'(x) = 2x$, we have $f'(0) = 0$.

(ii) For the constant function $f(x) = c$ with any $c \in \mathbb{R}$, f has a local minimum (and also a local maximum, though not strict) at every x. Since $f'(x) = 0$, the derivative is indeed identically zero.

(iii) For closed intervals, Theorem 17.2 does not hold if the extrema lie at the boundary of the interval. For example, the function $f(x) = x$ on the closed interval $[-1, 1]$ has a local (and global) minimum at $x = -1$ and a local (and global) maximum at $x = 1$. However, $f'(x) = 1$ for all x, so the derivative is not zero.

(iv) The condition in Theorem 17.2 is only necessary, not sufficient; that is, from $f'(x) = 0$ we cannot in general conclude that x is a minimum or maximum. For instance, the function $f(x) = x^3$ clearly has neither a local minimum nor a local maximum at $x = 0$, since $f(x) > 0$ for $x > 0$ and $f(x) < 0$ for $x < 0$. Nevertheless, since $f'(x) = 3x^2$, we have $f'(0) = 0$. ♣

How can we ensure that at a point $x \in D$ with $f'(x) = 0$ there actually exists a local maximum or minimum, and how can we distinguish between the two? For this, we require the Mean Value Theorem of differential calculus as a tool. This theorem establishes a relationship between the function values of f and the values of its derivative. The following theorem, due to Rolle, first provides a simplified version, from which we can easily derive the full theorem afterward.

THM
Rolle's Theorem

17.4. **Theorem.** *Let $f : [a, b] \to \mathbb{R}$ be a continuous function with $f(a) = f(b)$, which is differentiable on (a, b). Then there exists a $\xi \in (a, b)$ such that $f'(\xi) = 0$.*

Proof. If f is constant on $[a, b]$, then the derivative is zero on all of (a, b), and the statement follows immediately for any $\xi \in (a, b)$.

Otherwise, there exists some $x_0 \in (a, b)$ with $f(x_0) < f(a)$ or $f(x_0) > f(a)$. Since f is continuous, by Theorem 14.14 there exist points $p, q \in [a, b]$ such that

$$f(p) = \min\{f(x) \mid x \in [a, b]\} \quad \text{and} \quad f(q) = \max\{f(x) \mid x \in [a, b]\}.$$

In the case $f(x_0) < f(a) = f(b)$, we must have $p \in (a, b)$, and in the case $f(x_0) > f(a) = f(b)$, we must have $q \in (a, b)$. Since global minima/maxima are also local minima/maxima, it follows by Theorem 17.2 that $f'(p) = 0$ or $f'(q) = 0$. The claim follows in the first case with $\xi = p$ and in the second case with $\xi = q$. ❑

THM
Mean Value Theorem of Differential Calculus

17.5. **Theorem.** *Let $f : [a, b] \to \mathbb{R}$ be a continuous function that is differentiable on (a, b). Then there exists a $\xi \in (a, b)$ such that*

$$f'(\xi) = \frac{f(b) - f(a)}{b - a}.$$

Proof. We define the function

$$g(x) = f(x) + \frac{f(a) - f(b)}{b - a}(x - a).$$

Then

$$g(a) = f(a) + \frac{f(a) - f(b)}{b - a}(a - a) = f(a), \quad g(b) = f(b) + \frac{f(a) - f(b)}{b - a}(b - a) = f(a).$$

Since g, being the sum of continuous and differentiable functions, is continuous on $[a, b]$ and differentiable on (a, b), the assumptions of Rolle?s Theorem are satisfied. Thus, there exists some $\xi \in (a, b)$ with $g'(\xi) = 0$. Since

$$g'(x) = f'(x) + \frac{f(a) - f(b)}{b - a},$$

it follows that

$$f'(\xi) = g'(\xi) - \frac{f(a) - f(b)}{b - a} = \frac{f(b) - f(a)}{b - a}.$$

❑

To establish a connection with extrema, we consider the following consequence of the Mean Value Theorem.

17.6. **Theorem.** *Let $f : [a, b] \to \mathbb{R}$ be a continuous function that is differentiable on (a, b). If for all $x \in (a, b)$ the inequality*

$$f'(x) \geq 0 \quad (\text{respectively } f'(x) > 0,\ f'(x) \leq 0,\ f'(x) < 0)$$

holds, then f is monotonically increasing (respectively strictly increasing, monotonically decreasing, strictly decreasing) on $[a, b]$.

THM Monotonicity and derivative

Proof. We prove the case $f'(x) > 0$; the other cases are analogous. Assume that f is not strictly increasing. Then there exist $x_1, x_2 \in [a, b]$ with $x_1 < x_2$ and $f(x_1) \geq f(x_2)$. By the Mean Value Theorem, there exists some $\xi \in (x_1, x_2)$ with

$$f'(\xi) = \frac{f(x_2) - f(x_1)}{x_2 - x_1} \leq 0,$$

which contradicts the assumption that $f'(x) > 0$ for all $x \in (a, b)$. ❑

Now let us return to extrema. One way to determine whether a point $x \in D$ with $f'(x) = 0$ is actually a minimum or a maximum is to consider the second derivative.

17.7. **Theorem.** *Let $f : [a, b] \to \mathbb{R}$ be a continuous function that is twice differentiable on (a, b). Let $x \in (a, b)$ satisfy $f'(x) = 0$. Then the following implications hold:*

$$f''(x) > 0 \Rightarrow f \text{ has a strict local minimum at } x,$$

and

$$f''(x) < 0 \Rightarrow f \text{ has a strict local maximum at } x.$$

THM Sufficient conditions for minima and maxima

Proof. We prove the case $f''(x) > 0$. From

$$0 < f''(x) = \lim_{h \to 0} \frac{f'(x+h) - f'(x)}{h}$$

it follows that there exists $\varepsilon > 0$ such that

$$\frac{f'(x+h) - f'(x)}{h} > 0$$

for all $h \in \mathbb{R}$ with $|h| < \varepsilon$. Otherwise, there would exist arbitrarily small h with

$$\frac{f'(x+h) - f'(x)}{h} \leq 0,$$

which would imply $f''(x) \leq 0$ as $h \to 0$.

For $h \in (0, \varepsilon)$, this gives

$$f'(x+h) > f'(x) = 0,$$

so by Theorem 17.6, f is strictly increasing on $[x, x+\varepsilon]$. Thus $f(x) < f(y)$ for all $y \in (x, x+\varepsilon]$. Similarly, one shows that f is strictly decreasing on $[x-\varepsilon, x]$, from which $f(x) < f(y)$ for all $y \in [x-\varepsilon, x)$ follows. Hence, f has a strict local minimum at x. ❑

EXAMPLES
Minima and maxima

17.8. **Examples.** (i) For $f(x) = x^2$, we have $f'(x) = 2x$ and $f''(x) = 2$. Hence $f'(0) = 0$ and $f''(0) = 2$. Therefore, there is a local minimum at $x = 0$.

(ii) For $f(x) = x^3$, f has neither a local maximum nor a local minimum at $x = 0$, so $f''(0)$ can be neither positive nor negative and must thus equal zero. Since $f'(x) = 3x^2$, we have $f''(x) = 6x$, and indeed $f''(0) = 6 \cdot 0 = 0$.

(iii) The conditions in Theorem 17.7 are sufficient but not necessary, i.e. even if $f'(x) = 0$ and $f''(x) = 0$, f may still have a strict local minimum or maximum at x. An example is $f(x) = x^4$ at $x = 0$. Here, since $x^4 > 0$ for $x \neq 0$, f has a strict local minimum at $x = 0$. Nevertheless, since $f''(x) = 12x^2$, we have $f''(0) = 0$. ♣

For certain special functions, one can deduce the existence of a global minimum or maximum from $f'(x) = 0$. So-called convex functions satisfy, for all $x_1, x_2 \in D$ with $x_1 \neq x_2$ and all $\lambda \in (0,1)$, the inequality

$$f(\lambda x_1 + (1-\lambda)x_2) \leq \lambda f(x_1) + (1-\lambda) f(x_2).$$

For such functions, every x with $f'(x) = 0$ is a global minimum. Convexity can be checked using the second derivative: it holds if $f''(x) \geq 0$ for all $x \in D$.

The counterpart of convexity, obtained by replacing "$\leq$?? with "$\geq$??, is called concave and yields a global maximum. Both notions can also be formulated using strict inequalities, in which case the global minimum or maximum is also strict.

17.B Taylor Expansion.

For many applications—for example, when one wishes to evaluate complicated functions on a computer—it is useful to approximate them by simpler functions. Particularly nice "simple functions" are polynomials, that is, functions of the form

$$P(x) = a_n x^n + a_{n-1} x^{n-1} + \ldots + a_1 x + a_0,$$

where $n \in \mathbb{N}$ is the degree and the parameters $a_0, \ldots, a_n \in \mathbb{R}$ are the coefficients of the polynomial. Polynomials are convenient because, to store a polynomial of degree n, only the $n+1$ coefficients—i.e., only $n+1$ real numbers—need to be stored on the computer.

Now, the question arises: given a function f, how can we find a polynomial P of a prescribed degree $n \in \mathbb{N}$ that closely matches the function near a given point $x_0 \in \mathbb{R}$, that is, such that the values $P(x)$ for x near x_0 deviate little from the function values $f(x)$?

The idea of the Taylor expansion is to determine the polynomial in such a way that its function value and its first n derivatives at x_0 coincide with those of f. The necessary condition for this is, of course, that f is sufficiently differentiable at x_0. To simplify the computation, it is advisable to write the polynomial in the form

$$P(x) = a_n(x - x_0)^n + a_{n-1}(x - x_0)^{n-1} + \ldots + a_1(x - x_0) + a_0.$$

By expanding, one easily sees that this is again a polynomial of the above form (though with different coefficients). By successive differentiation, we obtain the formula

$$P^{(k)}(x_0) = k!\, a_k, \quad \text{for } k = 0, \ldots, n.$$

Thus, if we require that $P^{(k)}(x_0) = f^{(k)}(x_0)$ for $k = 0, \ldots, n$, we must set $a_k = f^{(k)}(x_0)/k!$. That is, we define

$$P(x) := \sum_{k=0}^{n} \frac{f^{(k)}(x_0)}{k!}(x - x_0)^k. \tag{17.1}$$

This polynomial is called the *Taylor polynomial* (of degree n) of f at x_0.

The following theorem gives information about the difference between $f(x)$ and $P(x)$.

THM Taylor approximation

17.9. **Theorem.** *Let D be an open interval, $f : D \to \mathbb{R}$ be $(n+1)$-times differentiable, and let P be defined by (17.1). Then for every $x \in D$ with $x > x_0$, there exists a $\xi \in (x_0, x)$ such that*

$$f(x) = P(x) + R_n, \quad \text{with } R_n := \frac{(x - x_0)^{n+1}}{(n+1)!} f^{(n+1)}(\xi).$$

The same statement holds for $x < x_0$ with $\xi \in (x, x_0)$. The term R_n is called the Lagrange remainder .

To prove Theorem 17.9, we need the following generalization of the Mean Value Theorem of differential calculus. Note that we can obtain the earlier form (Theorem 17.5) as a special case by choosing $g(x) = x$.

THM Generalized Mean Value Theorem of Differential Calculus

17.10. **Theorem.** *Let $f, g : [a, b] \to \mathbb{R}$ be continuous functions that are differentiable on (a, b) with $g'(x) \neq 0$ for all $x \in (a, b)$. Then $g(b) - g(a) \neq 0$, and there exists a $\xi \in (a, b)$ such that*

$$\frac{f'(\xi)}{g'(\xi)} = \frac{f(b) - f(a)}{g(b) - g(a)}.$$

Proof. By Rolle's theorem, we first note that $g(b) - g(a) \neq 0$, since otherwise $g(a) = g(b)$, and we would have $g'(x) = 0$ for some $x \in (a, b)$.

Now define

$$h(x) = (g(b) - g(a))f(x) - (f(b) - f(a))g(x).$$

Then

$$\begin{aligned} h(b) - h(a) &= (g(b) - g(a))f(b) - (f(b) - f(a))g(b) \\ &\quad -(g(b) - g(a))f(a) + (f(b) - f(a))g(a) = 0, \end{aligned}$$

so $h(b) = h(a)$. By Rolle's theorem, there exists a $\xi \in (a, b)$ such that

$$0 = h'(\xi) = (g(b) - g(a))f'(\xi) - (f(b) - f(a))g'(\xi),$$

from which the claim follows. ❑

Proof. (of Theorem 17.9): Define

$$g(x) := f(x) - P(x).$$

Since $g^{(k)}(x) = f^{(k)}(x) - P^{(k)}(x)$, this function satisfies:

(i) g is $(n+1)$-times differentiable on D;
(ii) $g(x_0) = g'(x_0) = \ldots = g^{(n)}(x_0) = 0$.

We show that for every function $g : D \to \mathbb{R}$ satisfying (i) and (ii), and for every $x \geq x_0$, there exists a $\xi \in (x_0, x)$ such that

$$g(x) = \frac{(x - x_0)^{n+1}}{(n+1)!} g^{(n+1)}(\xi). \tag{17.2}$$

From this, the theorem follows because P is a polynomial of degree n, hence $P^{(n+1)}(x) = 0$ for all $x \in \mathbb{R}$, and therefore $g^{(n+1)}(x) = f^{(n+1)}(x) - P^{(n+1)}(x) = f^{(n+1)}(x)$.

We prove (17.2) by induction on n.

For $n = 0$, since $g(x_0) = 0$, the usual Mean Value Theorem yields

$$\frac{g(x)}{x - x_0} = \frac{g(x) - g(x_0)}{x - x_0} = g'(\xi),$$

which is exactly (17.2) for $n = 0$.

For the induction step $n \to n+1$, consider a function that satisfies (i) and (ii) for $n+1$ instead of n. We must prove that (17.2) holds for this function with $n+1$ in place of n.

If g satisfies (i) and (ii) for $n+1$, then g' satisfies (i) and (ii) for n. By the induction hypothesis, (17.2) holds for g' and n. Choose any $\tilde{\xi} \in D$ with $\tilde{\xi} > x_0$, and apply (17.2) with $x = \tilde{\xi}$ and $g = g'$. Then there exists a $\xi \in (x_0, \tilde{\xi})$ such that

$$g'(\tilde{\xi}) = \frac{(\tilde{\xi} - x_0)^{n+1}}{(n+1)!} g^{(n+2)}(\xi), \tag{17.3}$$

where we have used $g'^{(n+1)} = g^{(n+2)}$.

Now we apply the generalized Mean Value Theorem to the functions g (instead of f) and $h(x) = (x - x_0)^{n+2}$ (instead of g) on the interval $[x_0, x]$. Then there exists a $\tilde{\xi} \in (x_0, x)$ such that

$$\frac{g'(\tilde{\xi})}{h'(\tilde{\xi})} = \frac{g(x) - g(x_0)}{h(x) - h(x_0)},$$

which, for our choice of g and h, gives

$$\frac{g(x)}{(x - x_0)^{n+2}} = \frac{g(x) - g(x_0)}{h(x) - h(x_0)} = \frac{g'(\tilde{\xi})}{(n+2)(\tilde{\xi} - x_0)^{n+1}}.$$

Substituting the expression for $g'(\tilde{\xi})$ from (17.3), we obtain

$$\frac{g(x)}{(x-x_0)^{n+2}} = \frac{(\tilde{\xi}-x_0)^{n+1}}{(n+1)!}\,\frac{g^{(n+2)}(\xi)}{(n+2)(\tilde{\xi}-x_0)^{n+1}} = \frac{g^{(n+2)}(\xi)}{(n+2)!},$$

which is precisely the desired equation (17.2). ❑

REMARK
Landau symbol

17.11. **Remark.** Analogous to the "small" Landau symbol $o(\cdot)$, which we introduced in Remark 16.5(i), we define the "big" Landau symbol $O(\cdot)$ as follows: For a function $r : \mathbb{R} \to \mathbb{R}$ for which there exist constants $\varepsilon, C > 0$ and a $p \in \mathbb{N}$ such that the inequality

$$\frac{|r(h)|}{h^p} \le C$$

holds for all $h \in (-\varepsilon, \varepsilon)$ with $h \neq 0$, we write simply

$$r(h) = O(h^p).$$

If the derivative $f^{(n+1)}$ is bounded in absolute value by some C on $(x-\varepsilon, x+\varepsilon)$ for some $\varepsilon > 0$ (which is always true if $f^{(n+1)}$ is continuous), then for the Taylor polynomial P under the assumptions of Theorem 17.9 we have

$$f(x) - P(x) = O((x-x_0)^{n+1}).$$

In this form, the Taylor expansion is often presented. ♠

EXAMPLES
Taylor expansions

17.12. **Examples.** (i) Let $f(x) = \sin x$ and $x_0 = 0$. Then

$$f(0) = \sin 0 = 0, \quad f'(0) = \cos 0 = 1, \quad f''(0) = -\sin 0 = 0,$$

$$f^{(3)}(0) = -\cos 0 = -1, \quad f^{(4)}(0) = \sin 0 = 0, \ldots$$

and thus (for odd n)

$$P(x) = x - \frac{x^3}{3!} + \frac{x^5}{5!} - \frac{x^7}{7!} \pm \ldots \pm \frac{x^n}{n!}.$$

We therefore obtain exactly the terms up to order n from the series representation of the sine in (15.6). For $n = 15$, f and P are shown graphically in Figure 11.

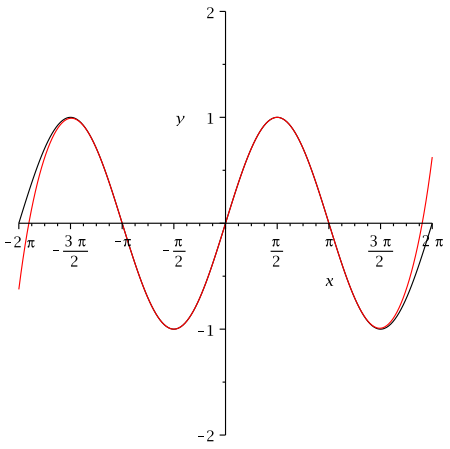


FIGURE 11. $\sin x$ (black) and the Taylor polynomial for $n = 15$ and $x_0 = 0$ (red)

Since for $f(x) = \sin x$ we have the inequality $|f^{(n)}(x)| \leq 1$ for all $x \in \mathbb{R}$ and all $n \in \mathbb{N}$ (as every derivative is of the form $\pm \sin x$ or $\pm \cos x$), Theorem 17.9 yields the estimate

$$|f(x) - P(x)| \leq \frac{|x|^{n+1}}{(n+1)!}.$$

Similarly, one verifies that the series representations of the exponential function (from Theorem 15.1) and of the cosine function (from (15.6)) coincide precisely with their respective Taylor polynomials.

(ii) The so-called *Runge function* is given by $f(x) = 1/(1+x^2)$. With a somewhat longer calculation, one finds

$$f^{(k)}(0) = (-1)^{k/2} k!$$

for even k, and $f^{(k)}(0) = 0$ for odd k. Hence (for even n),

$$P(x) = 1 - x^2 + x^4 - x^6 \pm \ldots \pm x^n.$$

Again, for $n = 15$, f and P are shown in Figure 12. ♣

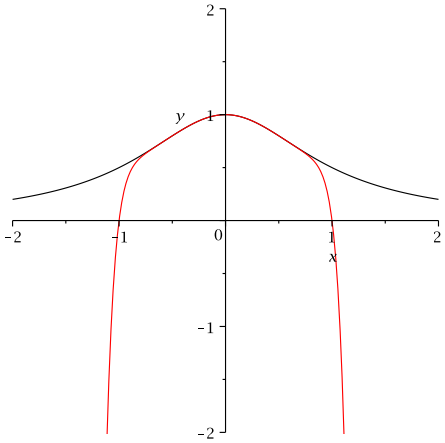


FIGURE 12. $1/(1+x^2)$ (black) and the Taylor polynomial for $n = 15$ and $x_0 = 0$ (red)

If we look at Examples 17.12(i) and (ii) for increasing degrees n of the Taylor polynomial, we notice that in the case of the sine, the polynomial P converges to the sine function on the entire interval ? in fact, on arbitrarily large intervals. In contrast, for the Runge function, we obtain no convergence at the boundaries or outside the interval $[-1, 1]$. Here, the polynomial values $P(x)$ provide a good approximation to $f(x)$ only in a neighborhood of x_0.

If $f : D \to \mathbb{R}$ is infinitely differentiable and the infinite series obtained from the Taylor polynomials,

$$\sum_{k=0}^{n} \frac{f^{(k)}(x_0)}{k!}(x - x_0)^k, \quad n \in \mathbb{N},$$

converges as $n \to \infty$, then this series is called the *Taylor series* of f at x_0. The Taylor series is a special case of a *power series* , which generally has the form

$$\sum_{k=0}^{n} c_k (x - x_0)^k, \quad n \in \mathbb{N},$$

for coefficients $c_k \in \mathbb{R}$.

Considering the derivation of the Taylor series, one might at first suspect that the equality

$$f(x) = \sum_{k=0}^{\infty} \frac{f^{(k)}(x_0)}{k!}(x - x_0)^k \tag{17.4}$$

holds. However, this is generally *not true*. More precisely, two problems may occur:

(1) The limit

$$\sum_{k=0}^{\infty} \frac{f^{(k)}(x_0)}{k!}(x - x_0)^k$$

may fail to exist for certain $x \in D$.

(2) Even if the limit exists, the equality (17.4) may fail to hold for some (or even many) $x \in D$.

In the remainder of this section, we examine these two cases in more detail, beginning with (1).

EXAMPLE Convergence of Taylor series for Runge function

17.13. **Example.** For the Runge function, the Taylor series (after appropriate reindexing) is

$$\sum_{k=0}^{n} (-1)^k x^{2k}.$$

Note that here $x_0 = 0$. For all $x \in \mathbb{R}$ with $|x| \geq 1$, we have $|(-1)^k x^{2k}| \geq 1$; thus, the terms of the Taylor series do not tend to zero, and by Theorem 9.5, the series does not converge. For $|x| < 1$, however, (absolute) convergence follows from the ratio test. Hence, the Taylor series converges for all $x \in \mathbb{R}$ with $|x| < 1$.

That it actually converges to $f(x)$ for these values can be shown by proving that the absolute value of the remainder term R_n is bounded by some expression b_n for all $|x| \leq a$ and $|\xi| \leq a$ with $a < 1$, where $b_n \to 0$ as $n \to \infty$. For brevity, we omit the explicit calculation of these bounds b_n for this example. ♣

Recalling the concept of the radius of convergence introduced after Equation (9.6), the Taylor series of the Runge function at $x_0 = 0$ therefore has a radius of convergence $r = 1$. In general, for each Taylor expansion one must separately examine ? typically using the remainder term ? whether the Taylor series actually converges to $f(x)$ and what its radius of convergence is. If a function $f : D \to \mathbb{R}$ can be represented by a power series (for example, by its Taylor series) for all $x \in D$, then f is called *analytic* on D.

17.C L'Hospital's Rules.

Another application of the generalized mean value theorem is given by L'Hospital's rules. These allow one to compute limits of the form $\lim_{x\to a} f(x)/g(x)$ when $\lim_{x\to a} g(x) = 0$ or $\lim_{x\to a} g(x) = \infty$, cases in which the usual quotient rule is not applicable.

THM L'Hospital's rule

17.14. **Theorem.** *Let $f, g : (a, b) \to \mathbb{R}$ be differentiable, and assume $g(x) \neq 0$ and $g'(x) \neq 0$ for all $x \in (a, b)$. Furthermore, suppose one of the following conditions holds:*

(i) $\lim_{x\to a} f(x) = 0$ *and* $\lim_{x\to a} g(x) = 0$,
(ii) $\lim_{x\to a} f(x) = \infty$ *and* $\lim_{x\to a} g(x) = \pm\infty$.

Then

$$\lim_{x\to a} \frac{f(x)}{g(x)} = \lim_{x\to a} \frac{f'(x)}{g'(x)},$$

provided the limit on the right-hand side (possibly improper) exists. In all limits, it is assumed that $x \in (a, b)$.

The analogous statement also holds for $x \to b$, and, in the case where $f, g : (a, \infty) \to \mathbb{R}$ or $f, g : (-\infty, b) \to \mathbb{R}$, likewise for $x \to \infty$ or $x \to -\infty$.

Proof. In case (i), we can extend f and g continuously at $x = a$ by setting their values to 0. By the generalized mean value theorem, for every $x \in (a, b)$ there exists a $\xi \in (a, x]$ such that

$$f'(\xi)g(x) = g'(\xi)f(x),$$

and thus

$$\frac{f'(\xi)}{g'(\xi)} = \frac{f(x)}{g(x)}.$$

As $x \to a$, we have $\xi \in (a, x]$ and therefore $\xi \to a$, from which the claim follows.

In case (ii), the argument is slightly more involved, since f and g cannot be extended continuously at $x = a$. For $a < y < x < b$, the generalized mean value theorem yields some $\xi \in [y, x]$ such that

$$\frac{f'(\xi)}{g'(\xi)} = \frac{f(y) - f(x)}{g(y) - g(x)}. \tag{17.5}$$

Since $f(y), g(y) \to \pm\infty$ as $y \to a$, both functions are nonzero for y sufficiently close to a. Hence,

$$\frac{f(y)}{g(y)} = \frac{f(y) - f(x)}{g(y) - g(x)} \frac{1 - g(x)/g(y)}{1 - f(x)/f(y)} = \frac{f'(\xi)}{g'(\xi)} \frac{1 - g(x)/g(y)}{1 - f(x)/f(y)}.$$

Because $f(x) \to \infty$ and $g(x) \to \infty$ as $x \to a$, there exist sequences $(x_n)_{n\in\mathbb{N}}$, $(y_n)_{n\in\mathbb{N}}$ with $x_n, y_n \in (a, b)$ and $x_n, y_n \to a$ as $n \to \infty$, such that $f(x_n)/f(y_n) \to 0$ and $g(x_n)/g(y_n) \to 0$. (Choose first (x_n), and then select y_n sufficiently close to a so that $f(y_n) > nf(x_n)$ and $g(y_n) > ng(x_n)$ hold.) Let ξ_n be the ξ corresponding to $x = x_n$ and $y = y_n$ in (17.5). Then

$$\lim_{y\to a} \frac{f(y)}{g(y)} = \lim_{n\to\infty} \frac{f(y_n)}{g(y_n)} = \lim_{n\to\infty} \frac{f'(\xi_n)}{g'(\xi_n)} \underbrace{\frac{1 - g(x_n)/g(y_n)}{1 - f(x_n)/f(y_n)}}_{\to 1} = \lim_{n\to\infty} \frac{f'(\xi_n)}{g'(\xi_n)} = \lim_{\xi\to a} \frac{f'(\xi)}{g'(\xi)}.$$

❑

18. Matrices

Date:
March 5, 2026

The results of Section 13 have shown that we can describe linear maps between two finite-dimensional K-vector spaces V and W of dimensions n and m by mn coefficients in K. To do this, we have to choose bases of V and of W; this gives us a basis of $\mathrm{Hom}(V,W)$ as in Corollary 13.23, and the relevant coefficients then are the coefficients of the given linear map $V \to W$ when written as a linear combination with respect to this basis. These coefficients are usually arranged in a special form, as follows.

DEF Matrix

18.1. **Definition.** Let K be a field and let $m, n \in \mathbb{N}$. A *$m \times n$ matrix with entries in K* (or short a *$m \times n$ matrix over K*) is a rectangular arrangement of mn elements of K written in the following way.

$$\begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{pmatrix}$$

To abbreviate this, we also write $(a_{ij})_{1\le i\le m, 1\le j\le n}$ (or just $(a_{ij})_{i,j}$ when m and n are known from the context) for this matrix. The matrix is *square* when $m = n$. For $i \in \{1, 2, \dots, m\}$ the n-tuple $(a_{i1}, a_{i2}, \dots, a_{in})$ is the *ith row* of the matrix; for $j \in \{1, 2, \dots, n\}$ the m-tuple $(a_{1j}, a_{2j}, \dots, a_{mj})$ is the *jth column* of the matrix.

We write $\mathrm{Mat}(m \times n, K)$ for the set of all $m \times n$ matrices with entries in K; if $m = n$, we usually abbreviate this to $\mathrm{Mat}(n, K)$. ♢

Another common notation for $\mathrm{Mat}(m \times n, K)$ is $K^{m\times n}$. In the literature one frequently also finds matrices written between square brackets,

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix}.$$

We will stick to (round) parentheses in these notes.

An $m \times n$ matrix over K is essentially nothing else than a family of elements of K with the index set $\{1, 2, \dots, m\} \times \{1, 2, \dots, n\}$, i.e.,

$$\mathrm{Mat}(m \times n, K) = K^{\{1,2,\dots,m\}\times\{1,2,\dots,n\}}$$

(this explains the notation $K^{m\times n}$ mentioned earlier). Since we have defined a structure as a K-vector space on arbitrary sets of the form K^I, we immediately obtain the following. (Note that $\#\big(\{1, 2, \dots, m\} \times \{1, 2, \dots, n\}\big) = mn$.)

LEMMA Vector space of $m \times n$-matrices

18.2. **Lemma.** *Let K be a field and let $m, n \in \mathbb{N}$. The set $\mathrm{Mat}(m \times n, K)$ with addition and scalar multiplication defined component-wise is a K-vector space of dimension mn.*

If $m = 0$ or $n = 0$ (or both), then $\mathrm{Mat}(m \times n, K)$ is a zero vector space; its only element is an empty matrix (with zero rows and n columns or with m rows and zero columns).

We therefore add matrices (that have the same number of rows and columns) as follows.

$$\begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{pmatrix} + \begin{pmatrix} b_{11} & b_{12} & \cdots & b_{1n} \\ b_{21} & b_{22} & \cdots & b_{2n} \\ \vdots & \vdots & & \vdots \\ b_{m1} & b_{m2} & \cdots & b_{mn} \end{pmatrix}$$
$$= \begin{pmatrix} a_{11}+b_{11} & a_{12}+b_{12} & \cdots & a_{1n}+b_{1n} \\ a_{21}+b_{21} & a_{22}+b_{22} & \cdots & a_{2n}+b_{2n} \\ \vdots & \vdots & & \vdots \\ a_{m1}+b_{m1} & a_{m2}+b_{m2} & \cdots & a_{mn}+b_{mn} \end{pmatrix}$$

Scalar multiplication is performed like this:

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

As noted at the beginning of this section, we can associate a matrix to a linear map. We first consider the case $V = K^n$ and $W = K^m$ with the standard bases $B = (\mathbf{e}_1, \dots, \mathbf{e}_n)$ of V and $B' = (\mathbf{e}'_1, \dots, \mathbf{e}'_m)$ of W (we write $\mathbf{e}'_i$ for the ith standard basis vector in K^m to distinguish it from the basis vectors $\mathbf{e}_j$ in K^n). By Corollary 13.23 we then obtain the basis $(\phi_{ij})_{1\le i\le m, 1\le j\le n}$ of $\operatorname{Hom}(K^n, K^m)$, where $\phi_{ij}(\mathbf{e}_k) = \mathbf{0}$ for $k \neq j$ and $\phi_{ij}(\mathbf{e}_j) = \mathbf{e}'_i$. If $\phi\colon K^n \to K^m$ is a linear map, then we can write ϕ as a linear combination

$$\phi = \sum_{i=1}^{m} \sum_{j=1}^{n} a_{ij}\phi_{ij} \qquad \text{with } a_{ij} \in K.$$

The associated matrix then is $A = (a_{ij})_{1\le i\le m, 1\le j\le n}$.

18.3. **Example.** We had seen in Example 13.24 that a linear map $\phi\colon \mathbb{R}^3 \to \mathbb{R}^2$ has the form $\phi(x, y, z) = (ax+by+cz, dx+ey+fz)$ with suitable $a, b, c, d, e, f \in \mathbb{R}$. Equivalently, $\phi = a\phi_{11} + b\phi_{12} + c\phi_{13} + d\phi_{21} + e\phi_{22} + f\phi_{23}$, giving the associated matrix

$$A = \begin{pmatrix} a & b & c \\ d & e & f \end{pmatrix}. \qquad \clubsuit$$

EXAMPLE
Matrix for $\phi\colon \mathbb{R}^3 \to \mathbb{R}^2$

The jth column of the $m \times n$ matrix associated to $\phi\colon K^n \to K^m$ consists exactly of the coefficients of the image of the jth basis vector $\mathbf{e}_j$ under ϕ; this is because

$$\phi(\mathbf{e}_j) = \sum_{i=1}^{m} \sum_{k=1}^{n} a_{ik}\phi_{ik}(\mathbf{e}_j) = \sum_{i=1}^{m} a_{ij}\mathbf{e}'_i = (a_{1j}, a_{2j}, \dots, a_{mj});$$

compare the proof of Corollary 13.23.

18.4. **Lemma.** *The association described above defines an isomorphism*

$$\operatorname{Hom}(K^n, K^m) \to \operatorname{Mat}(m \times n, K)$$

of K-vector spaces.

If we identify $\operatorname{Mat}(m \times n, K)$ with $K^{\{1,\dots,m\}\times\{1,\dots,n\}}$, then this isomorphism is the inverse of the linear combination map $K^{\{1,2,\dots,m\}\times\{1,2,\dots,n\}} \to \operatorname{Hom}(K^n, K^m)$ corresponding to the basis $(\phi_{ij})_{(i,j)\in\{1,\dots,m\}\times\{1,\dots,n\}}$ of $\operatorname{Hom}(K^n, K^m)$.

LEMMA
$\operatorname{Mat}(m \times n, K) \cong \operatorname{Hom}(K^n, K^m)$

Proof. The linear combination map mentioned in the statement,

$$\Phi\colon \operatorname{Mat}(m\times n,K)=K^{\{1,2,\dots,m\}\times\{1,2,\dots,n\}}\to\operatorname{Hom}(K^n,K^m)\,,$$

maps a matrix $(a_{ij})_{1\le i\le m,1\le j\le n}$ to the linear combination $\sum_{i,j}a_{ij}\phi_{ij}$. This map Φ is an isomorphism since $(\phi_{ij})_{1\le i\le m,1\le j\le n}$ is a basis of $\operatorname{Hom}(K^n,K^m)$ (compare Example 13.3). The map that associates to a linear map $\phi\colon K^n\to K^m$ its matrix clearly is the inverse map of Φ; in particular, it is also an isomorphism. ❑

What does the application of the linear map $\phi\colon K^n\to K^m$ look like in terms of the associated matrix $A=(a_{ij})_{i,j}$? We have that

$$\phi(x_1,x_2,\dots,x_n)=\sum_{i=1}^{m}\sum_{j=1}^{n}a_{ij}\phi_{ij}(x_1,x_2,\dots,x_n)=\sum_{i=1}^{m}\sum_{j=1}^{n}a_{ij}x_j\mathbf{e}'_i\,,$$

so the ith component of $\phi(x_1,x_2,\dots,x_n)$ is obtained as

$$\sum_{j=1}^{n}a_{ij}x_j=a_{i1}x_1+a_{i2}x_2+\ldots+a_{in}x_n\,.$$

This is then usually written as multiplication of the matrix A with the *column vector* corresponding to $(x_1,\dots,x_n)$. (So we identify K^n with $\operatorname{Mat}(n\times 1,K)$ and K^m with $\operatorname{Mat}(m\times 1,K)$.) Explicitly, application of ϕ then results in the following product.

$$\begin{pmatrix}a_{11}&a_{12}&\cdots&a_{1n}\\a_{21}&a_{22}&\cdots&a_{2n}\\\vdots&\vdots&&\vdots\\a_{m1}&a_{m2}&\cdots&a_{mn}\end{pmatrix}\begin{pmatrix}x_1\\x_2\\\vdots\\\vdots\\x_n\end{pmatrix}=\begin{pmatrix}a_{11}x_1+a_{12}x_2+\ldots+a_{1n}x_n\\a_{21}x_1+a_{22}x_2+\ldots+a_{2n}x_n\\\vdots\\a_{m1}x_1+a_{m2}x_2+\ldots+a_{mn}x_n\end{pmatrix}.$$

The result is again a column vector, of length m. Its ith component is obtained from the ith row of the matrix and the column vector corresponding to $(x_1,\dots,x_n)$ as the *scalar product*

DEF
Scalar product

$$a_{i1}x_1+a_{i2}x_2+\ldots+a_{in}x_n\,.$$

(The name comes from the fact that its value is a scalar:

"vector times vector = scalar".

Note that this is different from scalar multiplication

"scalar times vector = vector"!)

18.5. **Examples.** 2×3 matrices with entries in $\mathbb{R}$ correspond to linear maps $\mathbb{R}^3\to\mathbb{R}^2$. In this case the formula above specializes to

EXAMPLES

$$\begin{pmatrix}a&b&c\\d&e&f\end{pmatrix}\begin{pmatrix}x\\y\\z\end{pmatrix}=\begin{pmatrix}ax+by+cz\\dx+ey+fz\end{pmatrix}.$$

3×2 matrices over $\mathbb{R}$ correspond to linear maps $\mathbb{R}^2\to\mathbb{R}^3$. In this case we obtain

$$\begin{pmatrix}a&b\\c&d\\e&f\end{pmatrix}\begin{pmatrix}x\\y\end{pmatrix}=\begin{pmatrix}ax+by\\cx+dy\\ex+fy\end{pmatrix}.$$ ♣

Composition of linear maps corresponds to multiplication of matrices.

18.6. **Definition.** Let K be a field and let $l, m, n \in \mathbb{N}$. For two matrices $A \in \mathrm{Mat}(l \times m, K)$ and $B \in \mathrm{Mat}(m \times n, K)$ we define their *product*

$$A \cdot B \in \mathrm{Mat}(l \times n, K)$$

as the matrix associated to $f \circ g$, where $f\colon K^m \to K^l$ and $g\colon K^n \to K^m$ are the linear maps corresponding to A and B, respectively. As usual, we also write AB for $A \cdot B$. ♢

DEF
Matrix multiplication

In the same way that we can compose maps only when the codomain of the one map agrees with the domain of the other map, we can multiply matrices only when their sizes "match", which means that the number of columns of the left factor is the same as the number of rows of the right factor.

What does this multiplication of matrices look like concretely?

Let $A = (a_{ij})_{1\le i\le l, 1\le j\le m}$, $B = (b_{jk})_{1\le j\le m, 1\le k\le n}$ and $C = (c_{ik})_{1\le i\le l, 1\le k\le n} = AB$. Then c_{ik} must be the ith component of $f(g(\mathbf{e}_k))$. We have

$$f(g(\mathbf{e}_k)) = f(b_{1k}\mathbf{e}'_1 + b_{2k}\mathbf{e}'_2 + \ldots + b_{mk}\mathbf{e}'_m) = b_{1k}f(\mathbf{e}'_1) + b_{2k}f(\mathbf{e}'_2) + \ldots + b_{mk}f(\mathbf{e}'_m)\,,$$

and the ith component of $f(\mathbf{e}'_j)$ is a_{ij}. This shows that

$$c_{ik} = a_{i1}b_{1k} + a_{i2}b_{2k} + \ldots + a_{im}b_{mk} = \sum_{j=1}^{m} a_{ij}b_{jk}$$

is the **scalar product of the ith row of A with the kth column of B**:

The (i,k) entry of AB is "ih row of A times kth column of B".

The multiplication "matrix times column vector" introduced earlier is just a special case of this general multiplication of matrices.

18.7. **Example.** We compute the product of two matrices over $\mathbb{R}$. **EXAMPLE**

$$\begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{pmatrix} \begin{pmatrix} 1 & 2 \\ 3 & 4 \\ 5 & 6 \end{pmatrix} = \begin{pmatrix} 1\cdot 1 + 2\cdot 3 + 3\cdot 5 & 1\cdot 2 + 2\cdot 4 + 3\cdot 6 \\ 4\cdot 1 + 5\cdot 3 + 6\cdot 5 & 4\cdot 2 + 5\cdot 4 + 6\cdot 6 \end{pmatrix} = \begin{pmatrix} 22 & 28 \\ 49 & 64 \end{pmatrix}$$

$$\begin{pmatrix} 1 & 2 \\ 3 & 4 \\ 5 & 6 \end{pmatrix} \begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{pmatrix} = \begin{pmatrix} 1\cdot 1 + 2\cdot 4 & 1\cdot 2 + 2\cdot 5 & 1\cdot 3 + 2\cdot 6 \\ 3\cdot 1 + 4\cdot 4 & 3\cdot 2 + 4\cdot 5 & 3\cdot 3 + 4\cdot 6 \\ 5\cdot 1 + 6\cdot 4 & 5\cdot 2 + 6\cdot 5 & 5\cdot 3 + 6\cdot 6 \end{pmatrix} = \begin{pmatrix} 9 & 12 & 15 \\ 19 & 26 & 33 \\ 29 & 40 & 51 \end{pmatrix}$$

♣

The matrix associated to the identity map is quite important and has its own name.

18.8. **Definition.** Let K be a field and $n \in \mathbb{N}$. The matrix $I_n \in \mathrm{Mat}(n, K)$ that is associated to the identity map id_{K^n} is the *identity matrix (of size n over K)*. ♢

DEF
Identity matrix

The jth column of I_n must contain the jth standard basis vector. This means that I_n has the following form.

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

One also writes this as $I_n = (\delta_{ij})_{1\le i,j\le n}$, where

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

is the *Kronecker delta*.

DEF Kronecker delta

18.9. **Lemma.** *Let K be a field. The multiplication of matrices over K is associative and has the identity matrix as neutral element; it satisfies the distributive laws with respect to matrix addition. More precisely:*

LEMMA Properties of matrix mult.

(1) $\forall A \in \mathrm{Mat}(k\times l, K), B \in \mathrm{Mat}(l\times m, K), C \in \mathrm{Mat}(m\times n, K)$: $(AB)C = A(BC)$.

(2) $\forall A \in \mathrm{Mat}(m\times n, K)$: $I_m A = A = A I_n$.

(3) $\forall A \in \mathrm{Mat}(l\times m, K), B, C \in \mathrm{Mat}(m\times n, K)$: $A(B+C) = AB + AC$.

(4) $\forall A, B \in \mathrm{Mat}(l\times m, K), C \in \mathrm{Mat}(m\times n, K)$: $(A+B)C = AC + BC$.

(5) *Additionally,* $\forall \lambda \in K, A \in \mathrm{Mat}(l\times m, K), B \in \mathrm{Mat}(m\times n, K)$: $\lambda(AB) = (\lambda A)B = A(\lambda B)$.

In particular, $\mathrm{Mat}(n, K)$ *is a ring with the addition and multiplication of matrices.*

Proof. This is an immediate translation of the corresponding statements for linear maps; compare the proof of Theorem 13.25 (and note that the proofs there also work in the slightly more general situation needed here). Alternatively, it is easy (but somewhat tedious) to check the claims directly. ❑

DEF Matrix ring

18.10. **Definition.** The Ring $\mathrm{Mat}(n, K)$ is the *matrix ring (of size n over K)*. A matrix $A \in \mathrm{Mat}(n, K)$ is *invertible*, if there is a matrix $B \in \mathrm{Mat}(n, K)$ such that $AB = I_n$. Then we also have $BA = I_n$; we write A^{-1} for B and call A^{-1} the *inverse* of A. ♢

DEF invertible matrix

Let $f, g\colon K^n \to K^n$ be the linear maps corresponding to A, B. Then $AB = I_n$ means that $f \circ g = \mathrm{id}_{K^n}$. It follows that f is surjective and therefore an isomorphism (compare Corollary 13.14); then $g = f^{-1}$, which also implies $g \circ f = \mathrm{id}_{K^n}$, i.e., $BA = I_n$. The matrix $B = A^{-1}$ therefore is the matrix associated to f^{-1}.

EXAMPLE

18.11. **Example.** The matrix $A = \begin{pmatrix} 1 & \lambda \\ 0 & 1 \end{pmatrix} \in \mathrm{Mat}(2, K)$ (with $\lambda \in K$ arbitrary) is invertible since

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

We will learn in the next section how we can compute bases of the kernel and the image of a linear map $f\colon K^n \to K^m$ from the associated matrix. We will also see how we can determine whether a given matrix is invertible and how to compute its inverse when it is.

19. Normal Form and Linear Systems of Equations

Date: March 5, 2026

Let $f\colon K^n \to K^m$ be a linear map given by its matrix $A \in \mathrm{Mat}(m \times n, K)$. How can we determine a basis of $\ker(f)$? (This will also tell us what the dimension of the kernel is and therefore also determine the rank by the Rank-Nullity Theorem 13.17.)

The approach we will use is based on a general principle that is also useful in other parts of mathematics. The idea is to change the given matrix A to obtain another matrix A' such that

(1) both matrices (i.e., the corresponding linear maps) have the same kernel, and

(2) we can read off a basis of the kernel easily from A'.

We will first define the kind of matrix that allows us to read off a basis of the kernel. Then we will introduce small transformation steps that allow us to modify a matrix without changing its kernel, and finally, we will show that we can use a sequence of such steps to change any given matrix into one with the nice form.

DEF Row echelon form

19.1. **Definition.** Let K be a field, $m, n \in \mathbb{N}$, and $A = (a_{ij}) \in \mathrm{Mat}(m \times n, K)$. The matrix A is in *row echelon form*, if it has the following form (a star $*$ marks an arbitrary entry below).

$$A = \begin{pmatrix} 0\cdots0 & \mathbf{1} & *\cdots* & * & *\cdots* & * & *\cdots* \\ 0\cdots0 & 0 & 0\cdots0 & \mathbf{1} & *\cdots* & * & *\cdots* \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0\cdots0 & 0 & 0\cdots0 & 0 & 0\cdots0 & \mathbf{1} & *\cdots* \\ 0\cdots0 & 0 & 0\cdots0 & 0 & 0\cdots0 & 0 & 0\cdots0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0\cdots0 & 0 & 0\cdots0 & 0 & 0\cdots0 & 0 & 0\cdots0 \end{pmatrix}$$

More formally, this means that there exist $0 \le r \le m$ and indices $1 \le j_1 < j_2 < \cdots < j_r \le n$ such that

(1) $a_{ij} = 0$ if $i > r$,

(2) $a_{ij} = 0$ if $i \le r$ and $j < j_i$, and

(3) $a_{ij_i} = 1$ for all $i \in \{1, 2, \ldots, r\}$.

(So r is the number of rows that do not consist entirely of zeros and $j_1, j_2, \ldots, j_r$ are the indices of the columns that contain the *leading ones* of the first r rows.)

A is in *reduced row echelon form*, if in addition $a_{ij_k} = 0$ for all $1 \le i < k$ and all $k \in \{1, 2, \ldots, r\}$, i.e.,

$$A = \begin{pmatrix} 0\cdots0 & \mathbf{1} & *\cdots* & 0 & *\cdots* & 0 & *\cdots* \\ 0\cdots0 & 0 & 0\cdots0 & \mathbf{1} & *\cdots* & 0 & *\cdots* \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0\cdots0 & 0 & 0\cdots0 & 0 & 0\cdots0 & \mathbf{1} & *\cdots* \\ 0\cdots0 & 0 & 0\cdots0 & 0 & 0\cdots0 & 0 & 0\cdots0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0\cdots0 & 0 & 0\cdots0 & 0 & 0\cdots0 & 0 & 0\cdots0 \end{pmatrix}.$$

(The leading ones of the first r rows (in columns $j_1, j_2, \ldots, j_r$) are set in boldface above, and the rows containing them have been given a different color.) ♢

We transfer some notions from linear maps to their matrices.

19.2. **Definition.** Let $A \in \mathrm{Mat}(m \times n, K)$. Then the *image* of A, $\mathrm{im}(A)$, is the image of the corresponding linear map $f\colon K^n \to K^m$, the *rank* of A, $\mathrm{rk}(A)$, is the rank of f, and the *kernel* of A, $\ker(A)$, is the kernel of f. ♢

DEF Image, rank, kernel of a matrix

Since the images under f of the standard basis vectors of K^n, which generate the image of f, are given by the columns of A, the image $\mathrm{im}(A)$ of A is the linear hull of the columns of A. This is why $\mathrm{im}(A)$ is also called the *column space of A*. The *row space of A* is defined analogously as the linear hull of the rows of A in K^n.

DEF Column space row space

We now describe how we can obtain a basis of the kernel of A when A is in reduced row echelon form.

19.3. **Lemma.** *Let $A \in \mathrm{Mat}(m \times n, K)$ be in row echelon form with r and $j_1, j_2, \dots, j_r$ as in Definition 19.1. Then $\mathrm{rk}(A) = r$. If the matrix is in reduced row echelon form, then we obtain a basis of $\ker(A)$ as follows.*
Let $J = \{1, 2, \dots, n\} \setminus \{j_1, j_2, \dots, j_r\}$ be the set of indices of columns of A without a leading one. For $j \in J$ define the vector $b_j \in K^n$ as

$$b_j = \mathbf{e}_j - \sum_{i=1}^{r} a_{ij} \mathbf{e}_{j_i} .$$

Then $(b_j)_{j \in J}$ is a basis of $\ker(A)$.

LEMMA Rank and kernel of a matrix in row echelon form

Proof. The image of the linear map f corresponding to A is generated by the columns of A. The columns with numbers $j_1, j_2, \dots, j_r$ are linearly independent:

$$\lambda_1(1, 0, \dots, 0) + \lambda_2(*, 1, 0, \dots, 0) + \dots + \lambda_r(*, \dots, *, 1, 0, \dots, 0) = (0, \dots, 0)$$

successively implies $\lambda_r = 0$, …, $\lambda_2 = 0$, $\lambda_1 = 0$. Also, all columns are contained in the linear subspace $\langle \mathbf{e}'_1, \mathbf{e}'_2, \dots, \mathbf{e}'_r \rangle$ of K^m of dimension r. This implies that $\mathrm{im}(f) = \langle \mathbf{e}'_1, \mathbf{e}'_2, \dots, \mathbf{e}'_r \rangle$; therefore $\mathrm{rk}(A) = \mathrm{rk}(f) = r$.

Now assume that the matrix is in reduced row echelon form. By the Rank-Nullity Theorem 13.17, we know that $\dim \ker(A) = n - r = \#J$. It is therefore sufficient to show that the vectors b_j are in the kernel of A and that $(b_j)_{j \in J}$ is linearly independent. We write A_j for the jth column of A. Then $A_j = \sum_{i=1}^{r} a_{ij} \mathbf{e}'_i$ (here we use that $a_{ij} = 0$ for $i > r$), and $A_{j_i} = \mathbf{e}'_i$ for $1 \le i \le r$ (this is because A is in reduced row echelon form). Let $j \in J$. To see that b_j is in the kernel of A, we compute

$$f(b_j) = A_j - \sum_{i=1}^{r} a_{ij} A_{j_i} = \sum_{i=1}^{r} a_{ij} \mathbf{e}'_i - \sum_{i=1}^{r} a_{ij} \mathbf{e}'_i = \mathbf{0} .$$

To see that $(b_j)_{j \in J}$ is linearly independent, we consider a linear combination

$$\mathbf{0} = \sum_{j \in J} \lambda_j b_j = \sum_{j \in J} \lambda_j \mathbf{e}_j - \sum_{i=1}^{r} \Big(\sum_{j \in J} a_{ij} \lambda_j \Big) \mathbf{e}_{j_i} .$$

Since $(\mathbf{e}_1, \mathbf{e}_2, \dots, \mathbf{e}_n)$ is a basis of K^n and $\{1, 2, \dots, n\}$ is the disjoint union of J and $\{j_1, j_2, \dots, j_r\}$, we see by considering the coefficients of $\mathbf{e}_j$ for $j \in J$ that $\lambda_j = 0$ for all $j \in J$. ❑

A more down-to-earth way of obtaining the basis of the kernel is the following. The indices in J (corresponding to the columns without a leading one) denote positions in a vector in $\ker(A) \subset K^n$ such that we can choose the corresponding entries in any way we like; the remaining entries are then uniquely determined. To obtain a basis of $\ker(A)$, for each of these positions $j \in J$, we set the jth entry

to 1 and the others (indices in $J \setminus \{j\}$) to 0. We then solve the equations obtained from $Ab_j = \mathbf{0}$ for the remaining entries.

EXAMPLE Basis of kernel

19.4. **Example.** Let $K = \mathbb{R}$ and let A be the following matrix over $\mathbb{R}$:

$$A = \begin{pmatrix} 0 & \mathbf{1} & 2 & 0 & 0 & -2 \\ 0 & 0 & 0 & \mathbf{1} & 0 & 1 \\ 0 & 0 & 0 & 0 & \mathbf{1} & 5 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

Then A is in reduced row echelon form with $r = 3$ (the number of nonzero rows) and $j_1 = 2$, $j_2 = 4$, $j_3 = 5$. Therefore the rank is 3, $J = \{1, 3, 6\}$, and a basis of $\ker(A)$ is given by

$$b_1 = (\mathbf{1}, 0, \mathbf{0}, 0, 0, \mathbf{0}), \quad b_3 = (\mathbf{0}, -2, \mathbf{1}, 0, 0, \mathbf{0}), \quad b_6 = (\mathbf{0}, 2, \mathbf{0}, -1, -5, \mathbf{1}).$$

The components we can choose freely (positions 1, 3, 6) are marked in boldface. The remaining components of b_j are obtained from the negatives of the first r entries of the jth column of A.

It becomes clearer what happens behind the scenes when we write down the condition "$(x_1, x_2, \ldots, x_6) \in \ker(A)$" explicitly. This leads to the system of equations

$$\begin{aligned} x_2 + 2x_3 - 2x_6 &= 0 \\ x_4 + x_6 &= 0 \\ x_5 + 5x_6 &= 0 \end{aligned}$$

The fact that A is in reduced row echelon form implies that one can solve this system of equations for x_2, x_4, x_5:

$$\begin{aligned} x_2 &= -2x_3 + 2x_6 \\ x_4 &= -x_6 \\ x_5 &= -5x_6 \end{aligned}$$

We can therefore choose x_1, x_3, x_6 arbitrarily and then determine x_2, x_4, x_5. We obtain a basis of the kernel by letting (x_1, x_3, x_6) run through the standard basis of $K^3 = K^{n-j}$. ♣

Now we will introduce small modifications on a matrix that do not change its kernel. We can then chain many such small modifications; we will show afterwards that by doing this, we can transform any given matrix into reduced row echelon form.

DEF (Elementary) row operations

19.5. **Definition.** Let K be a field, $m, n \in \mathbb{N}$, and $A \in \mathrm{Mat}(m \times n, K)$.

(1) An *elementary row operation of type I* on the matrix A multiplies the ith row of A by λ. Here $i \in \{1, 2, \ldots, m\}$ and $\lambda \in K \setminus \{0\}$. We write $\mathbf{I}_i(\lambda)$ for this operation.

(2) An *elementary row operation of type II* on the matrix A adds the (scalar) λ-multiple of the jth row of A to the ith row of A. Here $i, j \in \{1, 2, \ldots, m\}$ with $i \neq j$ and $\lambda \in K$. We write $\mathbf{II}_{i,j}(\lambda)$ for this operation.

(3) An *elementary row operation of type III* on the matrix A swaps the ith and the jth rows of A. Here $i, j \in \{1, 2, \ldots, m\}$ with $i \neq j$. We write $\mathbf{III}_{i,j}$ for this operation.

A *row operation* on A is a chain of successive elementary row operations, beginning with the matrix A. ♢

The effect of an elementary row operation of type III can be obtained by a succession of suitable operations of types I and II (exercise). So introducing this type of operation is not strictly necessary, but it provides a convenient abbreviation.

19.6. **Lemma.** *Let K be a field, $m, n \in \mathbb{N}$, and $A \in \mathrm{Mat}(m \times n, K)$. Let further A' be a matrix that is obtained from A by an elementary row operation. Then $\ker(A') = \ker(A)$ and therefore also $\mathrm{rk}(A') = \mathrm{rk}(A)$. In addition, A' and A have the same row space.*

LEMMA Row operations preserve kernel, rank, row space

Proof. A vector $v = (x_1, x_2, \ldots, x_n) \in K^n$ is in the kernel of $A = (a_{ij})$ if and only if for all $i \in \{1, 2, \ldots, m\}$ we have $\sum_{j=1}^{n} a_{ij} x_j = 0$. An elementary row operation of type I replaces one of these equations by its λ-multiple for a $\lambda \neq 0$; this gives an equivalent equation. When we perform an elementary row operation of type II, then we add to one of the equations the λ-multiple of another equation. This implies that the new equations hold when the old ones did. Since we can undo this operation (by subtracting the λ-multiple of the other equation again), we see that in fact the new equations are equivalent to the old ones. (We do not need to consider operations of type III, but since they just change the order of the equations, it is clear that the solution set is unchanged.) It follows that $v \in \ker(A)$ if and only if $v \in \ker(A')$ as claimed. The equality of the ranks then follows from the Rank-Nullity Theorem 13.17.

Elementary row operations replace one row by a linear combination of some rows of the matrix. This linear combination is (by definition) in the row space of the original matrix, so the new row space is contained in the old one. Since we can undo elementary row operations by performing another elementary row operation, we also obtain the reverse inclusion and therefore equality of the row spaces. ❑

An easy induction (on the number of elementary row operations in the chain) then shows that kernel, rank, and row space are also preserved under arbitrary row operations.

Note that the *rank* is preserved under row operations, but the *image* of the matrix usually changes!

We now show that every matrix can be transformed by row operations into a matrix in reduced row echelon form.

19.7. **Theorem.** *Let K be a field, $m, n \in \mathbb{N}$, and $A \in \mathrm{Mat}(m \times n, K)$. Then A can be transformed into a matrix A' in reduced row echelon form by a sequence of elementary row operations.*

THM Normal form of matrices

Proof. We give an *algorithm* that produces a suitable sequence of elementary row operations. We proceed from left to right. At the end of the body of the main loop below, the submatrix consisting of the first k columns of the matrix will be in reduced row echelon form, with associated data r and $j_1, \ldots, j_r$ as in Definition 19.1.

1. Set $r \leftarrow 0$.
2. For $k = 1, \ldots, n$, do the following:
   if there is $r < i \leq m$ such that $a_{ik} \neq 0$, then

a. Set $r \leftarrow r+1$ (new nonzero row) and $j_r \leftarrow k$ (note the column index).
b. If $i \neq r$, then carry out row operation $\mathbf{III}_{r,i}$.
c. (Now $a_{rk} \neq 0$.) If $a_{rk} \neq 1$, then carry out row operation $\mathbf{I}_r(a_{rk}^{-1})$.
d. (Now $a_{rk} = 1$.) For $i \in \{1, 2, \ldots, m\} \backslash \{r\}$, carry out row operation $\mathbf{II}_{i,r}(-a_{ik})$ ("clean out column k"). (Now column k is $(0, \ldots, 0, 1, 0, \ldots, 0)$ with the 1 in position r.)

Assuming that at the start of the loop, the first $k-1$ columns are in reduced row echelon form with parameters r and $j_1, \ldots, j_r$ (note that this is trivially true when $k = 1$), we see that the operations carried out in the loop body will modify the matrix in such a way that now its first k columns are in reduced row echelon form, updating the parameters accordingly. Note that when $a_{ik} = 0$ for all $r < i \leq m$, nothing needs to be done, as in this case the first k columns are already in reduced row echelon form with the current parameters. After the last pass through the loop, $k = n$, and so the full matrix is in reduced row echelon form as desired. ❑

This algorithm is best understood by carrying it out for a concrete matrix.

EXAMPLE
Obtaining reduced row echelon form

19.8. **Example.** We determine the reduced row echelon form of the following matrix over $\mathbb{R}$.

$$A = \begin{pmatrix} 1 & 2 & 3 & 4 \\ 5 & 6 & 7 & 8 \\ 9 & 10 & 11 & 12 \end{pmatrix}$$

We begin with $k = r = 0$ in the algorithm.

- $k = 1$: We have $a_{11} = 1$, so we set $r \leftarrow 1$ and $j_1 \leftarrow 1$. We can skip steps 2b and 2c. To clean out the first column in step 2d, we subtract 5 times the first row from the second and 9 times the first row from the third. This gives the new matrix
$$\begin{pmatrix} \mathbf{1} & 2 & 3 & 4 \\ 0 & -4 & -8 & -12 \\ 0 & -8 & -16 & -24 \end{pmatrix}.$$
- $k = 2$: We have $a_{22} = -4$, so we set $r \leftarrow 2$ and $j_2 \leftarrow 2$. We skip step 2b; then we multiply the second row by $-1/4$ in step 2c. To clean out the second column, we subtract 2 times the second row from the first and add 8 times the second row to the third. We obtain the matrix
$$\begin{pmatrix} \mathbf{1} & 0 & -1 & -2 \\ 0 & \mathbf{1} & 2 & 3 \\ 0 & 0 & 0 & 0 \end{pmatrix}.$$
- $k = 3$: The only relevant entry a_{33} is zero, so we do nothing.
- $k = 4$: Again, the only relevant entry a_{34} is zero, and we do nothing.

So we obtain the reduced row echelon form

$$A' = \begin{pmatrix} \mathbf{1} & 0 & -1 & -2 \\ 0 & \mathbf{1} & 2 & 3 \\ 0 & 0 & 0 & 0 \end{pmatrix}.$$

We can now read off a basis of $\ker(A) = \ker(A')$ following Lemma 19.3:

$$b_3 = (1, -2, \mathbf{1}, \mathbf{0}) \qquad \text{and} \qquad b_4 = (2, -3, \mathbf{0}, \mathbf{1}). \qquad \clubsuit$$

Elementary row operations can be performed through multiplying by certain invertible matrices on the left. These matrices are the so-called *elementary matrices* $E_i(\lambda)$ with

$\lambda \in K^\times$ and $i \in \{1, 2, \dots, m\}$ and $E_{ij}(\lambda)$ with $\lambda \in K$ and $i, j \in \{1, 2, \dots, m\}$, $i \neq j$. To define them, we introduce $M_{kl} = (\delta_{ik}\delta_{jl})_{1 \leq i,j \leq m}$; this is a matrix all of whose entries are zero except the (k, l) entry, which is 1. (These matrices M_{kl} correspond to the basis elements ϕ_{kl} of $\mathrm{Hom}(K^m, K^m)$ as in Corollary 13.23.) Then we set

$$E_i(\lambda) = I_m + (\lambda - 1)M_{ii} \qquad \text{and} \qquad E_{ij}(\lambda) = I_m + \lambda M_{ij}\,.$$

The matrix $E_i(\lambda)$ differs from the identity matrix I_m just in the ith position on the diagonal, where it has the entry λ instead of 1. In $E_{ij}(\lambda)$ we have the entry λ at position (i, j) (away from the diagonal). Since

$$E_i(\lambda)E_i(\lambda^{-1}) = I_m \qquad \text{and} \qquad E_{ij}(\lambda)E_{ij}(-\lambda) = I_m\,,$$

these elementary matrices are invertible. Now what is the effect of multiplying on the left by an elementary matrix? To see this, note that

$$M_{kl}A = \Big(\sum_{h=1}^{m} \delta_{ik}\delta_{hl}a_{hj}\Big)_{1 \leq i \leq m, 1 \leq j \leq n} = (\delta_{ik}a_{lj})_{1 \leq i \leq m, 1 \leq j \leq n}\,;$$

in this matrix all rows are zero except for the kth row, which contains the lth row of A. So multiplying on the left by M_{kl} puts the lth row of A into the kth row and clears out all other rows.

We therefore see that the rows of $E_i(\lambda)A$ coincide with the corresponding rows of A, except for the ith row, which gets multiplied by λ. This is exactly the effect of elementary row operation $\mathbf{I}_i(\lambda)$. In the same way, the rows of $E_{ij}(\lambda)A$ agree with those of A, except for the ith row, which gets the λ-multiple of the jth row added to it. This is the effect of elementary row operation $\mathbf{II}_{i,j}(\lambda)$. We demonstrate this for the 2×3 matrix $A = \begin{pmatrix} a & b & c \\ d & e & f \end{pmatrix}$:

$$E_1(\lambda)A = \begin{pmatrix} \lambda & 0 \\ 0 & 1 \end{pmatrix}\begin{pmatrix} a & b & c \\ d & e & f \end{pmatrix} = \begin{pmatrix} \lambda a & \lambda b & \lambda c \\ d & e & f \end{pmatrix}$$

$$E_2(\lambda)A = \begin{pmatrix} 1 & 0 \\ 0 & \lambda \end{pmatrix}\begin{pmatrix} a & b & c \\ d & e & f \end{pmatrix} = \begin{pmatrix} a & b & c \\ \lambda d & \lambda e & \lambda f \end{pmatrix}$$

$$E_{12}(\lambda)A = \begin{pmatrix} 1 & \lambda \\ 0 & 1 \end{pmatrix}\begin{pmatrix} a & b & c \\ d & e & f \end{pmatrix} = \begin{pmatrix} a + \lambda d & b + \lambda e & c + \lambda f \\ d & e & f \end{pmatrix}$$

$$E_{21}(\lambda)A = \begin{pmatrix} 1 & 0 \\ \lambda & 1 \end{pmatrix}\begin{pmatrix} a & b & c \\ d & e & f \end{pmatrix} = \begin{pmatrix} a & b & c \\ d + \lambda a & e + \lambda b & f + \lambda c \end{pmatrix}$$

We can therefore interpret Theorem 19.7 as saying that for every matrix $A \in \mathrm{Mat}(m \times n, K)$ there is an invertible matrix $P \in \mathrm{Mat}(m, K)$ such that PA is in reduced row echelon form, where P is a product of elementary matrices.

Now consider an invertible matrix $A \in \mathrm{Mat}(m, K)$. If we apply this statement to A^{-1} and use the fact that the reduced row echelon form of an invertible $m \times m$ matrix is the identity matrix I_m (see Lemma 19.18 below), then we obtain a product P of elementary matrices such that $PA^{-1} = I_m$. So $A = P$. This proves the following.

Theorem. *Every invertible matrix is a product of elementary matrices.*

THM Elementary matrices generate invertible matrices

This implies:

Two matrices $A, B \in \mathrm{Mat}(m \times n, K)$ can be transformed into each other by row operations if and only if there is an invertible matrix $P \in \mathrm{Mat}(m, K)$ such that $B = PA$.

We can define *column operations* in a completely analogous way. They are performed through multiplying on the *right* by elementary matrices. One then obtains the following analogous statement.

Two matrices $A, B \in \mathrm{Mat}(m \times n, K)$ can be transformed into each other by column operations if and only if there is an invertible matrix $Q \in \mathrm{Mat}(n, K)$ such that $B = AQ$.

We have on occasion spoken of "the" reduced row echelon form of a matrix. Indeed, it turns out that the result of any procedure that transforms a matrix into reduced row echelon form using row operations is uniquely determined, as shown by the theorem below. So in the end, it does not matter at all which row operations one performs in which order, as long as one obtains a matrix in reduced row echelon form in the end (and did not make mistakes on the way).

Theorem. *If $A, B \in \mathrm{Mat}(m \times n, K)$ are two matrices in reduced row echelon form and with the same row space, then $A = B$.*

THM Uniqueness of reduced row echelon form

Proof. Let $U \subset K^n$ be the common row space of A and B. For $0 \leq k \leq n$ we define the linear subspace

$$V_k = \{(x_1, \ldots, x_n) \in K^n \mid x_1 = x_2 = \ldots = x_k = 0\} \subset K^n$$

and $d_k = \dim(U \cap V_k)$. Then $d_0 = r = \dim U$, $d_n = 0$ and $d_k - 1 \leq d_{k+1} \leq d_k$. So there must be exactly r "jumps" j_i such that $d_{j_i - 1} = r + 1 - i$ and $d_{j_i} = r - i$ for $i \in \{1, 2, \ldots, r\}$. The definition of row echelon form implies that j_i must be the position of the leading one in the ith row of A and of B. The linear map

$$\phi \colon U \longrightarrow K^r, \qquad (x_1, x_2, \ldots, x_n) \longmapsto (x_{j_1}, x_{j_2}, \ldots, x_{j_r})$$

then is an isomorphism, and the first r rows of A and of B must be the preimages $\phi^{-1}(\mathbf{e}_1), \phi^{-1}(\mathbf{e}_2), \ldots, \phi^{-1}(\mathbf{e}_r)$ of the standard basis vectors of K^r. So A and B have the same rows. ❑

We now consider linear equations and systems of equations.

19.9. **Definition.** Let V and W be two K-vector spaces and $f \colon V \to W$ a linear map. If $b \in W$ is a given vector, then the equation $f(x) = b$, whose solutions $x \in V$ are to be determined, is a *linear equation*. The equation is *homogeneous*, if $b = \mathbf{0}$, otherwise *inhomogeneous*.

DEF Linear equation

If $V = K^n$ and $W = K^m$, then the equation can also be written in terms of the matrix $A \in \mathrm{Mat}(m \times n, K)$ associated to f as $A\boldsymbol{x} = \boldsymbol{b}$ with column vectors $\boldsymbol{x} \in K^n$ and $\boldsymbol{b} \in K^m$. In this case one also calls this a *system of linear equations* (with m equations in n unknowns). ♢

We can already describe the structure of the solution set of a linear equation in some detail.

19.10. **Theorem.** *Let V and W be two K-vector spaces and $f \colon V \to W$ a linear map.*

THM Solution set of a linear equation

(1) *The solution set of a homogeneous linear equation $f(x) = \mathbf{0}$ is a linear subspace of V, namely, the kernel of f.*

(2) *Let $\mathbf{0} \neq b \in W$. If $b \notin \mathrm{im}(f)$, then the inhomogeneous linear equation $f(x) = b$ has no solution. Otherwise let $x_0 \in V$ with $f(x_0) = b$. Then the solution set is given by $x_0 + \ker(f) = \{x_0 + v \mid v \in \ker(f)\}$.*

Proof. The first claim follows directly from the definition of the kernel and the fact that $\ker(f)$ is a linear subspace of V. In the second claim it is clear that there are solutions if and only if $b \in \operatorname{im}(f)$ (by the definition of $\operatorname{im}(f)$). It remains to show the last statement. Let $x \in V$. Then

$$\begin{aligned} f(x) = b &\iff f(x) = f(x_0) \iff f(x - x_0) = \mathbf{0} \\ &\iff x - x_0 \in \ker(f) \iff x \in x_0 + \ker(f)\,. \end{aligned}$$

❑

The general recipe for solving a linear equation $f(x) = b$ therefore is as follows.

(1) Check whether $b \in \operatorname{im}(f)$. If this is not the case, there are no solutions.

(2) Determine a "particular solution" $x_0 \in V$.

(3) Determine $\ker(f)$.

(4) The solution set is $x_0 + \ker(f)$. If $\ker(f)$ is finite-dimensional with basis $(x_1, x_2, \ldots, x_n)$, then the "general solution" is

$$x = x_0 + \lambda_1 x_1 + \lambda_2 x_2 + \ldots + \lambda_n x_n$$

with $\lambda_1, \lambda_2, \ldots, \lambda_n \in K$.

The first two steps frequently go together, as determining that $b \in \operatorname{im}(f)$ will usually involve finding a preimage.

In the homogeneous case $b = \mathbf{0}$ we always have $b \in \operatorname{im}(f)$, and so we can take $x_0 = \mathbf{0}$; the solution set is simply $\ker(f)$.

EXAMPLE
Inhomogeneous linear differential equation

19.11. **Example.** We consider the following inhomogeneous linear first order differential equation.

$$y'(x) + y(x) = x\,,$$

to be solved for $y \in \mathcal{C}^1(\mathbb{R})$. Here $K = \mathbb{R}$, $V = \mathcal{C}^1(\mathbb{R})$, $W = \mathcal{C}(\mathbb{R})$ and $f\colon y \mapsto y' + y$. Written in the form as in Definition 19.9, the equation is $f(y) = \mathrm{id}_{\mathbb{R}}$. We first look for a particular solution. With some trial and error, we find $y_0(x) = x - 1$. (There are more systematic ways of doing that, but this is beyond this course.) The next step is to determine the kernel of f, which is the set of all functions y such that $y' + y = \mathbf{0}$. I claim that $\ker(f) = \langle x \mapsto e^{-x} \rangle$, i.e., the functions y satisfying $y' + y = \mathbf{0}$ are exactly those of the form $y(x) = Ce^{-x}$ with $C \in \mathbb{R}$. To see this, we consider $z(x) = e^x y(x)$, which should be constant, so we look at its derivative,

$$z'(x) = e^x y(x) + e^x y'(x) = e^x \big(y(x) + y'(x)\big) = \mathbf{0}\,.$$

As a function whose derivative is constant zero, z must be constant, so there is $C \in \mathbb{R}$ with $z(x) = C$ for all x, and so $y(x) = Ce^{-x}$. Conversely, it is easy to see that these functions are indeed solutions of $y' + y = \mathbf{0}$. The general solution of our differential equation therefore has the shape

$$y(x) = x - 1 + Ce^{-x}\,, \qquad C \in \mathbb{R}\,.$$

♣

What does the recipe described above look like concretely when we want to solve a linear system of equations? So let $A\boldsymbol{x} = \boldsymbol{b}$ be a linear system of equations with $A \in \operatorname{Mat}(m \times n, K)$. In the homogeneous case $\boldsymbol{b} = \mathbf{0}$ the solution set is the kernel of A, so we need to determine a basis of $\ker(A)$. We know how to do that: we transform A into reduced row echelon form and then we can read off a basis of the kernel as in Lemma 19.3. In the inhomogeneous case we let $A' = (A \mid \boldsymbol{b})$ be the *extended matrix* of the system; it is obtained by adding the column vector $\boldsymbol{b}$ as an $(n+1)$th column.

DEF
extended matrix

THM
Inhomogeneous linear system

19.12. **Theorem.** *Let K be a field, let $m, n \in \mathbb{N}$, let $A \in \mathrm{Mat}(m \times n, K)$, and let $\boldsymbol{b} \in K^m$ be a column vector. Let further $A' = (A \mid \boldsymbol{b})$ be the extended matrix of the system of linear equations $A\boldsymbol{x} = \boldsymbol{b}$. Then*

$$\boldsymbol{b} \in \mathrm{im}(A) \iff \mathrm{rk}(A') = \mathrm{rk}(A)\,.$$

This condition can be checked by transforming A' into reduced row echelon form $\tilde{A}'$. Then $\mathrm{rk}(A') = \mathrm{rk}(A)$ is equivalent with the property that the last column of $\tilde{A}'$ does not contain a leading one of a row (in terms of the notation from Definition 19.1, this means $j_r \leq n$). In this case we can read off a particular solution of $A\boldsymbol{x} = \boldsymbol{b}$ from $\tilde{A}'$ as follows. Assume that last column of $\tilde{A}'$ is $(\tilde{b}_1, \ldots, \tilde{b}_r, 0, \ldots, 0)$. Then

$$\boldsymbol{x}_0 = \sum_{i=1}^{r} \tilde{b}_i \mathbf{e}_{j_i}$$

is a solution of the linear system.

Proof. Let $A_1, \ldots, A_n$ be the columns of A. We have the following equivalences.

$$\begin{aligned} \boldsymbol{b} \in \mathrm{im}(A) = \langle A_1, \ldots, A_n \rangle &\iff \langle A_1, \ldots, A_n, \boldsymbol{b} \rangle = \langle A_1, \ldots, A_n \rangle \\ &\iff \mathrm{im}(A') = \mathrm{im}(A)\,. \end{aligned}$$

The last statement implies that $\mathrm{rk}(A') = \mathrm{rk}(A)$. We always have that $\mathrm{im}(A) \subset \mathrm{im}(A')$, so $\mathrm{rk}(A') = \mathrm{rk}(A)$ implies that both images agree. This shows the first claim.

Let now $\tilde{A}'$ be the reduced row echelon form of A'. Then the first n columns of $\tilde{A}'$ give the reduced row echelon form $\tilde{A}$ of A. The rank of A' is strictly larger than the rank of A if and only if $\tilde{A}'$ has more nonzero rows than $\tilde{A}$. But this exactly means that the last column of $\tilde{A}'$ contains a leading one. This shows the second claim. For the last claim, we observe that the row operations that are performed while transforming A' into its reduced row echelon form $\tilde{A}'$ replace the original equations by equivalent ones. Writing $\tilde{A}' = (\tilde{A} \mid \tilde{\boldsymbol{b}})$, we therefore see that the linear system $\tilde{A}\boldsymbol{x} = \tilde{\boldsymbol{b}}$ has the same solution set as the original linear system. Since the j_ith column of $\tilde{A}$ contains the standard basis vector $\mathbf{e}'_i$, we can check that, indeed,

$$\tilde{A}\boldsymbol{x}_0 = \sum_{i=1}^{r} \tilde{b}_i \tilde{A}\mathbf{e}_{j_i} = \sum_{i=1}^{r} \tilde{b}_i \mathbf{e}'_i = \tilde{\boldsymbol{b}}\,. \qquad \square$$

EXAMPLE
Linear system

19.13. **Example.** We solve the following linear system (for $K = \mathbb{Q}$ or $\mathbb{R}$).

$$\begin{array}{rcrcrcl} x_1 & & & - & x_3 & - & 2x_4 & = & 3 \\ -x_1 & + & x_2 & + & x_3 & & & = & -2 \\ & & x_2 & & & - & x_4 & = & 0 \end{array}$$

or in matrix notation,

$$\begin{pmatrix} 1 & 0 & -1 & -2 \\ -1 & 1 & 1 & 0 \\ 0 & 1 & 0 & -1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} = \begin{pmatrix} 3 \\ -2 \\ 0 \end{pmatrix}\,.$$

The extended matrix is

$$A' = \begin{pmatrix} 1 & 0 & -1 & -2 & 3 \\ -1 & 1 & 1 & 0 & -2 \\ 0 & 1 & 0 & -1 & 0 \end{pmatrix}\,.$$

Its reduced row echelon form is obtained as

$$\tilde{A}' = \begin{pmatrix} \mathbf{1} & 0 & -1 & 0 & 1 \\ 0 & \mathbf{1} & 0 & 0 & -1 \\ 0 & 0 & 0 & \mathbf{1} & -1 \end{pmatrix}.$$

This corresponds to the linear system

$$\begin{aligned} x_1 \quad - \quad x_3 \quad &= \quad 1 \\ x_2 \quad &= -1 \\ x_4 \quad &= -1 \end{aligned}$$

This gives us the particular solution $\boldsymbol{x}_0 = (\mathbf{1}, -\mathbf{1}, 0, -\mathbf{1})$. In addition, we read off that $\operatorname{rk}(A) = 3$, $\dim \ker(A) = 4 - 3 = 1$, and a basis of $\ker(A)$ is given by $\boldsymbol{x}_1 = (1, 0, \mathbf{1}, 0)$. The general solution of the linear system therefore is

$$\boldsymbol{x} = \boldsymbol{x}_0 + \lambda_1 \boldsymbol{x}_1 = (1 + \lambda_1, -1, \lambda_1, -1)$$

with $\lambda_1 \in K$. ♣

Because it is so important, we recap the solution recipe once again.

(1) Set up the extended matrix A'.

(2) Transform A' into reduced row echelon form $\tilde{A}'$. Let $r = \operatorname{rk}(\tilde{A}')$ and let $1 \leq j_1 < j_2 < \ldots < j_r \leq n+1$ be the positions of the leading ones of the first r rows of $\tilde{A}'$.

(3) $j_r = n + 1 \Rightarrow$ no solutions. Otherwise:

(4) Let $J = \{1, 2, \ldots, n\} \setminus \{j_1, j_2, \ldots, j_r\}$ be the set of "free" positions. Set $x_j = \lambda_j \in K$ arbitrary for $j \in J$ and solve the linear system corresponding to the matrix $\tilde{A}'$ for x_{j_i}, $i \in \{1, 2, \ldots, r\}$. This gives the general solution.

In the example above we have $r = 3$, $j_1 = 1$, $j_2 = 2$, $j_3 = 4 < 5 = n + 1$, $J = \{3\}$. We therefore set $x_3 = \lambda$ and solve for x_1, x_2, x_4.

This solution method for linear systems (and its variants) is known as *Gaussian elimination*, after Carl Friedrich Gauß. (But in fact this or equivalent methods were in use throughout Europe and Asia long before Gauß.) One variant is to first transform the matrix into (non-reduced) row echelon form and then successively solve the resulting system "from below" by substitution. This version is slightly more efficient in terms of the number or calculation steps (and so is preferred when implementing the algorithm), but it is a bit more complicated when performing computations by hand.

We have defined the rank of a matrix as the rank of the associated linear map, equivalently, as the dimension of its column space. One should therefore more precisely call it the "column rank" since we can as well consider the dimension of the row space or "row rank". Fortunately, there is no difference.

19.14. **Theorem.** *Let K be a field, let $m, n \in \mathbb{N}$ and $A \in \operatorname{Mat}(m \times n, K)$. Then the row space and the column space of A have the same dimension.*

THM
Row rank = column rank

Proof. According to Lemma 19.6 and Theorem 19.7 we can assume that A is in reduced row echelon form. Let $r = \operatorname{rk}(A)$ be the dimension of the column space of A. Then A has exactly r rows that are not zero rows, and these rows are linearly independent since the matrix $(a_{i,j_k})_{1 \leq i,k \leq r}$ (where $j_1, \ldots, j_r$ are, as usual, the column indices of the leading ones) is the identity matrix I_r. So the r non-zero

rows form a linearly independent spanning family of the row space, which shows that the row space also has dimension $r = \mathrm{rk}(A)$. ❑

We see that the Normal Form Algorithm from Theorem 19.7 also gives us the dimension and a basis of the row space of its input matrix. We can use this to determine the dimension and/or a basis of the linear subspace spanned by vectors $v_1, \ldots, v_m \in K^n$: we write these vectors as rows into a matrix and then determine its (reduced) row echelon form. The non-zero rows of the resulting matrix then are a basis of this space.

We can state Theorem 19.14 above in a very concise way using the following definition.

19.15. **Definition.** Let K be a field, let $m, n \in \mathbb{N}$ and $A = (a_{ij}) \in \mathrm{Mat}(m \times n, K)$. The *transpose* of A is $A^\top = (a_{ji})_{1 \le i \le n, 1 \le j \le m} \in \mathrm{Mat}(n \times m, K)$. ♢

DEF Transposed matrix

Recall how to read the notation

$$A^\top = (a_{ji})_{1 \le i \le n, 1 \le j \le m} \ :$$

The first index below after the closing parenthesis (here i) is the row index and the second index (here j) is the column index. The matrix $A^\top$ therefore has n rows and m columns. The entry in row i and column j is a_{ji}, which is the same entry as in *column* i and *row* j of the matrix A. Equivalently, we could write

$$A^\top = (a_{ij})_{1 \le j \le n, 1 \le i \le m} \, .$$

In this case j is the row index and i is the column index.

The effect is to "reflect the matrix about its main diagonal".

19.16. **Example.**

$$\begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{pmatrix}^\top = \begin{pmatrix} 1 & 4 \\ 2 & 5 \\ 3 & 6 \end{pmatrix} \, .$$ ♣

EXAMPLE Transposed matrix

The row space of A is the column space of $A^\top$ and conversely. Theorem 19.14 simply says

$$\mathrm{rk}(A^\top) = \mathrm{rk}(A) \, .$$

The matrix transpose plays well with the other operations on matrices.

19.17. **Lemma.** *Let K be a field and let $l, m, n \in \mathbb{N}$.*

LEMMA Rules for $A^\top$

(1) *The map $\mathrm{Mat}(m \times n, K) \to \mathrm{Mat}(n \times m, K)$, $A \mapsto A^\top$ is an isomorphism (in particular, $(A+B)^\top = A^\top + B^\top$ and $(\lambda A)^\top = \lambda A^\top$ for $A, B \in \mathrm{Mat}(m \times n, K)$, $\lambda \in K$; bijectivity follows from the last point below).*

(2) *For $A \in \mathrm{Mat}(l \times m, K)$ and $B \in \mathrm{Mat}(m \times n, K)$ we have $(AB)^\top = B^\top A^\top$.*

(3) *For $A \in \mathrm{Mat}(m \times n, K)$ we have $(A^\top)^\top = A$.*

Proof. Exercise. ❑

To finish this section, we want to give answers to the following two questions.

Question 1: How can we find out whether a (square) matrix $A \in \mathrm{Mat}(n, K)$ is invertible or not?

Question 2: How can we compute the inverse A^{-1} of an invertible matrix $A \in \mathrm{Mat}(n, K)$?

The following lemma answers the first question.

LEMMA Reduced row echelon form of invertible matrix

19.18. **Lemma.** *Let K be a field, $n \in \mathbb{N}$ and $A \in \mathrm{Mat}(n, K)$ be a square matrix. The matrix A is invertible if and only if its reduced row echelon form is the identity matrix I_n.*

Proof. A is invertible if and only if the associated linear map $f\colon K^n \to K^n$ is an isomorphism. Since domain and codomain of f have the same finite dimension n, this is equivalent with f being surjective, i.e., with $\mathrm{rk}(A) = \mathrm{rk}(f) = n$ (compare Corollary 13.14). This means exactly that the reduced row echelon form of A has no zero rows; we have $r = n$ and therefore $j_1 = 1$, $j_2 = 2$, …, $j_n = n$. So for each j, the jth column is the jth standard basis vector; such a matrix is the identity matrix. ❑

Now we consider the second question. Note that if A is invertible, then any linear system of the form $A\boldsymbol{x} = \boldsymbol{b}$ has a unique solution, which is given by $\boldsymbol{x} = A^{-1}\boldsymbol{b}$. (The converse is also true: if this system has a unique solution for every $\boldsymbol{b}$, then A must be invertible. This is an exercise.) Taking the standard basis vector $\mathbf{e}_j$ for $\boldsymbol{b}$, this gives us as solution exactly the jth column of A^{-1}. This means that we can find A^{-1} by solving the n linear systems $A\boldsymbol{x} = \mathbf{e}_j$ for $j \in \{1, 2, \dots, n\}$. This can be done in parallel, as described in the next theorem.

THM Computation of A^{-1}

19.19. **Theorem.** *Let K be a field, $n \in \mathbb{N}$ and $A \in \mathrm{Mat}(n, K)$ a square matrix. Let further $A' = (A \mid I_n) \in \mathrm{Mat}(n \times 2n, K)$ and $\tilde{A}'$ its reduced row echelon form. The matrix A is invertible if and only if $\tilde{A}'$ has the form $(I_n \mid B)$; then $B = A^{-1}$.*

Proof. Let $\tilde{A}' = (\tilde{A} \mid B)$, then $\tilde{A}$ is the reduced row echelon form of A. By Lemma 19.18, A is invertible if and only if $\tilde{A} = I_n$. This shows the first claim.

The matrix A' represents the linear system $A(\boldsymbol{x}_1 \mid \boldsymbol{x}_2 \mid \cdots \mid \boldsymbol{x}_n) = (\mathbf{e}_1 \mid \mathbf{e}_2 \mid \cdots \mid \mathbf{e}_n)$, or in short form $AX = I_n$, where $X = (\boldsymbol{x}_1 \mid \boldsymbol{x}_2 \mid \cdots \mid \boldsymbol{x}_n) \in \mathrm{Mat}(n, K)$. The row operations leading to the reduced row echelon form produce the equivalent system $I_n X = B$, which shows that $X = B$ is the solution of $AX = I_n$; this means that $B = A^{-1}$. ❑

EXAMPLE Computation of A^{-1}

19.20. **Example.** Let $K = \mathbb{Q}$ and

$$A = \begin{pmatrix} -1 & -2 & 2 & 2 \\ 2 & 5 & -4 & -4 \\ -1 & -1 & 1 & 1 \\ 1 & 1 & 0 & 1 \end{pmatrix}.$$

We compute the reduced row echelon form of

$$A' = \left(\begin{array}{cccc|cccc} -1 & -2 & 2 & 2 & 1 & 0 & 0 & 0 \\ 2 & 5 & -4 & -4 & 0 & 1 & 0 & 0 \\ -1 & -1 & 1 & 1 & 0 & 0 & 1 & 0 \\ 1 & 1 & 0 & 1 & 0 & 0 & 0 & 1 \end{array}\right):$$

$$A' \xrightarrow{\mathbf{I}_1(-1);\mathbf{II}_{2,1}(-2),\mathbf{II}_{3,1}(1),\mathbf{II}_{4,1}(-1)} \begin{pmatrix} \mathbf{1} & 2 & -2 & -2 & -1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 2 & 1 & 0 & 0 \\ 0 & 1 & -1 & -1 & -1 & 0 & 1 & 0 \\ 0 & -1 & 2 & 3 & 1 & 0 & 0 & 1 \end{pmatrix}$$

$$\xrightarrow{\mathbf{II}_{1,2}(-2),\mathbf{II}_{3,2}(-1),\mathbf{II}_{4,2}(1)} \begin{pmatrix} \mathbf{1} & 0 & -2 & -2 & -5 & -2 & 0 & 0 \\ 0 & \mathbf{1} & 0 & 0 & 2 & 1 & 0 & 0 \\ 0 & 0 & -1 & -1 & -3 & -1 & 1 & 0 \\ 0 & 0 & 2 & 3 & 3 & 1 & 0 & 1 \end{pmatrix}$$

$$\xrightarrow{\mathbf{I}_3(-1);\mathbf{II}_{1,3}(2),\mathbf{II}_{4,3}(-2)} \begin{pmatrix} \mathbf{1} & 0 & 0 & 0 & 1 & 0 & -2 & 0 \\ 0 & \mathbf{1} & 0 & 0 & 2 & 1 & 0 & 0 \\ 0 & 0 & \mathbf{1} & 1 & 3 & 1 & -1 & 0 \\ 0 & 0 & 0 & 1 & -3 & -1 & 2 & 1 \end{pmatrix}$$

$$\xrightarrow{\mathbf{II}_{3,4}(-1)} \begin{pmatrix} \mathbf{1} & 0 & 0 & 0 & 1 & 0 & -2 & 0 \\ 0 & \mathbf{1} & 0 & 0 & 2 & 1 & 0 & 0 \\ 0 & 0 & \mathbf{1} & 0 & 6 & 2 & -3 & -1 \\ 0 & 0 & 0 & \mathbf{1} & -3 & -1 & 2 & 1 \end{pmatrix}$$

We read off that

$$A^{-1} = \begin{pmatrix} 1 & 0 & -2 & 0 \\ 2 & 1 & 0 & 0 \\ 6 & 2 & -3 & -1 \\ -3 & -1 & 2 & 1 \end{pmatrix}.$$

♣

20. Matrices and linear maps

Date: March 5, 2026

So far, we have considered matrices as associated to linear maps $K^n \to K^m$. However, it is only relevant that we fixed a basis each of the domain and codomain (which we took to be the standard bases). So we can associate to a K-linear map $f\colon V \to V'$ a matrix when we have fixed bases $B = (b_1, b_2, \dots, b_n)$ of V and $B' = (b'_1, b'_2, \dots, b'_m)$ of V'. Then there are uniquely determined scalars $a_{ij} \in K$ such that

$$f(b_j) = a_{1j}b'_1 + a_{2j}b'_2 + \ldots + a_{mj}b'_m$$

holds for all $j \in \{1, 2, \dots, n\}$.

DEF $\mathrm{Mat}_{B,B'}(f)$

20.1. **Definition.** In the situation described above, the matrix

$$\mathrm{Mat}_{B,B'}(f) = (a_{ij})_{1\le i\le m, 1\le j\le n} \in \mathrm{Mat}(m \times n, K)$$

is the *matrix of f relative to the bases B and B'*. ♢

In the same way as before, the jth column of the matrix contains the coefficients of the image $f(b_j)$ of the jth basis vector in B when expressed as a linear combination of the basis B'.

Alternatively, we can describe $\mathrm{Mat}_{B,B'}(f)$ as the matrix associated to the linear map $\phi_{B'}^{-1} \circ f \circ \phi_B\colon K^n \to K^m$.

$$\begin{array}{ccc} V & \xleftarrow{\phi_B} & K^n \\ {\scriptstyle f}\downarrow & & \downarrow{\scriptstyle \mathrm{Mat}_{B,B'}(f)} \\ V' & \xleftarrow{\phi_{B'}} & K^m \end{array}$$

Recall that $\phi_B\colon K^n \to V$ is the linear combination map coming from B (similarly for $\phi_{B'}\colon K^m \to V'$):

$$\phi_B(\lambda_1, \lambda_2, \dots, \lambda_n) = \lambda_1 b_1 + \lambda_2 b_2 + \ldots + \lambda_n b_n\,.$$

Since B and B' are bases, these linear combination maps ϕ_B and $\phi_{B'}$ are isomorphisms; in particular, the inverse isomorphism $\phi_{B'}^{-1}$ exists.

EXAMPLE Matrix of a linear map

20.2. **Example.** We consider $V = P_{<3}$, the real vector space of polynomial functions of degree < 3 and the linear map $D\colon V \to V$, $f \mapsto f'$. Let further $B = (x \mapsto 1, x \mapsto x, x \mapsto x^2)$ and $B' = (x \mapsto 1, x \mapsto x-1, x \mapsto (x-1)(x-2))$ denote two bases of V. Then:

$$\mathrm{Mat}_{B,B}(D) = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 2 \\ 0 & 0 & 0 \end{pmatrix}, \quad \mathrm{Mat}_{B',B}(D) = \begin{pmatrix} 0 & 1 & -3 \\ 0 & 0 & 2 \\ 0 & 0 & 0 \end{pmatrix},$$

$$\mathrm{Mat}_{B,B'}(D) = \begin{pmatrix} 0 & 1 & 2 \\ 0 & 0 & 2 \\ 0 & 0 & 0 \end{pmatrix} \quad \text{and} \quad \mathrm{Mat}_{B',B'}(D) = \begin{pmatrix} 0 & 1 & -1 \\ 0 & 0 & 2 \\ 0 & 0 & 0 \end{pmatrix}. \qquad ♣$$

We see that one and the same linear map can be described by many different matrices. What is the relation between these various matrices? We begin with a simple statement on compositions of linear maps.

20.3. **Lemma.** *Let $g\colon V \to V'$ and $f\colon V' \to V''$ be two K-linear maps between finite-dimensional vector spaces. Let further bases B of V, B' of V' and B'' of V'' be given. Then we have*

LEMMA Matrix of $f \circ g$

$$\mathrm{Mat}_{B,B''}(f \circ g) = \mathrm{Mat}_{B',B''}(f)\,\mathrm{Mat}_{B,B'}(g)\,.$$

Proof. This is an immediate consequence of the definition of matrix multiplication. ❑

We obtain the following.

20.4. **Corollary.** *Let $f\colon V \to V'$ be a K-linear map between finite-dimensional vector spaces. Let B and $\tilde{B}$ be two bases of V and B' and $\tilde{B}'$ two bases of V'. Then*

COR Change of basis

$$\mathrm{Mat}_{\tilde{B},\tilde{B}'}(f) = \mathrm{Mat}_{B',\tilde{B}'}(\mathrm{id}_{V'})\,\mathrm{Mat}_{B,B'}(f)\,\mathrm{Mat}_{\tilde{B},B}(\mathrm{id}_V)\,.$$

Proof. This follows from Lemma 20.3 and $f = \mathrm{id}_{V'} \circ f \circ \mathrm{id}_V$. Sketch:

$$\begin{array}{ccc} (V,\tilde{B}) & \xrightarrow{f} & (V',\tilde{B}') \\ {\scriptstyle \mathrm{id}_V}\downarrow & & \uparrow{\scriptstyle \mathrm{id}_{V'}} \\ (V,B) & \xrightarrow{f} & (V',B') \end{array}$$

❑

Since id_V and $\mathrm{id}_{V'}$ are isomorphisms, the *change-of-basis matrices* $\mathrm{Mat}_{\tilde{B},B}(\mathrm{id}_V)$ and $\mathrm{Mat}_{B',\tilde{B}'}(\mathrm{id}_{V'})$ are invertible.

DEF Change-of-basis matrix

Conversely, any invertible matrix can occur as a change-of-basis matrix, where one of the two bases can be specified arbitrarily.

20.5. **Lemma.** *Let K be a field, let $n \in \mathbb{N}$, let V be a K-vector space with basis $B = (b_1, b_2, \ldots, b_n)$. Let further $A \in \mathrm{Mat}(n, K)$ be an invertible matrix. Then there are bases B' and B'' of V such that*

LEMMA Change-of-basis matrices

$$A = \mathrm{Mat}_{B,B'}(\mathrm{id}_V) = \mathrm{Mat}_{B'',B}(\mathrm{id}_V)\,.$$

Proof. Let $A = (a_{ij})$ and $B'' = (b''_1, b''_2, \ldots, b''_n)$. The statement $A = \mathrm{Mat}_{B'',B}(\mathrm{id}_V)$ means $b''_j = a_{1j}b_1 + \ldots + a_{nj}b_n$. We define b''_j by this equation for $j \in \{1, 2, \ldots, n\}$; then the desired relation holds (note that B'' is a basis since A is invertible: we can write the b_i as linear combinations of the b''_j whose coefficients are the entries of A^{-1}; this implies that B'' is a spanning family).

There is then also a basis B' such that $A^{-1} = \mathrm{Mat}_{B',B}(\mathrm{id}_V)$; then $A = \mathrm{Mat}_{B,B'}(\mathrm{id}_V)$ since

$$\mathrm{Mat}_{B,B'}(\mathrm{id}_V)\,\mathrm{Mat}_{B',B}(\mathrm{id}_V) = \mathrm{Mat}_{B',B'}(\mathrm{id}_V) = I_n$$

by Lemma 20.3. ❑

20.6. **Theorem.** *Let K be a field and $n \in \mathbb{N}$. The set of all invertible matrices in $\mathrm{Mat}(n, K)$ forms a group under matrix multiplication.*

THM Group of invertible matrices

Proof. Matrix multiplication is associative and the (invertible) identity matrix I_n is a neutral element. Every invertible matrix has by definition an inverse (which is itself invertible). It remains to be shown that the binary operation is well-defined, i.e., that the product of two invertible matrices is again invertible. This follows from

$$(AB)(B^{-1}A^{-1}) = A(BB^{-1})A^{-1} = AA^{-1} = I_n = (B^{-1}A^{-1})(AB)\,,$$

showing that AB has inverse $B^{-1}A^{-1}$. ❑

If M is a monoid, then the set of invertible elements of M forms a submonoid that is in fact a group. The proof is identical.

20.7. **Definition.** The group of invertible matrices in $\mathrm{Mat}(n, K)$ is the *general linear group*, denoted $\mathrm{GL}(n, K)$. ♢

DEF $\mathrm{GL}(n, K)$

The notation $\mathrm{GL}_n(K)$ is also quite common.

20.8. **Theorem.** *Let K be a field, let $m, n \in \mathbb{N}$, let V be an n-dimensional and V' an m-dimensional K-vector space, let $f\colon V \to V'$ be linear. Let further B be a basis of V, B' a basis of V' and $A = \mathrm{Mat}_{B,B'}(f)$. Then the set of all matrices of f relative to arbitrary bases of V and V' is exactly*

THM Matrices of the same linear map

$$\{PAQ \mid P \in \mathrm{GL}(m, K), Q \in \mathrm{GL}(n, K)\}\,.$$

Proof. By Corollary 20.4 and the discussion following it, every matrix of f has the form PAQ with invertible matrices P and Q. Conversely, fix invertible matrices $P \in \mathrm{GL}(m, K)$ and $Q \in \mathrm{GL}(n, K)$. By Lemma 20.5, there are bases $\tilde{B}$ of V and $\tilde{B}'$ of V' such that $P = \mathrm{Mat}_{B',\tilde{B}'}(\mathrm{id}_{V'})$ and $Q = \mathrm{Mat}_{\tilde{B},B}(\mathrm{id}_V)$. Then

$$PAQ = \mathrm{Mat}_{B',\tilde{B}'}(\mathrm{id}_{V'})\,\mathrm{Mat}_{B,B'}(f)\,\mathrm{Mat}_{\tilde{B},B}(\mathrm{id}_V) = \mathrm{Mat}_{\tilde{B},\tilde{B}'}(f)$$

is also a matrix of f. ❑

20.9. **Definition.** Let K be a field, $m, n \in \mathbb{N}$ and $A, B \in \mathrm{Mat}(m \times n, K)$. The matrices A and B are *equivalent*, if there are matrices $P \in \mathrm{GL}(m, K)$ and $Q \in \mathrm{GL}(n, K)$ such that $PAQ = B$. ♢

DEF Equivalence of matrices

So two matrices in $\mathrm{Mat}(m \times n, K)$ are equivalent if and only if they represent the same linear map (relative to usually different bases).

20.10. **Theorem.** *Let K be a field, $m, n \in \mathbb{N}$ and $A, B \in \mathrm{Mat}(m \times n, K)$. The matrices A and B are equivalent if and only if $\mathrm{rk}(A) = \mathrm{rk}(B)$. In this case let $r = \mathrm{rk}(A) = \mathrm{rk}(B)$; then both matrices are equivalent to the matrix*

$$M_r = \left(\begin{array}{c|c} I_r & \mathbf{0}_{r,n-r} \\ \hline \mathbf{0}_{m-r,r} & \mathbf{0}_{m-r,n-r} \end{array}\right).$$

Here $\mathbf{0}_{k,l}$ denotes a zero matrix with k rows and l columns.

THM Classification of matrices up to equivalence

Proof. Let $r = \mathrm{rk}(A)$. We show that A is equivalent to M_r. Let $f\colon K^n \to K^m$ be the associated linear map; it has rank r, which implies $\dim\ker(f) = n - r$. We choose a basis $B = (b_1, \dots, b_n)$ of K^n such that $(b_{r+1}, \dots, b_n)$ is a basis of $\ker(f)$. Setting $b'_i = f(b_i)$ for $i \in \{1, 2, \dots, r\}$, we obtain a basis $(b'_1, \dots, b'_r)$ of $\mathrm{im}(f)$: we have $\dim\mathrm{im}(f) = r$ and

$$\mathrm{im}(f) = \langle f(b_1), \dots, f(b_r), f(b_{r+1}), \dots, f(b_n)\rangle = \langle b'_1, \dots, b'_r, \mathbf{0}, \dots, \mathbf{0}\rangle = \langle b'_1, \dots, b'_r\rangle,$$

so $(b'_1, \dots, b'_r)$ is a spanning family of $\mathrm{im}(f)$ of the correct length r. We extend this basis of $\mathrm{im}(f)$ to a basis $B' = (b'_1, \dots, b'_m)$ of K^m. Then $\mathrm{Mat}_{B,B'}(f)$ is exactly M_r, and so A is equivalent to M_r.

If $\mathrm{rk}(B) = r$ as well, then by the same argument, B is also equivalent to M_r. This implies that A and B are equivalent: There are $P, P' \in \mathrm{GL}(m, K)$ and $Q, Q' \in \mathrm{GL}(n, K)$ such that $M_r = PAQ = P'BQ'$. This gives $B = (P'^{-1}P)A(QQ'^{-1})$.

Conversely, two equivalent matrices A and B must have the same rank; note that this is the rank of the linear map represented by both. ❑

Using the results from the small print on page 188 we can deduce the following from Theorem 20.10.

Corollary. *Every matrix $A \in \mathrm{Mat}(m \times n, K)$ can be transformed by row and column operations into the matrix M_r, where $r = \mathrm{rk}(A)$.*

COR Row and column operations

Equivalence of matrices is an example of an *equivalence relation*. A *relation* R between two sets X and Y is formally just a subset $R \subset X \times Y$. If for elements $x \in X$ and $y \in Y$ the pair (x, y) is an element of R, then we say that x and y *are in relation* R and sometimes write $x\,R\,y$ or similar. In the case $X = Y$ we also speak of a *relation on* X. Such a relation is

- *reflexive*, if $\forall x \in X\colon x\,R\,x$,
- *symmetric*, if $\forall x, y \in X\colon x\,R\,y \Rightarrow y\,R\,x$, and
- *transitive*, if $\forall x, y, z \in X\colon (x\,R\,y \ \wedge\ y\,R\,z) \Rightarrow x\,R\,z$.

A relation on X that is reflexive, symmetric and transitive is an *equivalence relation on* X. Examples are given by the equality relation $x = y$ (the "finest" equivalence relation on X) or the "all relation" $R = X \times X$ (the "coarsest" equivalence relation on X).

Lemma. *Let $A, B \in \mathrm{Mat}(m \times n, K)$. We write $A \sim B$, if A and B are equivalent, i.e., if there are $P \in \mathrm{GL}(m, K)$ and $Q \in \mathrm{GL}(n, K)$ such that $B = PAQ$.*

The relation $\sim$ is an equivalence relation on $\mathrm{Mat}(m \times n, K)$.

LEMMA Equivalence of matrices is an equiv. rel.

Proof.

- Reflexivity: $A \sim A$ since we can take $P = I_m$, $Q = I_n$.

- Symmetry: Assume that $A \sim B$; then there are $P \in \mathrm{GL}(m,K)$ and $Q \in \mathrm{GL}(n,K)$ such that $B = PAQ$. Then also $P^{-1} \in \mathrm{GL}(m,K)$ and $Q^{-1} \in \mathrm{GL}(n,K)$, and we have $B = P^{-1}AQ^{-1}$, so $B \sim A$.
- Transitivity: Assume $A \sim B$ and $B \sim C$. Then there are $P_1, P_2 \in \mathrm{GL}(m,K)$ and $Q_1, Q_2 \in \mathrm{GL}(n,K)$ such that $B = P_1AQ_1$ and $C = P_2BQ_2$. We have $P_2P_1 \in \mathrm{GL}(m,K)$ and $Q_1Q_2 \in \mathrm{GL}(n,K)$, so from $C = (P_2P_1)A(Q_1Q_2)$ we deduce $A \sim C$. ❑

The most important property of an equivalence relation on a set X is that it leads to a partition of X into *equivalence classes*. If $\sim$ is an equivalence relation on X and $x \in X$, then we write $[x]$ for the set $\{y \in X \mid x \sim y\}$ of all elements of X that are equivalent to x; $[x]$ is called the *equivalence class of* x. Every element of $[x]$ is a *representative* of the equivalence class.

LEMMA
Properties of equivalence relations

Lemma. *Let $\sim$ be an equivalence relation on a set X and let $x \in X$. Then for $y \in X$ the following statements are equivalent.*

(1) $x \sim y$.

(2) $y \in [x]$.

(3) $[y] \cap [x] \neq \emptyset$.

(4) $[y] = [x]$.

In particular, two equivalence classes $[x]$ and $[y]$ are either equal or disjoint.

Proof. "(1) $\Leftrightarrow$ (2)" follows from the definition of $[x]$.

"(2) $\Rightarrow$ (3)": The reflexivity of $\sim$ gives $y \in [y]$, so $y \in [x]$ implies that $y \in [y] \cap [x]$.

"(3) $\Rightarrow$ (4)": Let $z \in [y] \cap [x]$ and $w \in [y]$. Then we have $y \sim w$, $y \sim z$ and $x \sim z$; using symmetry and transitivity of $\sim$ we obtain $x \sim w$, hence $w \in [x]$. Since w was arbitrary, this shows $[y] \subset [x]$. We obtain $[x] \subset [y]$ in the same way.

"(4) $\Rightarrow$ (2)": From $y \in [y]$ and $[y] = [x]$ we obtain $y \in [x]$. ❑

We can form the *set of equivalence classes* $X/\sim = \{[x] \mid x \in X\}$. There is then a natural (or "canonical") surjective map $f\colon X \to X/\sim$, $x \mapsto [x]$. It follows from the lemma just proved that the preimage set $f^{-1}(\{[x]\})$ is exactly $[x]$. Conversely, every surjective map $f\colon X \to M$ leads to an equivalence relation on X given by $x \sim y \iff f(x) = f(y)$. One also says that f *induces* this equivalence relation.

The statement of Theorem 20.10 then says that the equivalence of $m \times n$ matrices is the same as the equivalence relation that is induced by $\mathrm{Mat}(m \times n, K) \to \{1, 2, \ldots, \min\{m,n\}\}$, $A \mapsto \mathrm{rk}(A)$, and that M_r is a representative of the equivalence class given by $\mathrm{rk}(A) = r$.

If instead of row and column operations one restricts to only row operations, then one considers two matrices A and B as equivalent, if $B = PA$ with an invertible matrix P. This is because row operations correspond to multiplication on the left by an invertible matrix; see the small print on page 188. The results there can be interpreted as saying that every equivalence class of this equivalence relation has a uniquely determined representative in reduced row echelon form.

21. The determinant

Date:
March 5, 2026

In this section we introduce the *determinant* of a square matrix. This is a scalar that tells us whether the matrix is invertible or not. In $\mathbb{R}^n$ with the usual notion of volume and orientation, the determinant of $A \in \mathrm{Mat}(n, \mathbb{R})$ can also be interpreted as the (signed) volume of the parallelotope spanned by the columns of A; this makes the determinant an important tool in integration theory. A slightly different, but related, interpretation is as the scaling factor of volumes under the linear map corresponding to A. We begin by defining the determinant recursively.

DEF
Determinant of a matrix

21.1. **Definition.** Let K be a field and $A = (a_{ij}) \in \mathrm{Mat}(n, K)$ with $n \in \mathbb{N}$. We define the *determinant* of A, $\det(A)$, recursively as follows.

(1) If $n = 0$, we set $\det(A) = 1$.

(2) Now assume $n > 0$. We write $A_{ij} \in \mathrm{Mat}(n-1, K)$ (where $i, j \in \{1, 2, \dots, n\}$) for the matrix obtained from A by removing the ith row and the jth column. Then we set

$$\begin{aligned}\det(A) &= \sum_{j=1}^{n} (-1)^{j-1} a_{1j} \det(A_{1j}) \\ &= a_{11}\det(A_{11}) - a_{12}\det(A_{12}) + \ldots + (-1)^{n-1} a_{1n}\det(A_{1n})\,.\end{aligned}$$

The determinant is also written in the following way.

$$\det\big((a_{ij})\big) = \begin{vmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{vmatrix}$$ ♢

EXAMPLES
Determinant

21.2. **Examples.** For small positive sizes n, unfolding the definition gives the following formulas.

$$\det\big((a)\big) = a$$

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

$$\begin{vmatrix} a & b & c \\ d & e & f \\ g & h & i \end{vmatrix} = aei - afh + bfg - bdi + cdh - ceg$$

The formula for a 3×3 determinant can be memorized using the "Rule of Sarrus": one repeats the first two columns at the end of the matrix and takes the sum of the products over the diagonals going from top to bottom minus the sum of the products over the diagonals going from bottom to top.

$$\begin{vmatrix} a & b & c & a & b \\ d & e & f & d & e \\ g & h & i & g & h \end{vmatrix}$$

There is no similar rule for larger determinants! ♣

We now investigate the properties of the determinant.

21.3. **Theorem.** *Let K be a field and $n \in \mathbb{N}$. We write $d = \det\colon \mathrm{Mat}(n, K) \to K$ for the determinant map on $n \times n$ matrices.* **THM** properties of the determinant

(1) *$d(A)$ is linear as a function of each row of A (considering the entries of the other rows as fixed).*

(2) *If A has two equal rows, then $d(A) = 0$.*

(3) *$d(I_n) = 1$.*

If $d\colon \mathrm{Mat}(n, K) \to K$ is any map that satisfies (1) and (2), then the following hold.

(4) *If A' is obtained from A by an elementary row operation $\mathbf{I}_i(\lambda)$, then $d(A') = \lambda d(A)$.*

(5) *If A' is obtained from A by an elementary row operation $\mathbf{II}_{i,j}(\lambda)$, then $d(A') = d(A)$.*

(6) *If A' is obtained from A by swapping two rows, then $d(A') = -d(A)$.*

(7) *We have $d(A) = \det(A) d(I_n)$ for all $A \in \mathrm{Mat}(n, K)$. In particular, $\det\colon \mathrm{Mat}(n, K) \to K$ is the unique map that satisfies (1), (2), and (3).*

Furthermore,

(8) $\det(A) \neq 0 \iff \mathrm{rk}(A) = n \iff$ *A invertible.*

Proof. We begin by showing (4)–(6).

(4) This is a special case of property (1) (which we assume here).

(5) We write $d(v_1, \dots, v_n)$ for $d(A)$ when $v_1, \dots, v_n \in K^n$ are the rows of A, and for $1 \leq i < j \leq n$ we set $d_{ij}(v_i, v_j) = d(v_1, \dots, v_n)$, keeping the v_k with $k \notin \{i, j\}$ fixed. Then we have

$$d(A') = d_{ij}(v_i + \lambda v_j, v_j) \overset{(1)}{=} d_{ij}(v_i, v_j) + \lambda d_{ij}(v_j, v_j) \overset{(2)}{=} d_{ij}(v_i, v_j) = d(A)\,.$$

(6) In the notation of the proof of part (5) we have

$$\begin{aligned} 0 &\overset{(2)}{=} d_{ij}(v_i + v_j, v_i + v_j) \\ &\overset{(1)}{=} d_{ij}(v_i, v_i) + d_{ij}(v_i, v_j) + d_{ij}(v_j, v_i) + d_{ij}(v_j, v_j) \\ &\overset{(2)}{=} d_{ij}(v_i, v_j) + d_{ij}(v_j, v_i)\,; \end{aligned}$$

this implies $d(A') = d_{ij}(v_j, v_i) = -d_{ij}(v_i, v_j) = -d(A)$.

The proof of the first three statements for $d = \det$ is done by induction over n. The statements are trivially true when $n = 0$. So we now assume $n > 0$ and that the statements are already shown for smaller values of n.

(1) $\det(A)$ is linear in the first row of A since by definition, $\det(A)$ is a linear combination of entries of the first row whose coefficients do not depend on the first row. Now let $k \in \{2, 3, \dots, n\}$. By the induction hypothesis, for each $j \in \{1, 2, \dots, n\}$, $\det(A_{1j})$ is linear in the $(k-1)$st row of A_{1j} and therefore linear in the kth row of A. We see that $\det(A)$ is a linear combination of maps that are linear as functions of the kth row of A, with coefficients that do not depend on the kth row of A, and so $\det(A)$ is linear in the kth row of A.

(2) Let A be a matrix, in which the kth and the lth row agree, for some $1 \leq k < l \leq n$. If $k > 1$, then in every matrix A_{1j} the $(k-1)$st and the $(l-1)$st rows agree; the induction hypothesis then implies $\det(A_{1j}) = 0$ for all j, which gives $\det(A) = 0$. It remains to deal with the case $k = 1$. If $l > 2$, then we swap the lth row with the second row. By the induction hypothesis, which implies the statement (6) for $d = \det$ on matrices of size $n-1$, this leads to a change of sign in all $\det(A_{1j})$, which does not affect whether $\det(A) = 0$ or not. We can therefore assume that the first two rows of A are the same. We write $d_{jm} = d_{mj}$ for the determinant of the matrix that is obtained from A by removing the first two rows and the columns j and m (with $j \neq m$). Taking into account that $a_{2j} = a_{1j}$ by assumption, we obtain

$$\begin{aligned}
\det(A) &= \sum_{j=1}^{n} (-1)^{j-1} a_{1j} \det(A_{1j}) \\
&= \sum_{j=1}^{n} (-1)^{j-1} a_{1j} \left(\sum_{m=1}^{j-1} (-1)^{m-1} a_{2m} d_{jm} + \sum_{m=j+1}^{n} (-1)^{m} a_{2m} d_{jm} \right) \\
&= \sum_{1 \leq m < j \leq n} (-1)^{j-m} a_{1j} a_{1m} d_{jm} + \sum_{1 \leq j < m \leq n} (-1)^{m-j-1} a_{1j} a_{1m} d_{jm} \\
&\overset{(*)}{=} \sum_{1 \leq j < m \leq n} (-1)^{m-j} a_{1j} a_{1m} d_{jm} + \sum_{1 \leq j < m \leq n} (-1)^{m-j-1} a_{1j} a_{1m} d_{jm} \\
&= \sum_{1 \leq j < m \leq n} \big((-1)^{m-j} + (-1)^{m-j-1} \big) a_{1j} a_{1m} d_{jm} \\
&= 0 \,.
\end{aligned}$$

(At $(*)$ we have swapped j and m in the first sum, using that $a_{1m} a_{1j} d_{mj} = a_{1j} a_{1m} d_{jm}$.)

(3) The recursive definition gives $\det(I_n) = 1 \cdot \det(I_{n-1}) = 1$.

Finally:

(7) Let A' be the reduced row echelon form of A. We see from (4), (5) and (6) that $d(A) = d_0(A) d(A')$, where $d_0(A) \neq 0$ depends only on A and not on d: $d_0(A)$ is the product of the scalars λ^{-1} and -1 that come from the elementary row operations $\mathbf{I}_i(\lambda)$ and $\mathbf{III}_{i,j}$ that are performed to obtain A' from A. If A is invertible, then $A' = I_n$ (Lemma 19.18), which gives $d(A) = d_0(A) d(I_n)$. When $d = \det$, we have $d_0(A) = \det(A)$. If A is not invertible, then A' has a zero row, which implies by property (1) that $d(A') = 0$. When $d = \det$, we have $\det(A') = 0$. In both cases we obtain $d(A) = \det(A) d(I_n)$ as claimed.

(8) It follows from statements (4), (5) and (6) that $\det(A) = 0$ if and only if $\det(A') = 0$, where A' is again the reduced row echelon form of A. If $\mathrm{rk}(A) = n$, then A is invertible and $A' = I_n$ by Lemma 19.18, so $\det(A') = \det(I_n) = 1 \neq 0$ by (3). If $\mathrm{rk}(A) < n$, then A' has a zero row, which implies $\det(A') = 0$ by (1). The second equivalence is a consequence of the fact that a linear map $K^n \to K^n$ is surjective if and only if it is an isomorphism; compare Corollary 13.14. ❑

If we consider the determinant of an $n \times n$ matrix A as a function of the n rows of A, which are vectors in K^n, then we obtain a so-called *alternating multilinear form*. This is a special case of a multilinear map.

Definition. Let K be a field and let $V_1, V_2, \ldots, V_m$ and W be K-vector spaces. A map $f\colon V_1 \times V_2 \times \ldots \times V_m \to W$ is *multilinear*, if f is K-linear in each argument, which means that

$$f(v_1, \ldots, v_{i-1}, \lambda v_i, v_{i+1}, \ldots, v_m) = \lambda f(v_1, \ldots, v_{i-1}, v_i, v_{i+1}, \ldots, v_m)$$

and

$$\begin{aligned} &f(v_1, \ldots, v_{i-1}, v_i + v_i', v_{i+1}, \ldots, v_m) \\ &\qquad = f(v_1, \ldots, v_{i-1}, v_i, v_{i+1}, \ldots, v_m) + f(v_1, \ldots, v_{i-1}, v_i', v_{i+1}, \ldots, v_m) \end{aligned}$$

for all $i \in \{1, 2, \ldots, m\}$, $v_j \in V_j$, $\lambda \in K$, $v_i' \in V_i$. If $W = K$, then f is also called a *multilinear form*.

A multilinear form $f\colon V^m = V \times V \times \cdots \times V \to K$ is *alternating*, if $f(v_1, \ldots, v_m) = 0$ whenever $v_i = v_j$ for some $1 \leq i < j \leq m$. ♢

DEF
multilinear map
alternating multilinear form

The statements (1) and (2) in Theorem 21.3 exactly say that $\det(A)$ is an alternating multilinear form of the rows of A. Since we can identify a K-vector space V with given basis $(b_1, \ldots, b_n)$ with the standard space K^n, statement (7) in Theorem 21.3 has the following interpretation.

Theorem. *Let V be a K-vector space with basis $(b_1, b_2, \ldots, b_n)$. Then there exists a unique alternating multilinear form $d\colon V^n \to K$ such that $d(b_1, b_2, \ldots, b_n) = 1$.*

THM
Existence and uniqueness of alternating multilinear forms

For practical purposes, the most important parts of Theorem 21.3 are those that show how the determinant changes (or does not change) under elementary row operations. This leads to a practical procedure for the computation of larger determinants.

21.4. **Example.**

EXAMPLE
Determinant computation

$$\begin{vmatrix} 1 & -1 & 1 & -1 \\ 1 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 \\ 1 & 2 & 4 & 8 \end{vmatrix} = \begin{vmatrix} 1 & -1 & 1 & -1 \\ 0 & 1 & -1 & 1 \\ 0 & 2 & 0 & 2 \\ 0 & 3 & 3 & 9 \end{vmatrix} = \begin{vmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & -1 & 1 \\ 0 & 0 & 2 & 0 \\ 0 & 0 & 6 & 6 \end{vmatrix}$$

$$= 2 \cdot \begin{vmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & -1 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 6 & 6 \end{vmatrix} = 2 \cdot \begin{vmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 6 \end{vmatrix} = 2 \cdot 6 \cdot \begin{vmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{vmatrix} = 12 \qquad ♣$$

The uniqueness statement (7) in Theorem 21.3 is important, because it can be used to show further properties of the determinant.

21.5. **Theorem.** *Let K be a field, let $n > 0$ and let $A = (a_{ij}) \in \mathrm{Mat}(n, K)$. Using the notation A_{ij} from Definition 21.1, we have for every $i \in \{1, 2, \ldots, n\}$ that*

$$\det(A) = \sum_{j=1}^{n} (-1)^{j-i} a_{ij} \det(A_{ij}) \,.$$

THM
Expansion along the ith row

Proof. In the same way as in the proof of Theorem 21.3, one shows that the right hand side has the properties (1), (2) and (3). By the uniqueness of the determinant, it follows that it must equal $\det(A)$.

(Alternatively, one can swap rows 1 and i and apply Theorem 21.3 (6).) ❑

21.6. **Example.** We can simplify the computation of the determinant in Example 21.4 if we expand along the second row. **EXAMPLE**

$$\begin{vmatrix} 1 & -1 & 1 & -1 \\ 1 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 \\ 1 & 2 & 4 & 8 \end{vmatrix} = - \begin{vmatrix} -1 & 1 & -1 \\ 1 & 1 & 1 \\ 2 & 4 & 8 \end{vmatrix} = \ldots = 12 \qquad \clubsuit$$

21.7. **Theorem.** *Let K be a field, let $n \in \mathbb{N}$ and let $A, B \in \mathrm{Mat}(n, K)$. Then* **THM** Multiplicativity of the determinant

$$\det(AB) = \det(A)\det(B)\,.$$

If A is invertible, then $\det(A^{-1}) = \det(A)^{-1}$.

Proof. We fix B and consider A as varying. Set $d_B\colon \mathrm{Mat}(n, K) \to K$, $A \mapsto \det(AB)$. The properties of matrix multiplication imply that the kth row of AB is obtained as the kth row of A (considered as a row vector) multiplied with B. In particular, the kth row of AB is a linear function of the kth row of A and does not depend on the entries in the other rows of A. This implies that d_B is linear in the rows of A. It also follows that when A has to equal rows, the same is true of AB. This shows that d_B also satisfies property (2) in Theorem 21.3. The uniqueness statement in Theorem 21.3 then gives $\det(AB) = d_B(A) = \det(A)d_B(I_n) = \det(A)\det(B)$. The last claim follows from $\det(A)\det(A^{-1}) = \det(I_n) = 1$. ❑

21.8. **Theorem.** *Let K be a field, $n \in \mathbb{N}$ and $A \in \mathrm{Mat}(n, K)$. Then* **THM** Symmetry of the determinant

$$\det(A^\top) = \det(A)\,.$$

Proof. We have to show that $\det(A^\top)$ has the properties (1), (2) and (3) in Theorem 21.3. That $\det(I_n^\top) = \det(I_n) = 1$ is clear. The other two statements are equivalent to saying that $\det(A)$ is linear in the *columns* of A and vanishes (i.e., takes the value zero) when A has two equal *columns*. The first claim follows easily by induction from the recursive definition of the determinant: for fixed k, every term in the sum is linear in the kth column of A (either via a_{1k} or via $\det(A_{1j})$). To see the second claim, note that when A has two equal columns, then $\mathrm{rk}(A) < n$, which implies $\det(A) = 0$ by part (8) of Theorem 21.3. ❑

This implies that one can also use *column* operations when computing determinants, even mixed with row operations. Similarly, we obtain a formula for expanding a determinant along a column.

21.9. **Corollary.** *Let K be a field, let $n > 0$ and let $A = (a_{ij}) \in \mathrm{Mat}(n, K)$. Using the notation A_{ij} from Definition 21.1, we have for every $j \in \{1, \ldots, n\}$ that* **COR** Expansion along the jth column

$$\det(A) = \sum_{i=1}^{n} (-1)^{j-i} a_{ij} \det(A_{ij})\,.$$

Proof. This follows from Theorem 21.5, applied with $A^\top$, and from $\det(A^\top) = \det(A)$. ❑

21.10. **Example.** A matrix $A \in \mathrm{Mat}(n, \mathbb{R})$ such that $AA^\top = I_n$ is *orthogonal*. What can we say about $\det(A)$? **EXAMPLE** Orthogonal matrix

We have

$$1 = \det(I_n) = \det(AA^\top) = \det(A)\det(A^\top) = \det(A)^2\,,$$

and so $\det(A) = \pm 1$. ♣

21.11. **Example.** We compute the determinant from Example 21.4 another time. **EXAMPLE** Determinant computation

$$\begin{vmatrix} 1 & -1 & 1 & -1 \\ 1 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 \\ 1 & 2 & 4 & 8 \end{vmatrix} = - \begin{vmatrix} -1 & 1 & -1 \\ 1 & 1 & 1 \\ 2 & 4 & 8 \end{vmatrix} = - \begin{vmatrix} -1 & 1 & 0 \\ 1 & 1 & 0 \\ 2 & 4 & 6 \end{vmatrix} = -6 \cdot \begin{vmatrix} -1 & 1 \\ 1 & 1 \end{vmatrix}$$

$$= -6((-1)\cdot 1 - 1 \cdot 1) = 12$$

(Expansion along the second row, elementary column operation $\mathbf{II}_{3,1}(-1)$, expansion along the third column, formula for 2×2 determinants). ♣

For completeness, we note how the determinant changes upon scaling the whole matrix.

21.12. **Corollary.** *Let K be a field, $n \in \mathbb{N}$, $\lambda \in K$ and $A \in \mathrm{Mat}(n, K)$. Then* **COR** $\det(\lambda A)$

$$\det(\lambda A) = \lambda^n \det(A)\,.$$

Proof. This follows from the fact that $\det(A)$ is linear in *every* column of A. Writing $A = (\boldsymbol{a}_1 \mid \boldsymbol{a}_2 \mid \cdots \mid \boldsymbol{a}_n)$, we have

$$\begin{aligned} \det(\lambda A) &= \det(\lambda \boldsymbol{a}_1 \mid \lambda \boldsymbol{a}_2 \mid \cdots \mid \lambda \boldsymbol{a}_n) \\ &= \lambda \det(\boldsymbol{a}_1 \mid \lambda \boldsymbol{a}_2 \mid \cdots \mid \lambda \boldsymbol{a}_n) = \ldots = \lambda^n \det(\boldsymbol{a}_1 \mid \boldsymbol{a}_2 \mid \cdots \mid \boldsymbol{a}_n)\,. \end{aligned}$$ ❑

The expansion of the determinant along rows and columns leads to a "formula" for the inverse of a matrix.

21.13. **Definition.** Let K be a field, $n > 0$ and $A \in \mathrm{Mat}(n, K)$. The matrix $\tilde{A} \in \mathrm{Mat}(n, K)$, whose entry in the ith row and jth column is given by $(-1)^{i-j}\det(A_{ji})$ (not A_{ij}!) is the *adjugate matrix* of A. ♢ **DEF** Adjugate matrix

21.14. **Theorem.** *Let K be a field, $n > 0$ and $A \in \mathrm{Mat}(n, K)$. Then* **THM** Adjugate matrix

$$A\tilde{A} = \tilde{A}A = \det(A) I_n\,.$$

If A is invertible, then $A^{-1} = \det(A)^{-1}\tilde{A}$.

Proof. The (i, k) entry in the product $A\tilde{A}$ is

$$\sum_{j=1}^{n} a_{ij}(-1)^{j-k}\det(A_{kj})\,.$$

If $k = i$, this gives $\det(A)$ by Theorem 21.5 on the expansion along the ith row. If $k \neq i$, we instead obtain the expansion along the kth row of the determinant of the matrix obtained from A by replacing the kth row by the contents of the ith row. Since this matrix has two equal rows, its determinant is zero. This shows $A\tilde{A} = \det(A)I_n$. The claim $\tilde{A}A = \det(A)I_n$ follows in the same way using Corollary 21.9. The last claim is obtained through multiplication by $\det(A)^{-1}A^{-1}$. ❑

21.15. **Example.** For $n = 2$ we obtain the formula

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix}^{-1} = \frac{1}{ad - bc} \begin{pmatrix} d & -b \\ -c & a \end{pmatrix} .$$

♣

EXAMPLE
Inverse of a 2×2 matrix

In a somewhat similar way to the explicit formula for the inverse given using the adjugate matrix, there is an explicit formula for the solution of a system of linear equations when this solution is unique, which is equivalent to the matrix of the system being invertible (hence in particular square). This is known as Cramer's Rule. So let $A \in \mathrm{Mat}(n, K)$ be invertible and let $b \in K^n$ be a column vector. For $j \in \{1, 2, \ldots, n\}$ let A_j be the matrix obtained from A by replacing its jth column with the vector b. Then the ith component of the vector $x = A^{-1}b$, which is the unique solution of the system $Ax = b$, is given by $\det(A_i)/\det(A)$. The proof is an exercise.

We give an example of a general formula for a special kind of determinant.

21.16. **Theorem.** *Let K be a field, let $n \in \mathbb{N}$, let $a_1, a_2, \ldots, a_n \in K$ and let A be the following "Vandermonde matrix" of $a_1, a_2, \ldots, a_n$:*

THM
Vandermonde determinant

$$A = (a_i^{j-1})_{1 \le i,j \le n} = \begin{pmatrix} 1 & a_1 & a_1^2 & \cdots & a_1^{n-1} \\ 1 & a_2 & a_2^2 & \cdots & a_2^{n-1} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & a_n & a_n^2 & \cdots & a_n^{n-1} \end{pmatrix} \in \mathrm{Mat}(n, K) .$$

Then

$$\det(A) = \prod_{1 \le i < j \le n} (a_j - a_i) .$$

Proof. By induction on n. The case $n = 0$ is clear ($\det(A) = 1$ and the product is empty). So assume that $n > 0$ and the claim is shown for $n - 1$. We subtract the first row from each of the other rows and then expand along the first column; this results in

$$\det(A) = \begin{vmatrix} a_2 - a_1 & a_2^2 - a_1^2 & \cdots & a_2^{n-1} - a_1^{n-1} \\ \vdots & \vdots & & \vdots \\ a_n - a_1 & a_n^2 - a_1^2 & \cdots & a_n^{n-1} - a_1^{n-1} \end{vmatrix} .$$

We have the general relation $x^m - y^m = (x-y)(x^{m-1} + x^{m-2}y + \ldots + xy^{m-2} + y^{m-1})$. We can therefore extract from the first, second, ..., $(n-1)$st row a factor $(a_2 - a_1)$, $(a_3 - a_1)$, ..., $(a_n - a_1)$ and obtain

$$\det(A) = \prod_{j=2}^{n} (a_j - a_1) \cdot \begin{vmatrix} 1 & a_2 + a_1 & \cdots & a_2^{n-2} + a_1 a_2^{n-3} + \ldots + a_1^{n-2} \\ \vdots & \vdots & & \vdots \\ 1 & a_n + a_1 & \cdots & a_n^{n-2} + a_1 a_n^{n-3} + \ldots + a_1^{n-2} \end{vmatrix} .$$

For $j = n-1, n-2, \ldots, 3, 2$, we successively subtract the a_1-multiple of column $j-1$ from column j. This does not change the determinant and leaves the Vandermonde matrix of $a_2, a_3, \ldots, a_n$. The claim now follows from the induction hypothesis. ❑

Note that the Vandermonde matrix is precisely the matrix of the linear map

$$\phi \colon P_{<n} \longrightarrow \mathbb{R}^n, \quad f \longmapsto \big(f(a_1), f(a_2), \ldots, f(a_n)\big)$$

(with $P_{<n}$ as usual the $\mathbb{R}$-vector space of polynomial functions of degree $< n$), which plays a role when interpolating given data points by polynomials, relative to the basis $(x \mapsto 1, x \mapsto x, x \mapsto x^2, \ldots, x \mapsto x^{n-1})$ of $P_{<n}$ and the standard basis of $\mathbb{R}^n$; compare Examples 12.28 and 13.15.

Another (unrelated) object named after Vandermonde is "Vandermonde's identity"

$$\sum_{j=0}^{k} \binom{m}{j}\binom{n}{k-j} = \binom{m+n}{k}.$$

We will now see how to generalize the formulas for small determinants in Example 21.2 to arbitrarily large determinants. These formulas are obtained from the recursive Definition 21.1 of the determinant. This gives a sum of terms of the form $\pm a_{1,\sigma(1)}a_{2,\sigma(2)}\cdots a_{n,\sigma(n)}$, where the column indices $\sigma(1),\sigma(2),\dots,\sigma(n)$ are pairwise distinct (this is because every column that was used is removed in the further development). The map $\sigma\colon \{1,2,\dots,n\} \to \{1,2,\dots,n\}$ is therefore bijective, i.e., a permutation. It is not hard to see that conversely, each permutation shows up in this way in the expansion of the determinant. Recall that the permutations of $\{1,2,\dots,n\}$ are the elements of the *symmetric group* S_n; the binary operation in this group is composition of maps. We have shown the following.

G.W. Leibniz (1646–1716)

21.17. **Theorem.** *Let $n \in \mathbb{N}$. There is a unique map*

$$\varepsilon\colon S_n \to \{-1,1\},$$

such that for every field K and every matrix $A = (a_{ij}) \in \mathrm{Mat}(n,K)$ we have

$$\det(A) = \sum_{\sigma\in S_n} \varepsilon(\sigma) a_{1,\sigma(1)}a_{2,\sigma(2)}\cdots a_{n,\sigma(n)}.$$

THM Leibniz formula

This formula has $\#S_n = n! = 1\cdot 2\cdot 3\cdots(n-1)\cdot n$ terms; it is therefore completely useless for computation unless n is very small! It is, however, useful for theoretical considerations. For example, it shows immediately that the determinant of a matrix whose entries are integers is itself an integer.

21.18. **Definition.** $\varepsilon(\sigma) \in \{-1,1\}$ is the *sign* of the permutation $\sigma \in S_n$. σ is *even*, if $\varepsilon(\sigma) = 1$ and *odd*, if $\varepsilon(\sigma) = -1$. ♢

DEF Sign of a permutation

To have a complete specification of the Leibniz formula, we need to figure out what the sign function is. To this end, we introduce the permutation matrix associated to σ.

21.19. **Definition.** Let K be a field, $n \in \mathbb{N}$ and $\sigma \in S_n$. Then we denote by $P(\sigma)$ the matrix $(\delta_{i,\sigma(j)})_{1\le i,j\le n} \in \mathrm{Mat}(n,K)$ and call it the *permutation matrix associated to σ*. ♢

DEF permutation matrix

The entries of $P(\sigma)$ are 1 at positions of the form $(\sigma(j),j)$ and 0 otherwise. This means that the jth column of $P(\sigma)$ contains the standard basis vector $\mathbf{e}_{\sigma(j)}$; the linear map $K^n \to K^n$ corresponding to $P(\sigma)$ therefore permutes the standard basis according to σ: $P(\sigma)\mathbf{e}_j = \mathbf{e}_{\sigma(j)}$, where $\mathbf{e}_j$ is considered as a column vector.

21.20. **Lemma.** *Let K be a field and $n \in \mathbb{N}$.*

LEMMA Properties of permutation matrices

(1) *For $\sigma,\tau \in S_n$ we have $P(\sigma\circ\tau) = P(\sigma)P(\tau)$.*

(2) *For $\sigma \in S_n$ we have $\varepsilon(\sigma) = \det\big(P(\sigma)\big)$.*

(3) *For $\sigma,\tau \in S_n$ we have $\varepsilon(\sigma\circ\tau) = \varepsilon(\sigma)\varepsilon(\tau)$.*

(4) *If σ is a* transposition *(this is a permutation that interchanges two elements and fixes all others), then $\varepsilon(\sigma) = -1$.*

DEF transposition

Proof.

(1) For all $j \in \{1, 2, \dots, n\}$ we have
$$P(\sigma)P(\tau)\mathbf{e}_j = P(\sigma)\mathbf{e}_{\tau(j)} = \mathbf{e}_{\sigma(\tau(j))} = \mathbf{e}_{(\sigma\circ\tau)(j)} = P(\sigma\circ\tau)\mathbf{e}_j\,;$$
this shows that $P(\sigma\circ\tau) = P(\sigma)P(\tau)$.

(2) The only nonzero term in the formula 21.17 for $\det(P(\sigma^{-1}))$ is $\varepsilon(\sigma)\delta_{1,1}\cdots\delta_{n,n} = \varepsilon(\sigma)$. This implies (since $P(\sigma)^{-1} = P(\sigma^{-1})$ by part (1) and $\varepsilon(\sigma) = \pm 1$)
$$\det\big(P(\sigma)\big) = \det\big(P(\sigma)^{-1}\big)^{-1} = \det\big(P(\sigma^{-1})\big)^{-1} = \varepsilon(\sigma)^{-1} = \varepsilon(\sigma)\,.$$

(3) This follows from (1) and (2) and the multiplicativity of the determinant.

(4) In this case $P(\sigma)$ is obtained from I_n by swapping two rows (or columns), hence $\varepsilon(\sigma) = \det(P(\sigma)) = -\det(I_n) = -1$. ❑

Since (as is fairly easy to see) every permutation can be written as a composition of transpositions, ε is uniquely determined by properties (3) and (4) in Lemma 21.20: if σ is a composition of an even number of transpositions, then σ is even, otherwise σ is odd.

There is a kind of formula for $\varepsilon(\sigma)$. First a little definition.

Definition. Let $\sigma \in S_n$. A pair (i, j) with $1 \le i < j \le n$ is an *inversion* of σ, if $\sigma(i) > \sigma(j)$. ♢

DEF Inversion

The we have the following result.

Theorem. *Let $\sigma \in S_n$ and let m be the number of inversions of σ. Then $\varepsilon(\sigma) = (-1)^m$.*

THM Sign and inversions

Proof. Let $\varepsilon'(\sigma)$ be the function defined by $(-1)^{\text{number of inversions of }\sigma}$. The transposition τ that swaps k and l (with $k < l$) has exactly $m = 1 + 2(l - k - 1)$ inversions (namely, (k, l), together with (k, j) and (j, l) for all $k < j < l$). Since m is odd, it follows that $\varepsilon'(\tau) = (-1)^m = -1 = \varepsilon(\tau)$.

We also have that
$$\varepsilon'(\sigma) = \prod_{1\le i<j\le n} \frac{\sigma(j) - \sigma(i)}{j - i}$$
since the product on the right has absolute value 1 (every factor in the denominator also shows up up to sign in the numerator), and $\sigma(j) - \sigma(i)$ is negative if and only if (i, j) is an inversion of σ. From this we obtain for all $\sigma, \tau \in S_n$ that
$$\begin{aligned}
\varepsilon'(\sigma\circ\tau) &= \prod_{1\le i<j\le n} \frac{\sigma(\tau(j)) - \sigma(\tau(i))}{j - i} \\
&= \prod_{1\le i<j\le n} \frac{\sigma(\tau(j)) - \sigma(\tau(i))}{\tau(j) - \tau(i)} \prod_{1\le i<j\le n} \frac{\tau(j) - \tau(i)}{j - i} \\
&= \prod_{1\le k<l\le n} \frac{\sigma(l) - \sigma(k)}{l - k} \prod_{1\le i<j\le n} \frac{\tau(j) - \tau(i)}{j - i} = \varepsilon'(\sigma)\varepsilon'(\tau)\,.
\end{aligned}$$
The function ε' therefore has the properties (3) and (4) from Lemma 21.20 and must therefore agree with ε. ❑

In the last part of this section we will study the geometric significance of the determinant. We work in $\mathbb{R}^n$. The first step is to define the notion of (positive or negative) orientation of a basis.

21.21. **Definition.** Let $(\boldsymbol{b}_1, \boldsymbol{b}_2, \ldots, \boldsymbol{b}_n)$ be a basis of $\mathbb{R}^n$ and let $A = \left(\boldsymbol{b}_1 \,\middle|\, \cdots \,\middle|\, \boldsymbol{b}_n\right)$ be the matrix whose columns are the basis vectors. We say that the basis is *positively oriented*, if $\det(A) > 0$, and *negatively oriented*, if $\det(A) < 0$. ♢

DEF Orientation of a basis

The standard basis is positively oriented. In the case $n = 2$, a basis is positively oriented if and only if the angle measured counterclockwise from the first to the second basis vector is less than π $(= 180°)$.

The comparison of the orientations of a basis and its image leads to the notion of orientation-preserving, respectively, orientation-reversing linear map.

21.22. **Definition.** Let $f\colon \mathbb{R}^n \to \mathbb{R}^n$ be an automorphism (i.e., an invertible endomorphism). Then f is *orientation-preserving*, if f maps positively oriented bases to positively oriented bases, and f is *orientation-reversing*, if f maps positively oriented bases to negatively oriented bases. ♢

DEF Orientation-preserving, reversing

It is easy to see that f is orientation-preserving (resp., reversing) if and only if $\det(A) > 0$ (resp., $\det(A) < 0$), where A is the matrix corresponding to f: if B is the matrix whose columns form the vectors of a positively oriented basis, then the columns of AB are the images of these basis vectors. Since $\det(B) > 0$ we then have

$$\det(AB) = \det(A)\det(B) > 0 \iff \det(A) > 0\,.$$

We will now consider the volume of "linearly distorted cubes". In the plane $\mathbb{R}^2$, these are parallelograms. The general definition is as follows.

21.23. **Definition.** A *parallelotope* in $\mathbb{R}^n$ is the set

$$P(\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n) = \{t_1\boldsymbol{x}_1 + t_2\boldsymbol{x}_2 + \ldots + t_n\boldsymbol{x}_n \mid t_1, t_2, \ldots, t_n \in [0,1]\} \subset \mathbb{R}^n\,,$$

where $(\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n)$ is an n-tuple of vectors in $\mathbb{R}^n$. The parallelotope is *degenerate*, if the vectors $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n$ spanning it are linearly dependent (then $P(\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n)$ is contained in the proper linear subspace $\langle \boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n \rangle$ of $\mathbb{R}^n$). ♢

DEF Parallelotope

We will now look into how one can define the "oriented volume" of such a parallelotope. It should have the following properties.

(1) $\operatorname{vol} P(\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_n) = 1$ (the n-dimensional unit cube has volume 1).

(2) $\operatorname{vol} P(\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n)$ is positive (resp., negative), if $(\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n)$ is a positively (resp., negatively) oriented basis of $\mathbb{R}^n$.

(3) $\operatorname{vol} P(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{j-1}, \lambda\boldsymbol{x}_j, \boldsymbol{x}_{j+1}, \ldots, \boldsymbol{x}_n) = \lambda \operatorname{vol} P(\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n)$ for $\lambda \in \mathbb{R}$ and $j \in \{1, 2, \ldots, n\}$.

(4) $\operatorname{vol} P(\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n) = 0$, if $P(\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n)$ is degenerate.

(5) $\operatorname{vol} P(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{j-1}, \boldsymbol{x}_j + \boldsymbol{x}'_j, \boldsymbol{x}_{j+1}, \ldots, \boldsymbol{x}_n)$
$= \operatorname{vol} P(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{j-1}, \boldsymbol{x}_j, \boldsymbol{x}_{j+1}, \ldots, \boldsymbol{x}_n)$
$+ \operatorname{vol} P(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{j-1}, \boldsymbol{x}'_j, \boldsymbol{x}_{j+1}, \ldots, \boldsymbol{x}_n)$.

This last property becomes plausible when recalling the formula "base times height" for the volume: the height of $\boldsymbol{x}_j + \boldsymbol{x}'_j$ above the "base" that is given by the parallelotope spanned by the remaining vectors is the sum of the (oriented) heights of $\boldsymbol{x}_j$ and $\boldsymbol{x}'_j$.

21.24. **Theorem.** *The only map* vol *from the set of all parallelotopes in* $\mathbb{R}^n$ *to* $\mathbb{R}$ *that has the properties listed above is*

$$\operatorname{vol} P(\boldsymbol{x}_1, \boldsymbol{x}_2, \dots, \boldsymbol{x}_n) = \det(\boldsymbol{x}_1, \boldsymbol{x}_2, \dots, \boldsymbol{x}_n)\,,$$

where $\det(\boldsymbol{x}_1, \boldsymbol{x}_2, \dots, \boldsymbol{x}_n)$ *is the determinant of the matrix whose columns are the vectors* $\boldsymbol{x}_1, \boldsymbol{x}_2, \dots, \boldsymbol{x}_n$.

THM Determinant is oriented volume

Proof. The determinant does have the required properties (Theorem 21.3 and Theorem 21.8, also Definition 21.21). The uniqueness statement in Theorem 21.3, together with Theorem 21.8 (which implies that the analogous statement is also true in terms of columns instead of rows), imply that the determinant is the only map satisfying the properties (1), (3), (4) and (5) above. ❑

This means that we can use the determinant to measure volumes.

21.25. **Example.** The area of the triangle with vertices (x_1, y_1), (x_2, y_2) and (x_3, y_3) is

$$\frac{1}{2}\left|\det \begin{pmatrix} x_1 & x_2 & x_3 \\ y_1 & y_2 & y_3 \\ 1 & 1 & 1 \end{pmatrix}\right|.$$

We can shift the triangle so that the first vertex is at the origin. Then the desired area is half the area of the parallelogram spanned by $(x_2 - x_1, y_2 - y_1)$ and $(x_3 - x_1, y_3 - y_1)$. The oriented area of this parallelogram is

$$\det \begin{pmatrix} x_2 - x_1 & x_3 - x_1 \\ y_2 - y_1 & y_3 - y_1 \end{pmatrix}.$$

The determinant at the beginning can be transformed into this form by the column operations $\mathbf{II}_{2,1}(-1)$ and $\mathbf{II}_{3,1}(-1)$, followed by expansion along the third row. By taking the absolute value, we obtain the area instead of the oriented area. ♣

EXAMPLE Area of a triangle

The multiplicativity of the determinant leads to an interpretation of the determinant of an endomorphism that is relevant for applications in Analysis (for example, the change of variables formula for higher-dimensional integrals).

21.26. **Theorem.** *Let* $f\colon \mathbb{R}^n \to \mathbb{R}^n$ *be linear with corresponding matrix* A *and let further* $\boldsymbol{x}_1, \boldsymbol{x}_2, \dots, \boldsymbol{x}_n \in \mathbb{R}^n$. *Then*

$$\operatorname{vol} f\big(P(\boldsymbol{x}_1, \boldsymbol{x}_2, \dots, \boldsymbol{x}_n)\big) = \det(A) \operatorname{vol} P(\boldsymbol{x}_1, \boldsymbol{x}_2, \dots, \boldsymbol{x}_n)\,.$$

THM Determinant is scaling of volume

The determinant of an endomorphism tells us with which factor the oriented volume gets multiplied when the endomorphism is applied.

Proof. Let X be the matrix whose columns are the vectors $\boldsymbol{x}_1, \boldsymbol{x}_2, \dots, \boldsymbol{x}_n$. Since f is linear, we have that $f\big(P(\boldsymbol{x}_1, \boldsymbol{x}_2, \dots, \boldsymbol{x}_n)\big) = P\big(f(\boldsymbol{x}_1), f(\boldsymbol{x}_2), \dots, f(\boldsymbol{x}_n)\big)$. This gives

$$\begin{aligned} \operatorname{vol} f\big(P(\boldsymbol{x}_1, \boldsymbol{x}_2, \dots, \boldsymbol{x}_n)\big) &= \det\big(f(\boldsymbol{x}_1), f(\boldsymbol{x}_2), \dots, f(\boldsymbol{x}_n)\big) \\ &= \det(A\boldsymbol{x}_1, A\boldsymbol{x}_2, \dots, A\boldsymbol{x}_n) \\ &= \det(AX) = \det(A)\det(X) \\ &= \det(A) \operatorname{vol} P(\boldsymbol{x}_1, \boldsymbol{x}_2, \dots, \boldsymbol{x}_n)\,. \end{aligned}$$

❑

22. Eigenvalues and eigenvectors

Date: March 5, 2026

We have recently proved a classification theorem (Theorem 20.10) that can be interpreted as saying that the only property that distinguishes linear maps between two given finite-dimensional vector spaces is their *rank*: If $f\colon V \to W$ is linear with $\operatorname{rk}(f) = r$, then we can choose bases of V and W in such a way that f is given by the matrix M_r, and M_r depends only on r (and the dimensions of V and W).

Instead of linear maps between two different vector spaces V and W, we will now consider *endomorphisms* $f\colon V \to V$. Since now there is only one vector space around, we can choose only *one* basis. This gives us quite a bit less room to play with, leading to a significantly harder classification problem.

Of course, one can also consider linear maps $f\colon V \to V$ as a special case of linear maps $V \to W$, where it "happens to be the case" that $W = V$. Then one will allow the choice of different bases of V on the domain and the codomain side. For example, this leads to the change-of-basis matrices $\operatorname{Mat}_{B,B'}(\mathrm{id}_V)$. On the other hand, in this way the information gets lost that we actually have *the same* vector space on both sides, and not just two vector spaces that happen to be isomorphic (i.e., have the same dimension). For the classification problem that we want to study here (and then in more detail in the second-semester Linear Algebra course), it is however essential that we consider f as an endomorphism of V. Otherwise a statement of the form $f(v) = \lambda v$ (see Definition 22.3 below) would not be meaningful (or at least would not carry over to the matrices describing f).

We begin by recording how the matrices of f with respect to different bases of V are related to each other.

THM Matrices of an endomorphism

22.1. **Theorem.** *Let V be a K-vector space with basis $B = (b_1, b_2, \dots, b_n)$ and let f be an endomorphism of V. Let further $A = \operatorname{Mat}_{B,B}(f) \in \operatorname{Mat}(n, K)$ be the matrix of f relative to B. Then*

$$\{\operatorname{Mat}_{B',B'}(f) \mid B' \text{ Basis of } V\} = \{PAP^{-1} \mid P \in \mathrm{GL}(n, K)\}\,.$$

Proof. We have $\operatorname{Mat}_{B',B'}(f) = \operatorname{Mat}_{B,B'}(\mathrm{id}_V)\operatorname{Mat}_{B,B}(f)\operatorname{Mat}_{B',B}(\mathrm{id}_V) = PAP^{-1}$, where $P = \operatorname{Mat}_{B,B'}(\mathrm{id}_V) \in \mathrm{GL}(n, K)$. Conversely, every matrix $P \in \mathrm{GL}(n, K)$ can be written as $P = \operatorname{Mat}_{B,B'}(\mathrm{id}_V)$; see Corollary 20.4. ❑

DEF Similarity of matrices

22.2. **Definition.** Let K be a field and $n \in \mathbb{N}$. Two matrices $A, A' \in \operatorname{Mat}(n, K)$ are *similar*, if there is a matrix $P \in \mathrm{GL}(n, K)$ such that $A' = PAP^{-1}$. ♢

One shows in essentially the same way as for equivalence of matrices that similarity of matrices is an equivalence relation.

If A is a matrix of an endomorphism f of V, then the matrices of f with respect to arbitrary bases of V are exactly the matrices similar to A.

The classification of matrices up to similarity (and with it the classification of endomorphisms of finite-dimensional vector spaces) is fairly complicated. It is provided by the *Jordan Normal Form* that we will discuss in the next semester. For now, we will restrict ourselves to studying simpler "invariants" (meaning data that depend only on f and not on the chosen basis).

The main idea is to compare the endomorphism $f\colon V \to V$ with other particularly simple endomorphisms. The simplest possible endomorphisms certainly are just multiplication by a scalar $\lambda \in K$, $v \mapsto \lambda v$. We can then ask whether there are

elements of V that have the same behavior under f and under this map. This leads to the following definition.

22.3. **Definition.** Let V be a K-vector space and $f \in \mathrm{End}(V)$. A scalar $\lambda \in K$ is an *eigenvalue* of f, if there exists a vector $\mathbf{0} \neq v \in V$ such that $f(v) = \lambda v$. Every such vector is an *eigenvector* of f for the eigenvalue λ. ♢

DEF
Eigenvalue
eigenvector

Note the condition $v \neq \mathbf{0}$! Without it, the definition makes no sense, as then every λ would be an eigenvalue (since $f(\mathbf{0}) = \mathbf{0} = \lambda\mathbf{0}$).

22.4. **Definition.** Let V be a K-vector space, $\lambda \in K$ and $f \in \mathrm{End}(V)$. The linear subspace

$$E_\lambda(f) = \{v \in V \mid f(v) = \lambda v\} = \ker(\lambda \,\mathrm{id}_V - f)$$

of V is the *λ-eigenspace* or the *eigenspace for the eigenvalue λ* of f.

The dimension $\dim E_\lambda(f)$ of the λ-eigenspace of f is the *geometric multiplicity* of the eigenvalue λ of f. ♢

DEF
Eigenspace
geometric
multiplicity

The eigenspace $E_\lambda(f)$ consists of the zero vector together with all eigenvectors for the eigenvalue λ. This shows that λ is an eigenvalue of f if and only if $E_\lambda(f) \neq \{\mathbf{0}\}$, i.e., the geometric multiplicity is (strictly) positive.

22.5. **Examples.**

EXAMPLES
Eigenvalues
eigenspaces

(1) $E_0(f)$ is precisely the kernel of f. Hence zero is an eigenvalue of f if and only if f is not injective. If V is finite-dimensional, then this is also equivalent with f not being an isomorphism.

(2) Let $f\colon \mathbb{R}^2 \to \mathbb{R}^2$, $(x, y) \mapsto (y, x)$. Then f has the eigenvalues 1 and -1 since $v_1 = (1, 1) \in E_1(f)$ and $v_{-1} = (1, -1) \in E_{-1}(f)$. It is easy to see that both eigenspaces are one-dimensional.

(3) Let $f\colon \mathcal{C}^\infty(\mathbb{R}) \to \mathcal{C}^\infty(\mathbb{R})$, $h \mapsto h'$. Then f has *every* $\lambda \in \mathbb{R}$ as an eigenvalue, and we have $E_\lambda(f) = \langle x \mapsto e^{\lambda x} \rangle$. (Proof as in Example 19.11.) ♣

We will see that the situation as in Example (2) above is fairly typical for endomorphisms of finite-dimensional vector spaces. We being by showing that there cannot be too many eigenvalues.

22.6. **Theorem.** *Let V be a K-vector space and $f \in \mathrm{End}(V)$. Let $\lambda_1, \ldots, \lambda_m \in K$ be pairwise distinct and for each $j \in \{1, 2, \ldots, m\}$ let $v_j \in V$ be an eigenvector of f for the eigenvalue λ_j. Then $(v_1, v_2, \ldots, v_m)$ is linearly independent.*

THM
Linear
independence
of
eigenvectors

Proof. We proceed by induction on m. The case $m = 0$ is clear (the empty family is linearly independent). So we now assume that $m > 0$ and the claim has been shown for $m-1$. As usual, we consider a linear combination $\alpha_1 v_1 + \alpha_2 v_2 + \ldots + \alpha_m v_m = \mathbf{0}$ with $\alpha_1, \alpha_2, \ldots, \alpha_m \in K$; we need to show that $\alpha_j = 0$ for all j. We have

$$\begin{aligned}
\mathbf{0} &= \lambda_m \mathbf{0} - f(\mathbf{0}) \\
&= \lambda_m(\alpha_1 v_1 + \alpha_2 v_2 + \ldots + \alpha_m v_m) - f(\alpha_1 v_1 + \alpha_2 v_2 + \ldots + \alpha_m v_m) \\
&= \lambda_m \alpha_1 v_1 + \lambda_m \alpha_2 v_2 + \ldots + \lambda_m \alpha_m v_m - \alpha_1 \lambda_1 v_1 - \alpha_2 \lambda_2 v_2 - \ldots - \alpha_m \lambda_m v_m \\
&= \alpha_1(\lambda_m - \lambda_1) v_1 + \alpha_2(\lambda_m - \lambda_2) v_2 + \ldots + \alpha_{m-1}(\lambda_m - \lambda_{m-1}) v_{m-1}\,.
\end{aligned}$$

From the induction hypothesis we obtain

$$\alpha_1(\lambda_m - \lambda_1) = \alpha_2(\lambda_m - \lambda_2) = \ldots = \alpha_{m-1}(\lambda_m - \lambda_{m-1}) = 0\,.$$

Since $\lambda_m \neq \lambda_1, \lambda_2, \ldots, \lambda_{m-1}$, this gives $\alpha_1 = \alpha_2 = \ldots = \alpha_{m-1} = 0$. The original equation is then reduced to $\alpha_m v_m = \mathbf{0}$. Since $v_m \neq \mathbf{0}$, this finally also implies $\alpha_m = 0$. ❑

22.7. **Corollary.** *Let V be a K-vector space and $f \in \operatorname{End}(V)$. Let further $\lambda_1, \lambda_2, \ldots, \lambda_m$ be pairwise distinct elements of K, and for $j \in \{1, 2, \ldots, m\}$ let $v_{j1}, v_{j2}, \ldots, v_{jn_j}$ (with suitable $n_j \in \mathbb{N}$) be linearly independent elements of $E_{\lambda_j}(f)$. Then the vectors v_{ji} (with $j \in \{1, 2, \ldots, m\}$ and $i \in \{1, 2, \ldots, n_j\}$) are linearly independent. In particular, we have*

$$\dim E_{\lambda_1}(f) + \dim E_{\lambda_2}(f) + \ldots + \dim E_{\lambda_m}(f) \leq \dim V\,.$$

COR Dimension of eigenspaces

If V is finite-dimensional, then f can therefore have at most $\dim V$ eigenvalues. More precisely, the sum of the geometric multiplicities of the eigenvalues is at most $\dim V$.

Proof. Let

$$\sum_{j=1}^{m} \sum_{i=1}^{n_j} \alpha_{ji} v_{ji} = \mathbf{0}$$

with $\alpha_{ji} \in K$. Let $v_j = \sum_{i=1}^{n_j} \alpha_{ji} v_{ji} \in E_{\lambda_j}(f)$; then $v_1 + v_2 + \ldots + v_m = \mathbf{0}$. Theorem 22.6 now implies $v_1 = v_2 = \ldots = v_m = \mathbf{0}$ since any nonzero vectors occurring would need to be linearly independent, contradicting the fact that their sum is zero. Since $(v_{j1}, v_{j2}, \ldots, v_{jn_j})$ is linearly independent, $v_j = \mathbf{0}$ then implies that $\alpha_{ji} = 0$ for all $i \in \{1, 2, \ldots, n_j\}$. Since $j \in \{1, 2, \ldots, m\}$ was arbitrary, all $\alpha_{ji} = 0$. This shows linear independence.

In the case that $\dim E_{\lambda_j}(f) = \infty$ for some j, the last claim follows from the fact that $E_{\lambda_j}(f)$ is a linear subspace of V, which implies $\dim V = \infty$. Otherwise, we can, for each j, choose the tuple $(v_{j1}, v_{j2}, \ldots, v_{jn_j})$ as a basis of $E_{\lambda_j}(f)$. The linear independence of the v_{ji} then implies that

$$\begin{aligned} \dim V &\geq \#\{v_{ji} \mid j \in \{1, 2, \ldots, m\}, i \in \{1, 2, \ldots, n_j\}\} \\ &= n_1 + n_2 + \ldots + n_m \\ &= \dim E_{\lambda_1}(f) + \dim E_{\lambda_2}(f) + \ldots + \dim E_{\lambda_m}(f)\,. \end{aligned}$$ ❑

Now how can we determine the eigenvalues (and then the eigenspaces) of a given endomorphism f? We choose a basis and find the matrix A of f relative to this basis. Then the question is how we can find the relevant information from the matrix A. We first transfer the notions of eigenvalue and so on to matrices.

22.8. **Definition.** Let K be a field, $n \in \mathbb{N}$ and $A \in \operatorname{Mat}(n, K)$. Let $\lambda \in K$. Then λ is an *eigenvalue* of A, if there is a column vector $\mathbf{0} \neq \boldsymbol{x} \in K^n$ such that $A\boldsymbol{x} = \lambda\boldsymbol{x}$. In this case, $\boldsymbol{x}$ is an *eigenvector* of A for the eigenvalue λ. The linear subspace

$$E_\lambda(A) = \{\boldsymbol{x} \in K^n \mid A\boldsymbol{x} = \lambda\boldsymbol{x}\} = \ker(\lambda I_n - A)$$

is the *λ-eigenspace* of A or the *eigenspace* of A for the eigenvalue λ; its dimension is the *geometric multiplicity* of the eigenvalue λ of A. ♢

DEF Eigenvalue etc. for matrices

The eigenvalues of A and their geometric multiplicities correspond to those of f.

The following simple observation is the key for the determination of the eigenvalues.

LEMMA Characterizing eigenvalues

22.9. **Lemma.** *Let K be a field, $\lambda \in K$, $n \in \mathbb{N}$ and $A \in \mathrm{Mat}(n, K)$. λ is an eigenvalue of A if and only if $\det(\lambda I_n - A) = 0$. The geometric multiplicity of the eigenvalue λ is $\dim \ker(\lambda I_n - A) = n - \mathrm{rk}(\lambda I_n - A)$.*

Proof. We have the following chain of equivalences.

$$\begin{aligned} \lambda \text{ is an eigenvalue of } A &\iff E_\lambda(A) \neq \{\mathbf{0}\} \\ &\iff \ker(\lambda I_n - A) \neq \{\mathbf{0}\} \\ &\iff \det(\lambda I_n - A) = 0 \end{aligned}$$

The last claim follows from the definition of the geometric multiplicity and the Rank-Nullity Theorem 13.17. ❑

EXAMPLE Determination of the eigenvalues

22.10. **Example.** We consider

$$A = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \in \mathrm{Mat}(2, \mathbb{R})\,.$$

(This is the matrix of $f\colon (x, y) \mapsto (y, x)$ as in Example 22.5 (2).) Then we have for $\lambda \in \mathbb{R}$

$$\det(\lambda I_2 - A) = \begin{vmatrix} \lambda & -1 \\ -1 & \lambda \end{vmatrix} = \lambda^2 - 1 = (\lambda - 1)(\lambda + 1)\,.$$

This vanishes exactly for $\lambda = 1$ and $\lambda = -1$, hence these are the eigenvalues of A. We can compute bases of the eigenspaces $E_\lambda(A)$ using the row echelon form algorithm, applied to $\lambda I_2 - A$. For $\lambda = 1$ we obtain

$$\lambda I_2 - A = I_2 - A = \begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix} \overset{\mathbf{II}_{2,1}(1)}{\longrightarrow} \begin{pmatrix} 1 & -1 \\ 0 & 0 \end{pmatrix};$$

this gives the (singleton) basis $\big((1, 1)\big)$ of $E_1(A)$. For $\lambda = -1$ we get instead

$$\lambda I_2 - A = -I_2 - A = \begin{pmatrix} -1 & -1 \\ -1 & -1 \end{pmatrix} \overset{\mathbf{I}_1(-1)}{\longrightarrow} \begin{pmatrix} 1 & 1 \\ -1 & -1 \end{pmatrix} \overset{\mathbf{II}_{2,1}(1)}{\longrightarrow} \begin{pmatrix} 1 & 1 \\ 0 & 0 \end{pmatrix};$$

the basis is $\big((-1, 1)\big)$. ♣

We see in this example that the determinant whose vanishing indicates that λ is an eigenvalue is a *polynomial* in λ (with coefficients in K). So before we can continue our investigations, we need to talk about polynomials more generally.

Interlude: Polynomials.

DEF
Polynomial
polynomial
ring
degree
leading
coefficient
monic

22.11. **Definition.** Let K be a field, A *polynomial* in the *indeterminate* (or *variable*) X over K is an expression of the form

$$p = a_n X^n + a_{n-1} X^{n-1} + \ldots + a_1 X + a_0 \,,$$

where $n \in \mathbb{N}$ and $a_0, a_1, \ldots, a_n \in K$. a_j is the *jth coefficient* of p or the *coefficient of X^j* in p. We set $a_j = 0$ for $j > n$. If $a_n \neq 0$, then the polynomial has *degree* n, $\deg(p) = n$. In this case, a_n is the *leading coefficient* of p. If $a_n = 1$, then p is *monic*. If all $a_j = 0$, then p is the *zero polynomial*; its degree is defined as $\deg(\mathbf{0}) = -\infty$. If $n = 0$, then p is *constant* (i.e., $p = \mathbf{0}$ or $\deg(p) = 0$). We write $K[X]$ for the set of all polynomials in X over K.

Let $q = b_m X^m + \ldots + b_1 X + b_0$ be another polynomial. The polynomials p and q are *equal*, $p = q$, if and only if their coefficients agree: $a_j = b_j$ for all $j \in \mathbb{N}$ (with the convention $a_j = 0$ for $j > n$ and $b_j = 0$ for $j > m$). The *sum* of p and q is

$$p + q = \sum_{j=0}^{\max\{m,n\}} (a_j + b_j) X^j \,;$$

we have $\deg(p+q) \leq \max\{\deg(p), \deg(q)\}$. The *product* of p and q is

$$p \cdot q = \sum_{k=0}^{m+n} \Big(\sum_{i,j\colon i+j=k} a_i b_j \Big) X^k \,;$$

we have $\deg(pq) = \deg(p) + \deg(q)$. We identify K with the subset of $K[X]$ consisting of constant polynomials, i.e., we consider K to be a subset of $K[X]$. The set $K[X]$ together with the addition and multiplication just defined is a commutative ring, the *polynomial ring* in X over K. The restriction of the multiplication to $K \times K[X]$ turns $K[X]$ into an infinite-dimensional K-vector space with basis $(1, X, X^2, X^3, \ldots)$. ♢

If you don't like that the definition above talks about an "expression of the form ..." without saying what this actually *is*, then you are welcome to continue reading.

We can put the definition on a stable formal footing by setting

$$K[X] = \{(a_n)_{n\in\mathbb{N}} \in K^{\mathbb{N}} \mid \exists N \in \mathbb{N}\ \forall n > N \colon a_n = 0\} \,.$$

This is the set of all finite sequences of elements of K in the sense that all terms except finitely many are zero. We further set $X = (0, 1, 0, 0, 0, \ldots) \in K[X]$ and define the map $i\colon K \to K[X]$, $a \mapsto (a, 0, 0, 0, \ldots)$. The addition in $K[X]$ is defined component-wise, multiplication by X is given by

$$X \cdot (a_0, a_1, a_2, \ldots) = (0, a_0, a_1, a_2, \ldots)$$

and multiplication by $i(a)$ is given by

$$i(a) \cdot (a_0, a_1, a_2, \ldots) = (aa_0, aa_1, aa_2, \ldots) \,.$$

Then

$$(a_0, a_1, \ldots, a_n, 0, 0, 0, \ldots) = i(a_0) + i(a_1)X + i(a_2)X^2 + \ldots + i(a_n)X^n \,,$$

and we define multiplication on $K[x]$ in such a way that the associative and the distributive laws hold. We identify K with its image under i; we therefore just write a in place of $i(a)$. One still has to verify the ring axioms (easy, but tedious). The linear structure of $K[X]$ is that as a linear subspace of $K^{\mathbb{N}}$.

This approach still works when we leave out the finiteness condition in the definition. One then obtains the ring $K[\![X]\!]$ of *formal power series* in X over K. For $K = \mathbb{R}$ or $\mathbb{C}$ these power series play an important role in (real and complex) Analysis.

We can evaluate polynomials.

22.12. **Definition.** Let K be a field and let $p = a_n X^n + \ldots + a_1 X + a_0 \in K[X]$ be a polynomial. For $\lambda \in K$ the *value of p at λ* is given by

$$p(\lambda) = a_n \lambda^n + \ldots + a_1 \lambda + a_0 \,.$$

λ is a *zero* or *root* of p, if $p(\lambda) = 0$. For $p, q \in K[X]$ and $\lambda \in K$ we then have $(p+q)(\lambda) = p(\lambda) + q(\lambda)$ and $(p \cdot q)(\lambda) = p(\lambda) \cdot q(\lambda)$. ♢

DEF Values and zeros of polynomials

So a polynomial $p \in K[X]$ leads to a *polynomial function* $K \to K$, $\lambda \mapsto p(\lambda)$. The map $K[X] \to \mathrm{Map}(K, K)$ that associates to a polynomial the corresponding polynomial function is injective when the field K is infinite. This follows from the following result.

22.13. **Theorem.** *Let K be a field, $n \in \mathbb{N}$ and $x_1, x_2, \ldots, x_n \in K$ pairwise distinct. Let further $y_1, y_2, \ldots, y_n \in K$. Then there exists a uniquely determined polynomial $p \in K[X]$ with $\deg(p) < n$ such that $p(x_j) = y_j$ for all $j \in \{1, 2, \ldots, n\}$.*

THM Uniqueness of polynomials

Proof. This statement is the generalization of the result in Example 13.15 from $\mathbb{R}$ to arbitrary fields. The proof is the same as before. ❑

22.14. **Corollary.** *A polynomial $p \in K[X]$ with $\deg(p) = n \in \mathbb{N}$ cannot have more than n zeros in K.*

COR Zeros of polynomials

Proof. Assume for a contradiction that p has $n+1$ zeros $x_1, x_2, \ldots, x_{n+1} \in K$. Then p must be the uniquely determined polynomial of degree $< n+1$ that satisfies $p(a_j) = 0$ for all $j \in \{1, 2, \ldots, n+1\}$. The zero polynomial has this property, so we must have $p = \mathbf{0}$. But this contradicts the assumption $\deg(p) = n$. ❑

22.15. **Corollary.** *If the field K has infinitely many elements, then the map $K[X] \to \mathrm{Map}(K, K)$ that maps a polynomial to the corresponding polynomial function is injective.*

COR Polynomials and polynomial functions

This says that a polynomial $p \in K[X]$ is uniquely determined by its values $p(\lambda)$ for $\lambda \in K$. For example, we can identify the vector space P of polynomial functions with $\mathbb{R}[X]$.

Proof. We write Φ for the map $K[X] \to \mathrm{Map}(K, K)$. Φ is linear, so it suffices to show that $\ker(\Phi) = \{\mathbf{0}\}$. So let $p \in \ker(\Phi)$. Then $p(\lambda) = 0$ for all $\lambda \in K$, hence p has infinitely many zeros in K (here we use that K is infinite). By Corollary 22.14, p must be the zero polynomial. ❑

For finite fields K the statement is wrong. If $\#K = q < \infty$, then $\dim_K \mathrm{Map}(K, K) = q$ (a map $f \colon K \to K$ is determined uniquely by the q values $f(a)$ for $a \in K$). On the other hand, we have $\dim_K K[X] = \infty$, hence there cannot exist am injective linear map $K[X] \to \mathrm{Map}(K, K)$.

In this case, the kernel of Φ consists of all polynomials that have all elements of K as zeros. One can show that

$$\prod_{a \in K} (X - a) = X^q - X \,;$$

the kernel then consists exactly of the multiples of $X^q - X$ (as polynomials).

In a similar way as we can divide integers with remainder, there is also a division with remainder for polynomials ("polynomial long division").

22.16. **Theorem.** *Let K be a field and let $f, g \in K[X]$ be such that g is monic. Then there are uniquely determined polynomials q ("quotient") and r ("remainder") in $K[X]$ such that $f = qg + r$ and $\deg(r) < \deg(g)$.* **THM** Polynomial division

Proof. We begin by showing existence. Let $\deg(g) = m$, so

$$g = X^m + b_{m-1}X^{m-1} + \ldots + b_1 X + b_0 \,.$$

We consider g as fixed and proceed by induction on $\deg(f)$. If $\deg(f) < m$, then $q = 0$ and $r = f$ satisfy the conditions. We can therefore assume that $n = \deg(f) \geq m$ and that the existence statement is already proved for $\deg(f) < n$. We have $f = a_n X^n + a_{n-1}X^{n-1} + \ldots + a_0$, so we obtain

$$\tilde{f} = f - a_n X^{n-m} g = (a_{n-1} - a_n b_{m-1})X^{n-1} + \ldots \,;$$

the degree of $\tilde{f}$ is therefore smaller than n. By the induction hypothesis there exist $\tilde{q}, r \in K[X]$ such that $\tilde{f} = \tilde{q}g + r$ and $\deg(r) < m$. We set $q = a_n X^{n-m} + \tilde{q}$; then

$$f = a_n X^{n-m} g + \tilde{f} = a_n X^{n-m} g + \tilde{q}g + r = qg + r$$

as desired.

Now we show uniqueness. Let $q_1, q_2, r_1, r_2 \in K[X]$ with $f = q_1 g + r_1 = q_2 g + r_2$ and $\deg(r_1), \deg(r_2) < \deg(g)$. This implies that

$$(q_1 - q_2)g = r_2 - r_1 \,.$$

The right hand side has degree $< \deg(g)$. If we had $q_1 \neq q_2$, then the left hand side would have degree $\deg(q_1 - q_2) + \deg(g) \geq \deg(g)$, a contradiction. This shows that $q_1 = q_2$ and then also $r_1 = r_2$. ❑

This proof translates directly into the usual algorithm for polynomial long division. We keep subtracting suitable multiples of g from f until we are left with a polynomial whose degree is less than the degree of g.

22.17. **Corollary.** *Let K be a field, $p \in K[X]$ and $\lambda \in K$. We write*

$$p = q \cdot (X - \lambda) + r$$

as in Theorem 22.16. Then $r = p(\lambda)$ is constant. In particular, λ is a zero of p if and only if $r = 0$. **COR** Zeros

Proof. r is constant since $\deg(r) < 1 = \deg(X - \lambda)$. Furthermore,

$$p(\lambda) = q(\lambda)(\lambda - \lambda) + r = r \,.$$ ❑

If λ is a zero of p, then this shows that $p = (X - \lambda) \cdot q$ with some polynomial $q \in K[X]$. If λ is also a zero of q, then $p = (X - \lambda)^2 \cdot \tilde{q}$ and so on. This leads to the following definition.

22.18. **Definition.** Let K be a field, $\mathbf{0} \neq p \in K[X]$ a polynomial and $\lambda \in K$. The *multiplicity* of the zero λ of p is the largest number $n \in \mathbb{N}$ such that p can be written as $p = (X - \lambda)^n \cdot q$ with some polynomial $q \in K[X]$. In this case, $q(\lambda) \neq 0$. ♢ **DEF** Multiplicity of a zero

So λ is a zero if and only if its multiplicity as a zero is positive.

EXAMPLES
Multiplicities

22.19. **Examples.** We consider $K = \mathbb{R}$. For $p = X^3 - X^2 - X + 1 \in \mathbb{R}[X]$ we have

$$p = (X-1)^2(X+1)\,,$$

hence p has the zeros 1 (with multiplicity 2: a "double" zero) and -1 (with multiplicity 1: a "simple" zero).

For $q = X^3 + X^2 + X + 1 \in \mathbb{R}[X]$, on the other hand, we have

$$q = (X+1)(X^2+1)\,,$$

hence q has only the (simple) root -1 in $\mathbb{R}$; the second factor X^2+1 takes only strictly positive values and so cannot have a real zero. If we instead consider q as a polynomial in $\mathbb{C}[X]$, then we have

$$q = (X+1)(X+\boldsymbol{i})(X-\boldsymbol{i})\,;$$

so q has the three complex (simple) roots -1, $\boldsymbol{i}$ and $-\boldsymbol{i}$. ♣

EXAMPLE
Zeros

22.20. **Example.** How can we determine the zeros of a polynomial p? This is easy when $\deg(p) = 1$, and in the case $\deg(p) = 2$, one can use the well-known quadratic formula. For general polynomials the determination of their zeros is a difficult problem. In simple cases (for example, in exam problems) it is frequently possible to guess a zero α or find it by trial and error. Then we can divide p by $X - \alpha$ to obtain a polynomial of smaller degree whose zeros have to be found.

As an example, we consider

$$p = X^4 + X^3 - 6X^2 - 2X + 4 \in \mathbb{R}[X]\,.$$

We try if 0, 1, -1 are zeros and find that $p(-1) = 0$. We divide p by $X+1$ to obtain

$$p = (X+1)(X^3 - 6X + 4)\,,$$

so we still have to find the zeros of $p_1 = X^3 - 6X + 4$. Further trying gives the zero 2. We divide p_1 by $X - 2$ and obtain

$$p_1 = (X-2)(X^2 + 2X - 2)\,.$$

We can apply the formula for quadratic equations to the remaining factor; this gives the remaining zeros

$$\frac{-2 \pm \sqrt{2^2 - 4\cdot(-2)}}{2} = -1 \pm \sqrt{3}\,.$$ ♣

Now back to eigenvalues and eigenspaces. We have seen that the expression

$$\det(\lambda I_n - A)$$

decides whether λ is an eigenvalue of A or not. If $A = (a_{ij})$, then this determinant has the following form.

$$\begin{vmatrix} \lambda - a_{11} & -a_{12} & -a_{13} & \cdots & -a_{1n} \\ -a_{21} & \lambda - a_{22} & -a_{23} & \cdots & -a_{2n} \\ -a_{31} & -a_{32} & \lambda - a_{33} & \cdots & -a_{3n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ -a_{n1} & -a_{n2} & -a_{n3} & \cdots & \lambda - a_{nn} \end{vmatrix}$$

Plugging this into the Leibniz formula, we obtain

$$\begin{aligned} &(\lambda - a_{11})(\lambda - a_{22}) \cdots (\lambda - a_{nn}) + \text{terms with} \le n-2 \text{ factors } \lambda - a_{jj} \\ &\quad = \lambda^n - (a_{11} + \ldots + a_{nn})\lambda^{n-1} + \ldots + (-1)^n \det(A)\,. \end{aligned}$$

This has the form $p(\lambda)$ with a monic polynomial $p \in K[X]$ of degree n.

22.21. **Definition.** Let K be a field, $n \in \mathbb{N}$ and $A \in \mathrm{Mat}(n, K)$. The polynomial $\chi_A = \det(XI_n - A) \in K[X]$ is the *characteristic polynomial* of A. ♢

DEF Characteristic polynomial

The Leibniz formula shows that one can define the determinant more generally for matrices with entries in a commutative ring R; the determinant then is also an element of R. The multiplicativity and symmetry of the determinant are also valid in this more general context. In the definition above, we use that with the ring $K[X]$ and the matrix $XI_n - A \in \mathrm{Mat}(n, K[X])$.

One frequently also finds the definition $\det(A - XI_n)$ for the characteristic polynomial in the literature. This definition differs from the one given here only by a factor of $(-1)^n$. This does not change what the roots are, so the difference is irrelevant for the computation of the eigenvalues and their algebraic multiplicities (see below). The disadvantage of this other version is that the polynomial is not monic when n is odd (but has leading coefficient -1).

We have seen that the eigenvalues of A are exactly the zeros of the characteristic polynomial of A.

22.22. **Example.** What are the eigenvalues of the matrix

EXAMPLE Eigenvalues

$$A = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 1 & 1 \\ 1 & 2 & 3 \end{pmatrix} \in \mathrm{Mat}(3, \mathbb{R}) \; ?$$

We determine the characteristic polynomial:

$$\begin{aligned} \det(XI_3 - A) &= \begin{vmatrix} X-1 & 0 & 0 \\ -1 & X-1 & -1 \\ -1 & -2 & X-3 \end{vmatrix} \\ &= (X-1) \begin{vmatrix} X-1 & -1 \\ -2 & X-3 \end{vmatrix} \\ &= (X-1)\big((X-1)(X-3) - 1 \cdot 2\big) \\ &= (X-1)(X^2 - 4X + 1) \end{aligned}$$

One eigenvalue is$\lambda_1 = 1$; we find the other two with the quadratic formula as

$$\lambda_2 = 2 + \sqrt{3} \qquad \text{and} \qquad \lambda_3 = 2 - \sqrt{3}\,. \qquad \clubsuit$$

We can extend the definitions of determinant and characteristic polynomial to endomorphisms.

22.23. **Definition.** Let K be a field, V a finite-dimensional K-vector space and $f \in \mathrm{End}(V)$. Let B be an arbitrary basis of V and let $A = \mathrm{Mat}_{B,B}(f)$. Then we define the *determinant* of f as $\det(f) = \det(A)$ and the *characteristic polynomial* χ_f of f as the characteristic polynomial of A. ♢

DEF Determinant, char. pol. of endomorphisms

This definition makes sense, because it does not depend on the choice of the basis B. If $A' = \mathrm{Mat}_{B',B'}(f)$ with a different basis B' of V, then there is a matrix $P \in \mathrm{GL}(n, K)$ (where $n = \dim V$), such that $A' = PAP^{-1}$. Then

$$\begin{aligned} \det(A') &= \det(PAP^{-1}) = \det(P)\det(A)\det(P^{-1}) \\ &= \det(A)\det(PP^{-1}) = \det(A)\det(I_n) = \det(A)\,. \end{aligned}$$

We also have that $P(XI_n - A)P^{-1} = XPI_nP^{-1} - PAP^{-1} = XI_n - A'$, and the same computation shows that A and A' have the same characteristic polynomial.

22.24. **Definition.** Let K be a field, $n \in \mathbb{N}$, $A \in \mathrm{Mat}(n, K)$ and $\lambda \in K$. The *algebraic multiplicity* of λ as an eigenvalue of A is the multiplicity of λ as a zero of the characteristic polynomial of A. We define the *algebraic multiplicity* of λ as an eigenvalue of an endomorphism f of a finite-dimensional K-vector space V in the same way. ♢

DEF Algebraic multiplicity

We now have defined two multiplicities of eigenvalues, the geometric one and the algebraic one. What is their relation? What we know so far is

$$\text{geom. multiplicity} > 0 \iff \text{eigenvalue} \iff \text{alg. multiplicity} > 0$$

Are the two multiplicities necessarily equal?

22.25. **Example.** Let

$$A = \begin{pmatrix} \alpha & 1 \\ 0 & \alpha \end{pmatrix} \in \mathrm{Mat}(2, K)\,.$$

The characteristic polynomial of A is $(X-\alpha)^2$, so α has the algebraic multiplicity 2. On the other hand, $E_\alpha(A) = \langle (1,0) \rangle$ (for $\mathrm{rk}(\alpha I_2 - A) = 1$), so α has the geometric multiplicity 1. ♣

EXAMPLE alg. $\neq$ geom. multiplicity

So the multiplicities can differ. There is, however, one relation.

22.26. **Theorem.** *Let K be a field, V a finite-dimensional K-vector space, $f \in \mathrm{End}(V)$ and $\lambda \in K$. Then the geometric multiplicity of λ as an eigenvalue of f is not larger than its algebraic multiplicity.*

THM geom. $\leq$ alg. multiplicity

The analogous statement clearly also holds for matrices $A \in \mathrm{Mat}(n, K)$.

Proof. Let $m = \dim E_\lambda(f)$ be the geometric multiplicity, let $n = \dim V$ and let $(b_1, b_2, \ldots, b_m)$ be a basis of $E_\lambda(f)$. We can extend this basis to a basis $B = (b_1, b_2, \ldots, b_n)$ of V. Then

$$A = \mathrm{Mat}_{B,B}(f) = \left(\begin{array}{c|c} \lambda I_m & D \\ \hline \mathbf{0}_{n-m,m} & C \end{array}\right)$$

with matrices $D \in \mathrm{Mat}(m \times (n-m), K)$ and $C \in \mathrm{Mat}(n-m, K)$ since for $j \in \{1, 2, \ldots, m\}$ we have $f(b_j) = \lambda b_j$; the jth column of A therefore contains the λ-multiple of the jth standard basis vector. The characteristic polynomial of f then has the form

$$\det(XI_n - A) = \det\left(\begin{array}{c|c} (X-\lambda)I_m & -D \\ \hline \mathbf{0}_{n-m,m} & XI_{n-m} - C \end{array}\right) = (X-\lambda)^m \det(XI_{n-m} - C)\,.$$

Here we have expanded the determinant m-times along the first column. This shows that the multiplicity of λ as a zero of the characteristic polynomial is at least m. ❑

One can extend the formula above to more general block matrices (the proof is an exercise): Given $A \in \mathrm{Mat}(m, K)$, $B \in \mathrm{Mat}(m \times n, K)$ and $C \in \mathrm{Mat}(n, K)$, we have

$$\det\left(\begin{array}{c|c} A & B \\ \hline \mathbf{0}_{n,m} & C \end{array}\right) = \det(A)\det(C)\,.$$