Analysis I:
Calculus of One Real Variable

Peter Philip

Lecture Notes
Created for the Class of Winter Semester 2015/2016 at LMU Munich

January 19, 2021

Contents

1 Foundations: Mathematical Logic and Set Theory 5
  1.1 Introductory Remarks ........................................... 5
  1.2 Propositional Calculus .......................................... 5
    1.2.1 Statements .................................................. 5
    1.2.2 Logical Operators ......................................... 6
    1.2.3 Rules ......................................................... 9
  1.3 Set Theory ....................................................... 13
  1.4 Predicate Calculus ............................................. 17

2 Functions and Relations 23
  2.1 Functions ....................................................... 23
  2.2 Relations ....................................................... 31

3 Natural Numbers, Induction, and the Size of Sets 36
  3.1 Induction and Recursion ....................................... 36
  3.2 Cardinality: The Size of Sets ................................ 43

4 Real Numbers 46
  4.1 The Real Numbers as a Complete Totally Ordered Field ... 46
  4.2 Important Subsets ............................................. 48

*E-Mail: philip@math.lmu.de
†Resources used in the preparation of this text include [Kön04, Kun80, Wal04].
C Algebraic Structures
   C.1 Groups ................................................. 207
   C.2 Rings ................................................. 211
   C.3 Fields ................................................. 212

D Construction of the Real Numbers
   D.1 Natural Numbers ........................................ 213
   D.2 Interlude: Orders on Groups ................................ 218
   D.3 Integers ................................................. 219
   D.4 Rational Numbers ....................................... 223
   D.5 Real Numbers ........................................... 227
   D.6 Uniqueness .............................................. 233

E Series: Additional Material
   E.1 Riemann Rearrangement Theorem ........................... 237
   E.2 $b$-Adic Representations of Real Numbers ............... 239

F Cardinality of $\mathbb{R}$ and Some Related Sets .......... 244

G Partial Fraction Decomposition .................................. 248

H Irrationality of $e$ and $\pi$ .................................. 253
   H.1 Irrationality of $e$ ..................................... 253
   H.2 Irrationality of $\pi$ ..................................... 254

I Trigonometric Functions ........................................ 256
   I.1 Additional Trigonometric Formulas ....................... 256

J Differential Calculus ........................................... 256
   J.1 Continuous, But Nowhere Differentiable Functions ....... 256

References ......................................................... 259
1 Foundations: Mathematical Logic and Set Theory

1.1 Introductory Remarks

The task of mathematics is to establish the truth or falsehood of (formalizable) statements using rigorous logic, and to provide methods for the solution of classes of (e.g. applied) problems, ideally including rigorous logical proofs verifying the validity of the methods (proofs that the method under consideration will, indeed, provide a correct solution).

The topic of this class is calculus, which is short for infinitesimal calculus, usually understood (as it is here) to mean differential and integral calculus of real and complex numbers (more generally, calculus may refer to any method or system of calculation guided by the symbolic manipulation of expressions, we will briefly touch on another example in Sec. 1.2 below). In that sense, calculus is the beginning part of the broader field of (mathematical) analysis, the section of mathematics concerned with the notion of a limit (for us, the most important examples will be limits of sequences (Def. 7.1 below) and limits of functions (Def. 8.17 below)).

Before we can properly define our first limit, however, it still needs some preparatory work. In modern mathematics, the objects under investigation are almost always so-called sets. So one aims at deriving (i.e. proving) true (and interesting and useful) statements about sets from other statements about sets known or assumed to be true. Such a derivation or proof means applying logical rules that guarantee the truth of the derived (i.e. proved) statement.

However, unfortunately, a proper definition of the notion of set is not easy, and neither is an appropriate treatment of logic and proof theory. Here, we will only be able to briefly touch on the bare necessities from logic and set theory needed to proceed to the core matter of this class. We begin with logic in Sec. 1.2, followed by set theory in Sec. 1.3, combining both in Sec. 1.4. The interested student can find an introductory presentation of axiomatic set theory in Appendix A and he/she should consider taking a separate class on set theory, logic, and proof theory at a later time.

1.2 Propositional Calculus

1.2.1 Statements

Mathematical logic is a large field in its own right and, as indicated above, a thorough introduction is beyond the scope of this class – the interested reader may refer to [EFT07], [Kun12], and references therein. Here, we will just introduce some basic concepts using common English (rather than formal symbolic languages – a concept touched on in Sec. A.2 of the Appendix and more thoroughly explained in books like [EFT07]).

As mentioned before, mathematics establishes the truth or falsehood of statements. By a statement or proposition we mean any sentence (any sequence of symbols) that can
reasonably be assigned a *truth value*, i.e. a value of either *true*, abbreviated T, or *false*, abbreviated F. The following example illustrates the difference between statements and sentences that are not statements:

**Example 1.1. (a)** Sentences that are statements:

- Every dog is an animal. (T)
- Every animal is a dog. (F)
- The number 4 is odd. (F)
- \(2 + 3 = 5\). (T)
- \(\sqrt{2} < 0\). (F)
- \(x + 1 > 0\) holds for each natural number \(x\). (T)

(b) Sentences that are *not* statements:

- Let’s study calculus!
- Who are you?
- \(3 \cdot 5 + 7\).
- \(x + 1 > 0\).
- All natural numbers are green.

The fourth sentence in Ex. 1.1(b) is not a statement, as it can not be said to be either true or false without any further knowledge on \(x\). The fifth sentence in Ex. 1.1(b) is not a statement as it lacks any meaning and can, hence, not be either true or false. It would become a statement if given a definition of what it means for a natural number to be green.

### 1.2.2 Logical Operators

The next step now is to *combine* statements into new statements using *logical operators*, where the truth value of the combined statements depends on the truth values of the original statements and on the type of logical operator facilitating the combination.

The simplest logical operator is *negation*, denoted \(\neg\). It is actually a so-called *unary* operator, i.e. it does not combine statements, but is merely applied to one statement. For example, if \(A\) stands for the statement “Every dog is an animal.”, then \(\neg A\) stands for the statement “Not every dog is an animal.”; and if \(B\) stands for the statement “The number 4 is odd.”, then \(\neg B\) stands for the statement “The number 4 is not odd.”, which can also be expressed as “The number 4 is even.”

To completely understand the action of a logical operator, one usually writes what is known as a *truth table*. For negation, the truth table is

\[
\begin{array}{c|c}
  A & \neg A \\
  \hline \\
  T & F \\
  F & T \\
\end{array}
\] (1.1)
that means if the input statement $A$ is true, then the output statement $\neg A$ is false; if the input statement $A$ is false, then the output statement $\neg A$ is true.

We now proceed to discuss binary logical operators, i.e. logical operators combining precisely two statements. The following four operators are essential for mathematical reasoning:

Conjunction: $A$ and $B$, usually denoted $A \land B$.

Disjunction: $A$ or $B$, usually denoted $A \lor B$.

Implication: $A$ implies $B$, usually denoted $A \Rightarrow B$.

Equivalence: $A$ is equivalent to $B$, usually denoted $A \Leftrightarrow B$.

Here is the corresponding truth table:

<table>
<thead>
<tr>
<th>$A$</th>
<th>$B$</th>
<th>$A \land B$</th>
<th>$A \lor B$</th>
<th>$A \Rightarrow B$</th>
<th>$A \Leftrightarrow B$</th>
</tr>
</thead>
<tbody>
<tr>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
</tr>
<tr>
<td>T</td>
<td>F</td>
<td>F</td>
<td>T</td>
<td>F</td>
<td>F</td>
</tr>
<tr>
<td>F</td>
<td>T</td>
<td>F</td>
<td>T</td>
<td>T</td>
<td>F</td>
</tr>
<tr>
<td>F</td>
<td>F</td>
<td>F</td>
<td>T</td>
<td>T</td>
<td>T</td>
</tr>
</tbody>
</table>

When first seen, some of the assignments of truth values in (1.2) might not be completely intuitive, due to the fact that logical operators are often used somewhat differently in common English. Let us consider each of the four logical operators of (1.2) in sequence:

For the use in subsequent examples, let $A_1, \ldots, A_6$ denote the six statements from Ex. 1.1(a).

Conjunction: Most likely the easiest of the four, basically identical to common language use: $A \land B$ is true if, and only if, both $A$ and $B$ are true. For example, using Ex. 1.1(a), $A_1 \land A_4$ is the statement “Every dog is an animal and $2 + 3 = 5.$”, which is true since both $A_1$ and $A_4$ are true. On the other hand, $A_1 \land A_3$ is the statement “Every dog is an animal and the number 4 is odd.”, which is false, since $A_3$ is false.

Disjunction: The disjunction $A \lor B$ is true if, and only if, at least one of the statements $A, B$ is true. Here one already has to be a bit careful – $A \lor B$ defines the inclusive or, whereas “or” in common English is often understood to mean the exclusive or (which is false if both input statements are true). For example, using Ex. 1.1(a), $A_1 \lor A_4$ is the statement “Every dog is an animal or $2 + 3 = 5.$”, which is true since both $A_1$ and $A_4$ are true. The statement $A_1 \lor A_3$, i.e. “Every dog is an animal or the number 4 is odd.” is also true, since $A_1$ is true. However, the statement $A_2 \lor A_5$, i.e. “Every animal is a dog or $\sqrt{2} < 0.$” is false, as both $A_2$ and $A_5$ are false.

As you will have noted in the above examples, logical operators can be applied to combine statements that have no obvious contents relation. While this might seem strange, introducing contents-related restrictions is unnecessary as well as undesirable, since it is often not clear which seemingly unrelated statements might suddenly appear in a common context in the future. The same occurs when considering implications and equivalences, where it might seem even more obscure at first.
Implication: Instead of \( A \) implies \( B \), one also says if \( A \) then \( B \), is a consequence of \( A \), is concluded or inferred from \( A \), is sufficient for \( B \), or is necessary for \( A \). The implication \( A \Rightarrow B \) is always true, except if \( A \) is true and \( B \) is false. At first glance, it might be surprising that \( A \Rightarrow B \) is defined to be true for \( A \) false and \( B \) true, however, there are many examples of incorrect statements implying correct statements. For instance, squaring the (false) equality of integers \(-1 = 1\), implies the (true) equality of integers \(1 = 1\). However, as with conjunction and disjunction, it is perfectly valid to combine statements without any obvious context relation: For example, using Ex. 1.1(a), the statement \( A_1 \Rightarrow A_6 \), i.e. “Every dog is an animal implies \( x + 1 > 0 \) holds for each natural number \( x \).” is true, since \( A_6 \) is true, whereas the statement \( A_4 \Rightarrow A_2 \), i.e. “\( 2 + 3 = 5 \) implies every animal is a dog.” is false, as \( A_4 \) is true and \( A_2 \) is false.

Of course, the implication \( A \Rightarrow B \) is not really useful in situations, where the truth values of both \( A \) and \( B \) are already known. Rather, in a typical application, one tries to establish the truth of \( A \) to prove the truth of \( B \) (a strategy that will fail if \( A \) happens to be false).

**Example 1.2.** Suppose we know Sasha to be a member of a group of students, taking a class in Analysis. Then the statement \( A \) “Sasha has taken a class in Analysis before.” implies the statement \( B \) “There is at least one student in the group, who has taken the class before”. A priori, we might not know if Sasha has taken the Analysis class before, but if we can establish that Sasha has, indeed, taken the class before, then we also know \( B \) to be true. If we find Sasha to be taking the class for the first time, then we do not know, whether \( B \) is true or false.

Equivalence: \( A \iff B \) means \( A \) is true if, and only if, \( B \) is true. Once again, using input statements from Ex. 1.1(a), we see that \( A_1 \iff A_4 \), i.e. “Every dog is an animal is equivalent to \( 2 + 3 = 5 \).”, is true as well as \( A_2 \iff A_3 \), i.e. “Every animal is a dog is equivalent to the number \( 4 \) is odd.”. On the other hand, \( A_4 \iff A_5 \), i.e. “\( 2 + 3 = 5 \) is equivalent to \( \sqrt{2} < 0 \), is false.

Analogous to the situation of implications, \( A \iff B \) is not really useful if the truth values of both \( A \) and \( B \) are known a priori, but can be a powerful tool to prove \( B \) to be true or false by establishing the truth value of \( A \). It is obviously more powerful than the implication as illustrated by the following example (compare with Ex. 1.2):

**Example 1.3.** Suppose we know Sasha has obtained the highest score among the students registered for the Analysis class. Then the statement \( A \) “Sasha has taken the Analysis class before.” is equivalent to the statement \( B \) “The student with the highest score has taken the class before.” As in Ex. 1.2, if we can establish Sasha to have taken the class before, then we also know \( B \) to be true. However, in contrast to Ex. 1.2, if we find Sasha to have taken the class for the first time, then we know \( B \) to be false.

**Remark 1.4.** In computer science, the truth value T is often coded as 1 and the truth value F is often coded as 0.
1.2.3 Rules

Note that the expressions in the first row of the truth table (1.2) (e.g. $A \land B$) are not statements in the sense of Sec. 1.2.1, as they contain the statement variables (also known as propositional variables) $A$ or $B$. However, the expressions become statements if all statement variables are substituted with actual statements. We will call expressions of this form propositional formulas. Moreover, if a truth value is assigned to each statement variable of a propositional formula, then this uniquely determines the truth value of the formula. In other words, the truth value of the propositional formula can be calculated from the respective truth values of its statement variables – a first justification for the name propositional calculus.

Example 1.5. (a) Consider the propositional formula $(A \land B) \lor (\neg B)$. Suppose $A$ is true and $B$ is false. The truth value of the formula is obtained according to the following truth table:

<p>| | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>$A$</td>
<td>$B$</td>
<td>$A \land B$</td>
<td>$\neg B$</td>
<td>$(A \land B) \lor (\neg B)$</td>
</tr>
<tr>
<td>T</td>
<td>F</td>
<td>F</td>
<td>T</td>
<td>T</td>
</tr>
</tbody>
</table>

(b) The propositional formula $A \lor (\neg A)$, also known as the law of the excluded middle, has the remarkable property that its truth value is T for every possible choice of truth values for $A$:

<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>$A$</td>
<td>$\neg A$</td>
<td>$A \lor (\neg A)$</td>
</tr>
<tr>
<td>T</td>
<td>F</td>
<td>T</td>
</tr>
<tr>
<td>F</td>
<td>T</td>
<td>T</td>
</tr>
</tbody>
</table>

Formulas with this property are of particular importance.

Definition 1.6. A propositional formula is called a tautology or universally true if, and only if, its truth value is T for all possible assignments of truth values to all the statement variables it contains.

Notation 1.7. We write $\phi(A_1, \ldots, A_n)$ if, and only if, the propositional formula $\phi$ contains precisely the $n$ statement variables $A_1, \ldots, A_n$.

Definition 1.8. The propositional formulas $\phi(A_1, \ldots, A_n)$ and $\psi(A_1, \ldots, A_n)$ are called equivalent if, and only if, $\phi(A_1, \ldots, A_n) \iff \psi(A_1, \ldots, A_n)$ is a tautology.

Lemma 1.9. The propositional formulas $\phi(A_1, \ldots, A_n)$ and $\psi(A_1, \ldots, A_n)$ are equivalent if, and only if, they have the same truth value for all possible assignments of truth values to $A_1, \ldots, A_n$.

Proof. If $\phi(A_1, \ldots, A_n)$ and $\psi(A_1, \ldots, A_n)$ are equivalent and $A_i$ is assigned the truth value $t_i$, $i = 1, \ldots, n$, then $\phi(A_1, \ldots, A_n) \iff \psi(A_1, \ldots, A_n)$ being a tautology implies it has truth value T. From (1.2) we see that either $\phi(A_1, \ldots, A_n)$ and $\psi(A_1, \ldots, A_n)$ both have truth value T or they both have truth value F.

If, on the other hand, we know $\phi(A_1, \ldots, A_n)$ and $\psi(A_1, \ldots, A_n)$ have the same truth value for all possible assignments of truth values to $A_1, \ldots, A_n$, then, given such an
assignment, either \( \phi(A_1, \ldots, A_n) \) and \( \psi(A_1, \ldots, A_n) \) both have truth value T or both have truth value F, i.e. \( \phi(A_1, \ldots, A_n) \iff \psi(A_1, \ldots, A_n) \) has truth value T in each case, showing it is a tautology.

For all logical purposes, two equivalent formulas are exactly the same – it does not matter if one uses one or the other. The following theorem provides some important equivalences of propositional formulas. As too many parentheses tend to make formulas less readable, we first introduce some precedence conventions for logical operators:

**Convention 1.10.** \( \neg \) takes precedence over \( \land, \lor \), which take precedence over \( \Rightarrow, \iff \). So, for example,

\[
(A \lor \neg B \Rightarrow \neg B \land \neg A) \iff \neg C \land (A \lor \neg D)
\]

is the same as

\[
\left((A \lor (\neg B)) \Rightarrow (\neg B) \land (\neg A)\right) \iff \left((\neg C) \land (A \lor (\neg D))\right).
\]

**Theorem 1.11.**

(a) \( (A \Rightarrow B) \iff \neg A \lor B \). This means one can actually define implication via negation and disjunction.

(b) \( (A \iff B) \iff ((A \Rightarrow B) \land (B \Rightarrow A)) \), i.e. \( A \) and \( B \) are equivalent if, and only if, \( A \) is both necessary and sufficient for \( B \). One also calls the implication \( B \Rightarrow A \) the converse of the implication \( A \Rightarrow B \). Thus, \( A \) and \( B \) are equivalent if, and only if, both \( A \Rightarrow B \) and its converse hold true.

(c) Commutativity of Conjunction: \( A \land B \iff B \land A \).

(d) Commutativity of Disjunction: \( A \lor B \iff B \lor A \).

(e) Associativity of Conjunction: \( (A \land B) \land C \iff A \land (B \land C) \).

(f) Associativity of Disjunction: \( (A \lor B) \lor C \iff A \lor (B \lor C) \).

(g) Distributivity I: \( A \land (B \lor C) \iff (A \land B) \lor (A \land C) \).

(h) Distributivity II: \( A \lor (B \land C) \iff (A \lor B) \land (A \lor C) \).

(i) De Morgan’s Law I: \( \neg (A \land B) \iff \neg A \lor \neg B \).

(j) De Morgan’s Law II: \( \neg (A \lor B) \iff \neg A \land \neg B \).

(k) Double Negative: \( \neg \neg A \iff A \).

(l) Contraposition: \( (A \Rightarrow B) \iff (\neg B \Rightarrow \neg A) \).

**Proof.** Each equivalence is proved by providing a truth table and using Lem. 1.9.
(a):

<table>
<thead>
<tr>
<th>$A$</th>
<th>$B$</th>
<th>$\neg A$</th>
<th>$A \Rightarrow B$</th>
<th>$\neg A \lor B$</th>
</tr>
</thead>
<tbody>
<tr>
<td>T</td>
<td>T</td>
<td>F</td>
<td>T</td>
<td>T</td>
</tr>
<tr>
<td>T</td>
<td>F</td>
<td>F</td>
<td>F</td>
<td>F</td>
</tr>
<tr>
<td>F</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
</tr>
<tr>
<td>F</td>
<td>F</td>
<td>T</td>
<td>T</td>
<td>T</td>
</tr>
</tbody>
</table>

(b) – (h): Exercise.

(i):

<table>
<thead>
<tr>
<th>$A$</th>
<th>$B$</th>
<th>$\neg A$</th>
<th>$A \land B$</th>
<th>$\neg (A \land B)$</th>
<th>$\neg A \lor \neg B$</th>
</tr>
</thead>
<tbody>
<tr>
<td>T</td>
<td>T</td>
<td>F</td>
<td>F</td>
<td>F</td>
<td>F</td>
</tr>
<tr>
<td>T</td>
<td>F</td>
<td>F</td>
<td>T</td>
<td>F</td>
<td>T</td>
</tr>
<tr>
<td>F</td>
<td>T</td>
<td>T</td>
<td>F</td>
<td>T</td>
<td>T</td>
</tr>
<tr>
<td>F</td>
<td>F</td>
<td>T</td>
<td>F</td>
<td>T</td>
<td>T</td>
</tr>
</tbody>
</table>

(j): Exercise.

(k):

<table>
<thead>
<tr>
<th>$A$</th>
<th>$\neg A$</th>
<th>$\neg \neg A$</th>
</tr>
</thead>
<tbody>
<tr>
<td>T</td>
<td>F</td>
<td>T</td>
</tr>
<tr>
<td>F</td>
<td>T</td>
<td>F</td>
</tr>
</tbody>
</table>

(l):

<table>
<thead>
<tr>
<th>$A$</th>
<th>$B$</th>
<th>$\neg A$</th>
<th>$\neg B$</th>
<th>$A \Rightarrow B$</th>
<th>$\neg B \Rightarrow \neg A$</th>
</tr>
</thead>
<tbody>
<tr>
<td>T</td>
<td>T</td>
<td>F</td>
<td>F</td>
<td>T</td>
<td>T</td>
</tr>
<tr>
<td>T</td>
<td>F</td>
<td>F</td>
<td>T</td>
<td>F</td>
<td>F</td>
</tr>
<tr>
<td>F</td>
<td>T</td>
<td>T</td>
<td>F</td>
<td>T</td>
<td>T</td>
</tr>
<tr>
<td>F</td>
<td>F</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
</tr>
</tbody>
</table>

Having checked all the rules completes the proof of the theorem.

The importance of the rules provided by Th. 1.11 lies in their providing proof techniques, i.e. methods for establishing the truth of statements from statements known or assumed to be true. The rules of Th. 1.11 will be used frequently in proofs throughout this class.

**Remark 1.12.** Another important proof technique is the so-called proof by contradiction, also called indirect proof. It is based on the observation, called the principle of contradiction, that $A \land \neg A$ is always false:

\[
\begin{array}{c|c|c}
A & \neg A & A \land \neg A \\
T & F & F \\
F & T & F \\
\end{array}
\]

Thus, one possibility of proving a statement $B$ to be true is to show $\neg B \Rightarrow A \land \neg A$ for some arbitrary statement $A$. Since the right-hand side of the implication is false, the left-hand side must also be false, proving $B$ is true.
Two more rules we will use regularly in subsequent proofs are the so-called transitivity of implication and the transitivity of equivalence (we will encounter equivalence again in the context of relations in Sec. 1.3 below). In preparation for the transitivity rules, we generalize implication to propositional formulas:

**Definition 1.13.** In generalization of the implication operator defined in (1.2), we say the propositional formula $\phi(A_1, \ldots, A_n)$ implies the propositional formula $\psi(A_1, \ldots, A_n)$ (denoted $\phi(A_1, \ldots, A_n) \Rightarrow \psi(A_1, \ldots, A_n)$) if, and only if, each assignment of truth values to the $A_1, \ldots, A_n$ that makes $\phi(A_1, \ldots, A_n)$ true, makes $\psi(A_1, \ldots, A_n)$ true as well.

**Theorem 1.14.** (a) Transitivity of Implication: $(A \Rightarrow B) \land (B \Rightarrow C) \Rightarrow (A \Rightarrow C)$.
(b) Transitivity of Equivalence: $(A \Leftrightarrow B) \land (B \Leftrightarrow C) \Rightarrow (A \Leftrightarrow C)$.

**Proof.** According to Def. 1.13, the rules can be verified by providing truth tables that show that, for all possible assignments of truth values to the propositional formulas on the left-hand side of the implications, either the left-hand side is false or both sides are true. (a):

<table>
<thead>
<tr>
<th>$A$</th>
<th>$B$</th>
<th>$C$</th>
<th>$A \Rightarrow B$</th>
<th>$B \Rightarrow C$</th>
<th>$(A \Rightarrow B) \land (B \Rightarrow C)$</th>
<th>$A \Rightarrow C$</th>
</tr>
</thead>
<tbody>
<tr>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
</tr>
<tr>
<td>T</td>
<td>F</td>
<td>T</td>
<td>T</td>
<td>F</td>
<td>F</td>
<td>T</td>
</tr>
<tr>
<td>F</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>F</td>
<td>T</td>
</tr>
<tr>
<td>F</td>
<td>F</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
</tr>
<tr>
<td>T</td>
<td>T</td>
<td>F</td>
<td>T</td>
<td>F</td>
<td>F</td>
<td>F</td>
</tr>
<tr>
<td>T</td>
<td>F</td>
<td>F</td>
<td>T</td>
<td>F</td>
<td>F</td>
<td>F</td>
</tr>
<tr>
<td>F</td>
<td>T</td>
<td>F</td>
<td>F</td>
<td>F</td>
<td>T</td>
<td>T</td>
</tr>
<tr>
<td>F</td>
<td>F</td>
<td>F</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
</tr>
</tbody>
</table>

(b):

<table>
<thead>
<tr>
<th>$A$</th>
<th>$B$</th>
<th>$C$</th>
<th>$A \Leftrightarrow B$</th>
<th>$B \Leftrightarrow C$</th>
<th>$(A \Leftrightarrow B) \land (B \Leftrightarrow C)$</th>
<th>$A \Leftrightarrow C$</th>
</tr>
</thead>
<tbody>
<tr>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
</tr>
<tr>
<td>T</td>
<td>F</td>
<td>T</td>
<td>F</td>
<td>F</td>
<td>F</td>
<td>T</td>
</tr>
<tr>
<td>F</td>
<td>T</td>
<td>T</td>
<td>F</td>
<td>T</td>
<td>F</td>
<td>F</td>
</tr>
<tr>
<td>F</td>
<td>F</td>
<td>T</td>
<td>F</td>
<td>F</td>
<td>F</td>
<td>F</td>
</tr>
<tr>
<td>T</td>
<td>T</td>
<td>F</td>
<td>T</td>
<td>F</td>
<td>F</td>
<td>F</td>
</tr>
<tr>
<td>T</td>
<td>F</td>
<td>F</td>
<td>T</td>
<td>F</td>
<td>F</td>
<td>F</td>
</tr>
<tr>
<td>F</td>
<td>T</td>
<td>F</td>
<td>F</td>
<td>F</td>
<td>T</td>
<td>T</td>
</tr>
<tr>
<td>F</td>
<td>F</td>
<td>F</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
</tr>
</tbody>
</table>

Having checked both rules, the proof is complete.

**Definition and Remark 1.15.** A proof of the statement $B$ is a finite sequence of statements $A_1, A_2, \ldots, A_n$ such that $A_1$ is true; for $1 \leq i < n$, $A_i$ implies $A_{i+1}$, and $A_n$ implies $B$. If there exists a proof for $B$, then Th. 1.14(a) guarantees that $B$ is true.

**Remark 1.16.** Principle of Duality: In Th. 1.11, there are several pairs of rules that have an analogous form: (c) and (d), (e) and (f), (g) and (h), (i) and (j). These
analogies are due to the general law called the principle of duality: If \( \phi(A_1, \ldots, A_n) \Rightarrow \psi(A_1, \ldots, A_n) \) and only the operators \( \land, \lor, \neg \) occur in \( \phi \) and \( \psi \), then the reverse implication \( \Phi(A_1, \ldots, A_n) \Leftarrow \Psi(A_1, \ldots, A_n) \) holds, where one obtains \( \Phi \) from \( \phi \) and \( \Psi \) from \( \psi \) by replacing each \( \land \) with \( \lor \) and each \( \lor \) with \( \land \). In particular, if, instead of an implication, we start with an equivalence (as in the examples from Th. 1.11), then we obtain another equivalence.

### 1.3 Set Theory

In the previous section, we have had a first glance at statements and corresponding truth values. In the present section, we will move our focus to the objects such statements are about. Reviewing Example 1.1(a), and recalling that this is a mathematics class rather than one in zoology, the first two statements of Example 1.1(a) are less relevant for us than statements 3–6. As in these examples, we will nearly always be interested in statements involving numbers or collections of numbers or collections of such collections etc.

In modern mathematics, the term one usually uses instead of “collection” is “set”. In 1895, Georg Cantor defined a set as “any collection into a whole \( M \) of definite and separate objects \( m \) of our intuition or our thought”. The objects \( m \) are called the elements of the set \( M \). As explained in Appendix A, without restrictions and refinements, Cantor’s set theory is not free of contradictions and, thus, not viable to be used in the foundation of mathematics. Axiomatic set theory provides these necessary restrictions and refinements and an introductory treatment can also be found in Appendix A. However, it is possible to follow and understand the rest of this class, without having studied Appendix A.

**Notation 1.17.** We write \( m \in M \) for the statement “\( m \) is an element of the set \( M \)”.

**Definition 1.18.** The sets \( M \) and \( N \) are equal, denoted \( M = N \), if, and only if, \( M \) and \( N \) have precisely the same elements.

Definition 1.18 means we know everything about a set \( M \) if, and only if, we know all its elements.

**Definition 1.19.** The set with no elements is called the empty set; it is denoted by the symbol \( \emptyset \).

**Example 1.20.** For finite sets, we can simply write down all its elements, for example, \( A := \{0\} \), \( B := \{0, 17.5\} \), \( C := \{5, 1, 5, 3\} \), \( D := \{3, 5, 1\} \), \( E := \{2, \sqrt{2}, -2\} \), where the symbolism “:=” is to be read as “is defined to be equal to”.

Note \( C = D \), since both sets contain precisely the same elements. In particular, the order in which the elements are written down plays no role and a set does not change if an element is written down more than once.
If a set has many elements, instead of writing down all its elements, one might use abbreviations such as \( F := \{-4, -2, \ldots, 20, 22, 24\} \), where one has to make sure the meaning of the dots is clear from the context.

**Definition 1.21.** The set \( A \) is called a *subset* of the set \( B \) (denoted \( A \subseteq B \) and also referred to as the *inclusion* of \( A \) in \( B \)) if, and only if, every element of \( A \) is also an element of \( B \) (one sometimes also calls \( B \) a *superset* of \( A \) and writes \( B \supseteq A \)). Please note that \( A = B \) is allowed in the above definition of a subset. If \( A \subseteq B \) and \( A \neq B \), then \( A \) is called a *strict* subset of \( B \), denoted \( A \subset B \).

If \( B \) is a set and \( P(x) \) is a statement about an element \( x \) of \( B \) (i.e., for each \( x \in B \), \( P(x) \) is either true or false), then we can define a subset \( A \) of \( B \) by writing

\[
A := \{ x \in B : P(x) \}.
\]

This notation is supposed to mean that the set \( A \) consists precisely of those elements of \( B \) such that \( P(x) \) is true (has the truth value \( T \) in the language of Sec. 1.2).

**Example 1.22.** (a) For each set \( A \), one has \( A \subseteq A \) and \( \emptyset \subseteq A \).

(b) If \( A \subseteq B \), then \( A = \{ x \in B : x \in A \} \).

(c) We have \( \{3\} \subseteq \{6, 7, 3, 0\} \). Letting \( A := \{-10,-8,\ldots,8,10\} \), we have \( \{-2,0,2\} = \{ x \in A : x^3 \in A \} \), \( \emptyset = \{ x \in A : x+21 \in A \} \).

**Remark 1.23.** As a consequence of Def. 1.18, the sets \( A \) and \( B \) are equal if, and only if, one has both inclusions, namely \( A \subseteq B \) and \( B \subseteq A \). Thus, when proving the equality of sets, one often divides the proof into two parts, first proving one inclusion, then the other.

**Definition 1.24.** (a) The *intersection* of the sets \( A \) and \( B \), denoted \( A \cap B \), consists of all elements that are in \( A \) and in \( B \). The sets \( A, B \) are said to be *disjoint* if, and only if, \( A \cap B = \emptyset \).

(b) The *union* of the sets \( A \) and \( B \), denoted \( A \cup B \), consists of all elements that are in \( A \) or in \( B \) (as in the logical disjunction in (1.2), the or is meant nonexclusively). If \( A \) and \( B \) are disjoint, one sometimes writes \( A \cup B \) and speaks of the *disjoint union* of \( A \) and \( B \).

(c) The *difference* of the sets \( A \) and \( B \), denoted \( A \setminus B \) (read “\( A \) minus \( B \)” or “\( A \) without \( B \)”), consists of all elements of \( A \) that are not elements of \( B \), i.e. \( A \setminus B := \{ x \in A : x \notin B \} \). If \( B \) is a subset of a given set \( A \) (sometimes called the *universe* in this context), then \( A \setminus B \) is also called the *complement* of \( B \) with respect to \( A \). In that case, one also writes \( B^c := A \setminus B \) (note that this notation suppresses the dependence on \( A \)).

**Example 1.25.** (a) Examples of Intersections:

\[
\{1, 2, 3\} \cap \{3, 4, 5\} = \{3\},
\]

\[
\{\sqrt{2}\} \cap \{1, 2, \ldots, 10\} = \emptyset,
\]

\[
\{-1, 2, -3, 4, 5\} \cap \{-10, -9, \ldots, -1\} \cap \{-1, 7, -3\} = \{-1, -3\}.
\]
(b) Examples of Unions:

\[ \{1, 2, 3\} \cup \{3, 4, 5\} = \{1, 2, 3, 4, 5\}, \]  
\[ \{1, 2, 3\} \cup \{4, 5\} = \{1, 2, 3, 4, 5\}, \]  
\[ \{-1, 2, -3, 4, 5\} \cup \{-99, -98, \ldots, -1\} \cup \{-1, 7, -3\} = \{-99, -98, \ldots, -2, -1, 2, 4, 5, 7\}. \]

(1.8a)  

(1.8b)  

(1.8c)

(c) Examples of Differences:

\[ \{1, 2, 3\} \setminus \{3, 4, 5\} = \{1, 2\}, \]  
\[ \{1, 2, 3\} \setminus \{3, 2, 1, \sqrt{5}\} = \emptyset, \]  
\[ \{-10, -9, \ldots, 9, 10\} \setminus \{0\} = \{-10, -9, \ldots, -1\} \cup \{1, 2, \ldots, 9, 10\}. \]

With respect to the universe \(\{1, 2, 3, 4, 5\}\), it is

\[ \{1, 2, 3\}^c = \{4, 5\}; \]

(1.9d)  

with respect to the universe \(\{0, 1, \ldots, 20\}\), it is

\[ \{1, 2, 3\}^c = \{0\} \cup \{4, 5, \ldots, 20\}. \]

(1.9e)

As mentioned earlier, it will often be unavoidable to consider sets of sets. Here are first examples: \(\emptyset, \{0\}, \{0, 1\}, \{\{0\}, \{1, 2\}\}\).

**Definition 1.26.** Given a set \(A\), the set of all subsets of \(A\) is called the power set of \(A\), denoted \(\mathcal{P}(A)\) (for reasons explained later (cf. Prop. 2.18), the power set is sometimes also denoted as \(2^A\)).

**Example 1.27.** Examples of Power Sets:

\[ \mathcal{P}(\emptyset) = \{\emptyset\}, \]  
\[ \mathcal{P}\{\emptyset\} = \{\emptyset, \{\emptyset\}\}, \]  
\[ \mathcal{P}\{\mathcal{P}\{\emptyset\}\} = \mathcal{P}\{\emptyset, \{\emptyset\}\} = \{\emptyset, \{\emptyset\}, \{\{\emptyset\}\}, \mathcal{P}\{\emptyset\}\}. \]

(1.10a)  

(1.10b)  

(1.10c)

So far, we have restricted our set-theoretic examples to finite sets. However, not surprisingly, many sets of interest to us will be infinite (we will have to postpone a mathematically precise definition of finite and infinite to Sec. 3.2). We will now introduce the most simple infinite set.

**Definition 1.28.** The set \(\mathbb{N} := \{1, 2, 3, \ldots\}\) is called the set of natural numbers (for a more rigorous construction of \(\mathbb{N}\), based on the axioms of axiomatic set theory, see Sec. A.3.4 of the Appendix, where Th. A.46 shows \(\mathbb{N}\) to be, indeed, infinite). Moreover, we define \(\mathbb{N}_0 := \{0\} \cup \mathbb{N}\).
The following theorem compiles important set-theoretic rules:

**Theorem 1.29.** Let $A, B, C, U$ be sets.

(a) **Commutativity of Intersections:** $A \cap B = B \cap A$.

(b) **Commutativity of Unions:** $A \cup B = B \cup A$.

(c) **Associativity of Intersections:** $(A \cap B) \cap C = A \cap (B \cap C)$.

(d) **Associativity of Unions:** $(A \cup B) \cup C = A \cup (B \cup C)$.

(e) **Distributivity I:** $A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$.

(f) **Distributivity II:** $A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$.

(g) **De Morgan’s Law I:** $U \setminus (A \cap B) = (U \setminus A) \cup (U \setminus B)$.

(h) **De Morgan’s Law II:** $U \setminus (A \cup B) = (U \setminus A) \cap (U \setminus B)$.

(i) **Double Complement:** If $A \subseteq U$, then $U \setminus (U \setminus A) = A$.

**Proof.** In each case, the proof results from the corresponding rule of Th. 1.11:

(a): $x \in A \cap B \iff x \in A \land x \in B \quad \text{Th. 1.11(c)} \quad x \in B \land x \in A \iff x \in B \cap A$.

(g): Under the general assumption of $x \in U$, we have the following equivalences:

$x \in U \setminus (A \cap B) \iff \neg(x \in A \cap B) \iff \neg(x \in A \land x \in B) \quad \text{Th. 1.11(i)} \quad \neg(x \in A) \lor \neg(x \in B)$

$
\iff x \in U \setminus A \lor x \in U \setminus B \iff x \in (U \setminus A) \cup (U \setminus B).
$

The proofs of the remaining rules are left as an exercise.

**Remark 1.30.** The correspondence between Th. 1.11 and Th. 1.29 is no coincidence. One can actually prove that, starting with an equivalence of propositional formulas $\phi(A_1, \ldots, A_n) \Leftrightarrow \psi(A_1, \ldots, A_n)$, where both formulas contain only the operators $\land$, $\lor$, $\neg$, one obtains a set-theoretic rule (stating an equality of sets) by reinterpreting all statement variables $A_1, \ldots, A_n$ as variables for sets, all subsets of a universe $U$, and replacing $\land$ by $\cap$, $\lor$ by $\cup$, and $\neg$ by $U \setminus$ (if there are no multiple negations, then we do not need the hypothesis that $A_1, \ldots, A_n$ are subsets of $U$). The procedure also works in the opposite direction – one can start with a set-theoretic formula for an equality of sets and translate it into two equivalent propositional formulas.
1.4 Predicate Calculus

Now that we have introduced sets in the previous section, we have to return to the subject of mathematical logic once more. As it turns out, propositional calculus, which we discussed in Sec. 1.2, does not quite suffice to develop the theory of calculus (nor most other mathematical theories). The reason is that we need to consider statements such as

\[ x + 1 > 0 \text{ holds for each natural number } x. \quad (T) \quad (1.11a) \]

All real numbers are positive. \( (F) \) \hspace{1cm} (1.11b)

There exists a natural number bigger than 10. \( (T) \) \hspace{1cm} (1.11c)

There exists a real number \( x \) such that \( x^2 = -1 \). \( (F) \) \hspace{1cm} (1.11d)

For all natural numbers \( n \), there exists a natural number bigger than \( n \). \( (T) \) \hspace{1cm} (1.11e)

That means we are interested in statements involving \textit{universal quantification} via the quantifier “for all” (one also often uses “for each” or “for every” instead), \textit{existential quantification} via the quantifier “there exists”, or both. The quantifier of universal quantification is denoted by \( \forall \) and the quantifier of existential quantification is denoted by \( \exists \). Using these symbols as well as \( \mathbb{N} \) and \( \mathbb{R} \) to denote the sets of natural and real numbers, respectively, we can restate (1.11) as

\[ \forall_{x \in \mathbb{N}} x + 1 > 0. \quad (T) \quad (1.12a) \]

\[ \forall_{x \in \mathbb{R}} x > 0. \quad (F) \quad (1.12b) \]

\[ \exists_{n \in \mathbb{N}} n > 10. \quad (T) \quad (1.12c) \]

\[ \exists_{x \in \mathbb{R}} x^2 = -1. \quad (F) \quad (1.12d) \]

\[ \forall_{n \in \mathbb{N}} \exists_{m \in \mathbb{N}} m > n. \quad (T) \quad (1.12e) \]

**Definition 1.31.** A \textit{universal statement} has the form

\[ \forall_{x \in A} P(x), \quad (1.13a) \]

whereas an \textit{existential statement} has the form

\[ \exists_{x \in A} P(x). \quad (1.13b) \]

In (1.13), \( A \) denotes a set and \( P(x) \) is a sentence involving the variable \( x \), a so-called \textit{predicate} of \( x \), that becomes a statement (i.e. becomes either true or false) if \( x \) is substituted with any concrete element of the set \( A \) (in particular, \( P(x) \) is allowed to contain further quantifiers, but it must not contain any other quantifier involving \( x \) – one says \( x \) must be a \textit{free} variable in \( P(x) \), not bound by any quantifier in \( P(x) \)).

The universal statement (1.13a) has the truth value \( T \) if, and only if, \( P(x) \) has the truth value \( T \) for \textit{all} elements \( x \in A \); the existential statement (1.13b) has the truth value \( T \) if, and only if, \( P(x) \) has the truth value \( T \) for \textit{at least one} element \( x \in A \).
Remark 1.32. Some people prefer to write $\bigwedge_{x \in A}$ instead of $\forall x \in A$ and $\bigvee_{x \in A}$ instead of $\exists x \in A$. Even though this notation has the advantage of emphasizing that the universal statement can be interpreted as a big logical conjunction and the existential statement can be interpreted as a big logical disjunction, it is significantly less common. So we will stick to $\forall$ and $\exists$ in this class.

Remark 1.33. According to Def. 1.31, the existential statement (1.13b) is true if, and only if, $P(x)$ is true for at least one $x \in A$. So if there is precisely one such $x$, then (1.13b) is true; and if there are several different $x \in A$ such that $P(x)$ is true, then (1.13b) is still true. Uniqueness statements are often of particular importance, and one sometimes writes

$$\exists!_{x \in A} P(x)$$

for the statement “there exists a unique $x \in A$ such that $P(x)$ is true”. This notation can be defined as an abbreviation for

$$\exists_{x \in A} \left( P(x) \land \forall_{y \in A} (P(y) \Rightarrow x = y) \right).$$

Example 1.34. Here are some examples of uniqueness statements:

$$\exists!_{n \in \mathbb{N}} n > 10. \quad (F) \quad (1.16a)$$
$$\exists!_{n \in \mathbb{N}} 12 > n > 10. \quad (T) \quad (1.16b)$$
$$\exists!_{n \in \mathbb{N}} 11 > n > 10. \quad (F) \quad (1.16c)$$
$$\exists!_{x \in \mathbb{R}} x^2 = -1. \quad (F) \quad (1.16d)$$
$$\exists!_{x \in \mathbb{R}} x^2 = 1. \quad (F) \quad (1.16e)$$
$$\exists!_{x \in \mathbb{R}} x^2 = 0. \quad (T) \quad (1.16f)$$

Remark 1.35. As for propositional calculus, we also have some important rules for predicate calculus:

(a) Consider the negation of a universal statement, $\neg \forall_{x \in A} P(x)$, which is true if, and only if, $P(x)$ does not hold for each $x \in A$, i.e. if, and only if, there exists at least one $x \in A$ such that $P(x)$ is false (such that $\neg P(x)$ is true). We have just proved the rule

$$\neg \forall_{x \in A} P(x) \iff \exists_{x \in A} \neg P(x). \quad (1.17a)$$

Similarly, consider the negation of an existential statement. We claim the corresponding rule is

$$\neg \exists_{x \in A} P(x) \iff \forall_{x \in A} \neg P(x). \quad (1.17b)$$

Indeed, we can prove (1.17b) from (1.17a):

$$\neg \exists_{x \in A} P(x) \quad \text{Th. 1.11(k)} \quad \neg \exists_{x \in A} \neg P(x) \quad (1.17a) \quad \neg \forall_{x \in A} \neg P(x) \quad \text{Th. 1.11(k)} \quad \forall_{x \in A} \neg P(x). \quad (1.18)$$
One can interpret (1.17) as a generalization of the De Morgan’s laws Th. 1.11(i),(j).

One can actually generalize (1.17) even a bit more: If a statement starts with several quantifiers, then one negates the statement by replacing each $\forall$ with $\exists$ and vice versa plus negating the predicate after the quantifiers (see the example in (1.21e) below).

(b) If $A, B$ are sets and $P(x, y)$ denotes a predicate of both $x$ and $y$, then $\forall x \in A \quad \forall y \in B \quad P(x, y)$
and $\forall y \in B \quad \forall x \in A \quad P(x, y)$ both hold true if, and only if, $P(x, y)$ holds true for each $x \in A$ and each $y \in B$, i.e. the order of two consecutive universal quantifiers does not matter:

$$\forall x \in A \quad \forall y \in B \quad P(x, y) \iff \forall y \in B \quad \forall x \in A \quad P(x, y) \quad (1.19a)$$

In the same way, we obtain the following rule:

$$\exists x \in A \quad \exists y \in B \quad P(x, y) \iff \exists y \in B \quad \exists x \in A \quad P(x, y). \quad (1.19b)$$

If $A = B$, one also uses abbreviations of the form

$$\forall x, y \in A \quad P(x, y) \quad \text{for} \quad \forall x \in A \quad \forall y \in A \quad P(x, y), \quad (1.20a)$$

$$\exists x, y \in A \quad P(x, y) \quad \text{for} \quad \exists x \in A \quad \exists y \in A \quad P(x, y). \quad (1.20b)$$

Generalizing rules (1.19), we can always commute identical quantifiers. Caveat: Quantifiers that are not identical must not be commuted (see Ex. 1.36(d) below).

**Example 1.36. (a)** Negation of universal and existential statements:

Negation of (1.12a) : $\exists x \in \mathbb{N} \quad x + 1 \leq 0. \quad (F) \quad (1.21a)$

Negation of (1.12b) : $\exists x \in \mathbb{R} \quad x \leq 0. \quad (T) \quad (1.21b)$

Negation of (1.12c) : $\forall n \in \mathbb{N} \quad n \leq 10. \quad (F) \quad (1.21c)$

Negation of (1.12d) : $\forall x \in \mathbb{R} \quad x^2 \neq -1. \quad (T) \quad (1.21d)$

Negation of (1.12e) : $\exists n \in \mathbb{N} \quad \forall m \in \mathbb{N} \quad m \leq n. \quad (F) \quad (1.21e)$

(b) As a more complicated example, consider the negation of the uniqueness statement
(1.14), i.e. of (1.15):

\[
\neg \exists!_{x \in A} P(x) \iff \neg \exists_{x \in A} \left( P(x) \land \forall_{y \in A} (P(y) \Rightarrow x = y) \right)
\]

(1.17b), Th. 1.11(a) \iff

\[
\forall_{x \in A} \neg \left( P(x) \land \forall_{y \in A} (\neg P(y) \lor x = y) \right)
\]

Th. 1.11(j) \iff

\[
\forall_{x \in A} \left( \neg P(x) \lor \exists_{y \in A} (\neg P(y) \lor x = y) \right)
\]

(1.17a) \iff

\[
\forall_{x \in A} \left( \neg P(x) \lor \exists_{y \in A} (P(y) \land x \neq y) \right)
\]

Th. 1.11(j), (k) \iff

\[
\forall_{x \in A} \left( P(x) \Rightarrow \exists_{y \in A} (P(y) \land x \neq y) \right).
\]  (1.22)

So how to decode the expression, we have obtained at the end? It states that if

\( P(x) \) holds for some \( x \in A \), then there must be at least a second, different, element

\( y \in A \) such that \( P(y) \) is true. This is, indeed, precisely the negation of \( \exists!_{x \in A} P(x) \).

(c) Identical quantifiers commute:

\[
\forall_{x \in \mathbb{R}} \forall_{n \in \mathbb{N}} x^{2n} \geq 0 \iff \forall_{n \in \mathbb{N}} \forall_{x \in \mathbb{R}} x^{2n} \geq 0,
\]

(1.23a)

\[
\forall_{x \in \mathbb{R}} \exists_{y \in \mathbb{R}} \exists_{n \in \mathbb{N}} ny > x^2 \iff \forall_{x \in \mathbb{R}} \exists_{y \in \mathbb{R}} \exists_{n \in \mathbb{N}} ny > x^2.
\]  (1.23b)

(d) The following example shows that different quantifiers do, in general, not commute (i.e. do not yield equivalent statements when commuted): While the statement

\[
\forall_{x \in \mathbb{R}} \exists_{y \in \mathbb{R}} y > x
\]

(1.24a)

is true (for each real number \( x \), there is a bigger real number \( y \), e.g. \( y := x + 1 \) will do the job), the statement

\[
\exists_{y \in \mathbb{R}} \forall_{x \in \mathbb{R}} y > x
\]

(1.24b)

is false (for example, since \( y > y \) is false). In particular, (1.24a) and (1.24b) are not equivalent.

(e) Even though (1.14) provides useful notation, it is better not to think of \( \exists! \) as a quantifier. It is really just an abbreviation for (1.15), and it behaves very differently from \( \exists \) and \( \forall \): The following examples show that, in general, \( \exists! \) commutes neither with \( \exists \), nor with itself:

\[
\exists_{n \in \mathbb{N}} \exists!_{m \in \mathbb{N}} m < n \not\iff \exists!_{m \in \mathbb{N}} \exists_{n \in \mathbb{N}} m < n
\]

(the statement on the left is true, as one can choose \( n = 2 \), but the statement on the right is false, as \( \exists_{n \in \mathbb{N}} m < n \) holds for every \( m \in \mathbb{N} \)). Similarly,

\[
\exists!_{m \in \mathbb{N}} \exists!_{n \in \mathbb{N}} m < n \not\iff \exists!_{m \in \mathbb{N}} \exists!_{n \in \mathbb{N}} m < n
\]
(the statement on the left is still true and the statement on the right is still false (there is no \( m \in \mathbb{N} \) such that \( \exists! n \in \mathbb{N} \) \( m < n \)).)

**Remark 1.37.** One can make the following observations regarding the strategy for proving universal and existential statements:

(a) To prove that \( \forall x \in A P(x) \) is true, one must check the truth of \( P(x) \) for every element \( x \in A \) – examples are not enough!

(b) To prove that \( \forall x \in A P(x) \) is false, it suffices to find one \( x \in A \) such that \( P(x) \) is false – such an \( x \) is then called a *counterexample* and one counterexample is always enough to prove \( \forall x \in A P(x) \) is false!

(c) To prove that \( \exists x \in A P(x) \) is true, it suffices to find one \( x \in A \) such that \( P(x) \) is true – such an \( x \) is then called an *example* and one example is always enough to prove \( \exists x \in A P(x) \) is true!

The subfield of mathematical logic dealing with quantified statements is called *predicate calculus*. In general, one does not restrict the quantified variables to range only over elements of sets (as we have done above). Again, we refer to [EFT07] for a deeper treatment of the subject.

As an application of quantified statements, let us generalize the notion of union and intersection:

**Definition 1.38.** Let \( I \neq \emptyset \) be a nonempty set, usually called an *index set* in the present context. For each \( i \in I \), let \( A_i \) denote a set (some or all of the \( A_i \) can be identical).

(a) The *intersection*

\[
\bigcap_{i \in I} A_i := \left\{ x : \forall_{i \in I} x \in A_i \right\} \quad (1.25a)
\]

consists of all elements \( x \) that belong to every \( A_i \).

(b) The *union*

\[
\bigcup_{i \in I} A_i := \left\{ x : \exists_{i \in I} x \in A_i \right\} \quad (1.25b)
\]

consists of all elements \( x \) that belong to at least one \( A_i \). The union is called *disjoint* if, and only if, for each \( i, j \in I \), \( i \neq j \) implies \( A_i \cap A_j = \emptyset \).

**Proposition 1.39.** Let \( I \neq \emptyset \) be an index set, let \( M \) denote a set, and, for each \( i \in I \), let \( A_i \) denote a set. The following set-theoretic rules hold:

(a) \( \bigcap_{i \in I} A_i \cap M = \bigcap_{i \in I} (A_i \cap M) \).
Proof. We prove (c) and (e) and leave the remaining proofs as an exercise.

(c): 
\[ x \in \left( \bigcap_{i \in I} A_i \right) \cup M \iff x \in M \lor \forall \ x \in A_i \ \overset{(*)}{\iff} \ \forall \ i \in I \ (x \in A_i \lor x \in M) \iff x \in \bigcap_{i \in I} (A_i \cup M). \]

To justify the equivalence at (*), we make use of Th. 1.11(b) and verify ⇒ and ⇐. For ⇒ note that the truth of \( x \in M \) implies \( x \in A_i \lor x \in M \) is true for each \( i \in I \). If \( x \in A_i \) is true for each \( i \in I \), then \( x \in A_i \lor x \in M \) is still true for each \( i \in I \). To verify ⇐, note that the existence of \( i \in I \) such that \( x \in M \) implies the truth of \( x \in M \lor \forall \ x \in A_i \).

If \( x \in M \) is false for each \( i \in I \), then \( x \in A_i \) must be true for each \( i \in I \), showing \( x \in M \lor \forall \ x \in A_i \) is true also in this case.

(e): 
\[ x \in M \setminus \bigcap_{i \in I} A_i \iff x \in M \land \neg \forall \ x \in A_i \iff x \in M \land \exists \ x \notin A_i \]
\[ \iff \exists \ x \in M \setminus A_i \iff x \in \bigcup_{i \in I} (M \setminus A_i), \]
completing the proof. □
Example 1.40. We have the following identities of sets:

\[
\bigcap_{x \in \mathbb{R}} N = N, \quad (1.26a)
\]
\[
\bigcap_{n \in \mathbb{N}} \{1, 2, \ldots, n\} = \{1\}, \quad (1.26b)
\]
\[
\bigcup_{x \in \mathbb{R}} N = N, \quad (1.26c)
\]
\[
\bigcup_{n \in \mathbb{N}} \{1, 2, \ldots, n\} = N, \quad (1.26d)
\]
\[
N \setminus \bigcup_{n \in \mathbb{N}} \{2n\} = \{1, 3, 5, \ldots\} = \bigcap_{n \in \mathbb{N}} (N \setminus \{2n\}). \quad (1.26e)
\]

2 Functions and Relations

2.1 Functions

Definition 2.1. Let \( A, B \) be sets. Given \( x \in A \), \( y \in B \), the set

\[
(x, y) := \left\{ \{x\}, \{x, y\} \right\}
\]

is called the ordered pair (often shortened to just pair) consisting of \( x \) and \( y \). The set of all such pairs is called the Cartesian product \( A \times B \), i.e.

\[
A \times B := \{ (x, y) : x \in A \land y \in B \}. \quad (2.2)
\]

Example 2.2. Let \( A \) be a set.

\[
A \times \emptyset = \emptyset \times A = \emptyset, \quad (2.3a)
\]
\[
\{1, 2\} \times \{1, 2, 3\} = \{(1, 1), (1, 2), (1, 3), (2, 1), (2, 2), (2, 3)\}, \quad (2.3b)
\]
\[
\neq \{1, 2, 3\} \times \{1, 2\} = \{(1, 1), (1, 2), (2, 1), (2, 2), (3, 1), (3, 2)\}. \quad (2.3c)
\]

Also note that, for \( x \neq y \),

\[
(x, y) = \left\{ \{x\}, \{x, y\} \right\} \neq \left\{ \{y\}, \{x, y\} \right\} = (y, x). \quad (2.4)
\]

Definition 2.3. Given sets \( A, B \), a function or map \( f \) is an assignment rule that assigns to each \( x \in A \) a unique \( y \in B \). One then also writes \( f(x) \) for the element \( y \). The set \( A \) is called the domain of \( f \), denoted \( \text{dom}(f) \), and \( B \) is called the codomain of \( f \), denoted \( \text{codom}(f) \). The information about a map \( f \) can be concisely summarized by the notation

\[
f : A \longrightarrow B, \quad x \mapsto f(x), \quad (2.5)
\]

where \( x \mapsto f(x) \) is called the assignment rule for \( f \), \( f(x) \) is called the image of \( x \), and \( x \) is called a preimage of \( f(x) \) (the image must be unique, but there might be several preimages). The set

\[
\text{graph}(f) := \{ (x, y) \in A \times B : y = f(x) \}. \quad (2.6)
\]
is called the graph of $f$ (not to be confused with pictures visualizing the function $f$, which are also called graph of $f$). If one wants to be completely precise, then one identifies the function $f$ with the ordered triple $(A,B,\text{graph}(f))$.

The set of all functions with domain $A$ and codomain $B$ is denoted by $\mathcal{F}(A,B)$ or $B^A$, i.e.

$$\mathcal{F}(A,B) := B^A := \{(f : A \to B) : A = \text{dom}(f) \land B = \text{codom}(f)\}. \quad (2.7)$$

Caveat: Some authors reserve the word map for continuous functions, but we use function and map synonymously.

**Definition 2.4.** Let $A,B$ be sets and $f : A \to B$ a function.

(a) If $T$ is a subset of $A$, then

$$f(T) := \{f(x) \in B : x \in T\} \quad (2.8)$$

is called the image of $T$ under $f$.

(b) If $U$ is a subset of $B$, then

$$f^{-1}(U) := \{x \in A : f(x) \in U\} \quad (2.9)$$

is called the preimage or inverse image of $U$ under $f$.

(c) $f$ is called injective or one-to-one if, and only if, every $y \in B$ has at most one preimage, i.e. if, and only if, the preimage of $\{y\}$ has at most one element:

$$f \text{ injective} \iff \forall_{y \in B} \left( f^{-1}\{y\} = \emptyset \lor \exists_{x \in A} f(x) = y \right)$$

$$\iff \forall_{x_1,x_2 \in A} (x_1 \neq x_2 \Rightarrow f(x_1) \neq f(x_2)). \quad (2.10)$$

(d) $f$ is called surjective or onto if, and only if, every element of the codomain of $f$ has a preimage:

$$f \text{ surjective} \iff \forall_{y \in B} \exists_{x \in A} f(x) \iff \forall_{y \in B} f^{-1}\{y\} \neq \emptyset. \quad (2.11)$$

(e) $f$ is called bijective if, and only if, $f$ is injective and surjective.

**Example 2.5.** Examples of Functions:

- $f : \{1,2,3,4,5\} \to \{1,2,3,4,5\}$, $f(x) := -x + 6$, \quad (2.12a)
- $g : \mathbb{N} \to \mathbb{N}$, $g(n) := 2n$, \quad (2.12b)
- $h : \mathbb{N} \to \{2,4,6,\ldots\}$, $h(n) := 2n$, \quad (2.12c)
- $\tilde{h} : \mathbb{N} \to \{2,4,6,\ldots\}$, $\tilde{h}(n) := \begin{cases} n & \text{for } n \text{ even}, \\ n+1 & \text{for } n \text{ odd}, \end{cases}$ \quad (2.12d)
- $G : \mathbb{N} \to \mathbb{R}$, $G(n) := n/(n+1)$, \quad (2.12e)
- $F : \mathcal{P}(\mathbb{N}) \to \mathcal{P}(\mathcal{P}(\mathbb{N}))$, $F(A) := \mathcal{P}(A)$. \quad (2.12f)
Instead of \( f(x) := -x + 6 \) in (2.12a), one can also write \( x \mapsto -x + 6 \) and analogously in the other cases. Also note that, in the strict sense, functions \( g \) and \( h \) are different, since their codomains are different (however, using the following Def. 2.4(a), they have the same image in the sense that \( g(\mathbb{N}) = h(\mathbb{N}) \)). Furthermore,

\[
f(\{1, 2\}) = \{5, 4\} = f^{-1}(\{1, 2\}), \quad \tilde{h}^{-1}(\{2, 4, 6\}) = \{1, 2, 3, 4, 5, 6\}, \tag{2.13}
\]

\( f \) is bijective; \( g \) is injective, but not surjective; \( h \) is bijective; \( \tilde{h} \) is surjective, but not injective. Can you figure out if \( G \) and \( F \) are injective and/or surjective?

**Example 2.6. (a)** For each nonempty set \( A \), the map \( \text{Id} : A \rightarrow A \), \( \text{Id}(x) := x \), is called the **identity** on \( A \). If one needs to emphasize that \( \text{Id} \) operates on \( A \), then one also writes \( \text{Id}_A \) instead of \( \text{Id} \). The identity is clearly bijective.

**(b)** Let \( A, B \) be nonempty sets. A map \( f : A \rightarrow B \) is called **constant** if, and only if, there exists \( c \in B \) such that \( f(x) = c \) for each \( x \in A \). In that case, one also writes \( f \equiv c \), which can be read as “\( f \) is identically equal to \( c \)”.

If \( f \equiv c \), \( \emptyset \neq T \subseteq A \), and \( U \subseteq B \), then

\[
f(T) = \{c\}, \quad f^{-1}(U) = \begin{cases} A & \text{for } c \in U, \\ \emptyset & \text{for } c \notin U. \end{cases} \tag{2.14}
\]

\( f \) is injective if, and only if, \( A = \{x\} \); \( f \) is surjective if, and only if, \( B = \{c\} \).

**(c)** Given \( A \subseteq X \), the map

\[
\iota : A \rightarrow X, \quad \iota(x) := x,
\]

is called **inclusion** (also *embedding* or *imbedding*). An inclusion is always injective; it is surjective if, and only if, \( A = X \), i.e. if, and only if, it is the identity on \( A \).

**(d)** Given \( A \subseteq X \) and a map \( f : X \rightarrow B \), the map \( g : A \rightarrow B, g(x) = f(x) \), is called the **restriction** of \( f \) to \( A \); \( f \) is called the **extension** of \( g \) to \( X \). In this situation, one also uses the notation \( f|_A \) for \( g \) (some authors prefer the notation \( f|_A \) or \( f|A \)).

**Theorem 2.7.** Let \( f : A \rightarrow B \) be a map, let \( \emptyset \neq I \) be an index set, and assume \( S, T, S_i, i \in I \), are subsets of \( A \), whereas \( U, V, U_i, i \in I \), are subsets of \( B \). Then we have the
following rules concerning functions and set-theoretic operations:

\[
\begin{align*}
  f(S \cap T) &\subseteq f(S) \cap f(T), \quad \text{(2.16a)} \\
  f\left(\bigcap_{i \in I} S_i \right) &\subseteq \bigcap_{i \in I} f(S_i), \quad \text{(2.16b)} \\
  f(S \cup T) &= f(S) \cup f(T), \quad \text{(2.16c)} \\
  f\left(\bigcup_{i \in I} S_i \right) &= \bigcup_{i \in I} f(S_i), \quad \text{(2.16d)} \\
  f^{-1}(U \cap V) &= f^{-1}(U) \cap f^{-1}(V), \quad \text{(2.16e)} \\
  f^{-1}\left(\bigcap_{i \in I} U_i \right) &= \bigcap_{i \in I} f^{-1}(U_i), \quad \text{(2.16f)} \\
  f^{-1}(U \cup V) &= f^{-1}(U) \cup f^{-1}(V), \quad \text{(2.16g)} \\
  f^{-1}\left(\bigcup_{i \in I} U_i \right) &= \bigcup_{i \in I} f^{-1}(U_i), \quad \text{(2.16h)} \\
  f(f^{-1}(U)) &\subseteq U, \quad f^{-1}(f(S)) \supseteq S, \quad \text{(2.16i)} \\
  f^{-1}(U \setminus V) &= f^{-1}(U) \setminus f^{-1}(V). \quad \text{(2.16j)}
\end{align*}
\]

**Proof.** We prove (2.16b) (which includes (2.16a) as a special case) and the second part of (2.16i), and leave the remaining cases as exercises.

For (2.16b), one argues

\[
y \in f\left(\bigcap_{i \in I} S_i \right) \iff \exists x \in A \forall i \in I \left(x \in S_i \land y = f(x) \Rightarrow \forall i \in I \ y \in f(S_i) \iff y \in \bigcap_{i \in I} f(S_i)\right).
\]

The observation

\[
x \in S \Rightarrow f(x) \in f(S) \iff x \in f^{-1}(f(S)).
\]

establishes the second part of (2.16i). \[\square\]

It is an exercise to find counterexamples that show one can not, in general, replace the four subset symbols in (2.16) by equalities (it is possible to find examples with sets that have at most 2 elements).

**Definition 2.8.** The composition of maps \(f\) and \(g\) with \(f : A \to B\), \(g : C \to D\), and \(f(A) \subseteq C\) is defined to be the map

\[
g \circ f : A \to D, \quad (g \circ f)(x) := g(f(x)).
\]

The expression \(g \circ f\) is read as “\(g\) after \(f\)” or “\(g\) composed with \(f\)”.
Example 2.9. Consider the maps

\[ f : \mathbb{N} \to \mathbb{R}, \quad n \mapsto n^2, \quad (2.18a) \]
\[ g : \mathbb{N} \to \mathbb{R}, \quad n \mapsto 2n. \quad (2.18b) \]

We obtain \( f(\mathbb{N}) = \{1, 4, 9, \ldots\} \subseteq \text{dom}(g) \), \( g(\mathbb{N}) = \{2, 4, 6, \ldots\} \subseteq \text{dom}(f) \), and the compositions

\[ (g \circ f) : \mathbb{N} \to \mathbb{R}, \quad (g \circ f)(n) = g(n^2) = 2n^2, \quad (2.19a) \]
\[ (f \circ g) : \mathbb{N} \to \mathbb{R}, \quad (f \circ g)(n) = f(2n) = 4n^2, \quad (2.19b) \]

showing that composing functions is, in general, not commutative, even if the involved functions have the same domain and the same codomain.

Proposition 2.10. Consider maps \( f : A \to B \), \( g : C \to D \), \( h : E \to F \), satisfying \( f(A) \subseteq C \) and \( g(C) \subseteq E \).

(a) Associativity of Compositions:

\[ h \circ (g \circ f) = (h \circ g) \circ f. \quad (2.20) \]

(b) One has the following law for forming preimages:

\[ \forall W \in \mathcal{P}(D) \quad (g \circ f)^{-1}(W) = f^{-1}(g^{-1}(W)). \quad (2.21) \]

Proof. (a): Both \( h \circ (g \circ f) \) and \( (h \circ g) \circ f \) map \( A \) into \( F \). So it just remains to prove \( (h \circ (g \circ f))(x) = ((h \circ g) \circ f)(x) \) for each \( x \in A \). One computes, for each \( x \in A \),

\[ (h \circ (g \circ f))(x) = h((g \circ f)(x)) = h(g(f(x))) = (h \circ g)(f(x)) = ((h \circ g) \circ f)(x), \quad (2.22) \]

establishing the case.

(b): Exercise.

Definition 2.11. A function \( g : B \to A \) is called a right inverse (resp. left inverse) of a function \( f : A \to B \) if, and only if, \( f \circ g = \text{Id}_B \) (resp. \( g \circ f = \text{Id}_A \)). Moreover, \( g \) is called an inverse of \( f \) if, and only if, it is both a right and a left inverse. If \( g \) is an inverse of \( f \), then one also writes \( f^{-1} \) instead of \( g \). The map \( f \) is called (right, left) invertible if, and only if, there exists a (right, left) inverse for \( f \).

Example 2.12. (a) Consider the map

\[ f : \mathbb{N} \to \mathbb{N}, \quad f(n) := 2n. \quad (2.23a) \]

The maps

\[ g_1 : \mathbb{N} \to \mathbb{N}, \quad g_1(n) := \begin{cases} n/2 & \text{if } n \text{ even}, \\ 1 & \text{if } n \text{ odd}, \end{cases} \quad (2.23b) \]
\[ g_2 : \mathbb{N} \to \mathbb{N}, \quad g_2(n) := \begin{cases} n/2 & \text{if } n \text{ even}, \\ 2 & \text{if } n \text{ odd}, \end{cases} \quad (2.23c) \]
both constitute left inverses of \( f \). It follows from Th. 2.13(c) below that \( f \) does not have a right inverse.

(b) Consider the map

\[
\begin{align*}
f : \mathbb{N} &\to \mathbb{N}, \quad f(n) := \begin{cases} n/2 & \text{for } n \text{ even}, \\ (n+1)/2 & \text{for } n \text{ odd}. \end{cases} 
\end{align*}
\]

(2.24a)

The maps

\[
\begin{align*}
g_1 : \mathbb{N} &\to \mathbb{N}, \quad g_1(n) := 2n, \\
g_2 : \mathbb{N} &\to \mathbb{N}, \quad g_2(n) := 2n - 1,
\end{align*}
\]

(2.24b) (2.24c)

both constitute right inverses of \( f \). It follows from Th. 2.13(c) below that \( f \) does not have a left inverse.

(c) The map

\[
\begin{align*}
f : \mathbb{N} &\to \mathbb{N}, \quad f(n) := \begin{cases} n - 1 & \text{for } n \text{ even}, \\ n + 1 & \text{for } n \text{ odd}, \end{cases} 
\end{align*}
\]

(2.25a)

is its own inverse, i.e. \( f^{-1} = f \). For the map

\[
\begin{align*}
g : \mathbb{N} &\to \mathbb{N}, \quad g(n) := \begin{cases} 2 & \text{for } n = 1, \\ 3 & \text{for } n = 2, \\ 1 & \text{for } n = 3, \\ n & \text{for } n \notin \{1, 2, 3\}, \end{cases} 
\end{align*}
\]

(2.25b)

the inverse is

\[
\begin{align*}
g^{-1} : \mathbb{N} &\to \mathbb{N}, \quad g^{-1}(n) := \begin{cases} 3 & \text{for } n = 1, \\ 1 & \text{for } n = 2, \\ 2 & \text{for } n = 3, \\ n & \text{for } n \notin \{1, 2, 3\}. \end{cases} 
\end{align*}
\]

(2.25c)

While Examples 2.12(a),(b) show that left and right inverses are usually not unique, they are unique provided \( f \) is bijective (see Th. 2.13(c)).

**Theorem 2.13.** Let \( A, B \) be nonempty sets.

(a) \( f : A \to B \) is right invertible if, and only if, \( f \) is surjective (where the implication “\( \Leftarrow \)” makes use of the axiom of choice (AC), see Appendix A.4).

(b) \( f : A \to B \) is left invertible if, and only if, \( f \) is injective.

(c) \( f : A \to B \) is invertible if, and only if, \( f \) is bijective. In this case, the right inverse and the left inverse are unique and both identical to the inverse.
Proof. (a): If \( f \) is surjective, then, for each \( y \in B \), there exists \( x_y \in f^{-1}\{y\} \) such that \( f(x_y) = y \). By AC, we can define the choice function
\[
g : B \rightarrow A, \quad g(y) := x_y.
\] (2.26)
Then, for each \( y \in B \), \( f(g(y)) = y \), showing \( g \) is a right inverse of \( f \). Conversely, if \( g : B \rightarrow A \) is a right inverse of \( f \), then, for each \( y \in B \), it is \( y = f(g(y)) \), showing that \( g(y) \in A \) is a preimage of \( y \), i.e. \( f \) is surjective.

(b): Fix \( a \in A \). If \( f \) is injective, then, for each \( y \in B \) with \( f^{-1}\{y\} \neq \emptyset \), let \( x_y \) denote the unique element in \( A \) satisfying \( f(x_y) = y \). Define
\[
g : B \rightarrow A, \quad g(y) := \begin{cases} x_y & \text{for } f^{-1}\{y\} \neq \emptyset, \\ a & \text{otherwise}. \end{cases}
\] (2.27)
Then, for each \( x \in A \), \( g(f(x)) = x \), showing \( g \) is a left inverse of \( f \). Conversely, if \( g : B \rightarrow A \) is a left inverse of \( f \) and \( x_1, x_2 \in A \) with \( f(x_1) = f(x_2) = y \), then \( x_1 = (g \circ f)(x_1) = g(f(x_1)) = g(f(x_2)) = (g \circ f)(x_2) = x_2 \), showing \( y \) has precisely one preimage and \( f \) is injective.

(c): Assume \( g \) to be a left inverse of \( f \) and \( h \) to be a right inverse of \( f \). Then, for each \( y \in B \),
\[
g(y) = (g \circ (f \circ h))(y) = ((g \circ f) \circ h)(y) = h(y),
\] (2.28)
showing \( g = h \). In particular, if \( f \) has an inverse \( f^{-1} \), then \( g = h = f^{-1} \). If \( f \) is invertible, then \( f \) is bijective by (a) and (b). If \( f \) is bijective, then, by (a) and (b), \( f \) has a left inverse \( g \) and a right inverse \( h \) (here, this follows without using AC, since, if \( f \) is both injective and surjective, then, for each \( y \in B \), the element \( x_y \in f^{-1}\{y\} \) is unique, and (2.26) can be defined without AC). By (2.28), \( g = h \), i.e. \( f \) is invertible.

**Theorem 2.14.** Consider maps \( f : A \rightarrow B, g : B \rightarrow C \). If \( f \) and \( g \) are both injective (resp. both surjective, both bijective), then so is \( g \circ f \). Moreover, in the bijective case, one has
\[
(g \circ f)^{-1} = f^{-1} \circ g^{-1}.
\] (2.29)

**Proof.** Exercise.

**Definition 2.15.** (a) Given an index set \( I \) and a set \( A \), a map \( f : I \rightarrow A \) is sometimes called a family (of elements in \( A \)), and is denoted in the form \( f = \langle a_i \rangle_{i \in I} \) with \( a_i := f(i) \). When using this representation, one often does not even specify \( f \) and \( A \), especially if the \( a_i \) are themselves sets.

(b) A sequence in a set \( A \) is a family of elements in \( A \), where the index set is the set of natural numbers \( \mathbb{N} \). In this case, one writes \( (a_n)_{n \in \mathbb{N}} \) or \( (a_1, a_2, \ldots) \). More generally, a family is called a sequence, given a bijective map between the index set \( I \) and a subset of \( \mathbb{N} \).
(c) Given a family of sets \((A_i)_{i \in I}\), we define the \textit{Cartesian product} of the \(A_i\) to be the set of functions

\[
\prod_{i \in I} A_i := \left\{ \left( f : I \rightarrow \bigcup_{j \in I} A_j \right) : \forall \, i \in I \, f(i) \in A_i \right\}.
\]  

(2.30)

If \(I\) has precisely \(n\) elements with \(n \in \mathbb{N}\), then the elements of the Cartesian product \(\prod_{i \in I} A_i\) are called (ordered) \(n\)-\textit{tuples}, (ordered) \textit{triples} for \(n = 3\).

**Example 2.16.** (a) Using the notion of family, we can now say that the intersection \(\bigcap_{i \in I} A_i\) and union \(\bigcup_{i \in I} A_i\) as defined in Def. 1.38 are the intersection and union of the family of sets \((A_i)_{i \in I}\), respectively. As a concrete example, let us revisit (1.26b), where we have

\[
(A_n)_{n \in \mathbb{N}}, \quad A_n := \{1, 2, \ldots, n\}, \quad \bigcap_{n \in \mathbb{N}} A_n = \{1\}. 
\]  

(2.31)

(b) Examples of Sequences:

- Sequence in \(\{0, 1\}\):

\[
(1, 0, 1, 0, 1, 0, \ldots),
\]  

(2.32a)

- Sequence in \(\mathbb{N}\):

\[
(n^2)_{n \in \mathbb{N}} = (1, 4, 9, 16, 25, \ldots),
\]  

(2.32b)

- Sequence in \(\mathbb{R}\):

\[
((-1)^n \sqrt{n})_{n \in \mathbb{N}} = (-1, \sqrt{2}, -\sqrt{3}, \ldots),
\]  

(2.32c)

- Sequence in \(\mathbb{R}\):

\[
(1/n)_{n \in \mathbb{N}} = \left(1, \frac{1}{2}, \frac{1}{3}, \frac{1}{4}, \ldots\right),
\]  

(2.32d)

- Finite Sequence in \(\mathcal{P}(\mathbb{N})\):

\[
(\{3, 2, 1\}, \{2, 1\}, \{1\}, \emptyset).
\]  

(2.32e)

(c) The Cartesian product \(\prod_{i \in I} A_i\), where all sets \(A_i = A\), is the same as \(A^I\), the set of all functions from \(I\) into \(A\). So, for example, \(\prod_{n \in \mathbb{N}} \mathbb{R} = \mathbb{R}^\mathbb{N}\) is the set of all sequences in \(\mathbb{R}\). If \(I = \{1, 2, \ldots, n\}\) with \(n \in \mathbb{N}\), then

\[
\prod_{i \in I} A = A^{\{1, 2, \ldots, n\}} =: \prod_{i=1}^n A =: A^n
\]  

(2.33)

is the set of all \(n\)-\textit{tuples} with entries from \(A\).

In the following, we explain the common notation \(2^A\) for the power set \(\mathcal{P}(A)\) of a set \(A\). It is related to a natural identification between subsets and their corresponding characteristic function.

**Definition 2.17.** Let \(A\) be a set and let \(B \subseteq A\) be a subset of \(A\). Then

\[
\chi_B : A \rightarrow \{0, 1\}, \quad \chi_B(x) := \begin{cases} 
1 & \text{if } x \in B, \\
0 & \text{if } x \notin B,
\end{cases}
\]  

(2.34)

is called the \textit{characteristic function} of the set \(B\) (with respect to the universe \(A\)). One also finds the notations \(1_B\) and \(1_B\) instead of \(\chi_B\) (note that all the notations suppress the dependence of the characteristic function on the universe \(A\)).
Proposition 2.18. Let $A$ be a set. Then the map
\[ \chi : \mathcal{P}(A) \longrightarrow \{0, 1\}^A, \quad \chi(B) := \chi_B, \] (2.35)
is bijective (recall that $\mathcal{P}(A)$ denotes the power set of $A$ and $\{0, 1\}^A$ denotes the set of all functions from $A$ into $\{0, 1\}$).

Proof. $\chi$ is injective: Let $B, C \in \mathcal{P}(A)$ with $B \neq C$. By possibly switching the names of $B$ and $C$, we may assume there exists $x \in B$ such that $x \notin C$. Then $\chi_B(x) = 1$, whereas $\chi_C(x) = 0$, showing $\chi(B) \neq \chi(C)$, proving $\chi$ is injective.

$\chi$ is surjective: Let $f : A \longrightarrow \{0, 1\}$ be an arbitrary function and define $B := \{x \in A : f(x) = 1\}$. Then $\chi(B) = \chi_B = f$, proving $\chi$ is surjective. $\blacksquare$

Proposition 2.18 allows one to identify the sets $\mathcal{P}(A)$ and $\{0, 1\}^A$ via the bijective map $\chi$. This fact together with the common practise of set theory to identify the number 2 with the set $\{0, 1\}$ (cf. the first paragraph of Sec. D.1 in the Appendix) explains the notation $2^A$ for $\mathcal{P}(A)$.

2.2 Relations

Definition 2.19. Given sets $A$ and $B$, a relation is a subset $R$ of $A \times B$ (if one wants to be completely precise, a relation is an ordered triple $(A, B, R)$, where $R \subseteq A \times B$). If $A = B$, then we call $R$ a relation on $A$. One says that $a \in A$ and $b \in B$ are related according to the relation $R$ if, and only if, $(a, b) \in R$. In this context, one usually writes $a R b$ instead of $(a, b) \in R$.

Example 2.20. (a) The relations we are probably most familiar with are = and $\leq$.

The relation $R$ of equality, usually denoted $=$, makes sense on every nonempty set $A$:

\[ R := \Delta(A) := \{(x, x) \in A \times A : x \in A\}. \] (2.36)

The set $\Delta(A)$ is called the diagonal of the Cartesian product, i.e., as a subset of $A \times A$, the relation of equality is identical to the diagonal:

\[ x = y \iff x R y \iff (x, y) \in R = \Delta(A). \] (2.37)

Similarly, the relation $\leq$ on $\mathbb{R}$ is identical to the set
\[ R_{\leq} := \{(x, y) \in \mathbb{R}^2 : x \leq y\}. \] (2.38)

(b) Every function $f : A \longrightarrow B$ is a relation, namely the relation
\[ R_f = \{(x, y) \in A \times B : y = f(x)\} = \text{graph}(f). \] (2.39)

Conversely, if $B \neq \emptyset$, then every relation $R \subseteq A \times B$ uniquely corresponds to the function
\[ f_R : A \longrightarrow \mathcal{P}(B), \quad f_R(x) = \{y \in B : x R y\}. \] (2.40)
**Definition 2.21.** Let $R$ be a relation on the set $A$.

(a) $R$ is called **reflexive** if, and only if,

$$\forall x \in A \ x \ R \ x,$$

i.e. if, and only if, every element is related to itself.

(b) $R$ is called **symmetric** if, and only if,

$$\forall x,y \in A \ (x \ R \ y \ \Rightarrow \ y \ R \ x),$$

i.e. if, and only if, each $x$ is related to $y$ if, and only if, $y$ is related to $x$.

(c) $R$ is called **antisymmetric** if, and only if,

$$\forall x,y \in A \ ((x \ R \ y \ \land \ y \ R \ x) \ \Rightarrow \ x = y),$$

i.e. if, and only if, the only possibility for $x$ to be related to $y$ at the same time that $y$ is related to $x$ is in the case $x = y$.

(d) $R$ is called **transitive** if, and only if,

$$\forall x,y,z \in A \ ((x \ R \ y \ \land \ y \ R \ z) \ \Rightarrow \ x \ R \ z),$$

i.e. if, and only if, the relatedness of $x$ and $y$ together with the relatedness of $y$ and $z$ implies the relatedness of $x$ and $z$.

**Example 2.22.** The relations $=$ and $\leq$ on $\mathbb{R}$ (or $\mathbb{N}$) are reflexive, antisymmetric, and transitive; $=$ is also symmetric, whereas $\leq$ is not; $<$ is antisymmetric (since $x < y \land y < x$ is always false) and transitive, but neither reflexive nor symmetric. The relation

$$R := \{(x,y) \in \mathbb{N}^2 : (x,y \text{ are both even}) \lor (x,y \text{ are both odd})\}$$

(2.45)

on $\mathbb{N}$ is not antisymmetric, but reflexive, symmetric, and transitive. The relation

$$S := \{(x,y) \in \mathbb{N}^2 : y = x^2\}$$

(2.46)

is not transitive (for example, $2 \mathcal{S} 4$ and $4 \mathcal{S} 16$, but not $2 \mathcal{S} 16$), not reflexive, not symmetric; it is only antisymmetric.

**Definition 2.23.** A relation $R$ on a set $A$ is called an **equivalence relation** if, and only if, $R$ is reflexive, symmetric, and transitive. If $R$ is an equivalence relations, then one often writes $x \sim y$ instead of $x \ R \ y$.

**Example 2.24.** (a) The equality relation $=$ is an equivalence relation on each $A \neq \emptyset$.

(b) The relation $R$ defined in (2.45) is an equivalence relation on $\mathbb{N}$. 
(c) Given a disjoint union \( A = \bigcup_{i \in I} A_i \) with every \( A_i \neq \emptyset \) (which is sometimes called a decomposition of \( A \)), an equivalence relation on \( A \) is defined by

\[
x \sim y \iff \exists_{i \in I} (x \in A_i \land y \in A_i).
\]  

(2.47)

Conversely, given an equivalence relation \( \sim \) on a nonempty set \( A \), we can construct a decomposition \( A = \bigcup_{i \in I} A_i \) such that (2.47) holds: For each \( x \in A \), define

\[
[x] := \{y \in A : x \sim y\},
\]

(2.48)
called the equivalence class of \( x \); each \( y \in [x] \) is called a representative of \( [x] \). One verifies that the properties of \( \sim \) guarantee

\[
([x] = [y] \iff x \sim y) \land ([x] \cap [y] = \emptyset \iff (x \sim y)).
\]

(2.49)
The set of all equivalence classes \( I := A/\sim := \{[x] : x \in A \} \) is called the quotient set of \( A \) by \( \sim \), and \( A = \bigcup_{i \in I} A_i \) with \( A_i := i \) for each \( i \in I \) is the desired decomposition of \( A \).

**Definition 2.25.** A relation \( R \) on a set \( A \) is called a partial order if, and only if, \( R \) is reflexive, antisymmetric, and transitive. If \( R \) is a partial order, then one usually writes \( x \leq y \) instead of \( x R y \). A partial order \( \leq \) is called a total or linear order if, and only if, for each \( x, y \in A \), one has \( x \leq y \) or \( y \leq x \).

**Notation 2.26.** Given a (partial or total) order \( \leq \) on \( A \neq \emptyset \), we write \( x < y \) if, and only if, \( x \leq y \) and \( x \neq y \), calling \( < \) the strict order corresponding to \( \leq \) (note that the strict order is never a partial order).

**Definition 2.27.** Let \( \leq \) be a partial order on \( A \neq \emptyset \), \( \emptyset \neq B \subseteq A \).

(a) \( x \in A \) is called lower (resp. upper) bound for \( B \) if, and only if, \( x \leq b \) (resp. \( b \leq x \)) for each \( b \in B \). Moreover, \( B \) is called bounded from below (resp. from above) if, and only if, there exists a lower (resp. upper) bound for \( B \); \( B \) is called bounded if, and only if, it is bounded from above and from below.

(b) \( x \in B \) is called minimum or just min (resp. maximum or max) of \( B \) if, and only if, \( x \) is a lower (resp. upper) bound for \( B \). One writes \( x = \min B \) if \( x \) is minimum and \( x = \max B \) if \( x \) is maximum.

(c) A maximum of the set of lower bounds of \( B \) (i.e. a largest lower bound) is called \( \infimum \) of \( B \), denoted \( \inf B \); a minimum of the set of upper bounds of \( B \) (i.e. a smallest upper bound) is called \( \supremum \) of \( B \), denoted \( \sup B \).

**Example 2.28.** (a) For each \( A \subseteq \mathbb{R} \), the usual relation \( \leq \) defines a total order on \( A \). For \( A = \mathbb{R} \), we see that \( \mathbb{N} \) has 0 and 1 as lower bound with \( 1 = \min \mathbb{N} = \inf \mathbb{N} \). On the other hand, \( \mathbb{N} \) is unbounded from above. The set \( M := \{1,2,3\} \) is bounded with \( \min M = 1 \), \( \max M = 3 \). The positive real numbers \( \mathbb{R}^+ := \{x \in \mathbb{R} : x > 0\} \) have \( \inf \mathbb{R}^+ = 0 \), but they do not have a minimum (if \( x > 0 \), then \( 0 < x/2 < x \)).
(b) Consider \( A := \mathbb{N} \times \mathbb{N} \). Then
\[
(m_1, m_2) \leq (n_1, n_2) \iff m_1 \leq n_1 \land m_2 \leq n_2,
\]
(2.50)
defines a partial order on \( A \) that is not a total order (for example, neither \((1, 2) \leq (2, 1)\) nor \((2, 1) \leq (1, 2)\)). For the set
\[
B := \{(1, 1), (2, 1), (1, 2)\},
\]
(2.51)
we have \( \inf B = \min B = (1, 1) \), \( B \) does not have a max, but \( \sup B = (2, 2) \) (if \((m, n) \in A \) is an upper bound for \( B \), then \((2, 1) \leq (m, n) \) implies \( 2 \leq m \) and \((1, 2) \leq (m, n) \) implies \( 2 \leq n \), i.e. \((2, 2) \leq (m, n) \); since \((2, 2) \) is clearly an upper bound for \( B \), we have proved \( \sup B = (2, 2) \)).

A different order on \( A \) is the so-called lexicographic order defined by
\[
(m_1, m_2) \leq (n_1, n_2) \iff m_1 < n_1 \lor (m_1 = n_1 \land m_2 \leq n_2).
\]
(2.52)
In contrast to the order from (2.50), the lexicographic order does define a total order on \( A \).

**Lemma 2.29.** Let \( \leq \) be a partial order on \( A \neq \emptyset, \emptyset \neq B \subseteq A \). Then the relation \( \geq \), defined by
\[
x \geq y \iff y \leq x,
\]
(2.53)
is also a partial order on \( A \). Moreover, using obvious notation, we have, for each \( x \in A \),
\[
x \leq \text{-lower bound for } B \iff x \geq \text{-upper bound for } B,
\]
(2.54a)
\[
x \leq \text{-upper bound for } B \iff x \geq \text{-lower bound for } B,
\]
(2.54b)
\[
x = \min_\leq B \iff x = \max_\geq B,
\]
(2.54c)
\[
x = \max_\leq B \iff x = \min_\geq B,
\]
(2.54d)
\[
x = \inf_\leq B \iff x = \sup_\geq B,
\]
(2.54e)
\[
x = \sup_\leq B \iff x = \inf_\geq B.
\]
(2.54f)

**Proof.** Reflexivity, antisymmetry, and transitivity of \( \leq \) clearly imply the same properties for \( \geq \), respectively. Moreover
\[
x \leq \text{-lower bound for } B \iff \forall_{b \in B} x \leq b \iff \forall_{b \in B} b \geq x \iff x \geq \text{-upper bound for } B,
\]
proving (2.54a). Analogously, we obtain (2.54b). Next, (2.54c) and (2.54d) are implied by (2.54a) and (2.54b), respectively. Finally, (2.54e) is proved by
\[
x = \inf_\leq B \iff x = \max_\leq \{y \in A : y \leq \text{-lower bound for } B\}
\]
\[
\iff x = \min_\geq \{y \in A : y \geq \text{-upper bound for } B\} \iff x = \sup_\geq B,
\]
and (2.54f) follows analogously. \( \blacksquare \)
Proposition 2.30. Let \( \leq \) be a partial order on \( A \neq \emptyset, \emptyset \neq B \subseteq A \). The elements \( \max B, \min B, \sup B, \inf B \) are all unique, provided they exist.

Proof. Exercise. ■

Definition 2.31. Let \( A, B \) be nonempty sets with partial orders, both denoted by \( \leq \) (even though they might be different). A function \( f : A \to B \), is called (strictly) isotone, order-preserving, or increasing if, and only if,

\[
\forall x, y \in A \ (x < y \Rightarrow f(x) \leq f(y) \ (\text{resp. } f(x) < f(y))); \quad (2.55a)
\]

\( f \) is called (strictly) antitone, order-reversing, or decreasing if, and only if,

\[
\forall x, y \in A \ (x < y \Rightarrow f(x) \geq f(y) \ (\text{resp. } f(x) > f(y))); \quad (2.55b)
\]

Functions that are (strictly) isotone or antitone are called (strictly) monotone.

Proposition 2.32. Let \( A, B \) be nonempty sets with partial orders, both denoted by \( \leq \).

(a) A (strictly) isotone function \( f : A \to B \) becomes a (strictly) antitone function and vice versa if precisely one of the relations \( \leq \) is replaced by \( \geq \).

(b) If the order \( \leq \) on \( A \) is total and \( f : A \to B \) is strictly isotone or strictly antitone, then \( f \) is one-to-one.

(c) If the order \( \leq \) on \( A \) is total and \( f : A \to B \) is invertible and strictly isotone (resp. antitone), then \( f^{-1} \) is also strictly isotone (resp. antitone).

Proof. (a) is immediate from (2.55).

(b): Due to (a), it suffices to consider the case that \( f \) is strictly isotone. If \( f \) is strictly isotone and \( x \neq y \), then \( x < y \) or \( y < x \) since the order on \( A \) is total. Thus, \( f(x) < f(y) \) or \( f(y) < f(x) \), i.e. \( f(x) \neq f(y) \) in every case, showing \( f \) is one-to-one.

(c): Again, due to (a), it suffices to consider the isotone case. If \( u, v \in B \) such that \( u < v \), then \( u = f(f^{-1}(u)), v = f(f^{-1}(v)) \), and the isotonicity of \( f \) imply \( f^{-1}(u) < f^{-1}(v) \) (we are using that the order on \( A \) is total – otherwise, \( f^{-1}(u) \) and \( f^{-1}(v) \) need not be comparable). ■

Example 2.33. (a) \( f : \mathbb{N} \to \mathbb{N}, f(n) := 2n \), is strictly increasing, every constant map on \( \mathbb{N} \) is both increasing and decreasing, but not strictly increasing or decreasing. All maps occurring in (2.25) are neither increasing nor decreasing.

(b) The map \( f : \mathbb{R} \to \mathbb{R}, f(x) := -2x \), is invertible and strictly decreasing, and so is \( f^{-1} : \mathbb{R} \to \mathbb{R}, f^{-1}(x) := -x/2 \).
(c) The following counterexamples show that the assertions of Prop. 2.32(b),(c) are no longer correct if one does not assume the order on $A$ is total. Let $A$ be the set from (2.51) (where it had been called $B$) with the (nontotal) order from (2.50). The map

$$f : A \rightarrow \mathbb{N}, \begin{cases} f(1,1) := 1, \\ f(1,2) := 2, \\ f(2,1) := 2, \\ \end{cases}$$

(2.56)

is strictly isotone, but not one-to-one. The map

$$f : A \rightarrow \{1, 2, 3\}, \begin{cases} f(1,1) := 1, \\ f(1,2) := 2, \\ f(2,1) := 3, \\ \end{cases}$$

(2.57)

is strictly isotone and invertible, however $f^{-1}$ is not isotone (since $2 < 3$, but $f^{-1}(2) = (1, 2)$ and $f^{-1}(3) = (2, 1)$ are not comparable, i.e. $f^{-1}(2) \leq f^{-1}(3)$ is not true).

3 Natural Numbers, Induction, and the Size of Sets

3.1 Induction and Recursion

One of the most useful proof techniques is the method of induction – it is used in situations, where one needs to verify the truth of statements $\phi(n)$ for each $n \in \mathbb{N}$, i.e. the truth of the statement

$$\forall n \in \mathbb{N} \phi(n).$$

(3.1)

Induction is based on the fact that $\mathbb{N}$ satisfies the so-called Peano axioms:

**P1:** $\mathbb{N}$ contains a special element called *one*, denoted 1.

**P2:** There exists an injective map $S : \mathbb{N} \rightarrow \mathbb{N} \setminus \{1\}$, called the *successor function* (for each $n \in \mathbb{N}$, $S(n)$ is called the *successor of n*).

**P3:** If a subset $A$ of $\mathbb{N}$ has the property that $1 \in A$ and $S(n) \in A$ for each $n \in A$, then $A$ is equal to $\mathbb{N}$. Written as a formula, the third axiom is:

$$\forall A \in \mathcal{P}(\mathbb{N}) \left( 1 \in A \land S(A) \subseteq A \Rightarrow A = \mathbb{N} \right).$$

**Remark 3.1.** In Def. 1.28, we had introduced the natural numbers $\mathbb{N} := \{1, 2, 3, \ldots \}$. The successor function is $S(n) = n + 1$. In axiomatic set theory, one starts with the Peano axioms and shows that the axioms of set theory allow the construction of a set $\mathbb{N}$ which satisfies the Peano axioms. One then defines $2 := S(1)$, $3 := S(2)$, $\ldots$, $n + 1 := S(n)$. The interested reader can find more details in Appendix D.1.
Theorem 3.2 (Principle of Induction). Suppose, for each \( n \in \mathbb{N} \), \( \phi(n) \) is a statement (i.e. a predicate of \( n \) in the language of Def. 1.31). If (a) and (b) both hold, where

(a) \( \phi(1) \) is true,

(b) \( \forall n \in \mathbb{N} \ (\phi(n) \Rightarrow \phi(n+1)) \),

then (3.1) is true, i.e. \( \phi(n) \) is true for every \( n \in \mathbb{N} \).

Proof. Let \( A := \{ n \in \mathbb{N} : \phi(n) \} \). We have to show \( A = \mathbb{N} \). Since \( 1 \in A \) by (a), and

\[ n \in A \Rightarrow \phi(n) \overset{(b)}{\Rightarrow} \phi(n+1) \Rightarrow S(n) = n + 1 \in A, \tag{3.2} \]

i.e. \( S(A) \subseteq A \), the Peano axiom P3 implies \( A = \mathbb{N} \). \( \blacksquare \)

Remark 3.3. To prove some \( \phi(n) \) for each \( n \in \mathbb{N} \) by induction according to Th. 3.2 consists of the following two steps:

(a) Prove \( \phi(1) \), the so-called base case.

(b) Perform the inductive step, i.e. prove that \( \phi(n) \) (the induction hypothesis) implies \( \phi(n+1) \).

Example 3.4. We use induction to prove the statement

\[ \forall n \in \mathbb{N} \left( 1 + 2 + \cdots + n = \frac{n(n+1)}{2} \right) : \tag{3.3} \]

Base Case \((n = 1)\): \( 1 = \frac{1+2}{2} \), i.e. \( \phi(1) \) is true.

Induction Hypothesis: Assume \( \phi(n) \), i.e. \( 1 + 2 + \cdots + n = \frac{n(n+1)}{2} \) holds.

Induction Step: One computes

\[ 1 + 2 + \cdots + n + (n+1) = \frac{n(n+1)}{2} + n + 1 = \frac{n(n+1) + 2n + 2}{2} = \frac{n^2 + 3n + 2}{2} = \frac{(n+1)(n+2)}{2}, \tag{3.4} \]

i.e. \( \phi(n+1) \) holds and the induction is complete.

Corollary 3.5. Theorem 3.2 remains true if (b) is replaced by

\[ \forall n \in \mathbb{N} \left( \left( \forall_{1 \leq m \leq n} \phi(m) \right) \Rightarrow \phi(n+1) \right) . \tag{3.5} \]
Proof. If, for each \( n \in \mathbb{N} \), we use \( \psi(n) \) to denote \( \forall 1 \leq m \leq n \phi(m) \), then (3.5) is equivalent to \( \forall n \in \mathbb{N} (\psi(n) \Rightarrow \psi(n + 1)) \), i.e. to Th. 3.2(b) with \( \phi \) replaced by \( \psi \). Thus, Th. 3.2 implies \( \psi(n) \) holds true for each \( n \in \mathbb{N} \), i.e. \( \phi(n) \) holds true for each \( n \in \mathbb{N} \).

\[ \blacksquare \]

**Corollary 3.6.** Let \( I \) be an index set. Suppose, for each \( i \in I \), \( \phi(i) \) is a statement. If there is a bijective map \( f : \mathbb{N} \rightarrow I \) and (a) and (b) both hold, where

(a) \( \phi(f(1)) \) is true,

(b) \( \forall n \in \mathbb{N} (\phi(f(n)) \Rightarrow \phi(f(n + 1))) \),

then \( \phi(i) \) is true for every \( i \in I \).

**Finite Induction:** The above assertion remains true if \( f : \{1, \ldots, m\} \rightarrow I \) is bijective for some \( m \in \mathbb{N} \) and \( \mathbb{N} \) in (b) is replaced by \( \{1, \ldots, m - 1\} \).

Proof. If, for each \( n \in \mathbb{N} \), we use \( \psi(n) \) to denote \( \phi(f(n)) \), then Th. 3.2 shows \( \psi(n) \) is true for every \( n \in \mathbb{N} \). Given \( i \in I \), we have \( n := f^{-1}(i) \in \mathbb{N} \) with \( f(n) = i \), showing that \( \phi(i) = \phi(f(n)) = \psi(n) \) is true.

For the finite induction, let \( \psi(n) \) denote \( (n \leq m \land \phi(f(n))) \lor n > m \). Then, for \( 1 \leq n < m \), we have \( \psi(n) \Rightarrow \psi(n + 1) \) due to (b). For \( n \geq m \), we also have \( \psi(n) \Rightarrow \psi(n + 1) \) due to \( n \geq m \Rightarrow n + 1 > m \). Thus, Th. 3.2 shows \( \psi(n) \) is true for every \( n \in \mathbb{N} \). Given \( i \in I \), it is \( n := f^{-1}(i) \in \{1, \ldots, m\} \) with \( f(n) = i \). Since \( n \leq m \land \psi(n) \Rightarrow \phi(f(n)) \), we obtain that \( \phi(i) \) is true.

Apart from providing a widely employable proof technique, the most important application of Th. 3.2 is the possibility to define sequences inductively, using so-called recursion:

**Theorem 3.7 (Recursion Theorem).** Let \( A \) be a nonempty set and \( x \in A \). Given a sequence of functions \((f_n)_{n \in \mathbb{N}}\), where \( f_n : A^n \rightarrow A \), there exists a unique sequence \((x_n)_{n \in \mathbb{N}}\) in \( A \) satisfying the following two conditions:

(i) \( x_1 = x \).

(ii) \( \forall n \in \mathbb{N} \) \( x_{n+1} = f_n(x_1, \ldots, x_n) \).

The same holds if \( \mathbb{N} \) is replaced by an index set \( I \) as in Cor. 3.6.

Proof. To prove uniqueness, let \((x_n)_{n \in \mathbb{N}}\) and \((y_n)_{n \in \mathbb{N}}\) be sequences in \( A \), both satisfying (i) and (ii), i.e.

\[ x_1 = y_1 = x \quad \text{and} \quad \forall n \in \mathbb{N} \ \left( x_{n+1} = f_n(x_1, \ldots, x_n) \land y_{n+1} = f_n(y_1, \ldots, y_n) \right). \]
We prove by induction (in the form of Cor. 3.5) that \((x_n)_{n \in \mathbb{N}} = (y_n)_{n \in \mathbb{N}}\), i.e.

\[
\forall_{n \in \mathbb{N}} x_n = y_n : \phi(n)
\]

(3.7)

Base Case \((n = 1)\): \(\phi(1)\) is true according to (3.6a).

Induction Hypothesis: Assume \(\phi(m)\) for each \(m \in \{1, \ldots, n\}\), i.e. \(x_m = y_m\) holds for each \(m \in \{1, \ldots, n\}\).

Induction Step: One computes

\[
x_{n+1} = f_n(x_1, \ldots, x_n) = f_n(y_1, \ldots, y_n) = y_{n+1},
\]

(3.8)

i.e. \(\phi(n+1)\) holds and the induction is complete.

To prove existence, we have to show that there is a function \(F : \mathbb{N} \rightarrow A\) such that the following two conditions hold:

\[
\forall \quad F(1) = x,
\]

\[
\forall_{n \in \mathbb{N}} F(n+1) = f_n(F(1), \ldots, F(n)).
\]

(3.9b)

To this end, let

\[
\mathcal{F} := \left\{ B \subseteq \mathbb{N} \times A : (1, x) \in B \land \forall_{n \in \mathbb{N}} \left( (n+1, f_n(a_1, \ldots, a_n)) \in B \right) \right\}
\]

(3.10)

and

\[
G := \bigcap_{B \in \mathcal{F}} B.
\]

(3.11)

Note that \(G\) is well-defined, as \(\mathbb{N} \times A \in \mathcal{F}\). Also, clearly, \(G \in \mathcal{F}\). We would like to define \(F\) such that \(G = \text{graph}(F)\). For this to be possible, we will show, by induction,

\[
\forall_{n \in \mathbb{N}} \exists!_{x_n \in A} (n, x_n) \in G.
\]

(3.12)

Base Case \((n = 1)\): From the definition of \(G\), we know \((1, x) \in G\). If \((1, a) \in G\) with \(a \neq x\), then \(H := G \setminus \{(1, a)\} \in \mathcal{F}\), implying \(G \subseteq H\) in contradiction to \((1, a) \notin H\). This shows \(a = x\) and proves \(\phi(1)\).

Induction Hypothesis: Assume \(\phi(m)\) for each \(m \in \{1, \ldots, n\}\).

Induction Step: From the induction hypothesis, we know

\[
\exists!_{(x_1, \ldots, x_n) \in A^n} (1, x_1), \ldots, (n, x_n) \in G.
\]

Thus, if we let \(x_{n+1} := f_n(x_1, \ldots, x_n)\), then \((n+1, x_{n+1}) \in G\) by the definition of \(G\). If \((n+1, a) \in G\) with \(a \neq x_{n+1}\), then \(H := G \setminus \{(n+1, a)\} \in \mathcal{F}\) (using the uniqueness of
the \((1,x_1), \ldots, (n,x_n) \in G\), implying \(G \subseteq H\) in contradiction to \((n + 1, a) \notin H\). This shows \(a = x_{n+1}\), proves \(\phi(n+1)\), and completes the induction.

Due to (3.12), we can now define \(F : \mathbb{N} \rightarrow A; F(n) := x_n\), and the definition of \(G\) then guarantees the validity of (3.9).

Example 3.8. In many applications of Th. 3.7, one has functions \(g_n : A \rightarrow A\) and uses

\[ \forall n \in \mathbb{N} \ (f_n : A^n \rightarrow A; f_n(a_1, \ldots, a_n) := g_n(a_n)). \tag{3.13} \]

Here are some important concrete examples:

(a) The factorial function \(F : \mathbb{N}_0 \rightarrow \mathbb{N}; n \mapsto n!\), is defined recursively by

\[ 0! := 1, \quad 1! := 1, \quad \forall n \in \mathbb{N} \ (n + 1)! := (n + 1) \cdot n!, \tag{3.14a} \]

i.e. we have \(A = \mathbb{N}\) and \(g_n(x) := (n + 1) \cdot x\). So we obtain

\[ (n!)_{n \in \mathbb{N}_0} = (1, 1, 2, 6, 24, 120, \ldots). \tag{3.14b} \]

(b) For each \(a \in \mathbb{R}\) and each \(d \in \mathbb{R}\), we define the following arithmetic progression (also called arithmetic sequence) recursively by

\[ a_1 := a, \quad \forall n \in \mathbb{N} \ a_{n+1} := a_n + d, \tag{3.15a} \]

i.e. we have \(A = \mathbb{R}\) and \(g_n = g\) with \(g(x) := x + d\). For example, for \(a = 2\) and \(d = -0.5\), we obtain

\[ (a_n)_{n \in \mathbb{N}} = (2, 1.5, 1, 0.5, 0, -0.5, -1, -1.5, \ldots). \tag{3.15b} \]

(c) For each \(a \in \mathbb{R}\) and each \(q \in \mathbb{R} \setminus \{0\}\), we define the following geometric progression (also called geometric sequence) recursively by

\[ x_1 := a, \quad \forall n \in \mathbb{N} \ x_{n+1} := x_n \cdot q, \tag{3.16a} \]

i.e. we have \(A = \mathbb{R}\) and \(g_n = g\) with \(g(x) := x \cdot q\). For example, for \(a = 3\) and \(q = -2\), we obtain

\[ (x_n)_{n \in \mathbb{N}} = (3, -6, 12, -24, 48, \ldots). \tag{3.16b} \]

For the time being, we will continue to always specify \(A\) and the \(g_n\) or \(f_n\) in subsequent recursive definitions, but in the literature, most of the time, the \(g_n\) or \(f_n\) are not provided explicitly.

Example 3.9. (a) The Fibonacci sequence consists of the Fibonacci numbers, defined recursively by

\[ F_0 := 0, \quad F_1 := 1, \quad \forall n \in \mathbb{N} \ F_{n+1} := F_n + F_{n-1}, \tag{3.17a} \]
i.e. we have $A = \mathbb{N}_0$ and

$$f_n : A^n \to A, \quad f_n(a_1, \ldots, a_n) := \begin{cases} 1 & \text{for } n = 1, \\ a_n + a_{n-1} & \text{for } n \geq 2. \end{cases} \quad (3.17b)$$

So we obtain

$$(F_n)_{n \in \mathbb{N}_0} = (0, 1, 1, 2, 3, 5, 8, 13, 21, 34, 55, 89, \ldots). \quad (3.17c)$$

(b) For $A := \mathbb{N}$, $x := 1$, and

$$f_n : A^n \to A, \quad f_n(a_1, \ldots, a_n) := a_1 + \cdots + a_n, \quad (3.18a)$$

one obtains

$$x_1 = 1, \quad x_2 = f_1(1) = 1, \quad x_3 = f_2(1, 1) = 2, \quad x_4 = f_3(1, 1, 2) = 4, \quad x_5 = f_4(1, 1, 2, 4) = 8, \quad x_6 = f_5(1, 1, 2, 4, 8) = 16, \ldots \quad (3.18b)$$

**Definition 3.10.**

(a) **Summation Symbol:** On $A = \mathbb{R}$ (or, more generally, on every set $A$, where an addition $+ : A \times A \to A$ is defined), define recursively, for each given (possibly finite) sequence $(a_1, a_2, \ldots)$ in $A$:

$$\sum_{i=1}^{n} a_i := a_1, \quad \sum_{i=1}^{n+1} a_i := a_{n+1} + \sum_{i=1}^{n} a_i \text{ for } n \geq 1, \quad (3.19a)$$

i.e.

$$f_n : A^n \to A, \quad f_n(x_1, \ldots, x_n) := x_n + a_{n+1}. \quad (3.19b)$$

In (3.19a), one can also use other symbols for $i$, except $a$ and $n$; for a finite sequence, $n$ needs to be less than the maximal index of the finite sequence.

More generally, if $I$ is an index set and $\phi : \{1, \ldots, n\} \to I$ a bijective map, then define

$$\sum_{i \in I} a_i := \sum_{i=1}^{n} a_{\phi(i)}. \quad (3.19c)$$

The commutativity of addition implies that the definition in (3.19c) is actually independent of the chosen bijective map $\phi$ (cf. Th. B.5). Also define

$$\sum_{i \in \emptyset} a_i := 0 \quad (3.19d)$$

(for a general $A$, 0 is meant to be an element such that $a + 0 = 0 + a = a$ for each $a \in A$ and we can even define this if $0 \notin A$).

(b) **Product Symbol:** On $A = \mathbb{R}$ (or, more generally, on every set $A$, where a multiplication $\cdot : A \times A \to A$ is defined), define recursively, for each given (possibly finite) sequence $(a_1, a_2, \ldots)$ in $A$:

$$\prod_{i=1}^{1} a_i := a_1, \quad \prod_{i=1}^{n+1} a_i := a_{n+1} \cdot \prod_{i=1}^{n} a_i \text{ for } n \geq 1, \quad (3.20a)$$
i.e. 

\[ f_n : A^n \rightarrow A, \quad f_n(x_1, \ldots, x_n) := a_{n+1} \cdot x_n. \]  

(3.20b)

In (3.20a), one can also use other symbols for \( i \), except \( a \) and \( n \); for a finite sequence, \( n \) needs to be less than the maximal index of the finite sequence.

More generally, if \( I \) is an index set and \( \phi : \{1, \ldots, n\} \rightarrow I \) a bijective map, then define

\[ \prod_{i \in I} a_i := \prod_{i=1}^{n} a_{\phi(i)}. \]  

(3.20c)

The commutativity of multiplication implies that the definition in (3.20c) is actually independent of the chosen bijective map \( \phi \) (cf. Th. B.5). Also define

\[ \prod_{i \in \emptyset} a_i := 1 \]  

(3.20d)

(for a general \( A \), 1 is meant to be an element such that \( a \cdot 1 = 1 \cdot a = a \) for each \( a \in A \) and we can even define this if \( 1 \notin A \)).

**Example 3.11.** (a) Given \( a, d \in \mathbb{R} \), let \((a_n)_{n \in \mathbb{N}}\) be the arithmetic sequence as defined in (3.15a). It is an exercise to prove by induction that

\[ \forall n \in \mathbb{N} \quad a_n = a + (n - 1)d, \]  

(3.21a)

\[ \forall n \in \mathbb{N} \quad S_n := \sum_{i=1}^{n} a_i = \frac{n}{2} (a_1 + a_n) = \frac{n}{2} (2a + (n - 1)d), \]  

(3.21b)

where the \( S_n \) are called arithmetic sums.

(b) Given \( a \in \mathbb{R} \) and \( q \in \mathbb{R} \setminus \{0\} \), let \((x_n)_{n \in \mathbb{N}}\) be the geometric sequence as defined in (3.16a). We will prove by induction that

\[ \forall n \in \mathbb{N} \quad x_n = a q^{n-1}, \]  

(3.22a)

\[ \forall n \in \mathbb{N} \quad S_n := \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} (a q^{i-1}) = a \sum_{i=0}^{n-1} q^i = \begin{cases} \frac{n a}{1 - q} & \text{for } q = 1, \\ a \frac{(1-q^n)}{1-q} & \text{for } q \neq 1, \end{cases} \]  

(3.22b)

where the \( S_n \) are called geometric sums.

For the induction proof of (3.22a), \( \phi(n) \) is \( x_n = a q^{n-1} \). The base case, \( \phi(1) \), is the statement \( x_1 = a q^0 = a \), which is true. For the induction step, we assume \( \phi(n) \) and compute

\[ x_{n+1} = x_n \cdot q = q a q^{n-1} = a q^n, \]  

(3.23)

showing \( \phi(n) \Rightarrow \phi(n+1) \) and completing the proof.

For \( q = 1 \), the sum \( S_n \) is actually arithmetic with \( d = 0 \), i.e. \( S_n = na \) can be obtained from (3.21b). For the induction proof of (3.22b) with \( q \neq 1 \), \( \phi(n) \) is
\[ S_n = \frac{a(1-q^n)}{1-q}. \] The base case, \( \phi(1) \), is the statement \( S_1 = \frac{a(1-q)}{1-q} = a \), which is true. For the induction step, we assume \( \phi(n) \) and compute

\[
S_{n+1} = S_n + x_{n+1} \left( \frac{\phi(n)}{1-q} \right) + aq^n = \frac{a(1-q^n) +aq^n(1-q)}{1-q} = \frac{a(1-q^{n+1})}{1-q},
\]

showing \( \phi(n) \Rightarrow \phi(n+1) \) and completing the proof.

### 3.2 Cardinality: The Size of Sets

Cardinality measures the size of sets. For a finite set \( A \), it is precisely the number of elements in \( A \). For an infinite set, it classifies the set’s degree or level of infinity (it turns out that not all infinite sets have the same size).

**Definition 3.12.** (a) The sets \( A, B \) are defined to have the same cardinality or the same size if, and only if, there exists a bijective map \( \varphi : A \rightarrow B \). One can show that this defines an equivalence relation on every set of sets (see Th. A.53 of the Appendix).

(b) The cardinality of a set \( A \) is \( n \in \mathbb{N} \) (denoted \( \#A = n \)) if, and only if, there exists a bijective map \( \varphi : A \rightarrow \{1, \ldots, n\} \). The cardinality of \( \emptyset \) is defined as 0, i.e. \( \#\emptyset := 0 \). A set \( A \) is called finite if, and only if, there exists \( n \in \mathbb{N}_0 \) such that \( \#A = n \); \( A \) is called infinite if, and only if, \( A \) is not finite, denoted \( \#A = \infty \) (in the strict sense, this is an abuse of notation, since \( \infty \) is not a cardinality – for example \( \#\mathbb{N} = \infty \) and \( \#\mathcal{P}(\mathbb{N}) = \infty \), but \( \mathbb{N} \) and \( \mathcal{P}(\mathbb{N}) \) do not have the same cardinality, since the power set \( \mathcal{P}(A) \) is always strictly bigger than \( A \) (see Th. A.69 of the Appendix) – \( \#A = \infty \) is merely an abbreviation for the statement “\( A \) is infinite”). The interested student finds additional material regarding the uniqueness of finite cardinality in Th. A.61 and Cor. A.62, and regarding characterizations of infinite sets in Th. A.54 of the Appendix.

(c) The set \( A \) is called countable if, and only if, \( A \) is finite or \( A \) has the same cardinality as \( \mathbb{N} \). Otherwise, \( A \) is called uncountable.

---

In the rest of the section, we present a number of important results regarding the natural numbers and countability.

**Theorem 3.13.** (a) Every nonempty finite subset of a totally ordered set has a minimum and a maximum.

(b) Every nonempty subset of \( \mathbb{N} \) has a minimum.

**Proof.** (a): Let \( A \) be a set and let \( \leq \) denote a total order on \( A \). Moreover, let \( \emptyset \neq B \subseteq A \). We show by induction

\[
\forall n \in \mathbb{N} \left( \#B = n \Rightarrow B \text{ has a min} \right). \]

\[ \phi(n) \]
3  NATURAL NUMBERS, INDUCTION, AND THE SIZE OF SETS

Base Case \((n = 1)\): For \(n = 1\), \(B\) contains a unique element \(b\), i.e. \(b = \min B\), proving \(\phi(1)\).

Induction Step: Suppose \(\phi(n)\) holds and consider \(B\) with \(#B = n + 1\). Let \(b\) be one element from \(B\). Then \(C := B \setminus \{b\}\) has cardinality \(n\) and, according to the induction hypothesis, there exists \(c \in C\) satisfying \(c = \min C\). If \(c \leq b\), then \(c \leq x\) for each \(x \in B\), proving \(c = \min B\). If \(b \leq c\), then \(b \leq x\) for each \(x \in B\), proving \(b = \min B\). In each case, \(B\) has a min, proving \(\phi(n + 1)\) and completing the induction.

(b): Let \(\emptyset \neq A \subseteq \mathbb{N}\). We have to show \(A\) has a min. If \(A\) is finite, then \(A\) has a min by (a). If \(A\) is infinite, let \(n\) be an element from \(A\). Then the finite set \(B := \{k \in A : k \leq n\}\) must have a min \(m\) by (a). Since \(m \leq x\) for each \(x \in B\) and \(m \leq n < x\) for each \(x \in A \setminus B\), we have \(m = \min A\).

**Proposition 3.14.** Every subset \(A\) of \(\mathbb{N}\) is countable.

**Proof.** Since \(\emptyset\) is countable, we may assume \(A \neq \emptyset\). From Th. 3.13(b), we know that every nonempty subset of \(\mathbb{N}\) has a min. We recursively define a sequence in \(A\) by

\[
a_1 := \min A, \quad a_{n+1} := \begin{cases} 
\min A_n & \text{if } A_n := A \setminus \{a_i : 1 \leq i \leq n\} \neq \emptyset, \\
\min A & \text{if } A_n = \emptyset.
\end{cases}
\]

This sequence is the same as the function \(f : \mathbb{N} \to A, f(n) = a_n\). An easy induction shows that, for each \(n \in \mathbb{N}\), \(a_n \neq a_{n+1}\) implies the restriction \(f |_{\{1, \ldots, n+1\}}\) is injective. Thus, if there exists \(n \in \mathbb{N}\) such that \(a_n = a_{n+1}\), then \(f |_{\{1, \ldots, k\}} : \{1, \ldots, k\} \to A\) is bijective, where \(k := \min \{n \in \mathbb{N} : a_n = a_{n+1}\}\), showing \(A\) is finite, i.e. countable. If there does not exist \(n \in \mathbb{N}\) with \(a_n = a_{n+1}\), then \(f\) is injective. Another easy induction shows that, for each \(n \in \mathbb{N}\), \(f(\{1, \ldots, n\}) \supseteq \{k \in A : k \leq n\}\), showing \(f\) is also surjective, proving \(A\) is countable.

**Proposition 3.15.** For each set \(A \neq \emptyset\), the following three statements are equivalent:

(i) \(A\) is countable.

(ii) There exists an injective map \(f : A \to \mathbb{N}\).

(iii) There exists a surjective map \(g : \mathbb{N} \to A\).

**Proof.** Directly from the definition of countable in Def. 3.12(c), one obtains (i)⇒(ii) and (i)⇒(iii). To prove (ii)⇒(i), let \(f : A \to \mathbb{N}\) be injective. Then \(f : A \to f(A)\) is bijective, and, since \(f(A) \subseteq \mathbb{N}\), \(f(A)\) is countable by Prop. 3.14, proving \(A\) is countable as well. To prove (iii)⇒(i), let \(g : \mathbb{N} \to A\) be surjective. Then \(g\) has a right inverse \(f : A \to \mathbb{N}\). One can obtain this from Th. 2.13(a), but, here, we can actually construct \(f\) without the axiom of choice: For \(a \in A\), let \(f(a) := \min g^{-1}(\{a\})\) (recall Th. 3.13(b)). Then, clearly, \(g \circ f = \text{Id}_A\). But this means \(g\) is a left inverse for \(f\), showing \(f\) is injective according to Th. 2.13(b). Then \(A\) is countable by an application of (ii).

**Theorem 3.16.** If \((A_1, \ldots, A_n), n \in \mathbb{N}\), is a finite family of countable sets, then \(\prod_{i=1}^n A_i\) is countable.
Proof. We first consider the special case \( n = 2 \) with \( A_1 = A_2 = \mathbb{N} \) and show the map

\[
\phi : \mathbb{N} \times \mathbb{N} \rightarrow \mathbb{N}, \quad \phi(m,n) := 2^m \cdot 3^n,
\]

is injective: If \( \phi(m,n) = \phi(p,q) \), then \( 2^m \cdot 3^n = 2^p \cdot 3^q \). Moreover \( m \leq p \) or \( p \leq m \).

If \( m \leq p \), then \( 3^n = 2^{p-m} \cdot 3^q \). Since \( 3^n \) is odd, \( 2^{p-m} \cdot 3^q \) must also be odd, implying \( p - m = 0 \), i.e. \( m = p \). Moreover, we now have \( 3^n = 3^q \), implying \( n = q \), showing \( (m,n) = (p,q) \), i.e. \( \phi \) is injective.

We now come back to the general case stated in the theorem. If at least one of the \( A_i \) is empty, then \( A \) is empty. So it remains to consider the case, where all \( A_i \) are nonempty. The proof is conducted by induction by showing

\[
\forall n \in \mathbb{N}, \prod_{i=1}^{n} A_i \text{ is countable.}
\]

Base Case (\( n = 1 \)): \( \phi(1) \) is merely the hypothesis that \( A_1 \) is countable.

Induction Step: Assuming \( \phi(n) \), Prop. 3.15(ii) provides injective maps \( f_1 : \prod_{i=1}^{n} A_i \rightarrow \mathbb{N} \) and \( f_2 : A_{n+1} \rightarrow \mathbb{N} \). To prove \( \phi(n+1) \), we provide an injective map \( h : \prod_{i=1}^{n+1} A_i \rightarrow \mathbb{N} \): Define

\[
h : \prod_{i=1}^{n+1} A_i \rightarrow \mathbb{N}, \quad h(a_1, \ldots, a_n, a_{n+1}) := \phi(f_1(a_1, \ldots, a_n), f_2(a_{n+1})).
\]

The injectivity of \( f_1 \), \( f_2 \), and \( \phi \) clearly implies the injectivity of \( h \), thereby proving \( \phi(n+1) \) and completing the induction. \( \square \)

**Theorem 3.17.** If \( (A_i)_{i \in I} \) is a countable family of countable sets (i.e. \( \emptyset \neq I \) is countable and each \( A_i, i \in I \), is countable), then the union \( A := \bigcup_{i \in I} A_i \) is also countable (this result makes use of AC, cf. Rem. 3.18 below).

Proof. It suffices to consider the case that all \( A_i \) are nonempty. Moreover, according to Prop. 3.15(iii), it suffices to construct a surjective map \( \varphi : \mathbb{N} \rightarrow A \). Also according to Prop. 3.15(iii), the countability of \( I \) and the \( A_i \) provides us with surjective maps \( f : \mathbb{N} \rightarrow I \) and \( g_i : \mathbb{N} \rightarrow A_i \) (here AC is used to select each \( g_i \) from the set of all surjective maps from \( \mathbb{N} \) onto \( A_i \)). Define

\[
F : \mathbb{N} \times \mathbb{N} \rightarrow A, \quad F(m,n) := g_{f(m)}(n).
\]

Then \( F \) is surjective: Given \( x \in A \), there exists \( i \in I \) such that \( x \in A_i \). Since \( f \) is surjective, there is \( m \in \mathbb{N} \) satisfying \( f(m) = i \). Moreover, since \( g_i \) is surjective, there exists \( n \in \mathbb{N} \) with \( g_i(n) = x \). Then \( F(m,n) = g_i(n) = x \), verifying that \( F \) is surjective.

As \( \mathbb{N} \times \mathbb{N} \) is countable by Th. 3.16, there exists a surjective map \( h : \mathbb{N} \rightarrow \mathbb{N} \times \mathbb{N} \). Thus, \( F \circ h \) is the desired surjective map from \( \mathbb{N} \) onto \( A \). \( \square \)

**Remark 3.18.** The axiom of choice is, indeed, essential for the proof of Th. 3.17. It is shown in [Jec73, Th. 10.6] that it is consistent with the axioms of ZF (i.e. with the axioms of Sec. A.3 of the Appendix) that, e.g., the uncountable sets \( \mathcal{P}(\mathbb{N}) \) and \( \mathbb{R} \) (cf. Th. F.2 of the Appendix) are countable unions of countable sets.
4 Real Numbers

4.1 The Real Numbers as a Complete Totally Ordered Field

The set of real numbers, denoted \( \mathbb{R} \), is a set with special properties, namely a so-called complete totally ordered field, which, after some preliminaries, will be defined in Def. 4.3 below.

**Definition 4.1.** A total order \( \leq \) on a nonempty set \( A \) is called complete if, and only if, every nonempty subset \( B \) of \( A \) that is bounded from above has a supremum, i.e.

\[
\forall_{B \in \mathcal{P}(A) \setminus \{\emptyset\}} \left( \left( \exists_{x \in A} \forall_{b \in B} b \leq x \right) \Rightarrow \exists_{s \in A} s = \sup B \right).
\]  

(4.1)

**Lemma 4.2.** A total order \( \leq \) on a nonempty set \( A \) is complete if, and only if, every nonempty subset \( B \) of \( A \) that is bounded from below has an infimum.

**Proof.** According to Lem. 2.29, it suffices to prove one implication. We show that (4.1) implies that every nonempty \( B \) bounded from below has an infimum: Define

\[
C := \{x \in A : x \text{ is lower bound for } B\}.
\]  

(4.2)

Then every \( b \in B \) is an upper bound for \( C \) and (4.1) implies there exists \( s = \sup C \in A \).

To verify \( s = \inf B \), it remains to show \( s \in C \), i.e. that \( s \) is a lower bound for \( B \). However, every \( b \in B \) is an upper bound for \( C \) and \( s = \sup C \) is the min of all upper bounds for \( C \), i.e. \( s \leq b \) for each \( b \in B \), showing \( s \in C \). \[\square\]

**Definition 4.3.** Let \((A,+,\cdot)\) be a field and let \( \leq \) be a total order on \( A \). Then \( A \) (or, more precisely, \((A,+,\cdot,\leq)\)) is called a totally ordered field if, and only if, the order is compatible with addition and multiplication, i.e. if, and only if,

\[
\forall_{x,y,z \in A} \left( x \leq y \Rightarrow x + z \leq y + z \right),
\]  

(4.3a)

\[
\forall_{x,y \in A} \left( 0 \leq x \land 0 \leq y \Rightarrow 0 \leq x y \right).
\]  

(4.3b)

Finally, \( A \) is called a complete totally ordered field if, and only if, \( A \) is a totally ordered field that is complete in the sense of Def. 4.1.

**Theorem 4.4.** There exists a complete totally ordered field \( \mathbb{R} \) (it is called the set of real numbers). Moreover, \( \mathbb{R} \) is unique up to isomorphism, i.e. if \( A \) is a complete totally ordered field, then there exists an isomorphism \( \phi : A \to \mathbb{R} \), i.e. a bijective map \( \phi : A \to \mathbb{R} \), satisfying

\[
\forall_{x,y \in A} \phi(x + y) = \phi(x) + \phi(y),
\]  

(4.4a)

\[
\forall_{x,y \in A} \phi(xy) = \phi(x)\phi(y),
\]  

(4.4b)

\[
\forall_{x,y \in A} \left( x < y \Rightarrow \phi(x) < \phi(y) \right).
\]  

(4.4c)

It also turns out that the isomorphism is unique.
Proof. To really prove the existence of the real numbers by providing a construction is tedious and not easy. One possible construction is provided in Appendix D (the existence proof is completed in Th. D.41, the results regarding the isomorphism can be found in Th. D.45).

Theorem 4.5. The following statements and rules are valid in the set of real numbers \( \mathbb{R} \) (and, more generally, in every totally ordered field):

(a) \( x \leq y \Rightarrow -x \geq -y \).

(b) \( x \leq y \land z \geq 0 \Rightarrow xz \leq yz \) holds as well as \( x \leq y \land z \leq 0 \Rightarrow xz \geq yz \).

(c) \( x \neq 0 \Rightarrow x^2 := x \cdot x > 0 \). In particular \( 1 > 0 \).

(d) \( x > 0 \Rightarrow 1/x > 0 \), whereas \( x < 0 \Rightarrow 1/x < 0 \).

(e) If \( 0 < x < y \), then \( x/y < 1 \), \( y/x > 1 \), and \( 1/x > 1/y \).

(f) \( x < y \land u < v \Rightarrow x + u < y + v \).

(g) \( 0 < x < y \land 0 < u < v \Rightarrow xu < yv \).

(h) \( x < y \land 0 < \lambda < 1 \Rightarrow x < \lambda x + (1 - \lambda)y < y \). In particular \( x < \frac{x+y}{2} < y \).

Proof. (a): Using (4.3a): \( x \leq y \Rightarrow 0 \leq y - x \Rightarrow -y \leq -x \).

(b): One argues, for \( z \geq 0 \),

\[
\text{if } x \leq y \Rightarrow 0 \leq y - x \overset{\text{(4.3b)}}{\Rightarrow} 0 \leq (y - x)z = yz - xz \Rightarrow xz \leq yz,
\]

and, for \( z \leq 0 \),

\[
\text{if } x \leq y \Rightarrow 0 \leq y - x \overset{\text{(4.3b)}}{\Rightarrow} 0 \leq (y - x)(-z) = xz - yz \Rightarrow xz \geq yz.
\]

(c): From (b), one obtains \( x^2 \geq 0 \). From Th. C.10(i), one then gets \( x^2 > 0 \).

(d): If \( x > 0 \), then \( x^{-1} < 0 \) implies the false statement \( 1 = xx^{-1} < 0 \), i.e. \( x^{-1} > 0 \). The case \( x < 0 \) is treated analogously.

(e): Using (d), we obtain from \( 0 < x < y \) that \( x/y = xy^{-1} < yy^{-1} = 1 \) and \( 1 = xx^{-1} < yy^{-1} = y/x \).

(f): \( x < y \Rightarrow x + u < y + u \) and \( u < v \Rightarrow y + u < y + v \); both combined yield \( x + u < y + v \).

(g): \( 0 < x < y \land 0 < u < v \Rightarrow xu < yu \land yu < yv \Rightarrow xu < yv \).

(h): Since \( 0 < \lambda \) and \( 1 - \lambda > 0 \), \( x < y \) implies

\[
\lambda x < \lambda y \land (1 - \lambda)x < (1 - \lambda)y.
\]

Using (4.3a), we obtain

\[
x = \lambda x + (1 - \lambda)x < \lambda x + (1 - \lambda)y < \lambda y + (1 - \lambda)y = y,
\]

completing the proof of the theorem.
Theorem 4.6. Let $\emptyset \neq A, B \subseteq \mathbb{R}$, $\lambda \in \mathbb{R}$, and define

$$A + B := \{a + b : a \in A \land b \in B\}, \quad (4.5a)$$

$$\lambda A := \{\lambda a : a \in A\}. \quad (4.5b)$$

If $A$ and $B$ are bounded, then

$$\sup(A + B) = \sup A + \sup B, \quad (4.6a)$$
$$\inf(A + B) = \inf A + \inf B, \quad (4.6b)$$

$$\sup(\lambda A) = \begin{cases} 
\lambda \cdot \sup A & \text{for } \lambda \geq 0, \\
\lambda \cdot \inf A & \text{for } \lambda < 0,
\end{cases} \quad (4.6c)$$

$$\inf(\lambda A) = \begin{cases} 
\lambda \cdot \inf A & \text{for } \lambda \geq 0, \\
\lambda \cdot \sup A & \text{for } \lambda < 0.
\end{cases} \quad (4.6d)$$

Proof. Exercise.

4.2 Important Subsets

Remark 4.7. We would like to recover the natural numbers $\mathbb{N}$ as a subset of $\mathbb{R}$. Indeed, if we start with 1 as the neutral element of multiplication and define $2 := 1 + 1$, $3 := 2 + 1$, $\ldots$, then $\mathbb{N} := \{1, 2, \ldots\}$ is a subset of $\mathbb{R}$, satisfying the Peano axioms P1, P2, P3 of Sec. 3.1. However, if one does actually construct $\mathbb{R}$ according to the axioms of axiomatic set theory, then one starts by constructing $\mathbb{N}$ first, constructing $\mathbb{R}$ from $\mathbb{N}$ in several steps (cf. Appendix D). Depending on the construction used, the original set of natural numbers will typically not be the same set as the natural numbers as a subset of $\mathbb{R}$. However, both sets will satisfy the Peano axioms and you will have a canonical bijection between the two sets. Which one you consider the “genuine” set of natural numbers depends on your personal taste and philosophy and is completely irrelevant. Any two models of $\mathbb{N}$ will always produce equivalent results, since they must both satisfy the three Peano axioms.

We now introduce a zoo of important subsets of $\mathbb{R}$ together with corresponding notation:

$$\mathbb{N} := \{1, 2, 3, \ldots\} \quad \text{(natural numbers)}, \quad (4.7a)$$

$$\mathbb{N}_0 := \mathbb{N} \cup \{0\}, \quad (4.7b)$$

$$\mathbb{Z}^- := \{-n : n \in \mathbb{N}\} \quad \text{(negative integers)}, \quad (4.7c)$$

$$\mathbb{Z} := \mathbb{Z}^- \cup \mathbb{N}_0 \quad \text{(integers)}, \quad (4.7d)$$

$$\mathbb{Q}^+ := \{m/n : m, n \in \mathbb{N}\} \quad \text{(positive rational numbers)}, \quad (4.7e)$$

$$\mathbb{Q}_0^+ := \mathbb{Q}^+ \cup \{0\} \quad \text{(nonnegative rational numbers)}, \quad (4.7f)$$

$$\mathbb{Q}^- := \{-q : q \in \mathbb{Q}^+\} \quad \text{(negative rational numbers)}, \quad (4.7g)$$
For $a, b \in \mathbb{R}$ with $a \leq b$, one also defines the following intervals:

\[
[a, b] := \{ x \in \mathbb{R} : a \leq x \leq b \} \quad \text{(bounded closed interval)},
\]
\[
]a, b[ := \{ x \in \mathbb{R} : a < x < b \} \quad \text{(bounded open interval)},
\]
\[
]a, b] := \{ x \in \mathbb{R} : a < x \leq b \} \quad \text{(bounded half-open interval)},
\]
\[
]a, b[ := \{ x \in \mathbb{R} : a \leq x < b \} \quad \text{(bounded half-open interval)},
\]
\[
] - \infty, b[ := \{ x \in \mathbb{R} : x \leq b \} \quad \text{(unbounded closed interval)},
\]
\[
] - \infty, b[ := \{ x \in \mathbb{R} : x < b \} \quad \text{(unbounded open interval)},
\]
\[
]a, \infty[ := \{ x \in \mathbb{R} : a \leq x \} \quad \text{(unbounded closed interval)},
\]
\[
]a, \infty[ := \{ x \in \mathbb{R} : a < x \} \quad \text{(unbounded open interval)}.\nu
\]

For $a = b$, one says that the intervals defined by (4.8a) – (4.8d) are degenerate or trivial, where $[a, a] = \{a\}$, $]a, a[ = ]a, a[ = \emptyset$ – it is sometimes convenient to have included the degenerate cases in the definition. It is sometimes also useful to abandon the restriction $a \leq b$, to let $c := \min\{a, b\}$, $d := \max\{a, b\}$, and to define

\[
[a, b] := [c, d], \quad ]a, b[ := ]c, d[, \quad ]a, b[ := ]c, d[ \setminus \{a\}, \quad ]a, b[ := ]c, d[ \setminus \{b\}.\nu
\]

**Theorem 4.8 (Archimedean Property).** Let $\epsilon, x$ be real numbers. If $\epsilon > 0$ and $x > 0$, then there exists $n \in \mathbb{N}$ such that $n \epsilon > x$.

**Proof.** We conduct the proof by contradiction: Suppose $x$ is an upper bound for the set $A := \{ n \epsilon : n \in \mathbb{N}\}$. Since the order $\leq$ on $\mathbb{R}$ is complete, according to (4.1), there exists $s \in \mathbb{R}$ such that $s = \sup A$. In particular, $s - \epsilon$ is not an upper bound for $A$, i.e. there exists $n \in \mathbb{N}$ satisfying $n \epsilon > s - \epsilon$. But then $(n + 1) \epsilon > s$ in contradiction to $s = \sup A$. This shows $x$ is not an upper bound for $A$, thereby establishing the case. $\blacksquare$

## 5 Complex Numbers

### 5.1 Definition and Basic Arithmetic

According to Th. 4.5(c), $x^2 \geq 0$ holds for every real number $x \in \mathbb{R}$, i.e. the equation $x^2 + 1 = 0$ has no solution in $\mathbb{R}$. This deficiency of the real numbers motivates the effort to try to extend the field of real numbers to a larger field $\mathbb{C}$, the so-called complex numbers. The two requirements that $\mathbb{C}$ is to be a field containing $\mathbb{R}$ and that there is to
be some complex number $i \in \mathbb{C}$ satisfying $i^2 = -1$ already dictates the following laws of addition and multiplication for complex numbers $z = x + iy$ and $w = u + iv$ with $x, y, u, v \in \mathbb{R}$:

\[
\begin{align*}
    z + w &= x + iy + u + iv = x + u + i(y + v), \\
    zw &= (x + iy)(u + iv) = xu - yv + i(xv + yu).
\end{align*}
\]

Moreover, if $x + iy = u + iv$, then $(x - u)^2 = -(v - y)^2$, i.e. $x - u = 0 = v - y$, implying $x = u$ and $y = v$. This suggests to try defining complex numbers as pairs of real numbers. Indeed, this works:

**Definition 5.1.** We define the set of complex numbers $\mathbb{C} := \mathbb{R} \times \mathbb{R}$, where, keeping in mind (5.1), addition on $\mathbb{C}$ is defined by

\[
+ : \mathbb{C} \times \mathbb{C} \longrightarrow \mathbb{C}, \quad ((x, y), (u, v)) \mapsto (x, y) + (u, v) := (x + u, y + v),
\]

and multiplication on $\mathbb{C}$ is defined by

\[
\cdot : \mathbb{C} \times \mathbb{C} \longrightarrow \mathbb{C}, \quad ((x, y), (u, v)) \mapsto (x, y) \cdot (u, v) := (xu - yv, xv + yu).
\]

**Theorem 5.2.** (a) The set of complex numbers $\mathbb{C}$ with addition and multiplication as defined in Def. 5.1 forms a field, where $(0,0)$ and $(1,0)$ are the neutral elements with respect to addition and multiplication, respectively,

\[
-z := (-x, -y)
\]

is the additive inverse to $z = (x, y)$, whereas

\[
z^{-1} := \frac{1}{z} := \left(\frac{x}{x^2 + y^2}, \frac{-y}{x^2 + y^2}\right)
\]

is the multiplicative inverse to $z = (x, y) \neq (0,0)$.

(b) Defining subtraction and division in the usual way, for each $z, w \in \mathbb{C}$, by $w - z := w + (-z)$, and $w/z := wz^{-1}$ for $z \neq (0,0)$, respectively, all the rules stated in Th. C.10 are valid in $\mathbb{C}$.

(c) The map

\[
\iota : \mathbb{R} \longrightarrow \mathbb{C}, \quad \iota(x) := (x, 0),
\]

is a monomorphism, i.e. it is injective and satisfies

\[
\begin{align*}
    \forall x, y \in \mathbb{R} \quad \iota(x + y) &= \iota(x) + \iota(y), \\
    \forall x, y \in \mathbb{R} \quad \iota(xy) &= \iota(x) \cdot \iota(y).
\end{align*}
\]

It is customary to identify $\mathbb{R}$ with $\iota(\mathbb{R})$, as it usually does not cause any confusion. One then just writes $x$ instead of $(x,0)$. 


Proof. All computations required for (a) and (c) are straightforward and are left as an exercise; (b) is a consequence of (a), since Th. C.10 is valid for every field.

Notation 5.3. The number $i := (0, 1)$ is called the imaginary unit (note that, indeed, $i^2 = i \cdot i = (0, 1) \cdot (0, 1) = (0 \cdot 0 - 1 \cdot 1, 0 \cdot 1 + 1 \cdot 0) = (-1, 0) = -1$). Using $i$, one obtains the commonly used representation of a complex number $z = (x, y) \in \mathbb{C}$:

$$z = (x, y) = x \cdot (1, 0) + y \cdot (0, 1) = x + iy,$$

where one calls $\text{Re} z := x$ the real part of $z$ and $\text{Im} z := y$ the imaginary part of $z$. Moreover, $z$ is called purely imaginary if, and only if, $\text{Re} z = 0$ (as a consequence of this convention, one has the (harmless) pathology that $0$ is both real and purely imaginary).

Remark 5.4. There does not exist a total order $\leq$ on $\mathbb{C}$ that makes $\mathbb{C}$ into a totally ordered field (i.e. no total order on $\mathbb{C}$ can be compatible with addition and multiplication in the sense of (4.3)): Indeed, if there were such a total order $\leq$ on $\mathbb{C}$, then all the rules of Th. 4.5 had to be valid with respect to that total order $\leq$. In particular, $0 < 1^2 = 1$ and $0 < i^2 = -1$ had to be valid by Th. 4.5(c), and, then, $0 < 1 + (-1) = 0$ had to be valid by Th. 4.5(f). However, $0 < 0$ is false, showing that there is no total order on $\mathbb{C}$ that satisfies (4.3). Caveat: Of course, there do exist total orders on $\mathbb{C}$, just none compatible with addition and multiplication – for example, the lexicographic order on $\mathbb{R} \times \mathbb{R}$ (defined as it was in (2.52) for $\mathbb{N} \times \mathbb{N}$) constitutes a total order on $\mathbb{C}$.

Definition and Remark 5.5. Conjugation: For each complex number $z = x + iy$, we define its complex conjugate or just conjugate to be the complex number $\bar{z} := x - iy$. We then have the following rules that hold for each $z = x + iy, w = u + iv \in \mathbb{C}$:

(a) $\bar{z} + \bar{w} = x+u-iy-iv = \bar{z} + \bar{w}$ and $\bar{z}w = xu-yv-(xv+yu)i = (x-iy)(u-iv) = \bar{z} \bar{w}$.

(b) $z + \bar{z} = 2x = 2 \text{Re} z$ and $z - \bar{z} = 2yi = 2i \text{Im} z$.

(c) $z = \bar{z} \iff x + iy = x - iy \iff y = 0 \iff z \in \mathbb{R}$.

(d) $z \bar{z} = (x + iy)(x - iy) = x^2 + y^2 \in \mathbb{R}_0^+$.

5.2 Sign and Absolute Value (Modulus)

We face a certain conundrum regarding the handling of square roots. The problem is that we will need the notion of a continuous function to prove the existence of a unique square root $\sqrt{x}$ for every nonnegative real number $x$ and, in consequence, we will have to wait until Section 7.2.5 below to carry out this proof. On the other hand, it is extremely desirable to present the theory of convergence simultaneously for real and for complex numbers, which requires the notion of the absolute value or modulus of a complex number, to be defined in Def. 5.7(b) below as the square root of a nonnegative real number.

Faced with this difficulty, we will introduce the notion of square root now, assuming the existence, until we can add the proof in Section 7.2.5. Some students might be worried
that this might lead to a circular argument, where our later proof of the existence of square roots would somehow make use of our previous assumption of that existence. Of course, we will be careful not to make such a circular (and, thereby, logically invalid) argument. The point is that for real numbers the notion of absolute value does in no way depend on the notion of a square root (see Lem. 5.8 below).

**Definition and Remark 5.6.** We define a nonnegative real number \( y \in \mathbb{R}_0^+ \) to be the square root of the nonnegative real number \( x \in \mathbb{R}_0^+ \) if, and only if, \( y^2 = x \). If \( y \) is the square root of \( x \), then one uses the notation \( \sqrt{x} := y \). We will see in Rem. and Def. 7.61 that every \( x \in \mathbb{R}_0^+ \) has a unique square root and that the function \( f : \mathbb{R}_0^+ \rightarrow \mathbb{R}_0^+ \), \( f(x) := \sqrt{x} \), is strictly increasing (in particular, injective).

**Definition 5.7.** (a) The sign function is defined by

\[
\text{sgn} : \mathbb{R} \longrightarrow \mathbb{R}, \quad \text{sgn}(x) := \begin{cases} 
1 & \text{for } x > 0, \\
0 & \text{for } x = 0, \\
-1 & \text{for } x < 0. 
\end{cases}
\]  

(5.8)

It is emphasized that the sign function is only defined for real numbers (cf. Rem. 5.4)!

(b) The absolute value or modulus function is defined by

\[
\text{abs} : \mathbb{C} \longrightarrow \mathbb{R}_0^+, \quad z = x + iy \mapsto |z| := \sqrt{\bar{z}z} = \sqrt{x^2 + y^2},
\]  

(5.9)

where the term absolute value is often preferred for real numbers \( z \in \mathbb{R} \) and the term modulus is often preferred if one also considers complex numbers \( z \notin \mathbb{R} \).

**Lemma 5.8.** For each \( x \in \mathbb{R} \), one has

\[
|x| = x \cdot \text{sgn}(x) = \begin{cases} 
x & \text{for } x \geq 0, \\
-x & \text{for } x < 0. 
\end{cases}
\]  

(5.10)

**Proof.** One has

\[
|x| = \sqrt{x^2} = \begin{cases} 
x & \text{for } x \geq 0, \\
-x & \text{for } x < 0, 
\end{cases}
\]  

(5.11)

as claimed.

**Theorem 5.9.** The following rules hold for each \( z, w \in \mathbb{C} \):

(a) \( z \neq 0 \Rightarrow |z| > 0 \).

(b) \( ||z|| = |z| \).

(c) \( |z| = |ar{z}| \).

(d) \( \max\{|\text{Re} z|, |\text{Im} z|\} \leq |z| \leq |\text{Re} z| + |\text{Im} z| \).
Inverse Triangle Inequality:
\[ |z + w| \leq |z| + |w|. \]  
\tag{5.12}

Proof. We carry out the proofs for \( z, w \in \mathbb{C} \). However, for \( z, w \in \mathbb{R} \), everything can easily be shown directly from (5.10), without making use of square roots.

Let \( z = x + iy \) with \( x, y \in \mathbb{R} \).

(a): If \( z \neq 0 \), then \( x \neq 0 \) or \( y \neq 0 \), i.e. \( x^2 > 0 \) or \( y^2 > 0 \) by Th. 4.5(c), implying \( x^2 + y^2 > 0 \) by Th. 4.5(f), i.e. \( |z| = \sqrt{x^2 + y^2} > 0 \).

(b): Since \( a := |z| \in \mathbb{R}^+ \), we have \( |a| = \sqrt{a^2} = a = |z| \).

(c): Since \( \bar{z} = x - iy \), we have \( |ar{z}| = \sqrt{x^2 + (-y)^2} = \sqrt{x^2 + y^2} = |z| \).

(d): It is \( x = \text{Re} z, y = \text{Im} z \). Let \( a := \max\{|x|, |y|\} \). As remarked in Def. and Rem. 5.6, the square root function is increasing and, thus, taking square roots in the chain of inequalities \( a^2 \leq x^2 + y^2 \leq (|x| + |y|)^2 \) implies \( a \leq |z| \leq |x| + |y| \) as claimed.

(e): As remarked in Def. and Rem. 5.6, the square root function is injective, and, thus, \( (e) \) follows from

\[ |zw|^2 = zw \overline{zw} \overset{\text{Def. and Rem. 5.5(a)}}{=} zw \bar{z} \bar{w} = z \bar{z} w \bar{w} = |z|^2 |w|^2. \]

(f): Let \( w = u + iv \) with \( u, v \in \mathbb{R} \). We first consider the special case \( z = 1 \). Applying the formula (5.4b) for the inverse to \( w \), one obtains

\[ |w^{-1}|^2 = \frac{u^2}{(u^2 + v^2)^2} + \frac{v^2}{(u^2 + v^2)^2} = \frac{1}{u^2 + v^2} = \left( |w|^{-1} \right)^2, \]

i.e. \( |w^{-1}| = |w|^{-1} \). Now (f) follows from (e): \( |\frac{z}{w}| = |zw^{-1}| = |z||w^{-1}| = |z||w|^{-1} = \frac{|z|}{|w|} \).

(g) follows from

\[ |z + w|^2 = (z + w)(\bar{z} + \bar{w}) = z\bar{z} + w\bar{z} + z\bar{w} + w\bar{w} \overset{\text{Def. and Rem. 5.5(b)}}{=} |z|^2 + 2 \text{Re}(z\bar{w}) + |w|^2 \overset{(d)}{\leq} |z|^2 + 2|z\bar{w}| + |w|^2 = (|z| + |w|)^2, \]

once again using that the square root function is increasing.

(h): Using (g), we obtain

\[ |z| = |z - w + w| \leq |z - w| + |w| \Rightarrow |z| - |w| \leq |z - w|, \]

\[ |w| = |w - z + z| \leq |z - w| + |z| \Rightarrow -(|z| - |w|) \leq |z - w|, \]

implying \( ||z| - |w|| \leq |z - w| \) by (5.10) (notice \( |z| - |w| \in \mathbb{R} \)).
Remark 5.10. Each complex number \((x, y) = x + iy\) can be visualized as a point in the so-called complex plane, where the horizontal \(x\)-axis represents real numbers and the vertical \(y\)-axis represents purely imaginary numbers. Then the addition of complex numbers is precisely the vector addition of 2-dimensional vectors in the complex plane, and conjugation is represented by reflection through the \(x\)-axis. Moreover, the modulus \(|z|\) of a complex number is precisely its distance from the origin \((0, 0)\), and \(|z - w|\) is the distance between the points \(z = (x, y)\) and \(w = (u, v)\) in the plane. Complex multiplication can also be interpreted geometrically in the plane: If \(\phi\) denotes the angle that the vector representing \(z = (x, y)\) forms with the \(x\)-axis, and, likewise, \(\psi\) denotes the angle that the vector representing \(w = (u, v)\) forms with the \(x\)-axis, then \(zw\) is the vector of length \(|zw|\) that forms the angle \(\phi + \psi\) with the \(x\)-axis (we will better understand this geometrical interpretation of complex multiplication later (see Def. and Rem. 8.29), when writing complex numbers in the polar form \(z = x + iy = |z|\exp(i\phi)\), making use of the exponential function \(\exp\)).

5.3 Sums and Products

Here we compile some important rules involving sums and products. We are mostly interested in applying them to real and complex numbers. However, most of the rules, without extra difficulty, can be proved to hold in more general structures. We will provide the more general statements, but the reader will not lose much by merely thinking of \(\mathbb{C}\) rather than a general ring in Th. 5.11, and of \(\mathbb{R}\) rather than a general totally ordered field in Th. 5.12. Some rules involve exponentiation as defined in (C.10a) of the Appendix (in particular, it is used that \(z^0 = 1\) for each \(z \in R\), where \(R\) is a ring with unity), and some proofs make use of the corresponding exponentiation rules of Th. C.6.

Theorem 5.11. Let \((R, +, \cdot)\) be a ring (cf. Def. C.7 of the Appendix).

(a) For each \(n \in \mathbb{N}\) and each \(\lambda, \mu, z_j, w_j \in R\), \(j \in \{1, \ldots, n\}\):

\[
\sum_{j=1}^{n} (\lambda z_j + \mu w_j) = \lambda \sum_{j=1}^{n} z_j + \mu \sum_{j=1}^{n} w_j.
\]

(b) If \(R\) is a ring with unity, then, for each \(n \in \mathbb{N}_0\) and each \(z \in R\):

\[
(1 - z)(1 + z + z^2 + \cdots + z^n) = (1 - z) \sum_{j=0}^{n} z^j = 1 - z^{n+1}.
\]

(c) If \(R\) is a commutative ring with unity, then, for each \(n \in \mathbb{N}_0\) and each \(z, w \in R\):

\[
w^{n+1} - z^{n+1} = (w - z) \sum_{j=0}^{n} z^j w^{n-j} = (w - z)(w^n + zw^{n-1} + \cdots + z^{n-1}w + z^n).
\]
Proof. In each case, the proof can be conducted by an easy induction. We carry out (c) and leave the other cases as exercises. For (c), the base case \((n = 0)\) is provided by the true statement \(w^{0+1} - z^{0+1} = w - z = (w - z)z^0 w^{0-0}\). For the induction step, one computes

\[
(w - z) \sum_{j=0}^{n+1} z^j w^{n+1-j} = (w - z) \left( z^{n+1} w^0 + \sum_{j=0}^{n} z^j w^{n+1-j} \right) = (w - z)z^{n+1} + (w - z) w \sum_{j=0}^{n} z^j w^{n-j}
\]

\[
\text{ind. hyp.} \quad (w - z)z^{n+1} + w(w^{n+1} - z^{n+1}) = w^{n+2} - z^{n+2},
\]

completing the induction. ■

Theorem 5.12. Let \((F, +, \cdot)\) be a totally ordered field (cf. Def. 4.3).

(a) For each \(n \in \mathbb{N}\) and each \(x_j, y_j \in F, j \in \{1, \ldots, n\}\):

\[
\left( \forall_{j \in \{1, \ldots, n\}} x_j \leq y_j \right) \Rightarrow \sum_{j=1}^{n} x_j \leq \sum_{j=1}^{n} y_j,
\]

where equality can only hold if \(x_j = y_j\) for each \(j \in \{1, \ldots, n\}\).

(b) For each \(n \in \mathbb{N}\) and each \(x_j, y_j \in F, j \in \{1, \ldots, n\}\):

\[
\left( \forall_{j \in \{1, \ldots, n\}} 0 < x_j \leq y_j \right) \Rightarrow \prod_{j=1}^{n} x_j \leq \prod_{j=1}^{n} y_j,
\]

where equality can only hold if \(x_j = y_j\) for each \(j \in \{1, \ldots, n\}\).

Proof. Both cases are proved by simple inductions, where Th. 4.5(f) is used in (a) and Th. 4.5(g) is used in (b). ■

Theorem 5.13. Triangle Inequality: For each \(n \in \mathbb{N}\) and each \(z_j \in \mathbb{C}, j \in \{1, \ldots, n\}\):

\[
\left| \sum_{j=1}^{n} z_j \right| \leq \sum_{j=1}^{n} |z_j|.
\]

Proof. Another very simple induction that is left to the reader. ■

5.4 Binomial Coefficients and Binomial Theorem

The goal in this section is to expand \((z + w)^n\) into a sum. This sum involves the so-called [binomial coefficients \(\binom{n}{k}\)], which are also useful in other contexts. To obtain an idea for
what to expect, let us compute the cases \( n = 0, 1, 2, 3 \): \((z + w)^0 = 1, (z + w)^1 = z + w, (z + w)^2 = z^2 + 2zw + w^2, (z + w)^3 = z^3 + 3z^2w + 3zw^2 + w^3\). One finds that the coefficients form what is known as Pascal’s triangle, which we write for \( n = 0, \ldots, 5 \):

\[
\begin{array}{cccccc}
n = 0 : & & & & & 1 \\
n = 1 : & & & 1 & & 1 \\
n = 2 : & & 1 & & 2 & & 1 \\
n = 3 : & & 1 & 3 & 3 & 1 \\
n = 4 : & & 1 & 4 & 6 & 4 & 1 \\
n = 5 : & & 1 & 5 & 10 & 10 & 5 & 1 \\
\end{array}
\] (5.14)

The entries of the \( n \)th row of Pascal’s triangle are denoted by \( \binom{n}{0}, \ldots, \binom{n}{n} \). One also observes that one obtains each entry of the \((n+1)\)st row, except the first and last entry, by adding the corresponding entries in row \( n \) to the left and to the right of the considered entry in row \( n + 1 \). The first and last entry of each row are always set to 1. This can be summarized as

\[
\forall n \in \mathbb{N}_0 \quad \left( \binom{n}{0} = \binom{n}{n} = 1, \quad \binom{n+1}{k} = \binom{n}{k-1} + \binom{n}{k} \text{ for } k \in \{1, \ldots, n\} \right). \quad (5.15)
\]

The following Def. 5.14 provides a different and more general definition of binomial coefficients. We will then prove in Prop. 5.15 that the binomial coefficients as defined in Def. 5.14 do, indeed, satisfy (5.15).

**Definition 5.14.** For each \( \alpha \in \mathbb{C} \) and each \( k \in \mathbb{N}_0 \), we define the binomial coefficient

\[
\binom{\alpha}{0} := 1, \quad \binom{\alpha}{k} := \prod_{j=1}^{k} \frac{\alpha + 1 - j}{j} = \frac{\alpha(\alpha-1) \cdots (\alpha-k+1)}{1 \cdot 2 \cdots k} \text{ for } k \in \mathbb{N}. \quad (5.16)
\]

**Proposition 5.15.** (a) For each \( \alpha \in \mathbb{C} \) and each \( k \in \mathbb{N} \):

\[
\binom{\alpha}{0} = 1, \quad \binom{\alpha + 1}{k} = \binom{\alpha}{k-1} + \binom{\alpha}{k}. \quad (5.17)
\]

(b) For each \( n \in \mathbb{N}_0 \):

\[
\binom{n}{n} = 1. \quad (5.18)
\]

The above statements include (5.15) as a special case.

**Proof.** (a): The first identity is part of the definition in (5.16). For the second identity, we first observe, for each \( k \in \mathbb{N} \),

\[
\binom{\alpha}{k} = \prod_{j=1}^{k} \frac{\alpha + 1 - j}{j} = \frac{\alpha + 1 - k}{k} \prod_{j=1}^{k-1} \frac{\alpha + 1 - j}{j} = \binom{\alpha}{k-1} \frac{\alpha + 1 - k}{k}, \quad (5.19)
\]
which implies
\[
\binom{\alpha}{k-1} + \binom{\alpha}{k} = \binom{\alpha}{k-1} \left(1 + \frac{\alpha + 1 - k}{k}\right) = \binom{\alpha}{k} \frac{\alpha + 1}{k} = \frac{\alpha + 1}{k} \prod_{j=1}^{k-1} \frac{\alpha + 1 - j}{j} = \frac{\alpha + 1}{k} \prod_{j=1}^{k} \frac{\alpha + 2 - j}{j} = \binom{\alpha + 1}{k}.
\] (5.20)

(b): \(\binom{0}{0} = 1\) according to (5.16). For \(n \in \mathbb{N}\), (5.18) is proved by induction. The base case \((n = 1)\) is provided by the true statement \(\binom{1}{1} = \frac{1+1-1}{1} = 1\). For the induction step, one computes
\[
\binom{n+1}{n+1} = \prod_{j=1}^{n+1} \frac{n+1+1-j}{j} = \frac{n+1}{n+1} \prod_{j=1}^{n} \frac{n+1-j}{j} = \binom{n}{n} \text{ ind. hyp.} = 1,
\] (5.21)
which completes the induction.

\textbf{Theorem 5.16 (Binomial Theorem).} Let \(R\) be a commutative ring with unity (cf. Def. C.7 of the Appendix – for us, \(R = \mathbb{C}\) is the most important example). For each \(z, w \in R\) and each \(n \in \mathbb{N}_0\), the following formula holds:
\[
(z + w)^n = \sum_{k=0}^{n} \binom{n}{k} z^{n-k} w^k = z^n + \binom{n}{1} z^{n-1} w + \cdots + \binom{n}{n-1} z w^{n-1} + w^n. \tag{5.22}
\]

\textit{Proof.} The proof is conducted via induction on \(n\). The base case \((n = 0)\) is provided by the correct statement \((z + w)^0 = 1 = \binom{0}{0} z^0 w^0\). For the induction step, we first observe
\[
(z + w)^{n+1} = (z + w)(z + w)^n = z (z + w)^n + w (z + w)^n. \tag{5.23}
\]
Using the induction hypothesis, we now further manipulate the two terms on the right-hand side of (5.23):
\[
z (z + w)^n \overset{\text{ind. hyp.}}{=} z \sum_{k=0}^{n} \binom{n}{k} z^{n-k} w^k = \sum_{k=0}^{n} \binom{n}{k} z^{n+1-k} w^k = \binom{n+1}{n+1} = 0 \sum_{k=0}^{n+1} \binom{n}{k} z^{n+1-k} w^k,
\] (5.24)
\[
w (z + w)^n \overset{\text{ind. hyp.}}{=} w \sum_{k=0}^{n} \binom{n}{k} z^{n-k} w^k = \sum_{k=0}^{n} \binom{n}{k} z^{n-k} w^{k+1} = \sum_{k=1}^{n+1} \binom{n}{k-1} z^{n+1-k} w^k.
\] (5.25)
Plugging (5.24) and (5.25) into (5.23) yields

\[(z + w)^{n+1} = \left(\binom{n}{0} z^{n+1} w^0 + \sum_{k=1}^{n+1} \binom{n}{k} \binom{n}{k-1} w^k \right)^{n+1} = \left(\binom{n+1}{0} z^{n+1} w^0 + \sum_{k=1}^{n+1} \binom{n+1}{k} w^k \right)^{n+1} = \sum_{k=0}^{n+1} \binom{n+1}{k} z^{n+1-k} w^k,\]

completing the induction. ■

The binomial theorem can now be used to infer a few more rules that hold for the binomial coefficients:

**Corollary 5.17.** One has the following identities:

\[
\forall n \in \mathbb{N}_0 \quad \sum_{k=0}^{n} \binom{n}{k} = \binom{n}{0} + \binom{n}{1} + \cdots + \binom{n}{n} = 2^n, \quad (5.27a)
\]
\[
\forall n \in \mathbb{N} \quad \sum_{k=0}^{n} \binom{n}{k} (-1)^k = \binom{n}{0} - \binom{n}{1} + \binom{n}{2} - \cdots + (-1)^n \binom{n}{n} = 0. \quad (5.27b)
\]

**Proof.** (5.27a) is just (5.22) with \(z = w = 1\); (5.27b) is just (5.22) with \(z = 1\) and \(w = -1\). ■

The formulas provided by the following proposition are also sometimes useful.

**Proposition 5.18.** (a) For each \(\alpha \in \mathbb{C}\) and each \(k \in \mathbb{N}_0\):

\[
\sum_{j=0}^{k} \binom{\alpha + j}{j} = \binom{\alpha}{0} + \binom{\alpha+1}{1} + \cdots + \binom{\alpha+k}{k} = \binom{\alpha+k+1}{k}. \quad (5.28)
\]

(b) For each \(n, k \in \mathbb{N}_0\) with \(k \leq n\):

\[
\binom{n}{k} = \frac{n!}{k!(n-k)!}. \quad (5.29)
\]

Moreover, for \(n \geq 1\), one has \(\binom{n}{k} = \# P_k(\{1, \ldots, n\})\), where

\[
P_k(A) := \{B \in P(A) : \#B = k\} \quad (5.30)
\]

denotes the set of all subsets of a set \(A\) that have precisely \(k\) elements.

(c) For each \(n, k \in \mathbb{N}_0\):

\[
\sum_{j=0}^{k} \binom{n+j}{n} = \binom{n}{n} + \binom{n+1}{n} + \cdots + \binom{n+k}{n} = \binom{n+k+1}{n+1}. \quad (5.31)
\]
Proof. The induction proofs of (a) and (b) are left as exercises. For (c), one computes
\[
\sum_{j=0}^{k} \binom{n+j}{j} = \sum_{j=0}^{k} \frac{(n+j)!}{n!(n+j-n)!} = \sum_{j=0}^{k} \binom{n+j}{j}
\]
\[
= \binom{n+k+1}{k} \quad \text{(5.29)}
\]
\[
= \binom{n+k+1}{n+1} (5.29)
\]
thereby establishing the case. 

6 Polynomials

6.1 Arithmetic of $K$-Valued Functions

Notation 6.1. We will write $K$ in situations, where we allow $K$ to be $\mathbb{R}$ or $\mathbb{C}$.

Notation 6.2. If $A$ is any nonempty set, then one can add and multiply arbitrary functions $f,g : A \rightarrow K$, and one can define several further operations to create new functions from $f$ and $g$:

\[
(f + g) : A \rightarrow K, \quad (f + g)(x) := f(x) + g(x),
\]
\[
(\lambda f) : A \rightarrow K, \quad (\lambda f)(x) := \lambda f(x) \quad \text{for each } \lambda \in K,
\]
\[
(fg) : A \rightarrow K, \quad (fg)(x) := f(x)g(x),
\]
\[
(f/g) : A \rightarrow K, \quad (f/g)(x) := f(x)/g(x) \quad (\text{assuming } g(x) \neq 0),
\]
\[
\text{Re} f : A \rightarrow \mathbb{R}, \quad \text{Re}(f)(x) := \text{Re}(f(x)),
\]
\[
\text{Im} f : A \rightarrow \mathbb{R}, \quad \text{Im}(f)(x) := \text{Im}(f(x)).
\]

For $K = \mathbb{R}$, we further define

\[
\max(f,g) : A \rightarrow \mathbb{R}, \quad \max(f,g)(x) := \max\{f(x), g(x)\},
\]
\[
\min(f,g) : A \rightarrow \mathbb{R}, \quad \min(f,g)(x) := \min\{f(x), g(x)\},
\]
\[
f^+ : A \rightarrow \mathbb{R}, \quad f^+ := \max(f,0),
\]
\[
f^- : A \rightarrow \mathbb{R}, \quad f^- := \max(-f,0).
\]

Finally, once again also allowing $K = \mathbb{C},$

\[
|f| : A \rightarrow \mathbb{R}, \quad |f|(x) := |f(x)|.
\]

One calls $f^+$ and $f^-$ the positive part and the negative part of $f$, respectively. For $\mathbb{R}$-valued functions $f$, we have

\[
|f| = f^+ + f^-.
\]
6.2 Polynomials

Definition 6.3. Let $n \in \mathbb{N}_0$. Each function from $K$ into $K$, $x \mapsto x^n$, is called a monomial. A function $P$ from $K$ into $K$ is called a polynomial if, and only if, it is a linear combination of monomials, i.e. if, and only if $P$ has the form

$$P : K \to K, \quad P(x) = \sum_{j=0}^{n} a_j x^j = a_0 + a_1 x + \cdots + a_n x^n, \quad a_j \in K.$$

(6.2)

The $a_j$ are called the coefficients of $P$. The largest number $d \leq n$ such that $a_d \neq 0$ is called the degree of $P$, denoted $\deg(P)$. If all coefficients are 0, then $P$ is called the zero polynomial; the degree of the zero polynomial is defined as $-1$ (in Th. 6.6(b) below, we will see that each polynomial of degree $n \in \mathbb{N}_0$ is uniquely determined by its coefficients $a_0, \ldots, a_n$ and vice versa).

Polynomials of degree $\leq 0$ are constant. Polynomials of degree $\leq 1$ have the form $P(x) = a + bx$ and are called affine functions (often they are also called linear functions, even though this is not really correct for $a \neq 0$, since every function $P$ that is linear (in the sense of linear algebra) must satisfy $P(0) = 0$). Polynomials of degree $\leq 2$ have the form $P(x) = a + bx + cx^2$ and are called quadratic functions.

Each $\xi \in K$ such that $P(\xi) = 0$ is called a zero or a root of $P$.

A rational function is a quotient $P/Q$ of two polynomials $P$ and $Q$.

Remark 6.4. Let $\lambda \in K$ and let $P, Q$ be polynomials. Then $\lambda P$, $P + Q$, and $PQ$ defined according to Not. 6.2 are polynomials as well. More precisely, if $\lambda = 0$ or $P \equiv 0$, then $\lambda P = 0$; if $P \equiv 0$, then $P + Q = Q$; if $Q \equiv 0$, then $P + Q = P$; if $P \equiv 0$ or $Q \equiv 0$, then $PQ = 0$. If $\lambda \neq 0$ and

$$P(x) = \sum_{j=0}^{n} a_j x^j, \quad Q(x) = \sum_{j=0}^{m} b_j x^j,$$

(6.3)

with $\deg(P) = n \geq 0$, $\deg(Q) = m \geq 0$, $n \geq m \geq 0$,

then, defining $b_j := 0$ for each $j \in \{m+1, \ldots, n\}$ in case $n > m$,

$$\begin{align*}
(\lambda P)(x) &= \sum_{j=0}^{n} (\lambda a_j) x^j, \quad \deg(\lambda P) = n, \\
(P + Q)(x) &= \sum_{j=0}^{n} (a_j + b_j) x^j, \quad \deg(P + Q) \leq n = \max\{m, n\}, \\
(PQ)(x) &= \sum_{j=0}^{m+n} c_j x^j, \quad \deg(PQ) = m + n,
\end{align*}$$

(6.4a, 6.4b, 6.4c)

where, setting $a_k := 0$ for each $k \in \{n+1, \ldots, m+n\}$ and $b_k := 0$ for each $k \in \{m+1, \ldots, m+n\}$,

$$c_j = \sum_{k=0}^{j} a_k b_{j-k}.$$

(6.4d)
Formula (6.4c) can be proved by induction on \( m = \deg(Q) \in \mathbb{N}_0 \) as follows: For \( m = 0 \), we compute

\[
(PQ)(x) = b_0 \sum_{j=0}^{n} a_j x^j = \sum_{j=0}^{n+0} b_0 a_j x^j,
\]
i.e. \( c_j = b_0 a_j = \sum_{k=0}^{j} a_k b_{j-k} \), which establishes the base case, remembering \( b_{j-k} = 0 \) for \( j > k \). For the induction step, we compute, for \( \deg(Q) = m + 1 \),

\[
(PQ)(x) = \sum_{j=0}^{n} a_j x^j \sum_{\alpha=0}^{m+1} b_{\alpha} x^\alpha = \sum_{j=0}^{n} a_j x^j \left( b_{m+1} x^{m+1} + \sum_{\alpha=0}^{m} b_{\alpha} x^\alpha \right)
\]

\[
= \sum_{j=0}^{m+n+1} a_j b_{m+1} x^{m+1+j} + \sum_{j=0}^{m+n} \left( \sum_{k=0}^{j} a_k b_{j-k} \right) x^j
\]

which completes the induction step. There is a notational issue in the second and third line in of the above computation, since, in both lines, the \( b_{m+1} \) in the first sum is the actual \( b_{m+1} \) from \( Q \), but \( b_{m+1} = 0 \) in the second sum in both lines, which is due to the induction hypothesis being applied for \( m < m+1 \). This is actually used when combining both sums in the last step, computing, for \( m + 1 \leq j \leq m + n \):

\[
a_j b_{m+1} x^{j+1} + a_{j-m-1} \cdot 0 \cdot x^j = a_{j-m-1} b_{m+1} x^j.
\]

For \( j = m + n + 1 \), one has \( \sum_{k=0}^{m+n+1} a_k b_{m+n+1-k} = a_n b_{m+1} \), since \( b_{m+n+1-k} = 0 \) for \( n > k \) and \( a_k = 0 \) for \( k > n \).

Finally, \( \deg(PQ) = m + n \) follows from \( c_{m+n} = a_m b_n \neq 0 \).

**Theorem 6.5. (a)** For each polynomial \( P \) given in the form of (6.3) and each \( \xi \in \mathbb{K} \), we have the identity

\[
P(x) = \sum_{j=0}^{n} b_j (x - \xi)^j,
\]

where

\[
\forall j \in \{0, \ldots, n\} \quad b_j = \sum_{k=j}^{n} a_k \binom{k}{j} \xi^{k-j}, \quad \text{in particular} \quad b_0 = P(\xi), \quad b_n = a_n.
\]

(b) If \( P \) is a polynomial with \( n := \deg(P) \geq 1 \), then, for each \( \xi \in \mathbb{K} \), there exists a polynomial \( Q \) with \( \deg(Q) = n - 1 \) such that

\[
P(x) = P(\xi) + (x - \xi) Q(x).
\]

In particular, if \( \xi \) is a zero of \( P \), then \( P(x) = (x - \xi) Q(x) \).
Proof. (a): For $\xi = 0$, there is nothing to prove. For $\xi \neq 0$, defining the auxiliary variable $\eta := x - \xi$, we obtain $x = \xi + \eta$ and

$$P(x) = \sum_{k=0}^{n} a_k (\xi + \eta)^k (5.22) = \sum_{k=0}^{n} \sum_{j=0}^{k} a_k \binom{k}{j} \xi^{k-j} \eta^j = \sum_{k=0}^{n} \sum_{j=k}^{n} a_k \binom{k}{j} \xi^{k-j} \eta^j,$$

which is (6.5).

(b): According to (a), we have

$$P(x) = P(\xi) + (x - \xi) Q(x), \quad \text{with} \quad Q(x) = \sum_{j=1}^{n} b_j (x - \xi)^{j-1} = \sum_{j=0}^{n} b_{j+1} (x - \xi)^j, \quad (6.9)$$

proving (b). \quad \blacksquare

**Theorem 6.6. (a)** If $P$ is a polynomial with $n := \deg(P) \geq 0$, then $P$ has at most $n$ zeros.

(b) Let $P, Q$ be polynomials as in (6.3) with $n = m$, $\deg(P) \leq n$, and $\deg(Q) \leq n$. If $P(x_j) = Q(x_j)$ at $n + 1$ distinct points $x_0, x_1, \ldots, x_n \in \mathbb{K}$, then $a_j = b_j$ for each $j \in \{0, \ldots, n\}$.

**Consequence 1:** If $P, Q$ with degree $\leq n$ agree at $n + 1$ distinct points, then $P = Q$.

**Consequence 2:** If we know $P = Q$, then they agree everywhere, in particular at $\max\{\deg(P), \deg(Q)\} + 1$ distinct points, which implies they have the same coefficients.

Proof. (a): For $n = 0$, $P$ is constant, but not the zero polynomial, i.e. $P \equiv a_0 \neq 0$ with no zeros as claimed. For $n \in \mathbb{N}$, the proof is conducted by induction. The base case ($n = 1$) is provided by the observation that $\deg(P) = 1$ implies $P$ is the affine function with $P(x) = a_0 + a_1 x$, $a_1 \neq 0$, i.e. $P$ has precisely one zero at $\xi = -a_0/a_1$. For the induction step, assume $\deg(P) = n + 1$. If $P$ has no zeros, then the assertion of (a) holds true. Otherwise, $P$ has at least one zero $\xi \in \mathbb{K}$, and, according to Th. 6.5(b), there exists a polynomial $Q$ such that $\deg(Q) = n$ and

$$P(x) = (x - \xi) Q(x). \quad (6.10)$$

From the induction hypothesis, we gather that $Q$ has at most $n$ zeros, i.e. (6.10) implies $P$ has at most $n + 1$ zeros, which completes the induction.

(b): If $P(x_j) = Q(x_j)$ at $n + 1$ distinct points $x_j$, then each of these points is a zero of $P - Q$. Thus $P - Q$ is a polynomial of degree $\leq n$ with at least $n + 1$ zeros. Then (a) implies $\deg(P - Q) = -1$, i.e. $P - Q$ is the zero polynomial, i.e. $a_j - b_j = 0$ for each $j \in \{0, \ldots, n\}$. \quad \blacksquare
\textbf{Remark 6.7.} Let $P$ be a polynomial with $n := \deg(P) \geq 0$. According to Th. 6.6(a), $P$ has at most $n$ zeros. Using Th. 6.5(b) for an induction shows there exists $k \in \{0, \ldots, n\}$ and a polynomial $Q$ of degree $n - k$ such that
\begin{equation}
P(x) = Q(x) \prod_{j=1}^{k}(x - \xi_j) = (x - \xi_1)(x - \xi_2) \cdots (x - \xi_k)Q(x), \tag{6.11a}
\end{equation}
where $Q$ does not have any zeros in $K$ and $\{\xi_1, \ldots, \xi_k\} = \{\xi \in K : P(\xi) = 0\}$ is the set of zeros of $P$. It can of course happen that $P$ does not have any zeros and $P = Q$ (no $\xi_j$ exist). It can also occur that some of the $\xi_j$ in (6.11a) are identical. Thus, we can rewrite (6.11a) as
\begin{equation}
P(x) = Q(x) \prod_{j=1}^{l}(x - \lambda_j)^{m_j} = (x - \lambda_1)^{m_1}(x - \lambda_2)^{m_2} \cdots (x - \lambda_l)^{m_l}Q(x), \tag{6.11b}
\end{equation}
where $\lambda_1, \ldots, \lambda_l$, $l \in \{0, \ldots, k\}$, are the distinct zeros of $P$, and $m_j \in \mathbb{N}$ with $\sum_{j=1}^{l} m_j = k$. Then $m_j$ is called the \textit{multiplicity} of the zero $\lambda_j$ of $P$.

\section{Limits and Convergence of Real and Complex Numbers}

\subsection{Sequences}

Recall from Def. 2.15(b) that a sequence in $K$ is a function $f : \mathbb{N} \to K$, in this context usually denoted as $f = (z_n)_{n \in \mathbb{N}}$ or $(z_1, z_2, \ldots)$ with $z_n := f(n)$. Sometimes the sequence also has the form $(z_n)_{n \in I}$, where $I \neq \emptyset$ is a countable index set (e.g. $I = \mathbb{N}_0$) different from $\mathbb{N}$ (in the context of convergence (see the following Def. 7.1), $I$ must be $\mathbb{N}$ or it must have the same cardinality as $\mathbb{N}$, i.e. finite $I$ are not permissible).

\textbf{Definition 7.1.} The sequence $(z_n)_{n \in \mathbb{N}}$ in $K$ is said to be \textit{convergent with limit} $z \in K$ if, and only if, for each $\epsilon > 0$, there exists an index $N \in \mathbb{N}$ such that $|z_n - z| < \epsilon$ for every index $n > N$. The notation for $(z_n)_{n \in \mathbb{N}}$ converging to $z$ is $\lim_{n \to \infty} z_n = z$ or $z_n \to z$ for $n \to \infty$. Thus, by definition,
\begin{equation}
\lim_{n \to \infty} z_n = z \iff \forall \epsilon \in \mathbb{R}^+ \exists N \in \mathbb{N} \forall n > N \: |z_n - z| < \epsilon. \tag{7.1}
\end{equation}
The sequence $(z_n)_{n \in \mathbb{N}}$ in $K$ is called \textit{divergent} if, and only if, it is not convergent.

\textbf{Example 7.2. (a)} For every constant sequence $(z_n)_{n \in \mathbb{N}} = (a)_{n \in \mathbb{N}}$ with $a \in K$, one has $\lim_{n \to \infty} z_n = \lim_{n \to \infty} a = a$: Since, for each $n \in \mathbb{N}$, $|z_n - a| = |a - a| = 0$, one can choose $N = 1$ for each $\epsilon > 0$.

(b) $\lim_{n \to \infty} \frac{1}{n+a} = 0$ for each $a \in \mathbb{C}$: Here $z_n := 1/(n+a)$ (if $n = -a$, then set $z_n := w$ with $w \in \mathbb{C}$ arbitrary). Given $\epsilon > 0$, choose an arbitrary $N \in \mathbb{N}$ with $N \geq \epsilon^{-1} + |a|$. Then, for each $n \geq N$, we compute $|n+a| = |n - (-a)| \geq |n - a| = n - |a| > N - |a| \geq \epsilon^{-1}$, and, thus, $|z_n| = |n + a|^{-1} < \epsilon$ as desired.
(c) \((-1)^n\) is not convergent: We have \(z_n = 1\) for each even \(n\) and \(z_n = -1\) for each odd \(n\). Thus, for each \(z \neq 1\) and each even \(n\), \(|z_n - z| = |1 - z| > |1 - z|/2 =: \epsilon > 0\), i.e. \(z\) is not a limit of \((z_n)\). However, \(z = 1\) is also not a limit of the sequence, since, for each odd \(n\), \(|z_n - 1| = |1 - 1| = 2 > 1 =: \epsilon > 0\), proving that the sequence has no limit.

**Theorem 7.3. (a)** Let \((z_n)\) be a sequence in \(\mathbb{C}\). Then \((z_n)\) is convergent in \(\mathbb{C}\) if, and only if, both \((\text{Re } z_n)\) and \((\text{Im } z_n)\) are convergent in \(\mathbb{R}\). Moreover, in that case,

\[
\lim_{n \to \infty} z_n = z \iff \lim_{n \to \infty} \text{Re } z_n = \text{Re } z \land \lim_{n \to \infty} \text{Im } z_n = \text{Im } z. \tag{7.2}
\]

(b) Let \((x_n)\) be a sequence in \(\mathbb{R}\) and \(z \in \mathbb{C}\). Then

\[
\lim_{n \to \infty} x_n = z \implies z \in \mathbb{R}. \tag{7.3}
\]

**Proof.** (a): Suppose \((z_n)\) converges to \(z \in \mathbb{C}\). Then, given \(\epsilon > 0\), there exists \(N \in \mathbb{N}\) such that, for each \(n > N\), \(|z_n - z| < \epsilon\). In consequence, for each \(n > N\),

\[
|\text{Re } z_n - \text{Re } z| = |\text{Re } (z_n - z)| \leq |z_n - z| < \epsilon, \tag{7.4}
\]

proving \(\lim_{n \to \infty} \text{Re } z_n = \text{Re } z\). The proof of \(\lim_{n \to \infty} \text{Im } z_n = \text{Im } z\) is completely analogous. Conversely, suppose there are \(x, y \in \mathbb{R}\) such that \(\lim_{n \to \infty} \text{Re } z_n = x\) and \(\lim_{n \to \infty} \text{Im } z_n = y\). Here we encounter, for the first time, what is sometimes called an \(\epsilon/2\) argument: Given \(\epsilon > 0\), there exists \(N \in \mathbb{N}\) such that, for each \(n > N\), \(|\text{Re } z_n - x| < \epsilon/2\) and \(|\text{Im } z_n - y| < \epsilon/2\), implying, for each \(n > N\),

\[
|z_n - (x + iy)| = |\text{Re } z_n + i \text{Im } z_n - (x + iy)|
\leq |\text{Re } z_n - x| + |i||\text{Im } z_n - y| < \epsilon/2 + \epsilon/2 = \epsilon, \tag{7.5}
\]

proving \(\lim_{n \to \infty} z_n = x + iy\).

(b) is a direct consequence of (a).

**Example 7.4. (a)** According to Th. 7.3(a), we have

\[
\lim_{n \to \infty} \left(\sqrt{2} + \frac{i}{n - 17}\right) = \sqrt{2} + 0i = \sqrt{2}. \tag{Ex. 7.2(a),(b)}
\]

(b) According to Th. 7.3(a) and Ex. 7.2(c), the sequence \(\left(\frac{1}{n} + (-1)^n i\right)\) is divergent.

Another important example relies on the following inequality:

**Proposition 7.5** (Bernoulli’s Inequality). For each \(n \in \mathbb{N}_0\) and each \(x \in [-1, \infty]\), we have

\[
(1 + x)^n \geq 1 + nx, \tag{7.6}
\]

with strict inequality whenever \(n > 1\) and \(x \neq 0\).
Proof. For \( n = 0 \), (7.6) reads \( 1 \geq 1 \), for \( n = 1 \), (7.6) reads \( 1 + x \geq 1 + x \), for \( n = 2 \), (7.6) reads \( (1 + x)^2 = 1 + 2x + x^2 \geq 1 + 2x \), all three statements being trivially true, in the case \( n = 2 \) with strict inequality for \( x \neq 0 \). We now proceed by induction for \( n \geq 2 \). For the induction step, one estimates

\[
(1 + x)^{n+1} = (1 + x)^n (1 + x) \quad \text{ind. hyp., } x \geq -1 \\
\geq (1 + nx)(1 + x) = 1 + (n + 1)x + nx^2
\]

with strict inequality for \( x \neq 0 \).

\[\square\]

Example 7.6. We have, for each \( q \in \mathbb{C} \),

\[
|q| < 1 \implies \lim_{n \to \infty} q^n = 0:
\]

(7.8)

For \( q = 0 \), there is nothing to prove. For \( 0 < |q| < 1 \), it is \( |q|^{-1} > 1 \), i.e. \( h := |q|^{-1} - 1 > 0 \). Thus, for each \( \epsilon > 0 \) and \( N \geq 1/(\epsilon h) \), we obtain

\[
n > N \implies |q|^{-n} = (1 + h)^n \geq 1 + nh > nh > 1/\epsilon \implies |q^n| = |q|^n < \epsilon.
\]

(7.9)

Definition 7.7. (a) Given \( z \in \mathbb{K} \) and \( \epsilon \in \mathbb{R}^+ \), we call the set \( B_\epsilon(z) := \{w \in \mathbb{K} : |w - z| < \epsilon\} \) the \( \epsilon \)-neighborhood of \( z \) or, in anticipation of Analysis II, the (open) \( \epsilon \)-ball with center \( z \) (in fact, for \( \mathbb{K} = \mathbb{C} \), \( B_\epsilon(z) \) represents an open disk in the complex plane with center \( z \) and radius \( \epsilon \), whereas, for \( \mathbb{K} = \mathbb{R} \), \( B_\epsilon(z) = ]z - \epsilon, z + \epsilon[ \) is the open interval with center \( z \) and length \( 2\epsilon \)). More generally, a set \( U \subseteq \mathbb{K} \) is called a neighborhood of \( z \) if, and only if, there exists \( \epsilon > 0 \) with \( B_\epsilon(z) \subseteq U \) (so, for example, for \( \epsilon > 0 \), \( B_\epsilon(z) \) is always a neighborhood of \( z \), whereas \( \mathbb{R} \) and \( [z - \epsilon, \infty[ \) are neighborhoods of \( z \) for \( \mathbb{K} = \mathbb{R} \), but not for \( \mathbb{K} = \mathbb{C} \) (\( ]z - \epsilon, \infty[ \) not even being defined for \( z \notin \mathbb{R} \)); the sets \( \{z\}, \{w \in \mathbb{K} : \text{Re } w \geq \text{Re } z\}, \{w \in \mathbb{K} : \text{Re } w \geq \text{Re } z + \epsilon\} \) are never neighborhoods of \( z \).

(b) If \( \phi(n) \) is a statement for each \( n \in \mathbb{N} \), then \( \phi(n) \) is said to be true for almost all \( n \in \mathbb{N} \) if, and only if, there exists a finite subset \( A \subseteq \mathbb{N} \) such that \( \phi(n) \) is true for each \( n \in \mathbb{N} \setminus A \), i.e. if, and only if, \( \phi(n) \) is always true, with the possible exception of finitely many cases.

Remark 7.8. In the language of Def. 7.7, the sequence \( (z_n)_{n \in \mathbb{N}} \) converges to \( z \) if, and only if, every neighborhood of \( z \) contains almost all \( z_n \).

Definition 7.9. The sequence \( (z_n)_{n \in \mathbb{N}} \) in \( \mathbb{K} \) is called bounded if, and only if, the set \( \{ |z_n| : n \in \mathbb{N} \} \) is bounded in the sense of Def. 2.27(a).

Proposition 7.10. Let \( (z_n)_{n \in \mathbb{N}} \) be a sequence in \( \mathbb{K} \).

(a) Limits are unique, that means if \( z, w \in \mathbb{K} \) such that \( \lim_{n \to \infty} z_n = z \) and \( \lim_{n \to \infty} z_n = w \), then \( z = w \).

(b) If \( (z_n)_{n \in \mathbb{N}} \) is convergent, then it is bounded.
Proof. (a): Exercise.

(b): If \( \lim_{n \to \infty} z_n = z \), then \( A := \{|z_n| : |z_n - z| \geq 1\} \cup \{|z_1|\} \) is nonempty and finite. According to Th. 3.13(a), \( A \) has an upper bound \( M \). Then \( \max\{|z| + 1\} \) is an upper bound for \( \{|z_n| : n \in \mathbb{N}\} \), and 0 is always a lower bound, showing that the sequence is bounded.

\[ \boxed{\text{Proposition 7.11. Let } (z_n)_{n \in \mathbb{N}} \text{ be a sequence in } \mathbb{C} \text{ with } \lim_{n \to \infty} z_n = 0.} \]

(a) If \((b_n)_{n \in \mathbb{N}}\) is a sequences in \( \mathbb{C} \) such that there exists \( C \in \mathbb{R}^+ \) with \( |b_n| \leq C|z_n| \) for almost all \( n \), then \( \lim_{n \to \infty} b_n = 0 \).

(b) If \((c_n)_{n \in \mathbb{N}}\) is a bounded sequence in \( \mathbb{C} \), then \( \lim_{n \to \infty} (c_n z_n) = 0 \).

Proof. (a): Given \( \epsilon > 0 \), there exists \( N \in \mathbb{N} \) such that \( |z_n| < \epsilon / C \) and \( |b_n| \leq C|z_n| \) for each \( n > N \). Then, for each \( n > N \), \( |b_n| \leq C|z_n| < \epsilon \), proving \( \lim_{n \to \infty} b_n = 0 \).

(b): If \((c_n)_{n \in \mathbb{N}}\) is bounded, then there exists \( C \in \mathbb{R}^+ \) such that \( |c_n| \leq C \) for each \( n \in \mathbb{N} \). Thus, \( |c_n z_n| \leq C |z_n| \) for each \( n \in \mathbb{N} \), implying \( \lim_{n \to \infty} (c_n z_n) = 0 \) via (a).

\[ \boxed{\text{Example 7.12. The sequences } ((-1)^n)_{n \in \mathbb{N}} \text{ and } (b)_{n \in \mathbb{N}} \text{ with } b \in \mathbb{C} \text{ are bounded. Since, for each } a \in \mathbb{C}, \lim_{n \to \infty} \frac{1}{n+a} = 0 \text{ by Example 7.2(b)}, \text{ we obtain}} \]

\[ \lim_{n \to \infty} \frac{(-1)^n}{n+a} = \lim_{n \to \infty} \frac{b}{n+a} = 0 \quad (7.10) \]

from Prop. 7.11(b).

\[ \boxed{\text{Theorem 7.13. (a) Let } (z_n)_{n \in \mathbb{N}} \text{ and } (w_n)_{n \in \mathbb{N}} \text{ be sequences in } \mathbb{C}. \text{ Moreover, let } z, w \in \mathbb{C} \text{ with } \lim_{n \to \infty} z_n = z \text{ and } \lim_{n \to \infty} w_n = w. \text{ We have the following identities:}} \]

\[ \lim_{n \to \infty} (\lambda z_n) = \lambda z \quad \text{for each } \lambda \in \mathbb{C}, \quad (7.11a) \]

\[ \lim_{n \to \infty} (z_n + w_n) = z + w, \quad (7.11b) \]

\[ \lim_{n \to \infty} (z_n w_n) = zw, \quad (7.11c) \]

\[ \lim_{n \to \infty} \frac{z_n}{w_n} = \frac{z}{w} \quad \text{given all } w_n \neq 0 \text{ and } w \neq 0, \quad (7.11d) \]

\[ \lim_{n \to \infty} |z_n| = |z|, \quad (7.11e) \]

\[ \lim_{n \to \infty} \bar{z}_n = \bar{z}, \quad (7.11f) \]

\[ \lim_{n \to \infty} z_n^p = z^p \quad \text{for each } p \in \mathbb{N}. \quad (7.11g) \]

(b) Let \((x_n)_{n \in \mathbb{N}} \text{ and } (y_n)_{n \in \mathbb{N}} \text{ be sequences in } \mathbb{R}. \text{ Moreover, let } x, y \in \mathbb{R} \text{ with } \lim_{n \to \infty} x_n = x \text{ and } \lim_{n \to \infty} y_n = y. \text{ Then}} \]

\[ \lim_{n \to \infty} \max\{x_n, y_n\} = \max\{x, y\}, \quad (7.12a) \]

\[ \lim_{n \to \infty} \min\{x_n, y_n\} = \min\{x, y\}. \quad (7.12b) \]
(c) If, in the situation of (b) (i.e. for real sequences), \( x_n \leq y_n \) holds for almost all \( n \in \mathbb{N} \), then \( x \leq y \). In particular, if almost all \( x_n \geq 0 \), then \( x \geq 0 \).

**Proof.** We start with the identities of (a).

(7.11a): For \( \lambda = 0 \), there is nothing to prove. For \( \lambda \neq 0 \) and \( \epsilon > 0 \), there exists \( N \in \mathbb{N} \) such that, for each \( n > N \), \( |z_n - z| < \epsilon/|\lambda| \), implying

\[
\forall \ n > N \ |\lambda z_n - \lambda z| = |\lambda| |z_n - z| < \epsilon. \tag{7.13a}
\]

(7.11b): Given \( \epsilon > 0 \), there exists \( N \in \mathbb{N} \) such that, for each \( n > N \), \( |z_n - z| < \epsilon/2 \) and \( |w_n - w| < \epsilon/2 \), implying

\[
\forall \ n > N \ |z_n + w_n - (z + w)| \leq |z_n - z| + |w_n - w| < \epsilon/2 + \epsilon/2 = \epsilon. \tag{7.13b}
\]

(7.11c): Let \( M_1 := \max\{|z|, 1\} \). According to Prop. 7.10(b), there exists \( M_2 \in \mathbb{R}^+ \) such that \( M_2 \) is an upper bound for \( \{|w_n| : n \in \mathbb{N}\} \). Moreover, given \( \epsilon > 0 \), there exists \( N \in \mathbb{N} \) such that, for each \( n > N \), \( |z_n - z| < \epsilon/(2M_2) \) and \( |w_n - w| < \epsilon/(2M_1) \), implying

\[
\forall \ n > N \ \left| z_n w_n - zw \right| = \left| (z_n - z)w_n + z(w_n - w) \right| \leq |w_n| \cdot |z_n - z| + |z| \cdot |w_n - w| < \frac{M_2 \epsilon}{2M_2} + \frac{M_1 \epsilon}{2M_1} = \epsilon. \tag{7.13c}
\]

(7.11d): We first consider the case, where all \( z_n = 1 \). Given \( \epsilon > 0 \), there exists \( N \in \mathbb{N} \) such that, for each \( n > N \), \( |w_n - w| < \epsilon |w|^2/2 \) and \( |w_n - w| < |w|/2 \) (since \( w \neq 0 \) for this case), implying \( |w| \leq |w - w_n| + |w_n| < |w|/2 + |w_n| \) and \( |w_n| > |w|/2 \). Thus,

\[
\forall \ n > N \ \left| \frac{1}{w_n} - \frac{1}{w} \right| = \left| \frac{w_n - w}{w_n w} \right| \leq \frac{2 |w_n - w|}{|w|^2} < \frac{2 \epsilon |w|^2}{2} = \epsilon. \tag{7.13d}
\]

The general case now follows from (7.11c).

(7.11e): This is a consequence of the inverse triangle inequality (5.13): Given \( \epsilon > 0 \), there exists \( N \in \mathbb{N} \) such that, for each \( n > N \), \( |z_n - z| < \epsilon \), implying

\[
\forall \ n > N \ |z_n| - |z| \leq |z_n - z| < \epsilon. \tag{7.13e}
\]

(7.11f): Write \( z_n = x_n + iy_n \) and \( z = x + iy \) with \( x_n, y_n, x, y \in \mathbb{R}, n \in \mathbb{N} \). Then we know \( \lim_{n \to \infty} x_n = x \) and \( \lim_{n \to \infty} y_n = y \) from (7.2), and

\[
\lim_{n \to \infty} z_n = \lim_{n \to \infty} (x_n - iy_n) \quad (7.11a),(7.11b) \Rightarrow x - iy = z, \tag{7.13f}
\]

which establishes the case.

(7.11g) follows by induction from (7.11c) (cf. (7.16b) below).

The proofs for the two identities of (b) are left as exercises.

(c): Proceeding by contraposition, assume \( x > y \) and set \( s := (x+y)/2 \). Then \( y < s < x \) and \( y_n < s < x_n \) holds for almost all \( n \), i.e. \( x_n \leq y_n \) does not hold for almost all \( n \). \( \blacksquare \)
Example 7.14. (a) $\lim_{n \to \infty} \frac{n+a}{n+b} = 1$ for each $a, b \in \mathbb{C}$: Here $z_n := (n+a)/(n+b)$ (if $n = -b$, then set $z_n := w$ with $w \in \mathbb{C}$ arbitrary). Using (7.11b) and (7.11d), one obtains
\[
\lim_{n \to \infty} \frac{n+a}{n+b} = \lim_{n \to \infty} \frac{1+a/n}{1+b/n} = \lim_{n \to \infty} 1 + \lim_{n \to \infty} \frac{a}{1+b/n} = 1 + 0 = 1.
\] (7.14)

(b) Using (7.11b), (7.11d), and (7.11g), one obtains
\[
\lim_{n \to \infty} \frac{2n^5 - 3in^3 + 2i}{3n^5 + 17n} = \lim_{n \to \infty} \frac{2 - 3i/n^2 + 2i/n^5}{3 + 17/n^4} = \frac{2 + 0 + 0}{3 + 0} = \frac{2}{3}.
\] (7.15)

Corollary 7.15. For $k \in \mathbb{N}$, let $(z^{(1)}_n)_{n \in \mathbb{N}}, \ldots, (z^{(k)}_n)_{n \in \mathbb{N}}$ be sequences in $\mathbb{C}$. Moreover, let $z^{(1)}, \ldots, z^{(k)} \in \mathbb{C}$ with $\lim_{n \to \infty} z^{(j)}_n = z^{(j)}$ for each $j \in \{1, \ldots, k\}$. Then
\[
\lim_{n \to \infty} \sum_{j=1}^{k} z^{(j)}_n = \sum_{j=1}^{k} z^{(j)},
\] (7.16a)
\[
\lim_{n \to \infty} \prod_{j=1}^{k} z^{(j)}_n = \prod_{j=1}^{k} z^{(j)}.
\] (7.16b)

Proof. (7.16) follows by simple inductions from (7.11b) and (7.11c), respectively. ■

Theorem 7.16 (Sandwich Theorem). Let $(x_n)_{n \in \mathbb{N}}, (y_n)_{n \in \mathbb{N}},$ and $(a_n)_{n \in \mathbb{N}}$ be sequences in $\mathbb{R}$. If $x_n \leq a_n \leq y_n$ holds for almost all $n \in \mathbb{N}$, then
\[
\lim_{n \to \infty} x_n = \lim_{n \to \infty} y_n = x \in \mathbb{R} \implies \lim_{n \to \infty} a_n = x.
\] (7.17)

Proof. Given $\epsilon > 0$, there exists $N \in \mathbb{N}$ such that, for each $n > N$, $x_n \leq a_n \leq y_n$, $|x_n - x| < \epsilon$, and $|y_n - x| < \epsilon$, implying
\[
\forall_{n > N} x - \epsilon < x_n \leq a_n \leq y_n < x + \epsilon,
\] (7.18)
which establishes the case. ■

Example 7.17. Since, $0 < \frac{1}{n!} \leq \frac{1}{n}$ holds for each $n \in \mathbb{N}$, the Sandwich Th. 7.16 implies
\[
\lim_{n \to \infty} \frac{1}{n!} = 0.
\] (7.19)

Definition 7.18. Let $(x_n)_{n \in \mathbb{N}}$ be a sequence in $\mathbb{R}$. The sequence is said to diverge to $\infty$ (resp. to $-\infty$), denoted $\lim_{n \to \infty} x_n = \infty$ (resp. $\lim_{n \to \infty} x_n = -\infty$) if, and only if, for each $K \in \mathbb{R}$, almost all $x_n$ are bigger (resp. smaller) than $K$. Thus,
\[
\lim_{n \to \infty} x_n = \infty \iff \forall_{K \in \mathbb{R}} \exists_{N \in \mathbb{N}} \forall_{n > N} x_n > K,
\] (7.20a)
\[
\lim_{n \to \infty} x_n = -\infty \iff \forall_{K \in \mathbb{R}} \exists_{N \in \mathbb{N}} \forall_{n > N} x_n < K.
\] (7.20b)
Theorem 7.19. Suppose $S := (x_n)_{n \in \mathbb{N}}$ is a monotone sequence in $\mathbb{R}$ (increasing or decreasing). Defining $A := \{x_n : n \in \mathbb{N}\}$, the following holds:

$$
\lim_{n \to \infty} x_n = \begin{cases} 
\sup A & \text{if } S \text{ is increasing and bounded}, \\
\infty & \text{if } S \text{ is increasing and not bounded}, \\
\inf A & \text{if } S \text{ is decreasing and bounded}, \\
-\infty & \text{if } S \text{ is decreasing and not bounded}.
\end{cases}
$$

(7.21)

Proof. We treat the increasing case; the decreasing case is proved completely analogously. If $A$ is bounded and $\epsilon > 0$, let $K := \sup A - \epsilon$; if $A$ is unbounded, then let $K \in \mathbb{R}$ be arbitrary. In both cases, since $K$ cannot be an upper bound, there exists $N \in \mathbb{N}$ such that $x_N > K$. Since the sequence is increasing, for each $n > N$, $x_N \leq x_n$, showing $|\sup A - x_n| < \epsilon$ in the bounded case, and $x_n > K$ in the unbounded case. \(\square\)

Example 7.20. Theorem 7.19 implies

$$
\forall k \in \mathbb{N} \quad \left( \lim_{n \to \infty} n^k = \infty, \quad \lim_{n \to \infty} (-n^k) = -\infty \right).
$$

(7.22)

It is sometimes necessary to consider so-called subsequences and reorderings of a given sequence. Here, we are interested in sequences in $\mathbb{R}$ or $\mathbb{C}$, but for subsequences and reorderings it is irrelevant in which set $A$ the sequence takes its values. As it presents virtually no extra difficulty to introduce the notions for general sequences, and since we will need to consider sequences with values in sets other than $\mathbb{R}$ or $\mathbb{C}$ in Analysis II, we admit general sequences in the following definition.

Definition 7.21. Let $A$ be an arbitrary nonempty set. Consider a sequence $\sigma : \mathbb{N} \to A$. Given a function $\phi : \mathbb{N} \to \mathbb{N}$ (that means $(\phi(n))_{n \in \mathbb{N}}$ constitutes a sequence of indices), the new sequence $(\sigma \circ \phi) : \mathbb{N} \to A$ is called a subsequence of $\sigma$ if, and only if, $\phi$ is strictly increasing (i.e. $1 \leq \phi(1) < \phi(2) < \ldots$). Moreover, $\sigma \circ \phi$ is called a reordering of $\sigma$ if, and only if, $\phi$ is bijective. One can write $\sigma$ in the form $(z_n)_{n \in \mathbb{N}}$ by setting $z_n := \sigma(n)$, and one can write $\sigma \circ \phi$ in the form $(w_n)_{n \in \mathbb{N}}$ by setting $w_n := (\sigma \circ \phi)(n) = z_{\phi(n)}$. Especially for a subsequence of $(z_n)_{n \in \mathbb{N}}$, it is also common to write $(z_{n_k})_{k \in \mathbb{N}}$. This notation corresponds to the one above if one lets $n_k := \phi(k)$. Analogous definitions work if the index set $\mathbb{N}$ of $\sigma$ is replaced by a general countable nonempty index set $I$.

Example 7.22. Consider the sequence $(1, 2, 3, \ldots)$. Then $(2, 4, 6, \ldots)$ constitutes a subsequence and $(2, 1, 4, 3, 6, 5, \ldots)$ constitutes a reordering. Using the notation of Def. 7.21, the original sequence is given by $\sigma : \mathbb{N} \to \mathbb{N}, \sigma(n) := n$; the subsequence is selected via $\phi_1 : \mathbb{N} \to \mathbb{N}, \phi_1(n) := 2n$; and the reordering is accomplished via $\phi_2 : \mathbb{N} \to \mathbb{N}, \phi_2(n) := \begin{cases} 
n + 1 & \text{if } n \text{ is odd}, \\
n - 1 & \text{if } n \text{ is even}.
\end{cases}$

Proposition 7.23. Let $(z_n)_{n \in \mathbb{N}}$ be a sequence in $\mathbb{C}$. If $\lim_{n \to \infty} z_n = z$, then every subsequence and every reordering of $(z_n)_{n \in \mathbb{N}}$ is also convergent with limit $z$. 

Proof. Let \((w_n)_{n \in \mathbb{N}}\) be a subsequence of \((z_n)_{n \in \mathbb{N}}\), i.e. there is a strictly increasing function \(\phi : \mathbb{N} \rightarrow \mathbb{N}\) such that \(w_n = z_{\phi(n)}\). If \(\lim_{n \rightarrow \infty} z_n = z\), then, given \(\epsilon > 0\), there is \(N \in \mathbb{N}\) such that \(z_n \in B_\epsilon(z)\) for each \(n > N\). For \(N\) choose any number from \(\mathbb{N}\) that is \(\geq N\) and in \(\phi(\mathbb{N})\). Take \(M := \phi^{-1}(N)\) (where \(\phi^{-1} : \phi(\mathbb{N}) \rightarrow \mathbb{N}\)). Then, for each \(n > M\), one has \(\phi(n) > N \geq N\), and, thus, \(w_n = z_{\phi(n)} \in B_\epsilon(z)\), showing \(\lim_{n \rightarrow \infty} w_n = z\).

Let \((w_n)_{n \in \mathbb{N}}\) be a reordering of \((z_n)_{n \in \mathbb{N}}\), i.e. there is a bijective function \(\phi : \mathbb{N} \rightarrow \mathbb{N}\) such that \(w_n = z_{\phi(n)}\). Let \(\epsilon\) and \(N\) be as before. Define

\[
M := \max\{\phi^{-1}(n) : n \leq N\}. \tag{7.23}
\]

As \(\phi\) is bijective, it is \(\phi(n) > N\) for each \(n > M\). Then, for each \(n > M\), one has \(w_n = z_{\phi(n)} \in B_\epsilon(z)\), showing \(\lim_{n \rightarrow \infty} w_n = z\). \(\blacksquare\)

Definition 7.24. Let \((z_n)_{n \in \mathbb{N}}\) be a sequence in \(\mathbb{K}\). A point \(z \in \mathbb{K}\) is called a cluster point or an accumulation point of the sequence if, and only if, for each \(\epsilon > 0\), \(B_\epsilon(z)\) contains infinitely many members of the sequence (i.e. \(#\{n \in \mathbb{N} : z_n \in B_\epsilon(z)\} = \infty\)).

Example 7.25. The sequence \((-1)^n)_{n \in \mathbb{N}}\) has cluster points 1 and \(-1\).

Proposition 7.26. A point \(z \in \mathbb{K}\) is a cluster point of the sequence \((z_n)_{n \in \mathbb{N}}\) in \(\mathbb{K}\) if, and only if, the sequence has a subsequence converging to \(z\).

Proof. If \((w_n)_{n \in \mathbb{N}}\) is a subsequence of \((z_n)_{n \in \mathbb{N}}\), \(\lim_{n \rightarrow \infty} w_n = z\), then every \(B_\epsilon(z)\), \(\epsilon > 0\), contains infinitely many \(w_n\), i.e. infinitely many \(z_n\), i.e. \(z\) is a cluster point of \((z_n)_{n \in \mathbb{N}}\). Conversely, if \(z\) is a cluster point of \((z_n)_{n \in \mathbb{N}}\), then, inductively, define \(\phi : \mathbb{N} \rightarrow \mathbb{N}\) as follows: For \(\phi(1)\), choose the index \(k\) of any point \(z_k \in B_1(z)\) (such a point exists, since \(z\) is a cluster point of the sequence). Now assume that \(n > 1\) and that \(\phi(m)\) have already been defined for each \(m < n\). Let \(M := \max\{\phi(m) : m < n\}\). Since \(B_{1/n}(z)\) contains infinitely many \(z_k\), there must be some \(z_k \in B_{1/n}(z)\) such that \(k > M\). Choose this \(k\) as \(\phi(n)\). Thus, by construction, \(\phi\) is strictly increasing, i.e. \((w_n)_{n \in \mathbb{N}}\) with \(w_n := z_{\phi(n)}\) is a subsequence of \((z_n)_{n \in \mathbb{N}}\). Moreover, for each \(\epsilon > 0\), there is \(N \in \mathbb{N}\) such that \(1/N < \epsilon\). Then, for each \(n > N\), \(w_n \in B_{1/n}(z) \subseteq B_{1/\sqrt{n}}(z) \subseteq B_\epsilon(z)\), showing \(\lim_{n \rightarrow \infty} w_n = z\). \(\blacksquare\)

Theorem 7.27 (Bolzano-Weierstrass). Every bounded sequence \(S := (x_n)_{n \in \mathbb{N}}\) in \(\mathbb{K}\) has at least one cluster point in \(\mathbb{K}\). Moreover, for \(\mathbb{K} = \mathbb{R}\), the set \(A := \{x \in \mathbb{R} : x\) is cluster point of \(S\}\) has a max \(x^* \in \mathbb{R}\) and a min \(x_* \in \mathbb{R}\), i.e. every bounded sequence in \(\mathbb{R}\) has a largest and a smallest cluster point. In addition, for each \(\epsilon > 0\), the inequality \(x_* - \epsilon < x_n < x^* + \epsilon\) holds for almost all \(n\).

Proof. We first consider the case \(\mathbb{K} = \mathbb{R}\). Define

\[
A^* := \{x \in \mathbb{R} : x_n \leq x \text{ for almost all } n\}, \tag{7.24a}
\]

\[
A_* := \{x \in \mathbb{R} : x_n \geq x \text{ for almost all } n\}. \tag{7.24b}
\]

We claim \(A^* \neq \emptyset\) is bounded from below and \(x^* = \max A = \inf A^*\); \(A_* \neq \emptyset\) is bounded from above and \(x_* = \min A = \sup A_*\). We prove the claim for \(A^*\) – the proof for \(A_*\) is
conducted completely analogous. Let \( m, M \in \mathbb{R} \) be a lower and an upper bound for \( S \), respectively. Then \( M \in A^* \), showing \( A^* \neq \emptyset \); and \( m \) is a lower bound for \( A^* \). Since \( A^* \) is bounded from below, \( a := \inf A^* \in \mathbb{R} \) by the completeness of \( \mathbb{R} \). Moreover, for each \( \epsilon > 0 \), \( a - \epsilon \notin A^* \), as \( a \) is a lower bound for \( A^* \), i.e. \( x_n > a - \epsilon \) holds for infinitely many \( n \in \mathbb{N} \). On the other hand, \( a + \epsilon/2 \in A^* \) follows from \( a \) being the largest lower bound of \( A^* \), i.e. \( x_n > a + \epsilon/2 \) holds for only finitely many \( n \) (if any). In particular, we have shown \( x_n < a + \epsilon \) holds for almost all \( n \), and \( a - \epsilon < x_n < a + \epsilon \) must hold for infinitely many \( n \), showing \( a \) is a cluster point of \( S \). To see that \( a \) is the largest cluster point of \( S \) (i.e. \( a = \max A \)), we have to show that \( x > a \) implies \( x \) is not a cluster point of \( S \). However, letting \( \epsilon := x - a > 0 \), we had seen above that \( x_n > a + \epsilon/2 \) holds for only finitely many \( n \), i.e. \( B_{\epsilon/2}(x) \) contains only finitely many \( x_n \), showing \( x \) is not a cluster point of \( S \).

It now remains to consider the complex case, i.e. a bounded sequence \( S := (z_n)_{n \in \mathbb{N}} \) in \( \mathbb{C} \). For each \( n \in \mathbb{N} \), let \( z_n = x_n + iy_n \) with \( x_n, y_n \in \mathbb{R} \). Due to Th. 5.9(d), we have \( |x_n| \leq |z_n| \) and \( |y_n| \leq |z_n| \), i.e. the boundedness of \( S \) implies the boundedness of both \( (x_n)_{n \in \mathbb{N}} \) and \( (y_n)_{n \in \mathbb{N}} \). Then we know that \( (x_n)_{n \in \mathbb{N}} \) has a cluster point \( x \) and, by Prop. 7.26, \( S \) has a subsequence \( (z_{n_j})_{j \in \mathbb{N}} \) such that \( x = \lim_{j \to \infty} x_{n_j} \). As the subsequence \( (y_{n_j})_{j \in \mathbb{N}} \) is still bounded, it must have a cluster point \( y \) and a subsequence \( (y_{n_{j_k}})_{k \in \mathbb{N}} \) such that \( y = \lim_{k \to \infty} y_{n_{j_k}} \). Since \( x = \lim_{k \to \infty} x_{n_{j_k}} \) as well, we now have \( \lim_{k \to \infty} z_{n_{j_k}} = x + iy =: z \), i.e. \( S \) has a subsequence converging to \( z \). According to Prop. 7.26, \( z \) is a cluster point of \( S \).

**Definition 7.28.** A sequence \((z_n)_{n \in \mathbb{N}}\) in \( \mathbb{C} \) is defined to be a Cauchy sequence if, and only if, for each \( \epsilon \in \mathbb{R}^+ \), there exists \( N \in \mathbb{N} \) such that \( |z_n - z_m| < \epsilon \) for each \( n, m > N \), i.e.

\[
\forall \epsilon \in \mathbb{R}^+ \exists N \in \mathbb{N} \forall n, m > N \left| z_n - z_m \right| < \epsilon.
\]

**Theorem 7.29.** The sequence \((z_n)_{n \in \mathbb{N}}\) in \( \mathbb{C} \) is convergent if, and only if, it is a Cauchy sequence.

**Proof.** Suppose the sequence is convergent with \( \lim_{n \to \infty} z_n = z \). Then, given \( \epsilon > 0 \), there is \( N \in \mathbb{N} \) such that \( z_n \in B_{\frac{\epsilon}{2}}(z) \) for each \( n > N \). If \( n, m > N \), then \( |z_n - z_m| \leq |z_n - z| + |z - z_m| < \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon \), establishing that \((z_n)_{n \in \mathbb{N}}\) is a Cauchy sequence.

Conversely, suppose the sequence is a Cauchy sequence. Using similar reasoning as in the proof of Prop. 7.10(b), we first show the sequence is bounded. If the sequence is Cauchy, then there exists \( N \in \mathbb{N} \) such that \( |z_n - z_m| < 1 \) for all \( n, m > N \). Thus, the set \( A := \{|z_n| : |z_n - z_{N+1}| \geq 1 \} \cup \{|z_1|\} \subseteq \mathbb{R}_0^+ \) is nonempty and finite. According to Th. 3.13(a), \( A \) has an upper bound \( M \). Then \( \max\{M, |z_{N+1}| + 1\} \) is an upper bound for \( \{|z_n| : n \in \mathbb{N}\} \), showing that the sequence is bounded. From Th. 7.27, we obtain that the sequence has a cluster point \( z \). It remains to show \( \lim_{n \to \infty} z_n = z \). Given \( \epsilon > 0 \), choose \( N \in \mathbb{N} \) such that \( |z_n - z_m| < \epsilon/2 \) for all \( n, m > N \). Since \( z \) is a cluster point, there exists \( k > N \) such that \( |z_k - z| < \epsilon/2 \). Thus,

\[
\forall n > N \left| z_n - z \right| \leq |z_n - z_k| + |z_k - z| < \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon,
\]

proving \( \lim_{n \to \infty} z_n = z \). \( \blacksquare \)

\( \blacksquare \)
Example 7.30. Consider the sequence \( S := (s_n)_{n \in \mathbb{N}} \) defined by

\[
s_n := \sum_{k=1}^{n} \frac{1}{k} = 1 + \frac{1}{2} + \cdots + \frac{1}{n}. \tag{7.27}
\]

We claim \( S \) is not a Cauchy sequence and, thus, not convergent by Th. 7.29: For each \( N \in \mathbb{N} \), we find \( n, m > N \) such that \( s_n - s_m > 1/2 \), namely \( m = N + 1 \) and \( n = 2(N+1) \):

\[
\begin{align*}
\sum_{k=N+2}^{2(N+1)} \frac{1}{k} &= \frac{1}{N+2} + \frac{1}{N+3} + \cdots + \frac{1}{2(N+1)} \\
&> (N+1) \cdot \frac{1}{2(N+1)} = \frac{1}{2}. \tag{7.28}
\end{align*}
\]

While we have just seen that \( S \) is not convergent, it is clearly increasing, i.e. Th. 7.19 implies \( S \) is unbounded and \( \lim_{n \to \infty} s_n = \infty \). Sequences defined by longer and longer sums are known as series and will be studied further in Sec. 7.3 below. The series of the present example is known as the harmonic series. It has become famous as the simplest example of a series that does not converge even though its summands converge to 0. In terms of the notation introduced in Sec. 7.3 below, we have shown

\[
\sum_{k=1}^{\infty} \frac{1}{k} = 1 + \frac{1}{2} + \frac{1}{3} + \cdots = \infty. \tag{7.29}
\]

### 7.2 Continuity

#### 7.2.1 Definitions and First Examples

Roughly, a function is continuous if a small change in its input results in a small change of its output. For functions defined on an interval, the notion of continuity makes precise the idea of a function having no jump – no discontinuity – at some point \( x \) in its domain. For example, we would say the sign function of (5.8) has precisely one jump – one discontinuity – at \( x = 0 \), whereas quadratic functions (or, more generally, polynomials) do not have any jumps – they are continuous.

**Definition 7.31.** Let \( M \subseteq \mathbb{C} \). If \( \zeta \in M \), then a function \( f : M \to \mathbb{K} \) is said to be **continuous** in \( \zeta \) if, and only if, for each \( \epsilon > 0 \), there is \( \delta > 0 \) such that the distance between the values \( f(z) \) and \( f(\zeta) \) is less than \( \epsilon \), provided the distance between \( z \) and \( \zeta \) is less than \( \delta \), i.e. if, and only if,

\[
\forall \epsilon \in \mathbb{R}^+, \exists \delta \in \mathbb{R}^+, \forall z \in M \left( |z - \zeta| < \delta \Rightarrow |f(z) - f(\zeta)| < \epsilon \right). \tag{7.30}
\]

Moreover, \( f \) is called **continuous** if, and only if, \( f \) is continuous in every \( \zeta \in M \). The set of all continuous functions from \( f : M \to \mathbb{K} \) is denoted by \( C(M, \mathbb{K}) \), \( C(M) := C(M, \mathbb{R}) \).

**Example 7.32.** (a) Every constant map \( f : M \to \mathbb{K}, \emptyset \neq M \subseteq \mathbb{C} \), is continuous: In this case, given \( \epsilon \), we can choose any \( \delta > 0 \) we want, say \( \delta := 42 \): If \( \zeta, z \in M \), then

\[
|f(\zeta) - f(z)| = 0 < \epsilon, \text{ which holds independently of } \delta, \text{ in particular, if } |\zeta - z| < \delta.
\]
(b) Every affine function \( f : \mathbb{K} \rightarrow \mathbb{K}, f(z) := az + b \) is continuous: For \( a = 0 \), this follows from (a). For \( a \neq 0 \), given \( \epsilon > 0 \), choose \( \delta := \epsilon/|a| \). Then,

\[
\forall \zeta, z \in \mathbb{K}, |z - \zeta| < \delta = \frac{\epsilon}{|a|} \Rightarrow |f(z) - f(\zeta)| = |az + b - a\zeta - b| = |a||z - \zeta| < |a|\frac{\epsilon}{|a|} = \epsilon.
\]

(c) The sign function of (5.8) is not continuous: It is continuous in each \( \xi \in \mathbb{R} \setminus \{0\} \), but not continuous in 0: If \( \xi \neq 0 \), then, given \( \epsilon > 0 \), choose \( \delta := |\xi| \). If \( |x - \xi| < \delta \), then \( \text{sgn}(x) = \text{sgn}(\xi) \), i.e. \( |\text{sgn}(x) - \text{sgn}(\xi)| = 0 < \epsilon \), proving continuity in \( \xi \). However, at 0, for \( \epsilon := 1/2 \), we have

\[
\forall \delta > 0 \left| \text{sgn}(0) - \text{sgn}(\delta/2) \right| = |0 - 1| = 1 > \frac{1}{2} = \epsilon,
\]

showing \( \text{sgn} \) is not continuous in 0.

Some subtleties arise from the possibility that \( f \) can be defined on subsets of \( \mathbb{C} \) with very different properties. The notions introduced in Def. 7.33 help to deal with these subtleties.

**Definition 7.33.** Let \( M \subseteq \mathbb{C} \).

(a) The point \( z \in \mathbb{C} \) is called a cluster point or accumulation point of \( M \) if, and only if, each \( \epsilon \)-neighborhood of \( z, \epsilon \in \mathbb{R}^+ \), contains infinitely many points of \( M \), i.e. if, and only if,

\[
\forall \epsilon \in \mathbb{R}^+ \quad \#(M \cap B_\epsilon(z)) = \infty.
\]

Note: A cluster point of \( M \) is not necessarily in \( M \).

(b) The point \( z \) is called an isolated point of \( M \) if, and only if, there is \( \epsilon \in \mathbb{R}^+ \) such that \( B_\epsilon(z) \cap M = \{z\} \). Note: An isolated point of \( M \) is always in \( M \).

**Proposition 7.34.** If \( M \subseteq \mathbb{C} \), then each point of \( M \) is either a cluster point or an isolated point of \( M \), i.e.

\[
M = \{z \in M : z \text{ cluster point of } M\} \cup \{z \in M : z \text{ isolated point of } M\}.
\]

**Proof.** Consider \( z \in M \) that is not a cluster point of \( M \). We have to show that \( z \) is an isolated point of \( M \). Since \( z \) is not a cluster point of \( M \), there exists \( \epsilon > 0 \) such that \( A := (M \cap B_\epsilon(z)) \setminus \{z\} \) is finite. Define

\[
\epsilon := \begin{cases} 
\min\{|a - z| : a \in A\} & \text{if } A \neq \emptyset, \\
\tilde{\epsilon} & \text{if } A = \emptyset.
\end{cases}
\]

Then \( B_\epsilon(z) \cap M = \{z\} \), showing \( z \) is an isolated point of \( M \). Finally, the union in (7.34) is clearly disjoint.
Lemma 7.35. Let $M \subseteq \mathbb{C}$, $f : M \rightarrow \mathbb{K}$. If $\zeta$ is an isolated point of $M$, then $f$ is always continuous in $\zeta$.

Proof. Independently of the concrete definition of $f$, we know there is $\delta > 0$ such that $B_\delta(\zeta) \cap M = \{\zeta\}$. In other words, if $z \in M$ with $|z - \zeta| < \delta$, then $z = \zeta$, implying $|f(z) - f(\zeta)| = 0 < \epsilon$ for each $\epsilon > 0$, showing $f$ to be continuous in $\zeta$. ■

Example 7.36. (a) The sign function restricted to the set $M := ]-\infty, -1] \cup \{0\} \cup [1, \infty[$, i.e.

$$\text{sgn}(x) = \begin{cases} 1 & \text{for } x \in [1, \infty[, \\ 0 & \text{for } x = 0, \\ -1 & \text{for } x \in ]-\infty, -1] \end{cases}$$

is continuous: As in Ex. 7.32(c), one sees that $\text{sgn}$ is continuous in each $\xi \in M \setminus \{0\}$. However, now it is also continuous in 0, since 0 is an isolated point of $M$.

(b) Every function $f : \mathbb{N} \rightarrow \mathbb{K}$ is continuous, since every $n \in \mathbb{N}$ is an isolated point of $\mathbb{N}$ (due to $\{n\} = \mathbb{N} \cap B_\frac{1}{2}(n)$).

7.2.2 Continuity, Sequences, and Function Arithmetic

To make available the power of the results on convergent sequences from Sec. 7.1 to investigations regarding the continuity of functions, we need to understand the relationship between both notions. The core of this relationship is the contents of the following Th. 7.37, which provides a criterion allowing one to test continuity in terms of convergent sequences:

Theorem 7.37. Let $M \subseteq \mathbb{C}$, $f : M \rightarrow \mathbb{K}$. If $\zeta \in M$, then $f$ is continuous in $\zeta$ if, and only if, for each sequence $(z_n)_{n \in \mathbb{N}}$ in $M$ with $\lim_{n \rightarrow \infty} z_n = \zeta$, the sequence $(f(z_n))_{n \in \mathbb{N}}$ converges to $f(\zeta)$, i.e.

$$\lim_{n \rightarrow \infty} z_n = \zeta \Rightarrow \lim_{n \rightarrow \infty} f(z_n) = f(\zeta). \quad (7.36)$$

Proof. If $\zeta \in M$ is an isolated point of $M$, then there is $\delta > 0$ such that $M \cap B_\delta(\zeta) = \{\zeta\}$. Then every $f : M \rightarrow \mathbb{K}$ is continuous in $\zeta$ according to Lem. 7.35. On the other hand, every sequence in $M$ converging to $\zeta$ must be finally constant and equal to $\zeta$, i.e. (7.36) is trivially valid at $\zeta$. Thus, the assertion of the theorem holds if $\zeta \in M$ is an isolated point of $M$.

If $\zeta \in M$ is not an isolated point of $M$, then $\zeta$ is a cluster point of $M$ according to Prop. 7.34. So, for the remainder of the proof, let $\zeta \in M$ be a cluster point of $M$. Assume that $f$ is continuous in $\zeta$ and $(z_n)_{n \in \mathbb{N}}$ is a sequence in $M$ with $\lim_{n \rightarrow \infty} z_n = \zeta$. For each $\epsilon > 0$, there is $\delta > 0$ such that $z \in M$ and $|z - \zeta| < \delta$ implies $|f(z) - f(\zeta)| < \epsilon$. Since $\lim_{n \rightarrow \infty} z_n = \zeta$, there is also $N \in \mathbb{N}$ such that, for each $n > N$, $|z_n - \zeta| < \delta$. Thus, for each $n > N$, $|f(z_n) - f(\zeta)| < \epsilon$, proving $\lim_{n \rightarrow \infty} f(z_n) = f(\zeta)$. Conversely, assume that $f$ is not continuous in $\zeta$. We have to construct a sequence $(z_n)_{n \in \mathbb{N}}$ in $M$ with
\[ \lim_{n \to \infty} z_n = \zeta, \text{ but } (f(z_n))_{n \in \mathbb{N}} \text{ does not converge to } f(\zeta). \] Since \( f \) is not continuous in \( \zeta \), there must be some \( \epsilon_0 > 0 \) such that, for each \( 1/n, n \in \mathbb{N} \), there is at least one \( z_n \in M \) satisfying \( |z_n - \zeta| < 1/n \) and \( |f(z_n) - f(\zeta)| \geq \epsilon_0 \). Then \( (z_n)_{n \in \mathbb{N}} \) is a sequence in \( M \) with \( \lim_{n \to \infty} z_n = \zeta \) and \( (f(z_n))_{n \in \mathbb{N}} \) does not converge to \( f(\zeta) \).

We can now apply the rules of Th. 7.13 to see that all the arithmetic operations defined in Not. 6.2 preserve continuity:

**Theorem 7.38.** Let \( M \subseteq \mathbb{C}, f, g : M \to \mathbb{K}, \lambda \in \mathbb{K}, \zeta \in M \). If \( f, g \) are both continuous in \( \zeta \), then \( \lambda f, f + g, f g, f/g \) for \( g \neq 0, |f|, \text{ Re } f, \text{ and } \text{ Im } f \) are all continuous in \( \zeta \). If \( \mathbb{K} = \mathbb{R} \), then \( \max(f, g), \min(f, g), f^+ \text{ and } f^- \), are also all continuous in \( \zeta \).

**Proof.** Let \((z_n)_{n \in \mathbb{N}}\) be a sequence in \( M \) such that \( \lim_{n \to \infty} z_n = \zeta \). Then the continuity of \( f \) and \( g \) in \( \zeta \) yields \( \lim_{n \to \infty} f(z_n) = f(\zeta) \) and \( \lim_{n \to \infty} g(z_n) = g(\zeta) \). Then

\[
\begin{align*}
(7.11a) & \quad \lim_{n \to \infty} (\lambda f)(z_n) = (\lambda f)(\zeta), \\
(7.11b) & \quad \lim_{n \to \infty} (f + g)(z_n) = (f + g)(\zeta), \\
(7.11c) & \quad \lim_{n \to \infty} (fg)(z_n) = (fg)(\zeta), \\
(7.11d) & \quad \lim_{n \to \infty} (f/g)(z_n) = (f/g)(\zeta), \\
(7.11e) & \quad \lim_{n \to \infty} |f|(z_n) = |f|(\zeta), \\
(7.2) & \quad \lim_{n \to \infty} (\text{Re } f)(z_n) = (\text{Re } f)(\zeta), \\
(7.2) & \quad \lim_{n \to \infty} (\text{Im } f)(z_n) = (\text{Im } f)(\zeta).
\end{align*}
\]

If \( f, g \) are both \( \mathbb{R} \)-valued, then we also have

\[
\begin{align*}
(7.12a) & \quad \lim_{n \to \infty} \max(f, g)(z_n) = \max(f, g)(\zeta), \\
(7.12b) & \quad \lim_{n \to \infty} \min(f, g)(z_n) = \min(f, g)(\zeta),
\end{align*}
\]

and, finally, the continuity of \( f^+ \) and \( f^- \) follows from the continuity of \( \max(f, g) \).

**Corollary 7.39.** A function \( f : M \to \mathbb{C}, M \subseteq \mathbb{C} \), is continuous in \( \zeta \in M \) if, and only if, both \( \text{Re } f \) and \( \text{Im } f \) are continuous in \( \zeta \).

**Proof.** If \( f \) is continuous in \( \zeta \), then \( \text{Re } f \) and \( \text{Im } f \) are both continuous in \( \zeta \) by Th. 7.38. If \( \text{Re } f \) and \( \text{Im } f \) are both continuous in \( \zeta \), then, as

\[ f = \text{Re } f + i \text{ Im } f, \] (7.37)

\( f \) is continuous in \( \zeta \), once again, by Th. 7.38.

**Example 7.40.** (a) The continuity of the absolute value function \( z \mapsto |z| \) on \( \mathbb{K} \) can be concluded directly from (7.11e) and, alternatively, from combining the continuity of \( f : \mathbb{K} \to \mathbb{K}, f(z) = z \), according to Ex. 7.32(b), with the continuity of \( |f| \) according to Th. 7.38.
(b) Every polynomial $P : \mathbb{K} \to \mathbb{K}$, $P(x) = \sum_{j=0}^{n} a_j x^j$, $a_j \in \mathbb{K}$, is continuous: First note that every monomial $x \mapsto x^j$ is continuous on $\mathbb{K}$ by (7.11g). Then Th. 7.38 implies the continuity of $x \mapsto a_j x^j$ on $\mathbb{K}$. Now the continuity of $P$ follows from (7.16a) or, alternatively, by an induction from the $f+g$ part of Th. 7.38.

(c) Let $P,Q : \mathbb{K} \to \mathbb{K}$, be polynomials and let $A := Q^{-1}\{0\}$ the set of all zeros of $Q$ (if any). Then the rational function $(P/Q) : \mathbb{K} \setminus A \to \mathbb{K}$ is continuous as a consequence of (b) plus the $f/g$ part of Th. 7.38.

**Theorem 7.41.** Let $D_f,D_g \subseteq \mathbb{C}$, $f : D_f \to \mathbb{C}$, $g : D_g \to \mathbb{K}$, $f(D_f) \subseteq D_g$. If $f$ is continuous in $\zeta \in D_f$ and $g$ is continuous in $f(\zeta) \in D_g$, then $g \circ f : D_f \to \mathbb{K}$ is continuous in $\zeta$. In consequence, if $f$ and $g$ are both continuous, then the composition $g \circ f$ is also continuous.

**Proof.** Let $\zeta \in D_f$ and assume $f$ is continuous in $\zeta$ and $g$ is continuous in $f(\zeta)$. If $(z_n)_{n \in \mathbb{N}}$ is a sequence in $D_f$ such that $\lim_{n \to \infty} z_n = \zeta$, then the continuity of $f$ in $\zeta$ implies that $\lim_{n \to \infty} f(z_n) = f(\zeta)$. Then the continuity of $g$ in $f(\zeta)$ implies $\lim_{n \to \infty} g(f(z_n)) = g(f(\zeta))$, thereby establishing the continuity of $g \circ f$ in $\zeta$. ■

### 7.2.3 Bounded, Closed, and Compact Sets

Subsets $A$ of $\mathbb{C}$ (and even subsets of $\mathbb{R}$) can be extremely complicated. If the set $A$ has one or more of the benign properties defined in the following, then this can often be exploited in some useful way (we will see an important example in Th. 7.54 below).

**Definition 7.42.** Consider $A \subseteq \mathbb{C}$.

(a) $A$ is called **bounded** if, and only if, $A = \emptyset$ or the set $\{|z| : z \in A\}$ is bounded in $\mathbb{R}$ in the sense of Def. 2.27(a), i.e. if, and only if,

$$\exists_{M \in \mathbb{R}^+} A \subseteq B_M(0).$$

(b) $A$ is called **closed** if, and only if, every sequence in $A$ that converges in $\mathbb{C}$ has its limit in $A$ (note that $\emptyset$ is, thus, closed).

(c) $A$ is called **compact** if, and only if, $A$ is both closed and bounded.

**Example 7.43.** (a) Clearly, $\emptyset$ and sets containing single points $\{z\}$, $z \in \mathbb{C}$ are compact. The sets $\mathbb{C}$ and $\mathbb{R}$ are simple examples of closed sets that are not bounded.

(b) Let $a,b \in \mathbb{R}$, $a < b$. Each bounded interval $]a,b[$, $[a,b]$, $[a,b[ [a,b]$ is, indeed, bounded (e.g. by $M := e + \max\{|a|,|b|\}$ for each $e \in \mathbb{R}^+$). If $(x_n)_{n \in \mathbb{N}}$ is a sequence in $[a,b]$, converging to $x \in \mathbb{R}$, then Th. 7.13(c) shows $a \leq x \leq b$, i.e. $x \in [a,b]$ and $[a,b]$ is, indeed, closed. Analogously, one sees that the unbounded intervals $[a,\infty[$ and $]-\infty,a]$ are also closed. On the other hand, open and half-open intervals are not closed: For sufficiently large $n$, the convergent sequence $(b - \frac{1}{n})_{n \in \mathbb{N}}$ is in $[a,b[ [a,b]$ but $\lim_{n \to \infty}(b - \frac{1}{n}) = b \notin [a,b]$ and the other cases are treated analogously. In particular, only intervals of the form $[a,b]$ (and trivial intervals) are compact.
For each $\epsilon > 0$ and each $z \in \mathbb{C}$, the set $B_\epsilon(z)$ is bounded (since $B_\epsilon(z) \subseteq B_{\epsilon + |z|}(0)$ by the triangle inequality), but not closed (since, for sufficiently large $n \in \mathbb{N}$, $(z + \epsilon - \frac{1}{n})_{n \in \mathbb{N}}$ is a sequence in $B_\epsilon(z)$, converging to $z + \epsilon \notin B_\epsilon(z)$). In particular, $B_\epsilon(z)$ is not compact.

**Proposition 7.44.** (a) Finite unions of bounded (resp. closed, resp. compact) sets are bounded (resp. closed, resp. compact), i.e. if $A_1, \ldots, A_n \subseteq \mathbb{C}$, $n \in \mathbb{N}$, are bounded (resp. closed, resp. compact), then $A := \bigcup_{j=1}^n A_j$ is also bounded (resp. closed, resp. compact).

(b) Arbitrary (i.e. finite or infinite) intersections of bounded (resp. closed, resp. compact) sets are bounded (resp. closed, resp. compact), i.e. if $I \neq \emptyset$ is an arbitrary index set and, for each $j \in I$, $A_j \subseteq \mathbb{C}$ is bounded (resp. closed, resp. compact), then $A := \bigcap_{j \in I} A_j$ is also bounded (resp. closed, resp. compact).

**Proof.** (a): Exercise.

(b): Fix $j_0 \in I$. If all $A_j$, $j \in I$, are bounded, then, in particular, there is $M \in \mathbb{R}^+$ such that $A_{j_0} \subseteq B_M(0)$. Thus, $A = \bigcap_{j \in I} A_j \subseteq A_{j_0} \subseteq B_M(0)$ shows $A$ is also bounded. If all $A_j$, $j \in I$, are closed and $(a_n)_{n \in \mathbb{N}}$ is a sequence in $A$ that converges to some $z \in \mathbb{C}$, then $(a_n)_{n \in \mathbb{N}}$ is a sequence in each $A_j$, $j \in I$, and, since each $A_j$ is closed, $z \in A_j$ for each $j \in I$, i.e. $z \in A = \bigcap_{j \in I} A_j$. If all $A_j$, $j \in I$, are compact, then they are all closed and bounded and, thus, $A$ is closed and bounded, i.e. $A$ is compact.

**Example 7.45.** (a) According to Prop. 7.44(a), all finite subsets of $\mathbb{C}$ are compact.

(b) $\mathbb{N} = \bigcup_{n \in \mathbb{N}} \{n\}$ shows that infinite unions of compact sets can be unbounded, and $]0, 1[ = \bigcup_{n \in \mathbb{N}} [\frac{1}{n}, 1 - \frac{1}{n}]$ shows that infinite unions of compact sets are not always closed.

Many more examples of closed sets can be obtained as preimages of closed sets under continuous maps according to the following remark:

**Remark 7.46.** In Analysis II, it will be shown in the more general context of maps $f$ between topological spaces that a map $f$ is continuous if, and only if, all preimages $f^{-1}(A)$ under $f$ of closed sets $A$ are closed. Here, we will only prove the following special case:

$$f : \mathbb{C} \longrightarrow \mathbb{K} \text{ continuous and } A \subseteq \mathbb{K} \text{ closed } \implies f^{-1}(A) \subseteq \mathbb{C} \text{ closed.} \quad (7.38)$$

Indeed, suppose $f$ is continuous and $A \subseteq \mathbb{K}$ is closed. If $(z_n)_{n \in \mathbb{N}}$ is a sequence in $f^{-1}(A)$ with $\lim_{n \to \infty} z_n = z \in \mathbb{C}$, then $(f(z_n))_{n \in \mathbb{N}}$ is a sequence in $A$. The continuity of $f$ then implies $\lim_{n \to \infty} f(z_n) = f(z)$ and, then, $f(z) \in A$, since $A$ is closed. Thus, $z \in f^{-1}(A)$, showing $f^{-1}(A)$ is closed.

**Example 7.47.** (a) For each $z \in \mathbb{C}$ and each $r > 0$, the closed disk $\overline{B}_r(z) := \{w \in \mathbb{C} : |z - w| \leq r\}$ with radius $r$ and center $z$ is, indeed, closed by (7.38), since

$$\overline{B}_r(z) = f^{-1}[0, r], \quad (7.39)$$
where \( f \) is the continuous map \( f : \mathbb{C} \rightarrow \mathbb{R}, f(w) := |z - w| \). Since \( \overline{B}_r(z) \) is clearly bounded, it is also compact.

(b) For each \( z \in \mathbb{C} \) and each \( r > 0 \), the circle (also called a 1-sphere) \( S_r(z) := \{ w \in \mathbb{C} : |z - w| = r \} \) with radius \( r \) and center \( z \) is closed by (7.38), since \( S_r(z) = f^{-1}\{r\} \), where \( f \) is the same map as in (7.39). Moreover, \( S_r(z) \) is also clearly bounded, and, thus, compact.

(c) According to (7.38), for each \( x \in \mathbb{R} \), the closed half-spaces \( \{ z \in \mathbb{C} : \text{Re} z \geq x \} = \text{Re}^{-1}[x, \infty[ \) and \( \{ z \in \mathbb{C} : \text{Im} z \geq x \} = \text{Im}^{-1}[x, \infty[ \) are, indeed, closed.

**Theorem 7.48.** A subset \( K \) of \( \mathbb{C} \) is compact if, and only if, every sequence in \( K \) has a subsequence that converges to some limit \( z \in K \).

**Proof.** If \( K \) is closed and bounded, and \( (z_n)_{n \in \mathbb{N}} \) is a sequence in \( K \), then the boundedness, the Bolzano-Weierstrass Th. 7.27, and Prop. 7.26 yield a subsequence that converges to some \( z \in \mathbb{C} \). However, since \( K \) is closed, \( z \in K \).

Conversely, assume every sequence in \( K \) has a subsequence that converges to some limit \( z \in K \). Let \( (z_n)_{n \in \mathbb{N}} \) be a sequence in \( K \) that converges to some \( w \in \mathbb{C} \). Then this sequence must have a subsequence that converges to some \( z \in K \). However, according to Prop. 7.23, it must be \( w = z \in K \), showing \( K \) is closed. If \( K \) is not bounded, then there exists a sequence \( (z_n)_{n \in \mathbb{N}} \) in \( K \) such that \( \lim_{n \to \infty} |z_n| = \infty \). Every subsequence \( (z_{n_k})_{k \in \mathbb{N}} \) then still has the property that \( \lim_{k \to \infty} |z_{n_k}| = \infty \), in particular, each subsequence is unbounded and cannot converge to some \( z \in \mathbb{C} \) (let alone in \( K \)).

**Caveat 7.49.** In Analysis II, we will generalize the notion of compactness to subsets of so-called metric spaces. In metric spaces, it is still true that a set \( K \) is compact if, and only if, every sequence in \( K \) has a subsequence that converges to some limit in \( K \). However, while it remains true that every compact set is closed and bounded, the converse does not hold in general metric spaces (in general, even in closed sets, there exist bounded sequences that do not have convergent subsequences).

One reason that compact sets are useful is that real-valued continuous functions on compact sets assume a maximum and a minimum, which is the contents of Th. 7.54 below. In preparation, we now define maxima and minima for real-valued functions.

**Definition 7.50.** Let \( M \subseteq \mathbb{C} \), \( f : M \rightarrow \mathbb{R} \).

(a) Given \( z \in M \), \( f \) has a (strict) global min at \( z \) if, and only if, \( f(z) \leq f(w) \) \( (f(z) < f(w)) \) for each \( w \in M \setminus \{z\} \). Analogously, \( f \) has a (strict) global max at \( z \) if, and only if, \( f(z) \geq f(w) \) \( (f(z) > f(w)) \) for each \( w \in M \setminus \{z\} \). Moreover, \( f \) has a (strict) global extreme value at \( z \) if, and only if, \( f \) has a (strict) global min or a (strict) global max at \( z \).
(b) Given \( z \in M \), \( f \) has a (strict) local min at \( z \) if, and only if, there exists \( \epsilon > 0 \) such that \( f(z) \leq f(w) \) (\( f(z) < f(w) \)) for each \( w \in \{ w \in M : |z - w| < \epsilon \} \setminus \{ z \} \).

Analogously, \( f \) has a (strict) local max at \( z \) if, and only if, there exists \( \epsilon > 0 \) such that \( f(z) \geq f(w) \) (\( f(z) > f(w) \)) for each \( w \in \{ w \in M : |z - w| < \epsilon \} \setminus \{ z \} \).

Moreover, \( f \) has a (strict) local extreme value at \( z \) if, and only if, \( f \) has a (strict) local min or a (strict) local max at \( z \).

**Remark 7.51.** In the context of Def. 7.50, it is immediate from the respective definitions that \( f \) has a (strict) global min at \( z \in M \) if, and only if, \( -f \) has a (strict) global max at \( z \). Moreover, the same holds if “global” is replaced by “local”. It is equally obvious that every (strict) global min/max is a (strict) local min/max.

**Theorem 7.52.** If \( K \subseteq \mathbb{C} \) is compact, and \( f : K \rightarrow \mathbb{C} \) is continuous, then \( f(K) \) is compact.

**Proof.** If \((w_n)_{n \in \mathbb{N}}\) is a sequence in \( f(K) \), then, for each \( n \in \mathbb{N} \), there is some \( z_n \in K \) such that \( f(z_n) = w_n \). As \( K \) is compact, there is a subsequence \((a_n)_{n \in \mathbb{N}} \) of \((z_n)_{n \in \mathbb{N}}\) with \( \lim_{n \to \infty} a_n = a \) for some \( a \in K \). Then \((f(a_n))_{n \in \mathbb{N}}\) is a subsequence of \((w_n)_{n \in \mathbb{N}}\) and the continuity of \( f \) yields \( \lim_{n \to \infty} f(a_n) = f(a) \in f(K) \), showing that \((w_n)_{n \in \mathbb{N}}\) has a convergent subsequence with limit in \( f(K) \). By Th. 7.48, we have therefore established that \( f(K) \) is compact. \( \blacksquare \)

**Lemma 7.53.** If \( K \) is a nonempty compact subset of \( \mathbb{R} \), then \( K \) contains a smallest and a largest element, i.e. there exist \( m, M \in K \) such that \( m \leq x \leq M \) for each \( x \in K \).

**Proof.** Since the compact set \( K \) is bounded, we know that

\[-\infty < m := \inf K \leq \sup K := M < \infty.\]

According to the definition of the inf and sup as largest lower bound and smallest upper bound, respectively, for each \( n \in \mathbb{N} \), there must be elements \( x_n, y_n \in K \) such that \( m \leq x_n \leq m + \frac{1}{n} \) and \( M - \frac{1}{n} \leq y_n \leq M \). Since the compact set \( K \) is also closed, we get \( m = \lim_{n \to \infty} x_n \in K \) and \( M = \lim_{n \to \infty} y_n \in K \). \( \blacksquare \)

**Theorem 7.54.** If \( K \subseteq \mathbb{C} \) is compact, and \( f : K \rightarrow \mathbb{R} \) is continuous, then \( f \) assumes its max and its min, i.e. there are \( z_m \in K \) and \( z_M \in K \) such that \( f \) has a global min at \( z_m \) and a global max at \( z_M \). In particular, the continuous function \( f \) assumes its max and min on each compact interval \( K = [a, b] \subseteq \mathbb{R} \), \( a, b \in \mathbb{R} \).

**Proof.** Since \( K \) is compact and \( f \) is continuous, \( f(K) \subseteq \mathbb{R} \) is compact according to Th. 7.52. Then, by Lem. 7.53, \( f(K) \) contains a smallest element \( m \) and a largest element \( M \). This, in turn, implies that there are \( z_m, z_M \in K \) such that \( f(z_m) = m \) and \( f(z_M) = M \). \( \blacksquare \)

**Example 7.55.** On an unbounded set, a continuous function does not necessarily have a global max or a global min, as one can already see from \( x \mapsto x \). An example for a continuous function on a bounded, but not closed, interval, that does not have a global max is \( f : [0, 1] \rightarrow \mathbb{R}, f(x) := 1/x \), which is continuous by Th. 7.38.
7.2.4 Intermediate Value Theorem

**Theorem 7.56** (Bolzano’s Theorem). Let \( a, b \in \mathbb{R} \) with \( a < b \). If \( f : [a, b] \rightarrow \mathbb{R} \) is continuous with \( f(a) > 0 \) and \( f(b) < 0 \), then \( f \) has at least one zero in \( [a, b] \). More precisely, the set \( A := f^{-1}(0) \) has a min \( \xi_1 \) and a max \( \xi_2 \), \( a < \xi_1 \leq \xi_2 < b \), where \( f > 0 \) on \( [a, \xi_1[ \) and \( f < 0 \) on \( ]\xi_2, b] \).

**Proof.** Let \( \xi_1 := \inf f^{-1}(\mathbb{R}_0^+) \).

(a): \( f(\xi_1) \leq 0 \): This is clear if \( \xi_1 = b \). If \( \xi_1 < b \), then, for each \( n \in \mathbb{N} \) sufficiently large, there exists \( x_n \in [\xi_1, \xi_1 + 1/n] \subseteq [a, b] \) such that \( f(x_n) \leq 0 \). Then \( \lim_{n \to \infty} x_n = \xi_1 \) and the continuity of \( f \) implies \( \lim_{n \to \infty} f(x_n) = f(\xi_1) \). Now \( f(\xi_1) \leq 0 \) is a consequence of Th. 7.13(c). In particular, (a) yields \( a < \xi_1 \) and \( f > 0 \) on \( [a, \xi_1[ \).

(b): \( f(\xi_1) \geq 0 \): The continuity of \( f \) implies \( \lim_{n \to \infty} f(\xi_1 - 1/n) = f(\xi_1) \) and, since we have already seen \( f(\xi_1 - 1/n) > 0 \) for each \( n \in \mathbb{N} \) sufficiently large, \( f(\xi_1) \geq 0 \) is again a consequence of Th. 7.13(c). In particular, we have \( \xi_1 < b \).

Combining (a) and (b), we have \( f(\xi_1) = 0 \) and \( a < \xi_1 < b \).

Defining \( \xi_2 := \sup f^{-1}(\mathbb{R}_0^+) \), \( \xi_2 = 0 \) and \( a < \xi_2 < b \) is shown completely analogous. Then \( f < 0 \) on \( ]\xi_2, b] \) is also clear as well as \( \xi_1 \leq \xi_2 \).

**Theorem 7.57** (Intermediate Value Theorem). Let \( a, b \in \mathbb{R} \) with \( a < b \). If \( f : [a, b] \rightarrow \mathbb{R} \) is continuous, then \( f \) assumes every value between \( f(a) \) and \( f(b) \), i.e.

\[
\left[ \min\{f(a), f(b)\}, \max\{f(a), f(b)\} \right] \subseteq f([a, b]).
\]  

(7.40)

**Proof.** If \( f(a) = f(b) \), then there is nothing to prove. If \( f(a) < f(b) \) and \( \eta \in ]f(a), f(b)[ \), then consider the auxiliary function \( g : [a, b] \rightarrow \mathbb{R}, g(x) := \eta - f(x) \). Then \( g \) is continuous with \( g(a) = \eta - f(a) > 0 \) and \( g(b) = \eta - f(b) < 0 \). According to Bolzano’s Th. 7.56, there exists \( \xi \in ]a, b[ \) such that \( g(\xi) = \eta - f(\xi) = 0 \), i.e. \( f(\xi) = \eta \) as claimed. If \( f(b) < f(a) \) and \( \eta \in ]f(b), f(a)[ \), then consider the auxiliary function \( g : [a, b] \rightarrow \mathbb{R}, g(x) := f(x) - \eta \). Then \( g \) is continuous with \( g(a) = f(a) - \eta > 0 \) and \( g(b) = f(b) - \eta < 0 \). Once again, according to Bolzano’s Th. 7.56, there exists \( \xi \in ]a, b[ \) such that \( g(\xi) = f(\xi) - \eta = 0 \), i.e. \( f(\xi) = \eta \).

**Theorem 7.58.** If \( I \subseteq \mathbb{R} \) is an interval (of one of the 8 types listed in (4.8)) and \( f : I \rightarrow \mathbb{R} \) is continuous, then \( f(I) \) is also an interval (it can degenerate to a single point if \( f \) is constant). More precisely, if \( \emptyset \neq I = [a, b] \) is a compact interval, then \( \emptyset \neq f(I) = [\min f(I), \max f(I)] \); if \( I \) is not a compact interval, then one of the following 9 cases occurs:

\[
\begin{align*}
f(I) &= \mathbb{R}, \\
f(I) &= -\infty, \sup f(I), \\
f(I) &= -\infty, \sup f(I)[, \quad (7.41b) \\
f(I) &= \inf f(I), \infty[ \\
f(I) &= \inf f(I), \sup f(I], \quad (7.41d) \\
f(I) &= \min f(I), \max f(I) \\
\end{align*}
\]
7 LIMITS AND CONVERGENCE OF REAL AND COMPLEX NUMBERS

\[ f(I) = [\inf f(I), \sup f(I)], \quad (7.41f) \]
\[ f(I) = \inf f(I), \infty[, \quad (7.41g) \]
\[ f(I) = \inf f(I), \sup f(I)], \quad (7.41h) \]
\[ f(I) = \inf f(I), \sup f(I][. \quad (7.41i) \]

Proof. If \( I \) is a compact interval, then we merely combine Th. 7.54 with Th. 7.57. Otherwise, let \( \eta \in f(I) \). If \( f(I) \) has an upper bound, then Th. 7.57 implies \([\eta, \sup f(I)] \subseteq f(I) \) and \( f(I) \cap [\eta, \infty[ \subseteq [\eta, \sup f(I)] \). If \( f(I) \) does not have an upper bound, then Th. 7.57 implies \( f(I) \cap [\eta, \infty[ = [\eta, \infty[ \). Analogously, one obtains \( f(I) \cap ]-\infty, \eta[ = ]-\infty, \eta[ \) or \( f(I) \cap ]-\infty, \eta] = [\inf f(I), \eta[ \) or \( f(I) \cap ]-\infty, \eta] = [\inf f(I), \eta[ \), showing that there are precisely the 9 possibilities of (7.41) for \( f(I) = (f(I) \cap ]-\infty, \eta[) \cup (f(I) \cap [\eta, \infty[) \). \( \blacksquare \)

The above results will have striking consequences in the following Sec. 7.2.5.

Example 7.59. The piecewise affine function

\[ f : [0,1] \longrightarrow \mathbb{R}, \quad f(x) := \begin{cases} \end{cases} \]
\[ (-1)^n \cdot n - \frac{2n+1}{n+1} \cdot \frac{1}{n} \quad \text{for } x \in \left[\frac{1}{n}, \frac{1}{n-1}\right], \text{ } n \text{ even,} \]
\[ (-1)^n \cdot n + \frac{2n+1}{n+1} \cdot \frac{1}{n} \quad \text{for } x \in \left[\frac{1}{n}, \frac{1}{n-1}\right], \text{ } n \text{ odd,} \]

satisfies \( f(1/n) = (-1)^n \cdot n \) for each \( n \in \mathbb{N} \) and is an example of a continuous function on the bounded half-open interval \( I := [0,1) \) with \( f(I) = \mathbb{R} \).

7.2.5 Inverse Functions, Existence of Roots, Exponential Function, Logarithm

Theorem 7.60. Let \( I \subseteq \mathbb{R} \) be an interval (of one of the 8 types listed in (4.8)). If \( f : I \longrightarrow \mathbb{R} \) is strictly increasing (resp. decreasing), then \( f \) has an inverse function \( f^{-1} \) defined on \( J := f(I) \), i.e. \( f^{-1} : J \longrightarrow I \), and \( f^{-1} \) is continuous and strictly increasing (resp. decreasing). If \( f \) is also continuous, then \( J \) must be an interval.

Proof. From Prop. 2.32(b), we know \( f : I \longrightarrow \mathbb{R} \) is one-to-one. Then \( f : I \longrightarrow f(I) \) is invertible and Prop. 2.32(c) shows \( f^{-1} \) is strictly monotone in the same sense as \( f \). We need to prove the continuity of \( f^{-1} \). We assume \( f \) to be strictly increasing (the case where \( f \) is strictly decreasing then follows by considering \(-f\)). Let \( \eta \in J, \epsilon > 0 \), and \( \xi \in I \) with \( f(\xi) = \eta \). If \( I = \{\xi\} \), then \( J = \{\eta\} \), and there is nothing to prove. It remains to consider three cases:

(a) \( \xi = \min I \), i.e. \( \xi \) is the left endpoint of \( I \) (and \( \xi \neq \max I \)),
(b) \( \xi = \max I \), i.e. \( \xi \) is the right endpoint of \( I \) (and \( \xi \neq \min I \)),
(c) \( \xi \) is neither the min nor the max of \( I \).

We carry out the proof for (c) and leave the (very similar) cases (a) and (b) as exercises. In case (c), there are points \( \xi_1, \xi_2 \in B_\epsilon(\xi) \cap I \) such that

\[ \xi - \epsilon < \xi_1 < \xi < \xi_2 < \xi + \epsilon. \quad (7.42) \]
As $f$ is strictly increasing, this implies

$$f(\xi_1) < \eta < f(\xi_2).$$

Choose $\delta > 0$ such that

$$f(\xi_1) < \eta - \delta < \eta < \eta + \delta < f(\xi_2).$$

Then

$$\forall y \in J \cap B_\delta(\eta) \quad \left(f(\xi_1) < y < f(\xi_2) \quad \Rightarrow \quad f^{-1}(y) \in B_\epsilon(\xi)\right),$$

proving the continuity of $f^{-1}$ in $\eta$. That $J$ must be an interval if $f$ is continuous was already shown in Th. 7.58. 

**Remark and Definition 7.61 (Roots).** We are now in a position to fulfill the promise made in Def. and Rem. 5.6, i.e. to prove the existence of unique roots for nonnegative real numbers: For each $n \in \mathbb{N}$, the function $f : \mathbb{R}^+_0 \rightarrow \mathbb{R}$, $f(x) := x^n$, is continuous and strictly increasing with $J := f(\mathbb{R}^+_0) = \mathbb{R}^+_0$. Then Th. 7.60 implies the existence of a continuous and strictly increasing inverse function $f^{-1} : \mathbb{R}^+_0 \rightarrow \mathbb{R}^+_0$. For each $x \in \mathbb{R}^+_0$, we call $f^{-1}(x)$ the $n$th root of $x$ and write $\sqrt[n]{x} := x^{1/n} := f^{-1}(x)$. Then $(\sqrt[n]{x})^n = (x^{1/n})^n = x$ is immediate from the definition. Caveat: By definition, roots are always nonnegative and they are only defined for nonnegative numbers (when studying complex numbers and $\mathbb{C}$-valued functions more deeply in the field of Complex Analysis, one typically extends the notion of root, but we will not pursue this route in this class). As anticipated in Def. and Rem. 5.6, one also writes $\sqrt{x}$ instead of $\sqrt[2]{x}$ and calls $\sqrt{x}$ the square root of $x$.

**Remark and Definition 7.62.** It turns out that $\sqrt{2}$ (and many other roots) are not rational numbers, i.e. $\sqrt{2} \notin \mathbb{Q}$. This is easily proved by contradiction: If $\sqrt{2} \in \mathbb{Q}$, then there exist natural numbers $m, n \in \mathbb{N}$ such that $\sqrt{2} = m/n$. Moreover, by canceling possible factors of 2, we may assume at least one of the numbers $m, n$ is odd. Now $\sqrt{2} = m/n$ implies $m^2 = 2n^2$, i.e. $m^2$ and, thus, $m$ must be even. In consequence, there exists $p \in \mathbb{N}$ such that $m = 2p$, implying $2n^2 = m^2 = 4p^2$ and $n^2 = 2p^2$. Thus $n^2$ and $n$ must also be even, in contradiction to $m, n$ not both being even.

The elements of $\mathbb{R} \setminus \mathbb{Q}$ are called irrational numbers. It turns out that most real numbers are irrational numbers – one can show that $\mathbb{Q}$ is countable, whereas $\mathbb{R} \setminus \mathbb{Q}$ is not countable (actually, every interval contains countably many rational and uncountably many irrational numbers, see Appendix F, in particular, Th. F.1(c) and Cor. F.4).

**Theorem 7.63 (Inequality Between the Arithmetic Mean and the Geometric Mean).** If $n \in \mathbb{N}$ and $x_1, \ldots, x_n \in \mathbb{R}^+_0$, then

$$\sqrt[n]{x_1 \cdots x_n} \leq \frac{x_1 + \cdots + x_n}{n},$$

(7.43)

where the left-hand side is called the geometric mean and the right-hand side is called the arithmetic mean of the numbers $x_1, \ldots, x_n$. Equality occurs if, and only if, $x_1 = \cdots = x_n$. 

Proof. If at least one of the $x_j$ is 0, then (7.43) becomes the true statement $0 \leq \frac{x_1 + \cdots + x_n}{n}$ with strict equality if at least one $x_j > 0$. If $x_1 = \cdots = x_n = x$, then (7.43) also holds since both sides are equal to $x$. Thus, for the remainder of the proof, we assume all $x_j > 0$ and not all $x_j$ are equal. First, we consider the special case, where $\frac{x_1 + \cdots + x_n}{n} = 1$. Since not all $x_j$ are equal, there exists $k$ with $x_k \neq 1$. We prove (7.43) by induction for $n \in \{2, 3, \ldots\}$ in the form

$$\left(\sum_{j=1}^{n} x_j = n \land \exists k \in \{1, \ldots, n\}, x_k \neq 1\right) \Rightarrow \prod_{j=1}^{n} x_j < 1.$$ 

Base Case ($n = 2$): Since $x_1 + x_2 = 2$, $0 < x_1, x_2$ and not both $x_1$ and $x_2$ are equal to 1, there is $\epsilon > 0$ such that $x_1 = 1 + \epsilon$ and $x_2 = 1 - \epsilon$, i.e. $x_1 x_2 = 1 - \epsilon^2 < 1$, which establishes the base case. Induction Step: We now have $n \geq 2$ and $0 < x_1, \ldots, x_{n+1}$ with $\sum_{j=1}^{n+1} x_j = n + 1$ plus the existence of $k, l \in \{1, \ldots, n+1\}$ such that $x_k = 1 + \alpha$, $x_l = 1 - \beta$ with $\alpha, \beta > 0$. Then define $y := x_k + x_l - 1 = 1 + \alpha - \beta$. One observes $y > 0$ (since $\beta < 1$) and

$$y + \sum_{\substack{j=1, j \neq k, l}}^{n+1} x_j = -1 + \sum_{j=1}^{n+1} x_j = n \quad \text{ind. hyp.} \quad y \prod_{j=1, j \neq k, l}^{n+1} x_j \leq 1$$

(we can not exclude equality as $y$ and all the remaining $x_j$ might be equal to 1). Since $x_k x_l = (1 + \alpha)(1 - \beta) = 1 + \alpha - \beta - \alpha \beta = y - \alpha \beta < y$, we now infer $\prod_{j=1}^{n+1} x_j < 1$, concluding the induction proof. It remains to consider the case $\frac{x_1 + \cdots + x_n}{n} = \lambda > 0$, not all $x_j$ equal. One estimates

$$\sqrt[n]{x_1 \cdots x_n} = \lambda \sqrt[n]{\frac{x_1}{\lambda} \cdots \frac{x_n}{\lambda}} \quad \text{special case} \quad \lambda \frac{x_1 + \cdots + x_n}{\lambda n} = \frac{x_1 + \cdots + x_n}{n},$$

completing the proof of the theorem. ■

Corollary 7.64. For each $a \in \mathbb{R}_0^+ \setminus \{1\}$, $n \in \{2, 3, \ldots\}$, $p \in \{1, \ldots, n - 1\}$:

$$\sqrt[n]{a^p} < 1 + \frac{p}{n}(a - 1); \quad p = 1 \text{ yields } \sqrt[n]{a} < 1 + \frac{a - 1}{n}. \quad (7.44)$$

Proof. The simple application

$$\sqrt[n]{a^p} = \sqrt[n]{a^p \cdot \prod_{j=1}^{n-p} 1} \quad \text{Th. 7.63} \quad \frac{p a + n - p}{n} = 1 + \frac{p}{n}(a - 1) \quad (7.45)$$

of Th. 7.63 establishes the case. ■

Example 7.65. We use (7.43) to show

$$\lim_{n \to \infty} \sqrt[n]{n} = 1. \quad (7.46)$$
First note $0 < x < 1 \Rightarrow 0 < x^n < 1$, i.e., $\sqrt[n]{n} > 1$ for each $(\sqrt[n]{n})^n = n > 1$. Now write $n$ as the product of $n$ factors $n = \sqrt[n]{n}\sqrt[n]{n} \cdot \prod_{k=1}^{n-2} 1$. Then, for $n > 1$,

$$\sqrt[n]{n} = \frac{\sqrt[n]{n}\sqrt[n]{n} \cdot \prod_{k=1}^{n-2}}{1} \leq \frac{2\sqrt[n]{n} + n - 2}{n} < 1 + \frac{2}{\sqrt[n]{n}}. \quad (7.47)$$

It is an exercise to show

$$\lim_{n \to \infty} \frac{1}{\sqrt[n]{n}} = 0. \quad (7.48)$$

Now this together with $1 \leq \sqrt[n]{n} \leq 1 + \frac{2}{\sqrt[n]{n}}$ and the Sandwich Th. 7.16 proves (7.46).

**Example 7.66** (Euler’s Number). We use Th. 7.63 to prove the limit

$$e := \lim_{n \to \infty} \left(1 + \frac{1}{n}\right)^n \quad (7.49)$$

exists. It is known as Euler’s number. One can show it is an irrational number (see Appendix H.1) and its first digits are $e \approx 2.71828 \ldots$ It is of exceptional importance for analysis and mathematics in general, as it pops up in all kinds of different mathematical contexts. From Th. 7.63, we obtain

$$\forall n \in \mathbb{N} \forall x \in [-n, \infty[, \quad \left(1 + \frac{x}{n}\right)^n = 1 \cdot \left(1 + \frac{x}{n}\right)^n < \left(1 + \frac{x}{n + 1}\right)^{n+1}, \quad (7.50)$$

where we have used that, on both sides of the inequality in (7.50), there are $n+1$ factors having the same sum, namely $n + 1 + x$; and the inequality in (7.43) is strict, unless all factors are equal. We now apply (7.50) to the sequences $(a_n)_{n \in \mathbb{N}}, (b_n)_{n \in \mathbb{N}}, (c_n)_{n \in \mathbb{N}}$, where

$$\forall n \in \mathbb{N} \quad \begin{cases} a_n := \left(1 + \frac{1}{n}\right)^n, & b_n := \left(1 - \frac{1}{n}\right)^n, \\ c_n := b_{n+1}^{-1} = \left(\left(1 - \frac{1}{n + 1}\right)^{-1}\right)^{n+1} = \left(1 + \frac{1}{n}\right)^{n+1} \end{cases} \quad (7.51)$$

Applying (7.50) with $x = 1$ and $x = -1$, respectively, yields $(a_n)_{n \in \mathbb{N}}$ and $(b_n)_{n \in \mathbb{N}}$ are strictly increasing, and $(c_n)_{n \in \mathbb{N}}$ is strictly decreasing. On the other hand, $a_n < c_n$ holds for each $n \in \mathbb{N}$, showing $(a_n)_{n \in \mathbb{N}}$ is bounded from above by $c_1$, and $(c_n)_{n \in \mathbb{N}}$ is bounded from below by $a_1$. In particular, Th. 7.19 implies the convergence of both $(a_n)_{n \in \mathbb{N}}$ and $(c_n)_{n \in \mathbb{N}}$. Moreover, $\lim_{n \to \infty} c_n = \lim_{n \to \infty} (a_n(1+1/n)) = e \cdot 1 = e$, which, together with $a_n < e < c_n$ for each $n \in \mathbb{N}$, can be used to compute $e$ to an arbitrary precision.

**Definition 7.67.** Let $A \subseteq \mathbb{R}$ be a subset of the real numbers. Then $A$ is called dense in $\mathbb{R}$ if, and only if, every $\varepsilon$-neighborhood of every real number contains a point from $A$, i.e. if, and only if,

$$\forall x \in \mathbb{R} \forall \varepsilon \in \mathbb{R}^+ \quad A \cap B_\varepsilon(x) \neq \emptyset.$$

**Theorem 7.68.** (a) $\mathbb{Q}$ is dense in $\mathbb{R}$. 

(b) \( \mathbb{R} \setminus \mathbb{Q} \) is dense in \( \mathbb{R} \).

(c) For each \( x \in \mathbb{R} \), there exist sequences \((r_n)_{n \in \mathbb{N}}\) and \((s_n)_{n \in \mathbb{N}}\) in the rational numbers \(\mathbb{Q}\) such that \( x = \lim_{n \to \infty} r_n = \lim_{n \to \infty} s_n \), \((r_n)_{n \in \mathbb{N}}\) is strictly increasing and \((s_n)_{n \in \mathbb{N}}\) is strictly decreasing.

**Proof.** (a): Since each \( B_\epsilon(x) \) is an interval, it suffices to prove that every interval \([a, b[\), \(a < b\), contains a rational number. If \(0 \in [a, b[\), then there is nothing to prove. Suppose \(0 < a < b\) and set \(\delta := b - a > 0\). Choose \(n \in \mathbb{N}\) such that \(\frac{1}{n} < \delta\) and let
\[
q := \max \left\{ \frac{k}{n} : k \in \mathbb{N} \wedge \frac{k}{n} < b \right\}.
\]
Then \(q \in \mathbb{Q}\) and \(a < q < b\). If \(a < b < 0\), choose \(\delta\) and \(n\) as above, but let
\[
q := \min \left\{ -\frac{k}{n} : k \in \mathbb{N} \wedge -\frac{k}{n} > a \right\}.
\]
Then, once again, \(q \in \mathbb{Q}\) and \(a < q < b\).

(b): Analogous to (a), we show that every interval \([a, b[\), \(a < b\), contains an irrational number: According to (a), we choose \(q \in \mathbb{Q} \cap [a, b[\), \(\delta := b - q > 0\) and \(n \in \mathbb{N}\) such that \(\sqrt{2}/n < \delta\). Then \(a < \lambda := q + \sqrt{2}/n < b\) and also \(\lambda \in \mathbb{R} \setminus \mathbb{Q}\) (otherwise, \(\sqrt{2} = n(\lambda - q) \in \mathbb{Q}\)).

(c): Using (a), for each \(n \in \mathbb{N}\), we choose rational numbers \(r_n\) and \(s_n\) such that
\[
r_n \in \left] \frac{x - 1}{n}, \frac{x - 1}{n + 1} \right[ , \quad s_n \in \left] \frac{x + 1}{n}, \frac{x + 1}{n + 1} \right[ .
\]
Then, clearly, \((r_n)_{n \in \mathbb{N}}\) is strictly increasing, \((s_n)_{n \in \mathbb{N}}\) is strictly decreasing, and the Sandwich Th. 7.16 implies \(x = \lim_{n \to \infty} r_n = \lim_{n \to \infty} s_n\).

**Definition and Remark 7.69 (Exponentiation).** We have already used that Not. C.5 of the Appendix defines \(a^x\) for \((a, x) \in \mathbb{C} \times \mathbb{N}_0\) and for \((a, x) \in (\mathbb{C} \setminus \{0\}) \times \mathbb{Z}\) (using \(G = \mathbb{C}\) and \(G = \mathbb{C} \setminus \{0\}\), respectively). We will now extend the definition to \((a, x) \in \mathbb{R}^+ \times \mathbb{R}\) (later, we will further extend the definition to \((a, z) \in \mathbb{R}^+ \times \mathbb{C}\). The present extension to \((a, x) \in \mathbb{R}^+ \times \mathbb{R}\) is accomplished in two steps – first, in (a), for rational \(x\), then, in (b), for irrational \(x\).

(a) For rational \(x = k/n\) with \(k \in \mathbb{Z}\) and \(n \in \mathbb{N}\), define
\[
a^x := a^{k/n} := \sqrt[n]{a^k}.
\]
(7.52)

For this definition to make sense, we have to check it does not depend on the special representation of \(x\), i.e., we have to verify \(x = k/n = km/nm\) with \(k \in \mathbb{Z}\) and \(m, n \in \mathbb{N}\) implies \(a^{k/n} = a^{km/nm}\). To this end, observe, using Rem. and Def. 7.61,
\[
(a^{k/n})^{nm} = (\sqrt[n]{a^k})^{nm} = a^{km} \quad \text{and} \quad (a^{km/nm})^{nm} = (\sqrt[m]{a^{km}})^{nm} = a^{km},
\]
(7.53)
proving \(a^b = a^{\frac{k}{m}}\) (here, as in Rem. and Def. 7.61, we used that \(\lambda \mapsto \lambda^N\) is one-to-one on \(\mathbb{R}_0^+\) for each \(N \in \mathbb{N}\)). The exponentiation rules of Th. C.6 now extend to rational exponents in a natural way, i.e., for each \(a, b > 0\) and each \(x, y \in \mathbb{Q}\):

\[
\begin{align*}
    a^{x+y} &= a^{x} a^{y}, \quad (7.54a)\\
    (a^{x})^{y} &= a^{x y}, \quad (7.54b)\\
    a^{x} b^{x} &= (ab)^{x}. \quad (7.54c)
\end{align*}
\]

For the proof, by possibly multiplying numerator and denominator by some natural number, we can assume \(x = k/n\) and \(y = l/n\) with \(k, l \in \mathbb{Z}\) and \(n \in \mathbb{N}\). Then

\[
(a^{x+y})^{n} = (a^{k/l})^{n} = a^{k+l} \quad \text{Th. C.6(a)}
\]

\[
\overset{\text{Th. C.6(b)}}{=} a^{k} a^{l} = (a^{k/n})^{n} (a^{l/n})^{n} \quad \text{Th. C.6(c)}
\]

\[
(a^{x})^{y} = a^{x y} = (a^{x/n})^{n} \quad \text{Th. C.6(b)}
\]

proving (7.54a);

\[
((a^{x})^{y})^{n} = (a^{k/l})^{n} ((a^{k/n})^{n})^{n} \quad \text{Th. C.6(b)}
\]

\[
\overset{\text{Th. C.6(b)}}{=} a^{k l} = (a^{x y/n})^{n^{2}}
\]

proving (7.54b);

\[
(a^{x} b^{x})^{n} = (a^{k/n})^{n} (b^{l/n})^{n} a^{k} b^{k} \quad \text{Th. C.6(c)}
\]

\[
\overset{\text{Th. C.6(c)}}{=} (ab)^{k} = (ab)^{k/n} \quad \text{Th. C.6(b)}
\]

proving (7.54c).

Moreover, we obtain the following monotonicity rules for each \(a, b \in \mathbb{R}^+\) and each \(x, y \in \mathbb{Q}\):

\[
\begin{align*}
   \forall \ x > 0 \quad (a < b & \implies a^{x} < b^{x}), \quad (7.55a)\\
   \forall \ x < 0 \quad (a < b & \implies a^{x} > b^{x}), \quad (7.55b)\\
   \forall \ a > 1 \quad (x < y & \implies a^{x} < a^{y}), \quad (7.55c)\\
   \forall \ a > 0 \quad (x < y & \implies a^{x} > a^{y}), \quad (7.55d)
\end{align*}
\]

If \(x = k/n\) with \(k, n \in \mathbb{N}\) and \(a < b\), then \(a^{1/n} < b^{1/n}\) according to Rem. and Def. 7.61, which, in turn, implies \(a^{x} = (a^{1/n})^{k} < (b^{1/n})^{k} = b^{x}\), proving (7.55a); and \(a^{-1} > b^{-1}\) implies \(a^{-x} = (a^{-1})^{x} > (b^{-1})^{x} = b^{-x}\), proving (7.55b). If \(x < y\), set \(q := y - x > 0\). Then \(1 < a\) and (7.55a) imply \(1^{q} < a^{q}\), i.e. \(a^{x} < a^{x} a^{q} = a^{y}\), proving (7.55c). Similarly, \(0 < a < 1\) and (7.55a) imply \(a^{q} < 1^{q} = 1\), i.e. \(a^{y} = a^{x} a^{q} < a^{x}\), proving (7.55d).

The following estimates will also come in handy: For \(a \in \mathbb{R}^+\) and \(x, y \in \mathbb{Q}\):

\[
\forall \ m \in \mathbb{N} \quad (x, y \in [-m, m] \implies |a^{x} - a^{y}| \leq L |x - y|), \quad (7.57)
\]

\[
\text{where } L := \max\{a^{m+1}, (1/a)^{m+1}\}.
\]
For \( x \geq 1 \), (7.56) is proved by \( a^x < a^{x+1} < x \cdot a^{x+1} + 1 \); for \( x < 1 \), write \( x = p/n \) with \( p, n \in \mathbb{N} \) and \( p < n \), and apply (7.44) to obtain \( a^x < 1 + x(a - 1) < 1 + xa < 1 + x \cdot a^{x+1} \). For the proof of (7.57), first consider \( a > 1 \). Moreover, by possibly renaming \( x \) and \( y \), we may assume \( x < y \), i.e. \( z := y - x > 0 \). Thus, (7.57) holds with \( x \) replaced by \( z \). Multiplying the resulting inequality by \( a^x \) yields
\[
a^x a^z - a^x = a^y - a^x < z \cdot a^x a^{z+1} = (y - x) a^{y+1} \leq (y - x) a^{m+1},
\]
proving (7.57) for \( a > 1 \). For \( a = 1 \), it is clearly true, and for \( a < 1 \), it is \( a^{-1} > 1 \), i.e.
\[
|a^x - a^y| = |(a^{-1})^{-x} - (a^{-1})^{-y}| \leq |y - x| (a^{-1})^{m+1},
\]
finishing the proof of (7.57).

(b) We now define \( a^x \) for irrational \( x \) by letting
\[
a^x := \lim_{n \to \infty} a^{q_n}, \quad \text{where} \ (q_n)_{n \in \mathbb{N}} \ \text{is a sequence in} \ \mathbb{Q} \ \text{with} \ \lim_{n \to \infty} q_n = x. \quad (7.58)
\]
For this definition to make sense, we have to know such sequences \( (q_n)_{n \in \mathbb{N}} \) exist, which we do know from Th. 7.68(c). We also know from Th. 7.68(c) that there exists an increasing sequence \( (q_n)_{n \in \mathbb{N}} \) in \( \mathbb{Q} \) converging to \( x \), in particular, bounded by \( x \). Then, by (7.55c) and (7.55d), respectively, \( (a^{q_n})_{n \in \mathbb{N}} \) is increasing for \( a > 1 \) and decreasing for \( 0 < a < 1 \). Moreover, the sequence is bounded from above by \( a^N \) with \( N \in \mathbb{N}, N > x \), for \( a > 1 \); and bounded from below by 0 for \( 0 < a < 1 \). In both cases, Th. 7.19 implies convergence of the sequence to some limit that we may call \( a^x \). However, we still need to verify that, for each sequence \( (r_n)_{n \in \mathbb{N}} \) in \( \mathbb{Q} \) with \( \lim_{n \to \infty} r_n = x \), the sequence \( (a^{r_n})_{n \in \mathbb{N}} \) converges to the same limit \( a^x \) in \( \mathbb{R} \). If \( \lim_{n \to \infty} r_n = x \), then \( \lim_{n \to \infty} |q_n - r_n| = 0 \). Since \( (r_n)_{n \in \mathbb{N}} \) and \( (q_n)_{n \in \mathbb{N}} \) are bounded, (7.57) implies
\[
\exists L \in \mathbb{R}^+ \quad \forall n \in \mathbb{N} \quad |a^{q_n} - a^{r_n}| \leq L |q_n - r_n|, \quad (7.59)
\]
such that Prop. 7.11(a) implies \( \lim_{n \to \infty} |a^{q_n} - a^{r_n}| = 0 \) and
\[
\lim_{n \to \infty} a^{r_n} = \lim_{n \to \infty} (a^{r_n} - a^{q_n} + a^{q_n}) = 0 + a^x = a^x, \quad (7.60)
\]
showing (7.58) does not depend on the chosen sequence.

**Proposition 7.70.** The exponentiation rules (7.54), the monotonicity rules (7.55), and the estimates (7.56) and (7.57) remain valid if \( x, y \in \mathbb{Q} \) is replaced by \( x, y \in \mathbb{R} \). Moreover, for each \( a > 0 \) and each sequence \( (x_n)_{n \in \mathbb{N}} \) in \( \mathbb{R} \):
\[
\lim_{n \to \infty} x_n = x \in \mathbb{R} \quad \Rightarrow \quad \lim_{n \to \infty} a^{x_n} = a^x. \quad (7.61)
\]

**Proof.** Given \( x, y \in \mathbb{R} \), let \( (p_n)_{n \in \mathbb{N}} \) and \( (q_n)_{n \in \mathbb{N}} \) be sequences in \( \mathbb{Q} \) such that \( \lim_{n \to \infty} p_n = x \) and \( \lim_{n \to \infty} q_n = y \).
We start by verifying (7.57). As we can assume \( (p_n)_{n \in \mathbb{N}} \) and \( (q_n)_{n \in \mathbb{N}} \) to be monotone, we may also assume \( p_n, q_n \in [-m, m] \) for each \( n \in \mathbb{N} \). Then the rational case of (7.57) implies
\[
\forall n \in \mathbb{N} \quad |a^{p_n} - a^{q_n}| \leq L |p_n - q_n|,
\]

and Th. 7.13(c) establishes the case. Then (7.61) also follows, since
\[ 0 \leq |a^{x_n} - a^x| \leq L |x_n - x| \to 0. \]

We deal with (7.54) next. For each \( a, b > 0 > 0 \):
\[ a^x b^y = \lim_{n \to \infty} a^{p_n} b^{q_n} = \lim_{n \to \infty} (a^p b^q) = (ab)^x, \]
\[ \forall k \in \mathbb{N} \ (a^x)^{q_k} = \lim_{n \to \infty} (a^{p_n})^{q_k} = \lim_{n \to \infty} a^{p_n q_k} = a^{x q_k}, \]
\[ \Rightarrow (a^y)^x = \lim_{n \to \infty} (a^{x q_n}) = \lim_{n \to \infty} a^{x q_n} = a^{x y}, \]
thereby proving (7.54).

Proceeding to (7.55c), let \( a > 1 \) and \( h > 0 \). If \((q_n)_{n \in \mathbb{N}}\) is a strictly increasing sequence in \( \mathbb{Q}^+ \) with \( \lim_{n \to \infty} q_n = h \), then \( a^h = \lim_{n \to \infty} a^{q_n} > a^{q_1} > 1 \). Thus, if \( x < y \), let \( h := y - x > 0 \) to obtain \( a^y = a^x a^h > a^x \), i.e. (7.55c). If \( 0 < a < 1 \) and \( x < y \), then \((1/a)^x < (1/a)^y\), yielding (7.55d). For (7.55a), consider \( x > 0 \) and \( 0 < a < b \). Then
\[ \frac{b}{a} > 1 \Rightarrow \frac{b^x}{a^x} = \left( \frac{b}{a} \right)^x > 1 \Rightarrow b^x > a^x, \]
proving (7.55a). If \( x < 0 \) and \( 0 < a < b \), then \( a^x = (1/a)^{-x} > (1/b)^{-x} = b^x \), proving (7.55b).

Finally, it remains to verify (7.56). For \( x \geq 1 \), the proof for rational \( x \) still works for irrational \( x \). For \( 0 < x < 1 \), one uses the usual sequence \((q_n)_{n \in \mathbb{N}}\) in \( \mathbb{Q} \) with \( \lim_{n \to \infty} q_n = x \) and obtains (recalling \( a > 1 \))
\[ a^x = \lim_{n \to \infty} a^{q_n} \leq \lim_{n \to \infty} (1 + q_n(a - 1)) = 1 + x(a - 1) < 1 + x \cdot a^{x+1}, \]
proving (7.56).

**Definition 7.71 (Exponential and Power Functions).** (a) Each function of the form
\[ f : \mathbb{R}^+ \to \mathbb{R}, \quad f(x) := x^\alpha, \quad \alpha \in \mathbb{R}, \quad (7.62) \]
is called a power function. For \( \alpha > 0 \), the power function is extended to \( x = 0 \) by setting \( 0^\alpha := 0 \); for \( \alpha \in \mathbb{Z} \), it is defined on \( \mathbb{R} \setminus \{0\} \); for \( \alpha \in \mathbb{N}_0 \) even on \( \mathbb{R} \).

(b) Each function of the form
\[ f : \mathbb{R} \to \mathbb{R}^+, \quad f(x) := a^x, \quad a > 0, \quad (7.63) \]
is called a (general) exponential function. The case where \( a = e \) with \( e \) being Euler’s number from (7.49) is of particular interest and importance. Most of the time, when referring to an exponential function, one actually means \( x \mapsto e^x \). It is also common to write \( \exp(x) \) instead of \( e^x \).
Theorem 7.72. (a) Every power function as defined in Def. 7.71(a) is continuous on its respective domain. Moreover, for each \( \alpha > 0 \), it is strictly increasing on \([0, \infty[\); for each \( \alpha < 0 \), it is strictly decreasing on \( ]0, \infty[\).

(b) Every exponential function as defined in Def. 7.71(b) is continuous. Moreover, for each \( a > 1 \), it is strictly increasing; for each \( 0 < a < 1 \), it is strictly decreasing.

**Proof.** (a): The monotonicity claims are provided by (7.55a) and (7.55b), respectively. For each \( \alpha \in \mathbb{N}_0 \), the power function is a polynomial, for each \( \alpha \in \mathbb{Z} \), a rational function, i.e. continuity is provided by Ex. 7.40(b) and Ex. 7.40(c), respectively. For a general \( \alpha \in \mathbb{R} \), the continuity proof on \( \mathbb{R}^+ \) will be postponed to Ex. 7.76(a) below, where it can be accomplished more easily. So it remains to show the continuity in \( x = 0 \) for \( \alpha > 0 \).

However, if \((x_n)_{n \in \mathbb{N}}\) is a sequence in \( \mathbb{R}^+ \) with \( \lim_{n \to \infty} x_n = 0 \) and \( k \in \mathbb{N} \) with \( 1/k \leq \alpha \), then, at least for \( n \) sufficiently large such that \( x_n \leq 1 \), \( 0 < x_{n}^{\alpha} \leq x_{n}^{1/k} \) by (7.55d). Then the continuity of \( x \mapsto x^{1/k} \) implies \( \lim_{n \to \infty} x_{n}^{1/k} = 0 \) and the Sandwich Th. 7.16 implies \( \lim_{n \to \infty} x_{n}^{\alpha} = 0 \), proving continuity in \( x = 0 \).

(b): Everything has already been proved – continuity is provided by (7.61), monotonicity is provided by (7.55c) and (7.55d).

**Remark and Definition 7.73 (Logarithm).** According to Th. 7.72(b), for each \( a \in \mathbb{R}^+ \setminus \{1\} \), the exponential function \( f : \mathbb{R} \to \mathbb{R}^+ \), \( f(x) := a^x \), is continuous and strictly monotone with \( f(\mathbb{R}) = \mathbb{R}^+ \) (verify that the image is all of \( \mathbb{R}^+ \) as an exercise). Then Th. 7.60 implies the existence of a continuous and strictly monotone inverse function \( f^{-1} : \mathbb{R}^+ \to \mathbb{R} \). For each \( x \in \mathbb{R}^+ \), we call \( f^{-1}(x) \) the logarithm of \( x \) to base \( a \) and write \( \log_a x := f^{-1}(x) \). The most important special case is where the base is Euler’s number, \( a = e \). This is called the **natural** logarithm. Bases \( a = 2 \) and \( a = 10 \) also carry special names, **binary** and **common** logarithm, respectively. The notation is

\[
\ln x := \log_e x, \quad \log_2 x, \quad \log_{10} x, \quad (7.64)
\]

however, the notation in the literature varies – one finds \( \log \) used instead of \( \ln \), \( \log_2 \), and \( \log_{10} \); one also finds \( \lg \) instead of \( \log_2 \). So you always need to verify what precisely is meant by either notation.

**Corollary 7.74.** For each \( a \in \mathbb{R}^+ \setminus \{1\} \), the logarithm function \( f : \mathbb{R}^+ \to \mathbb{R} \), \( f(x) = \log_a x \) is continuous. For \( a > 1 \), it is strictly increasing; for \( 0 < a < 1 \), it is strictly decreasing.

**Theorem 7.75.** One obtains the following logarithm rules:

\[
\forall a \in \mathbb{R}^+ \setminus \{1\} \quad \log_a 1 = 0, \quad (7.65a)
\]

\[
\forall a \in \mathbb{R}^+ \setminus \{1\} \quad \log_a a = 1, \quad (7.65b)
\]

\[
\forall a \in \mathbb{R}^+ \setminus \{1\} \quad \forall x \in \mathbb{R}^+ \quad a^{\log_a x} = x, \quad (7.65c)
\]

\[
\forall a \in \mathbb{R}^+ \setminus \{1\} \quad \forall x \in \mathbb{R} \quad \log_a a^x = x, \quad (7.65d)
\]
∀ \ a ∈ \mathbb{R}^+, \{1\} \quad ∀ \ x, y ∈ \mathbb{R}^+ \quad \log_a(xy) = \log_a x + \log_a y, \quad (7.65e)

∀ \ a ∈ \mathbb{R}^+, \{1\} \quad ∀ \ x ∈ \mathbb{R}^+ \quad \log_a(x^y) = y \log_a x, \quad (7.65f)

∀ \ a ∈ \mathbb{R}^+, \{1\} \quad ∀ \ x, y ∈ \mathbb{R}^+ \quad \log_a(x/y) = \log_a x - \log_a y, \quad (7.65g)

∀ \ a ∈ \mathbb{R}^+, \{1\} \quad ∀ \ x ∈ \mathbb{R}^+ \quad ∀ \ n ∈ \mathbb{N} \quad \log_a \sqrt[n]{x} = \frac{1}{n} \log_a x, \quad (7.65h)

∀ \ a, b ∈ \mathbb{R}^+, \{1\} \quad ∀ \ x ∈ \mathbb{R}^+ \quad \log_b x = (\log_b a) \log_a x. \quad (7.65i)

Proof. All the rules are easy consequences of the logarithm being defined as the inverse function to \( f : \mathbb{R} \rightarrow \mathbb{R}^+, f(x) := a^x \).

(7.65a): It is \( \log_a 1 = f^{-1}(1) = 0 \), as \( f(0) = a^0 = 1 \).

(7.65b): It is \( \log_a a = f^{-1}(a) = 1 \), as \( f(1) = a^1 = a \).

(7.65c): It is \( a^{\log_a x} = f(f^{-1}(x)) = x \).

(7.65d): It is \( \log_a a^x = f^{-1}(f(x)) = x \).

(7.65e): It is \( \log_a(xy) = f^{-1}(xy) = f^{-1}(f(\log_a x + \log_a y)) = \log_a x + \log_a y \), since

\[
\log_a x + \log_a y = a^{\log_a x + \log_a y} = a^{\log_a x} a^{\log_a y} = xy.
\]

(7.65f): It is \( \log_a(x^y) = f^{-1}(x^y) = f^{-1}(f(y \log_a x)) = y \log_a x \), since

\[
f(y \log_a x) = a^{y \log_a x} = (a^{\log_a x})^y = x^y.
\]

(7.65g) is just a combination of (7.65e) and (7.65f): \( \log_a(x/y) = \log_a(x y^{-1}) = \log_a x - \log_a y \).

(7.65h) is just a special case of (7.65f): \( \log_a \sqrt[n]{x} = \log_a x^{1/n} = \frac{1}{n} \log_a x \).

(7.65i): One computes

\[
(log_b a) \log_a x = (log_b a) log_b a^{\log_a x} = \log_b x.
\]

Thus, we have verified all the rules and concluded the proof. ■

Example 7.76. (a) For each \( \alpha ∈ \mathbb{R} \), the power function

\[
f : \mathbb{R}^+ \rightarrow \mathbb{R}, \quad f(x) := x^\alpha = e^{\alpha \ln x}, \quad (7.66)
\]

is continuous, which follows from Th. 7.41, since \( f = \exp \circ (\alpha \ln) \), \( \ln \) is continuous by Cor. 7.74, and \( \exp \) is continuous by Th. 7.72(b).
As a consequence of Th. 7.41, each of the following functions $f_1, f_2, f_3$, where

- $f_1 : \mathbb{R} \rightarrow \mathbb{R}$, $f_1(x) := \left(\exp(\lambda + x^2)\right)^\alpha$,
- $f_2 : \mathbb{R} \rightarrow \mathbb{R}$, $f_2(x) := \frac{1}{e^{\alpha x} + \lambda}$,
- $f_3 : \mathbb{R} \rightarrow \mathbb{R}$, $f_3(x) := \frac{x^5}{(\lambda + |x|)^\alpha}$

is continuous for each $\alpha \in \mathbb{R}$ and each $\lambda \in \mathbb{R}^+$.

### 7.3 Series

#### 7.3.1 Definition and Convergence

Series are a special type of sequences, namely sequences whose members arise from summing up the members of another sequence. We have, on occasion, already encountered series, for example the harmonic series $(s_n)_{n \in \mathbb{N}}$, whose members $s_n$ were defined in (7.27). In the present section, we will study series more systematically.

**Definition 7.77.** Given a sequence $(a_n)_{n \in \mathbb{N}}$ in $\mathbb{K}$ (or, more generally, in any set $A$, where an addition is defined), the sequence $(s_n)_{n \in \mathbb{N}}$, where

$$\forall \, n \in \mathbb{N} \quad s_n := \sum_{j=1}^{n} a_j,$$

is called an (infinite) series and is denoted by

$$\sum_{j=1}^{\infty} a_j := \sum_{j \in \mathbb{N}} a_j := (s_n)_{n \in \mathbb{N}}.$$

The $a_n$ are called the *summands* of the series, the $s_n$ its *partial sums*. Moreover, each series $\sum_{j=k}^{\infty} a_j$ with $k \in \mathbb{N}$ is called a *remainder (series)* of the series $(s_n)_{n \in \mathbb{N}}$.

The example of the remainder series already shows that it is useful to allow countable index sets other than $\mathbb{N}$. Thus, if $(a_j)_{j \in I}$, where $I$ is a countable index set and $\phi : \mathbb{N} \rightarrow I$ a bijective map, then define

$$\sum_{j \in I} a_j := \sum_{j=1}^{\infty} a_{\phi(j)}$$

(compare the definition in (3.19c) for finite sums). Note that the definition depends on $\phi$, which is suppressed in the notation $\sum_{j \in I} a_j$.

For sequences in $\mathbb{K}$, the notion of convergence is available, and, thus, it is also available for series arising from real or complex sequences (as such series are, again, sequences in $\mathbb{K}$).
Definition 7.78. If \((s_n)_{n \in \mathbb{N}}\) is a series with the \(s_n\) defined as in (7.67) and with summands \(a_j \in \mathbb{K}\), then the series is called convergent with limit \(s \in \mathbb{K}\) if, and only if, \(\lim_{n \to \infty} s_n = s\) in the sense of (7.1). In that case, one writes
\[
\sum_{j=1}^{\infty} a_j = s
\]
(7.70)
and calls \(s\) the sum of the series. The series is called divergent if, and only if, it is not convergent. We write \(\sum_{j=1}^{\infty} a_j = \infty\) (resp. \(\sum_{j=1}^{\infty} a_j = -\infty\)) if, and only if, \((s_n)_{n \in \mathbb{N}}\) diverges to \(\infty\) (resp. \(-\infty\)) in the sense of Def. 7.18.

Caveat 7.79. One has to use care as the symbol \(\sum_{j=1}^{\infty} a_j\) is used with two completely different meanings. If it is used according to (7.68), then it means a sequence; if it is used according to (7.70), then it means a real or complex number (or, possibly, \(\infty\) or \(-\infty\)). It should always be clear from the context, if it means a sequence or a number. For example, in the statement “the series \(\sum_{j=1}^{\infty} 2^{-j}\) is convergent”, it means a sequence; whereas in the statement “\(\sum_{j=1}^{\infty} 2^{-j} = 1\)”, it means a number.

Example 7.80. (a) For each \(q \in \mathbb{C}\) with \(|q| < 1\), \(\sum_{j=0}^{\infty} q^j\) is called a geometric series.
From (3.22b) (the reader is asked to go back and check that (3.22b) and its proof, indeed, remain valid for each \(q \in \mathbb{C}\)), we obtain the partial sums \(s_n = \sum_{j=0}^{n} q^j = \frac{1-q^{n+1}}{1-q}\). Since \(|q| < 1\), we know \(\lim_{n \to \infty} q^{n+1} = 0\) from Ex. 7.6. Thus, the series is convergent with
\[
\forall_{|q|<1} \sum_{j=0}^{\infty} q^j = \lim_{n \to \infty} s_n = \lim_{n \to \infty} \frac{1-q^{n+1}}{1-q} = \frac{1}{1-q}. \tag{7.71}
\]
(b) In Ex. 7.30, we obtained the divergence of the harmonic series:
\[
\sum_{k=1}^{\infty} \frac{1}{k} = \infty. \tag{7.72}
\]
Corollary 7.81. Let \(\sum_{j=1}^{\infty} a_j\) and \(\sum_{j=1}^{\infty} b_j\) be convergent series in \(\mathbb{C}\).

(a) Linearity:
\[
\forall_{\lambda, \mu \in \mathbb{C}} \sum_{j=1}^{\infty} (\lambda a_j + \mu b_j) = \lambda \sum_{j=1}^{\infty} a_j + \mu \sum_{j=1}^{\infty} b_j. \tag{7.73}
\]
(b) Complex Conjugation:
\[
\sum_{j=1}^{\infty} \overline{a_j} = \sum_{j=1}^{\infty} a_j. \tag{7.74}
\]
(c) Monotonicity:
\[
\left( \forall_{j \in \mathbb{N}} a_j, b_j \in \mathbb{R} \land a_j \leq b_j \right) \quad \Rightarrow \quad \sum_{j=1}^{\infty} a_j \leq \sum_{j=1}^{\infty} b_j. \tag{7.75}
\]
Each remainder series \( \sum_{j=n+1}^{\infty} a_j \), \( n \in \mathbb{N} \), converges, and, letting \( S := \sum_{j=1}^{\infty} a_j \), \( s_n := \sum_{j=1}^{n} a_j \), \( r_n := \sum_{j=n+1}^{\infty} a_j \), one has
\[
\forall n \in \mathbb{N} \quad S = s_n + r_n, \quad \lim_{n \to \infty} a_n = \lim_{n \to \infty} r_n = 0. \tag{7.76}
\]

Proof. (a) follows from the first two identities of Th. 7.13(a), (b) is due to
\[
\sum_{j=1}^{\infty} a_j = \lim_{n \to \infty} \sum_{j=1}^{n} a_j \quad \text{Def. and Rem. 5.5(a)} \quad = \lim_{n \to \infty} \sum_{j=1}^{n} a_j \quad = \lim_{n \to \infty} \sum_{j=1}^{\infty} a_j = \sum_{j=1}^{\infty} a_j,
\]
(c) follows from Th. 7.13(c), and, for (d), one computes
\[
\forall n \in \mathbb{N} \quad r_n = \lim_{k \to \infty} \sum_{j=n+1}^{k} a_j = \lim_{k \to \infty} (s_k - s_n) = S - s_n,
\]
\[
\lim_{n \to \infty} r_n = \lim_{n \to \infty} (S - s_n) = S - S = 0,
\]
completing the proof. \( \square \)

7.3.2 Convergence Criteria

**Corollary 7.82.** Let \( \sum_{j=1}^{\infty} a_j \) be series such that all \( a_j \in \mathbb{R}_0^+ \). If \( s_n := \sum_{j=1}^{n} a_j \) are the partial sums of \( \sum_{j=1}^{\infty} a_j \), then
\[
\lim_{n \to \infty} s_n = \begin{cases} 
\sup\{s_n : n \in \mathbb{N}\} & \text{if } (s_n)_{n \in \mathbb{N}} \text{ is bounded}, \\
\infty & \text{if } (s_n)_{n \in \mathbb{N}} \text{ is not bounded.}\end{cases} \tag{7.77}
\]
Proof. Since \((s_n)_{n \in \mathbb{N}}\) is increasing, (7.77) is a consequence of (7.21). \( \square \)

**Theorem 7.83.** Let \( \sum_{j=1}^{\infty} a_j \) and \( \sum_{j=1}^{\infty} b_j \) be series in \( \mathbb{C} \) such that \( |a_j| \leq |b_j| \) holds for each \( j \geq k \) for some fixed \( k \in \mathbb{N} \).

(a) If \( \sum_{j=1}^{\infty} |b_j| \) is convergent, then \( \sum_{j=1}^{\infty} a_j \) is convergent as well, and, moreover,
\[
\left| \sum_{j=1}^{\infty} a_j \right| \leq \sum_{j=1}^{\infty} |b_j|. \tag{7.78}
\]
(b) If \( \sum_{j=1}^{\infty} a_j \) is divergent, then \( \sum_{j=1}^{\infty} |b_j| \) is divergent as well.
Proof. Since (b) is merely the contraposition of (a), it suffices to prove (a). To this end, let \( s_n := \sum_{j=1}^{n} a_j \) and \( t_n := \sum_{j=1}^{n} |b_j| \) be the partial sums of \( \sum_{j=1}^{\infty} a_j \) and \( \sum_{j=1}^{\infty} |b_j| \), respectively. Since \( (t_n)_{n \in \mathbb{N}} \) converges, it must be a Cauchy sequence by Th. 7.29. Thus,

\[
\forall \epsilon \in \mathbb{R}^+ \quad \exists \quad \forall \quad t_n - t_m = |b_{m+1}| + \cdots + |b_n| < \epsilon
\]

and the triangle inequality for finite sums implies

\[
\forall \epsilon \in \mathbb{R}^+ \quad \exists \quad \forall \quad |s_n - s_m| = |a_{m+1} + \cdots + a_n| \leq |a_{m+1}| + \cdots + |a_n| \leq |b_{m+1}| + \cdots + |b_n| < \epsilon,
\]

showing \( (s_n)_{n \in \mathbb{N}} \) is a Cauchy sequence as well, i.e. convergent by Th. 7.29. Since the triangle inequality for finite sums also implies \(|\sum_{j=k}^{n} a_j| \leq \sum_{j=k}^{n} |b_j|\) for each \( n \geq k \), (7.78) is now a consequence of Th. 7.13(c).

**Definition 7.84.** A series \( \sum_{j=1}^{\infty} a_j \) in \( \mathbb{R} \) is called alternating if, and only if, its summands alternate between positive and negative signs, i.e. if \( \text{sgn}(a_{j+1}) = -\text{sgn}(a_j) \neq 0 \) for each \( j \in \mathbb{N} \).

**Theorem 7.85** (Leibniz Criterion). Let \( \sum_{j=1}^{\infty} a_j \) be an alternating series. If the sequence \((|a_n|)_{n \in \mathbb{N}}\) of absolute values is strictly decreasing and \( \lim_{n \to \infty} a_n = 0 \), then the series is convergent and

\[
\forall \quad n \in \mathbb{N} \quad \exists \quad 0 < \theta_n < 1 \quad r_n := \sum_{j=n+1}^{\infty} a_j = \theta_n a_{n+1}, \tag{7.79}
\]

that means the error made when approximating the limit by the partial sum \( s_n \) has the same sign as the first neglected summand \( a_{n+1} \), and its absolute value is less than \(|a_{n+1}|\).

Proof. We first consider the case where \( a_1 > 0 \), i.e. where there exists a strictly decreasing sequence of positive numbers \((b_n)_{n \in \mathbb{N}}\) such that \( a_n = (-1)^{n+1} b_n \). As the \( b_n \) are strictly decreasing, we obtain \( b_n - b_{n+1} > 0 \) for each \( n \in \mathbb{N} \), such that the sequences \((u_n)_{n \in \mathbb{N}}\) and \((v_n)_{n \in \mathbb{N}}\), defined by

\[
\forall \quad n \in \mathbb{N} \quad u_n := s_{2n} = \sum_{j=1}^{n} (b_{2j-1} - b_{2j}) = (b_1 - b_2) + (b_3 - b_4) + \cdots + (b_{2n-1} - b_{2n}),
\]

\[
\forall \quad n \in \mathbb{N} \quad v_n := s_{2n+1} = b_1 - \sum_{j=1}^{n} (b_{2j} - b_{2j+1})
\]

\[
= b_1 - (b_2 - b_3) - (b_4 - b_5) - \cdots - (b_{2n} - b_{2n+1}),
\]

are strictly monotone, namely \((u_n)_{n \in \mathbb{N}}\) strictly increasing and \((v_n)_{n \in \mathbb{N}}\) strictly decreasing. Since, 0 < \( u_n < u_n + b_{2n+1} = v_n < b_1 \) for each \( n \in \mathbb{N} \), both sequences \((u_n)_{n \in \mathbb{N}}\) and \((v_n)_{n \in \mathbb{N}}\) are also bounded, and, thus, convergent by Th. 7.19, i.e. \( U := \lim_{n \to \infty} u_n \in \mathbb{R} \) and \( V := \lim_{n \to \infty} v_n \in \mathbb{R} \). Since

\[
V - U = \lim_{n \to \infty} (v_n - u_n) = \lim_{n \to \infty} (s_{2n+1} - s_{2n}) = \lim_{n \to \infty} a_{2n+1} = 0,
\]
we obtain $U = V$ and $\lim_{n \to \infty} s_n = U$ and $0 < U < b_1 = a_1$. In particular, there is $\theta \in ]0, 1[\ satisfying$ $\sum_{j=1}^{\infty} a_j = \theta a_1$.

In the case $a_1 < 0$, the above proof yields convergence of $-\sum_{j=1}^{\infty} a_j = \sum_{j=1}^{\infty} (-a_j)$ with $\sum_{j=1}^{\infty} (-a_j) = \theta (-a_1)$ for a suitable $\theta \in ]0, 1[\$. However, this then yields, as before, $\sum_{j=1}^{\infty} a_j = \theta a_1$.

Applying the above result to each remainder series $\sum_{j=n+1}^{\infty} a_j$, $n \in \mathbb{N}$, completes the proof of (7.79) and the theorem. 

Example 7.86. (a) Each of the following alternating series clearly converges, as the Leibniz criterion of Th. 7.85 clearly applies in each case:

$$\sum_{j=1}^{\infty} \frac{(-1)^{j+1}}{j} = 1 - \frac{1}{2} + \frac{1}{3} - + \ldots, \quad (7.80a)$$

$$\sum_{j=1}^{\infty} \frac{(-1)^{j+1}}{2j-1} = 1 - \frac{1}{3} + \frac{1}{5} - + \ldots, \quad (7.80b)$$

$$\sum_{j=1}^{\infty} \frac{(-1)^{j+1}}{\ln(j+1)} = \frac{1}{\ln 2} - \frac{1}{\ln 3} + \frac{1}{\ln 4} - + \ldots \quad (7.80c)$$

(b) To see that Th. 7.85 is false without its monotonicity requirement, take any divergent series with $\sum_{j=1}^{\infty} a_j = \infty, 0 < a_j, \lim_{j \to \infty} a_j = 0$ (for example the harmonic series), any convergent series with $\sum_{j=1}^{\infty} c_j = s \in \mathbb{R}^+$ and $0 < c_j$ (for example any geometric series with $0 < q < 1$), and define

$$d_n := \begin{cases} a(n+1)/2 & \text{for } n \text{ odd}, \\ -c_{n/2} & \text{for } n \text{ even}. \end{cases}$$

It is an exercise to show that $\sum_{j=1}^{\infty} d_j$ is an alternating series with $\lim_{n \to \infty} d_n = 0$ and $\sum_{j=1}^{\infty} d_j = \infty$.

Definition 7.87. The series $\sum_{j=1}^{\infty} a_j$ in $\mathbb{C}$ is said to be absolutely convergent if, and only if, $\sum_{j=1}^{\infty} |a_j|$ is convergent.

Corollary 7.88. Every absolutely convergent series $\sum_{j=1}^{\infty} a_j$ is also convergent and satisfies the triangle inequality for infinite series:

$$\left| \sum_{j=1}^{\infty} a_j \right| \leq \sum_{j=1}^{\infty} |a_j|. \quad (7.81)$$

Proof. The corollary is given by the special case $a_j = b_j$ for each $j \in \mathbb{N}$ of Th. 7.83(a).

Theorem 7.89. We consider the series $\sum_{j=1}^{\infty} a_j$ in $\mathbb{C}$.

(a) If $\sum_{j=1}^{\infty} c_j$ is a convergent series such that $c_j \in \mathbb{R}_0^+$ and $|a_j| \leq c_j$ for each $j \in \mathbb{N}$, then $\sum_{j=1}^{\infty} a_j$ is absolutely convergent.
(b) Root Test:

\[ \exists 0 < q < 1 \] \quad \left( \sqrt[n]{|a_n|} \leq q \text{ for almost all } n \in \mathbb{N} \right) \Rightarrow \sum_{j=1}^{\infty} a_j \text{ is absolutely convergent}, \quad (7.82a) \]

\[ \# \left\{ n \in \mathbb{N} : \sqrt[n]{|a_n|} \geq 1 \right\} = \infty \Rightarrow \sum_{j=1}^{\infty} a_j \text{ is divergent.} \quad (7.82b) \]

(c) Ratio Test: If all \( a_n \neq 0 \), then

\[ \exists 0 < q < 1 \] \quad \left( \left| \frac{a_{n+1}}{a_n} \right| \leq q \text{ for almost all } n \in \mathbb{N} \right) \Rightarrow \sum_{j=1}^{\infty} a_j \text{ is absolutely convergent,} \quad (7.83a) \]

\[ \left| \frac{a_{n+1}}{a_n} \right| \geq 1 \text{ for almost all } n \in \mathbb{N} \Rightarrow \sum_{j=1}^{\infty} a_j \text{ is divergent.} \quad (7.83b) \]

Proof. (a) is just another special case of Th. 7.83(a).

(b): If there is \( q \in ]0,1[ \) and \( N \in \mathbb{N} \) such that \( \sqrt[n]{|a_n|} \leq q \) for each \( n > N \), i.e. \( |a_n| \leq q^n \) for each \( n > N \), then, by (7.71), \( \sum_{j=1}^{\infty} |a_j| \) is bounded by \( \frac{1}{1-q} + \sum_{j=1}^{N} |a_j| \) and, thus, convergent. If \( \sqrt[n]{|a_n|} \geq 1 \) for infinitely many \( n \in \mathbb{N} \), then \( |a_n| \geq 1 \) for infinitely many \( n \in \mathbb{N} \), showing that \( (a_n)_{n \in \mathbb{N}} \) does not converge to 0, proving the divergence of \( \sum_{j=1}^{\infty} a_j \).

(c): If there is \( q \in ]0,1[ \) and \( N \in \mathbb{N} \) such that \( \left| \frac{a_{n+1}}{a_n} \right| \leq q \) for each \( n > N \), then, letting \( C := |a_{N+1}| \), an induction shows \( |a_{N+1+k}| \leq Cq^k \) for each \( k \in \mathbb{N} \), i.e., by (7.71), \( \sum_{j=1}^{\infty} |a_j| \) is bounded by \( \frac{C}{1-q} + \sum_{j=1}^{N+1} |a_j| \) and, thus, convergent. If there is \( N \in \mathbb{N} \) such that \( \left| \frac{a_{n+1}}{a_n} \right| \geq 1 \) for each \( n > N \), then \( |a_n| \geq |a_{N+1}| > 0 \) for each \( n > N \), showing \( (a_n)_{n \in \mathbb{N}} \) does not converge to 0 and proving the divergence of \( \sum_{j=1}^{\infty} a_j \).

Caveat 7.90. In (7.82a), it does not suffice to have \( \sqrt[n]{|a_n|} < 1 \) to conclude convergence, and, likewise, \( \left| \frac{a_{n+1}}{a_n} \right| < 1 \) does not suffice in (7.83a): As a counterexample, consider the harmonic series, which does not converge, but \( \sqrt[n]{1/n} < 1 \) for each \( n \geq 2 \) and \( \frac{1/(n+1)}{1/n} = \frac{n}{n+1} < 1 \) for each \( n \in \mathbb{N} \).

Example 7.91. (a) For each \( z \in \mathbb{C} \) with \( |z| < 1 \) and each \( p \in \mathbb{N}_0 \), the series \( \sum_{n=1}^{\infty} n^p z^n \) is absolutely convergent: We have \( \lim_{n \to \infty} \sqrt[n]{n^p} = 1 \) as a consequence of Ex. 7.65. This implies \( \lim_{n \to \infty} \sqrt[n]{|a_n|} = \lim_{n \to \infty} \sqrt[n]{n^p |z|^n} = |z| < 1 \). Thus, the root test of (7.82a) applies and proves convergence of the series.
(b) Let \( z \in \mathbb{C} \). The series \( \sum_{n=1}^{\infty} \frac{z^n n!}{n^n} \) is absolutely convergent for \( |z| < e \) and divergent for \( |z| \geq e \), where \( e \) is Euler’s number from (7.49). We have, for each \( n \in \mathbb{N} \),

\[
\left| \frac{a_{n+1}}{a_n} \right| = \left| \frac{|z| (n + 1)^n}{(n + 1)^n} \right| = \frac{|z|}{(1 + \frac{1}{n})^n} \to \frac{|z|}{e} \quad \text{for} \quad n \to \infty.
\]

(7.84)

Thus, the ratio test of (7.83a) applies and proves absolute convergence of the series for \( |z| < e \). For \( |z| > e \), (7.83b) applies and proves divergence. Since, according to Ex. 7.66, \( (1 + \frac{1}{n})^n < e \) for each \( n \in \mathbb{N} \), (7.83b) applies to prove divergence also for \( |z| = e \).

### 7.3.3 Absolute Convergence and Rearrangements

In general, one has to use care when dealing with infinite series, as convergence properties and even the limit in case of convergence can depend on the order of the summands (in obvious contrast to the situation of finite sums). For real series that are convergent, but not absolutely convergent, one has the striking Riemann rearrangement Th. 7.93, that states one can choose an arbitrary number \( S \in \mathbb{R} \cup \{-\infty, \infty\} \) and reorder the summands such that the new series converges to \( S \) (actually, Th. 7.93 says even more, namely that one can prescribe an entire interval of cluster points for the rearranged series). However, the situation is better for absolutely convergent series. In Th. 7.95, we will see that the sum of absolutely convergent series does not depend on the order of the summands.

**Proposition 7.92.** Let \( \sum_{j=1}^{\infty} a_j \) be a series in \( \mathbb{R} \). Defining

\[
\forall \quad j \in \mathbb{N} \quad \left( a_j^+ := \max\{a_j, 0\}, \quad a_j^- := \max\{-a_j, 0\} \right),
\]

(7.85)

the following assertions (a) and (b) hold:

(a) \( \sum_{j=1}^{\infty} a_j \) is absolutely convergent if, and only if, both series \( \sum_{j=1}^{\infty} a_j^+ \) and \( \sum_{j=1}^{\infty} a_j^- \) are convergent.

(b) If \( \sum_{j=1}^{\infty} a_j \) is convergent, but not absolutely convergent, then

\[
\sum_{j=1}^{\infty} a_j^+ = \sum_{j=1}^{\infty} a_j^- = \infty.
\]

(7.86)

**Proof.** The key observation is that (7.85) implies, for each \( j \in \mathbb{N} \),

\[
a_j^+ + a_j^- = |a_j|, \tag{7.87a}
\]

\[
a_j^+ - a_j^- = a_j, \tag{7.87b}
\]

\[
0 \leq a_j^+ - a_j^- \leq |a_j|. \tag{7.87c}
\]

(a): If \( \sum_{j=1}^{\infty} a_j^+ \) and \( \sum_{j=1}^{\infty} a_j^- \) are convergent, then

\[
\sum_{j=1}^{\infty} |a_j| \quad \overset{(7.87a), (7.73)}{=} \quad \sum_{j=1}^{\infty} a_j^+ + \sum_{j=1}^{\infty} a_j^-.
\]

(7.88)
and, in particular, \( \sum_{j=1}^{\infty} a_j \) is absolutely convergent. Conversely, if \( \sum_{j=1}^{\infty} a_j \) is absolutely convergent, then \( \sum_{j=1}^{\infty} a_j^+ \) and \( \sum_{j=1}^{\infty} a_j^- \) are convergent by (7.87c) and Th. 7.83(a).

(b): If \( \sum_{j=1}^{\infty} a_j \) and \( \sum_{j=1}^{\infty} a_j^+ \) are convergent, then (7.87b) implies that \( \sum_{j=1}^{\infty} a_j^- \) is also convergent and, thus, \( \sum_{j=1}^{\infty} a_j \) absolutely convergent by (a). Likewise, if \( \sum_{j=1}^{\infty} a_j \) and \( \sum_{j=1}^{\infty} a_j^- \) are convergent, then (7.87b) implies that \( \sum_{j=1}^{\infty} a_j^+ \) is also convergent and, once again, \( \sum_{j=1}^{\infty} a_j \) absolutely convergent by (a). Therefore, if \( \sum_{j=1}^{\infty} a_j \) is convergent, but not absolutely convergent, then (7.86) must hold by (7.77).

**Theorem 7.93** (Riemann Rearrangement Theorem a.k.a. Riemann Series Theorem). Let \( \sum_{j=1}^{\infty} a_j \) be a series in \( \mathbb{R} \). If \( \sum_{j=1}^{\infty} a_j \) is convergent, but not absolutely convergent, then, given \( x, y \in \mathbb{R} \cup \{-\infty, \infty\} \) with \( x \leq y \), there exists a rearrangement \( \sum_{j=1}^{\infty} b_j \) of the series (i.e. a reordering \( (b_j)_{j \in \mathbb{N}} \) of \( (a_j)_{j \in \mathbb{N}} \) such that \( \sum_{j=1}^{\infty} b_j \) has precisely all elements of \([x, y]\) as cluster points (where we call \(-\infty\) (resp. \(\infty\)) a cluster point of the real sequence \((t_n)_{n \in \mathbb{N}}\) if, and only if, \#\{\(n \in \mathbb{N}: t_n < -N\} = \infty\) (resp. \#\{\(n \in \mathbb{N}: t_n > N\} = \infty\}) for each \( N \in \mathbb{N} \). In particular, choosing \( S := x = y \in \mathbb{R} \cup \{-\infty, \infty\} \), one can prescribe an arbitrary limit \( S \) such that \( \sum_{j=1}^{\infty} b_j = S \).

**Sketch of Proof.** Here we just give a sketch of the proof to convey its fairly simple idea; a detailed proof is provided in Appendix E.1. According to Prop. 7.92(b), (7.86) must hold, where the \( a_j^+ \) and \( a_j^- \) are as defined in (7.85). Thus, we can define

\[
\forall \; k \in \mathbb{N} \quad x_k := \begin{cases} 
-k & \text{for } x = -\infty, \\
x & \text{for } x \in \mathbb{R}, \\
k & \text{for } x = \infty,
\end{cases}
\]

and,

\[
y_k := \begin{cases} 
-k & \text{for } y = -\infty, \\
y & \text{for } y \in \mathbb{R}, \\
k & \text{for } y = \infty,
\end{cases}
\]

and, noting \( x_k \leq y_k \) for almost all \( k \in \mathbb{N} \), alternate between adding summands \( a_j^+ \) until the partial sum exceeds \( y_k \) and subtracting summands \( a_j^- \) until the partial sum falls below \( x_k \). If \( k \) is sufficiently large such that \( x_k \leq y_k \), then, at each switching point (from adding to subtracting or vice versa), the absolute value of the difference between the last partial sum and \( x_k \) or \( y_k \), respectively, is less than the value of the last contributing nonzero summand. Since

\[
\lim_{j \to \infty} a_j^+ = \lim_{j \to \infty} a_j^- = 0,
\]

the partial sums corresponding to the switching points converge to the respective endpoints \( x \) or \( y \), respectively, and precisely all points between \( x \) and \( y \) are cluster points. ■

We will now study the more benign situation of absolutely convergent series.

**Theorem 7.94.** Let \( \sum_{j=1}^{\infty} a_j \) and \( \sum_{j=1}^{\infty} b_j \) be series in \( \mathbb{C} \) such that \( (b_n)_{n \in \mathbb{N}} \) is a reordering of \( (a_n)_{n \in \mathbb{N}} \) in the sense of Def. 7.21. If \( \sum_{j=1}^{\infty} a_j \) is absolutely convergent, then so is \( \sum_{j=1}^{\infty} b_j \) and \( \sum_{j=1}^{\infty} a_j = \sum_{j=1}^{\infty} b_j \).

**Proof.** Let \( s_n := \sum_{j=1}^{n} a_j, \; \tilde{s}_n := \sum_{j=1}^{n} |a_j|, \) and \( t_n := \sum_{j=1}^{n} b_j \) denote the respective partial sums. We will show that \( \lim_{n \to \infty} (s_n - t_n) = 0 \). Given \( \epsilon > 0 \), since \( (\tilde{s}_n)_{n \in \mathbb{N}} \) is a Cauchy sequence by Th. 7.29, there exists \( N \in \mathbb{N} \), such that

\[
\forall \; n, m > N \quad |\tilde{s}_n - \tilde{s}_m| = |a_{m+1}| + \cdots + |a_n| < \epsilon.
\]
Since \((b_n)_{n \in \mathbb{N}}\) is a reordering of \((a_n)_{n \in \mathbb{N}}\), there exists a bijective map \(\phi : \mathbb{N} \to \mathbb{N}\) such that \(b_n = a_{\phi(n)}\) for each \(n \in \mathbb{N}\). Since \(\phi\) is bijective, there exists \(M \in \mathbb{N}\) such that \(\{1, 2, \ldots, N + 1\} \subseteq \phi\{1, 2, \ldots, M\}\). Then \(n > M\) implies \(\phi(n) > N + 1\), and

\[
\forall n \in \mathbb{N}, \exists \epsilon > 0 \text{ such that } |s_n - t_n| \leq |a_{N+2}| + \cdots + |a_{N+k}| < \epsilon,
\]

since all \(a_j\) with \(j \leq N+1\) occur in both \(s_n\) and \(t_n\) and cancel in \(s_n - t_n\) (i.e. all \(a_j\) that do not cancel must have an index \(j > N+1\)). So we have shown that \(\lim_{n \to \infty} (s_n - t_n) = 0\), which, in turn, implies

\[
\sum_{j=1}^{\infty} b_j = \lim_{n \to \infty} t_n = \lim_{n \to \infty} (t_n - s_n + s_n) = 0 + \sum_{j=1}^{\infty} a_j = \sum_{j=1}^{\infty} a_j.
\]

Applying this to \(\tilde{s}_n := \sum_{j=1}^{n} |a_j|\) yields \(\sum_{j=1}^{\infty} |b_j| = \sum_{j=1}^{\infty} |a_j|\), proving absolute convergence of \(\sum_{j=1}^{\infty} b_j\).

**Theorem 7.95.** Let \(I\) be an arbitrary infinite countable index set and let

\[
I = \bigcup_{n \in \mathbb{N}} I_n
\]

be a disjoint decomposition of \(I\) into (empty, finite, or infinite) countable index sets \(I_n\).

(a) If the series \(\sum_{j \in I} a_j\) (cf. (7.69)) is absolutely convergent, then

\[
\sum_{j \in I} a_j = \sum_{n=1}^{\infty} \sum_{\alpha \in I_n} a_{\alpha}.
\]

(b) The following statements are equivalent:

(i) \(\sum_{j \in I} a_j\) is absolutely convergent.

(ii) There exists a constant \(C \in \mathbb{R}^+_0\) such that \(\sum_{j \in J} |a_j| \leq C\) for each finite subset \(J\) of \(I\).

(iii) \(\sum_{n=1}^{\infty} \sum_{\alpha \in I_n} |a_{\alpha}| < \infty\).

**Proof.** (a): First note that Th. 7.94 implies that, for absolutely convergent \(\sum_{j \in I} a_j\), the limit \(\sum_{j \in I} a_j = \sum_{j=1}^{\infty} a_{\phi(j)}\) does not depend on the bijective map \(\phi : \mathbb{N} \to I\): For each bijective map \(\psi : \mathbb{N} \to I\), \((a_{\psi(j)})_{j \in \mathbb{N}}\) is a reordering of \((a_{\phi(j)})_{j \in \mathbb{N}}\) and, thus, \(\sum_{j=1}^{\infty} a_{\psi(j)} = \sum_{j=1}^{\infty} a_{\phi(j)}\).

Analogously, the sums \(\sum_{\alpha \in I_n} a_{\alpha}\) do not depend on the order of the indices in \(I_n\).

**Claim 1.** If \(M \subseteq I\), then \(S(I) = S(M) + S(I \setminus M)\), where \(S(J) := \sum_{j \in J} a_j\) for each \(J \subseteq I\).
Proof. If $M = \emptyset$, then there is nothing to prove. If $\#M = n \in \mathbb{N}$, then let $\phi_1 : \{1, \ldots, n\} \rightarrow M$ and $\phi_2 : \{n + 1, n + 2, \ldots\} \rightarrow I \setminus M$ be bijective maps. Then

$$\phi : \mathbb{N} \rightarrow I, \quad \phi(j) := \begin{cases} \phi_1(j) & \text{for } j \leq n, \\ \phi_2(j) & \text{for } j > n, \end{cases}$$

is a bijective map. Moreover,

$$S(I) = \sum_{j=1}^{\infty} a_{\phi(j)} = \sum_{j=1}^{n} a_{\phi(j)} + \sum_{j=n+1}^{\infty} a_{\phi(j)} = S(M) + S(I \setminus M),$$

establishing the case.

If $\#M = \#(I \setminus M) = \#\mathbb{N}$, then let $\phi_1 : \{1, 3, 5, \ldots\} \rightarrow M$ and $\phi_2 : \{2, 4, 6, \ldots\} \rightarrow I \setminus M$ be bijective maps. Then

$$\phi : \mathbb{N} \rightarrow I, \quad \phi(j) := \begin{cases} \phi_1(j) & \text{for odd } j, \\ \phi_2(j) & \text{for even } j, \end{cases}$$

is a bijective map. Define,

$$b_{\phi(j)} := \begin{cases} a_{\phi(j)} & \text{for odd } j, \\ 0 & \text{for even } j, \end{cases} \quad c_{\phi(j)} := \begin{cases} a_{\phi(j)} & \text{for even } j, \\ 0 & \text{for odd } j. \end{cases}$$

One then obtains

$$S(I) = \sum_{j=1}^{\infty} a_{\phi(j)} = \sum_{j=1}^{n} b_{\phi(j)} + \sum_{j=1}^{\infty} c_{\phi(j)} = S(M) + S(I \setminus M),$$

establishing the case. \hfill \blacksquare

Claim 2. If $I = \bigcup_{n=1}^{k} M_n$ with $k \in \mathbb{N}$ is a decomposition of $I$, then, using the notation introduced in Cl. 1, $S(I) = \sum_{n=1}^{k} S(M_n)$.

Proof. Follows by an induction from Cl. 1. \hfill \blacksquare

Coming back to (7.89), Cl. 2 implies

$$\forall k \in \mathbb{N} \left( S(I) = S(I_1) + S(I_2) + \cdots + S(I_k) + S(M_k), \quad \text{where } M_k := I \setminus \bigcup_{j=1}^{k} I_j \right).$$

To prove the equality in (7.90), fix a bijective $\phi : \mathbb{N} \rightarrow I$, and let $\epsilon > 0$. Due to Cor. 7.81(d), the sums $r_n := \sum_{j=n+1}^{\infty} |a_{\phi(j)}|$ of the remainder series converge to 0, i.e. there exists $N \in \mathbb{N}$ such that $r_n < \epsilon$ for each $n > N$. More generally, for each (empty, finite, or infinite) subset $J \subseteq \{N + 2, N + 3, \ldots\}$,

$$\sum_{j \in J} |a_{\phi(j)}| \leq \sum_{j=N+2}^{\infty} |a_{\phi(j)}| = r_{N+1} < \epsilon.$$
Next, we choose $M \in \mathbb{N}$ sufficiently large such that \( \{\phi(1), \ldots, \phi(N+1)\} \subseteq I_1 \cup \cdots \cup I_M \).

Then, for each $k > M$,\[ |S(M_k)| = \left| \sum_{j \in M_k} a_j \right| \overset{(7.81)}{\leq} \sum_{j \in M_k} |a_j| \leq \sum_{j=N+2}^{\infty} |a_{\phi(j)}| = r_{N+1} < \epsilon, \]
proving
\[ \sum_{j \in I} a_j = S(I) = \lim_{k \to \infty} \sum_{n=1}^{k} S(I_n) = \sum_{n=1}^{\infty} \sum_{\alpha \in I_n} a_\alpha, \]
which is (7.90).

(b): (i) implies (ii) with $C := \sum_{j \in I} |a_j|$ using Cl. 1 (with $a_j$ replaced by $|a_j|$). (i) implies (iii) using (7.90) (with $a_j$ replaced by $|a_j|$). (ii) implies (i) via (7.77), as $C$ is an upper bound for \( \sum_{j=1}^{n} |a_{\phi(j)}| \) for each bijection $\phi : \mathbb{N} \to I$. Finally, (iii) implies (ii) with $C := \sum_{n=1}^{\infty} \sum_{\alpha \in I_n} |a_\alpha|$, since, given a finite $J \subseteq I$, there exists $k \in \mathbb{N}$ such that $J \subseteq I_1 \cup \cdots \cup I_k$, i.e.
\[ \sum_{j \in J} |a_j| \leq \sum_{n=1}^{k} \sum_{\alpha \in I_n} |a_\alpha| \leq \sum_{n=1}^{\infty} \sum_{\alpha \in I_n} |a_\alpha| = C, \]
thereby completing the proof. \[\blacksquare\]

Example 7.96. We apply Th. 7.95 to so-called double series, i.e. to series with index set $I := \mathbb{N} \times \mathbb{N}$. The following notation is common:
\[ \sum_{m,n=1}^{\infty} a_{mn} := \sum_{(m,n) \in \mathbb{N} \times \mathbb{N}} a_{(m,n)}, \tag{7.91} \]
where one writes $a_{mn}$ (also $a_{m,n}$) instead of $a_{(m,n)}$. Recall from Th. 3.16 that $\mathbb{N} \times \mathbb{N}$ is countable. In general, the convergence properties of the double series and, if it exists, the value of the sum, will depend on the chosen bijection $\phi : \mathbb{N} \to I$.

However, we will now assume our double series to be absolutely convergent. Then Th. 7.94 guarantees the sum does not depend on the chosen bijection and we can apply Th. 7.95. Applying Th. 7.95 to the decompositions
\[ \mathbb{N} \times \mathbb{N} = \bigcup_{m \in \mathbb{N}} \{ (m,n) : n \in \mathbb{N} \}, \tag{7.92a} \]
\[ \mathbb{N} \times \mathbb{N} = \bigcup_{n \in \mathbb{N}} \{ (m,n) : m \in \mathbb{N} \}, \tag{7.92b} \]
\[ \mathbb{N} \times \mathbb{N} = \bigcup_{k \in \mathbb{N}} \{ (m,n) \in \mathbb{N} \times \mathbb{N} : m + n = k \}, \tag{7.92c} \]
yields

\[
\sum_{(m,n) \in \mathbb{N} \times \mathbb{N}} a_{(m,n)} = \sum_{m=1}^{\infty} \sum_{n=1}^{\infty} a_{mn} = \sum_{n=1}^{\infty} \sum_{m=1}^{\infty} a_{mn} = \sum_{k=2}^{\infty} \sum_{m+n=k} a_{mn} := \sum_{k=2}^{\infty} \sum_{m=1}^{k-1} a_{m,k-m}. \tag{7.92c}
\]

**Theorem 7.97.** It is possible to compute the product of two absolutely convergent (real or complex) series \( \sum_{m=1}^{\infty} a_m \) and \( \sum_{m=1}^{\infty} b_m \) as a double series:

\[
\left( \sum_{m=1}^{\infty} a_m \right) \left( \sum_{m=1}^{\infty} b_m \right) = \sum_{m,n=1}^{\infty} a_m b_n = \sum_{k=2}^{\infty} \sum_{m=1}^{k-1} a_m b_{k-m} = \sum_{k=2}^{\infty} c_k,
\]

where \( c_k := \sum_{m=1}^{k-1} a_m b_{k-m} = a_1 b_{k-1} + a_2 b_{k-2} + \cdots + a_{k-1} b_1. \) \( \tag{7.94} \)

This form of computing the product is known as a Cauchy product.

**Proof.** We first show that \( \sum_{m,n=1}^{\infty} a_m b_n \) is absolutely convergent: By letting \( A := \sum_{m=1}^{\infty} |a_m| \) and \( B := \sum_{m=1}^{\infty} |b_m| \), we obtain

\[
\sum_{m=1}^{\infty} \sum_{n=1}^{\infty} |a_m b_n| = \sum_{m=1}^{\infty} (|a_m| B) = AB < \infty,
\]

i.e. \( \sum_{m,n=1}^{\infty} a_m b_n \) is absolutely convergent according to Th. 7.95(b)(iii). Now the second equality in (7.94) is just the third equality in (7.93), and the first equality in (7.94) also follows from (7.93):

\[
\sum_{m,n=1}^{\infty} a_m b_n = \sum_{m=1}^{\infty} \sum_{n=1}^{\infty} a_m b_n = \sum_{m=1}^{\infty} a_m \sum_{n=1}^{\infty} b_n = \left( \sum_{m=1}^{\infty} a_m \right) \left( \sum_{m=1}^{\infty} b_m \right),
\]

completing the proof. \( \square \)

Theorem 7.97 will be useful in Sec. 8.2 below.

### 7.3.4 \( b \)-Adic Representations of Real Numbers

We are mostly used to representing real numbers in the decimal system. For example, we write

\[
x = \frac{395}{3} = 131.\overline{6} = 1 \cdot 10^2 + 3 \cdot 10^1 + 1 \cdot 10^0 + \sum_{n=1}^{\infty} 6 \cdot 10^{-n}, \tag{7.95a}
\]

where

\[
\sum_{n=1}^{\infty} 6 \cdot 10^{-n} = 6 \cdot \left( \frac{1}{1 - \frac{1}{10}} - 1 \right) = 6 \cdot \frac{1}{9} = \frac{2}{3}. \tag{7.71}
\]
The decimal system represents real numbers as, in general, infinite series of decimal fractions. Digital computers represent numbers in the dual system, using base 2 instead of 10. For example, the number from (7.95a) has the dual representation

$$x = 10000011.10 = 2^7 + 2^1 + 2^0 + \sum_{n=0}^{\infty} 2^{-(2n+1)},$$  
5

(7.95b)

where it is an exercise to verify

$$\sum_{n=0}^{\infty} 2^{-(2n+1)} = \frac{2}{3}.$$  
5

Representations with base 16 (hexadecimal) and 8 (octal) are also of importance when working with digital computers. More generally, each natural number $b \geq 2$ can be used as a base.

**Definition 7.98.** Let $b \geq 2$ be a natural number.

(a) Given an integer $N \in \mathbb{Z}$ and a sequence $(d_N, d_{N-1}, d_{N-2}, \ldots)$ in $\{0, \ldots, b-1\}$, the series

$$\sum_{\nu=0}^{\infty} d_{N-\nu} b^{N-\nu}$$  
5

is called a $b$-adic series. The number $b$ is called the base or the radix, and the numbers $d_{\nu}$ are called digits.

(b) If $x \in \mathbb{R}_0^+$ is the sum of the $b$-adic series given by (7.96), than one calls the $b$-adic series a $b$-adic representation or a $b$-adic expansion of $x$.

**Theorem 7.99.** Given a natural number $b \geq 2$ and a nonnegative real number $x \in \mathbb{R}_0^+$, there exists a $b$-adic series representing $x$, i.e. there is $N \in \mathbb{Z}$ and a sequence $(d_N, d_{N-1}, d_{N-2}, \ldots)$ in $\{0, \ldots, b-1\}$ such that

$$x = \sum_{\nu=0}^{\infty} d_{N-\nu} b^{N-\nu}. $$  
5

(7.97)

If one introduces the additional requirement that $0 \neq d_N$, then each $x > 0$ has either a unique $b$-adic representation or precisely two $b$-adic representations. More precisely, for $0 \neq d_N$ and $x > 0$, the following statements are equivalent:

(i) The $b$-adic representation of $x$ is not unique.

(ii) There are precisely two $b$-adic representations of $x$.

(iii) There exists a $b$-adic representation of $x$ such that $d_n = 0$ for each $n \leq n_0$ for some $n_0 < N$.

(iv) There exists a $b$-adic representation of $x$ such that $d_n = b-1$ for each $n \leq n_0$ for some $n_0 \leq N$. 
Proof. The proof is a bit lengthy and is provided in Appendix E.2.

Example 7.100. Every natural number has precisely two decimal (i.e. 10-adic) representations. For instance,

\[ 2 = 2.0 = 1 + \sum_{n=1}^{\infty} 9 \cdot 10^{-n} = 1 + 9 \left( \frac{1}{1 - \frac{1}{10}} - 1 \right) \]

and analogously for all other natural numbers.

8 Convergence of \(K\)-Valued Functions

8.1 Pointwise and Uniform Convergence

So far we have studied the convergence of sequences in \(K\). We will now also need to study the convergence of sequences \((f_n)_{n \in \mathbb{N}}\), where each member \(f_n\) of the sequence is a function \(f_n : M \rightarrow \mathbb{K}, M \subseteq \mathbb{C}\). Here, for the first time, we encounter the situation that there exist different useful notions of convergence for such sequences.

Definition 8.1. Let \((f_n)_{n \in \mathbb{N}}\) be a sequence of functions, \(f_n : M \rightarrow \mathbb{K}, \emptyset \neq M \subseteq \mathbb{C}\).

(a) We say \((f_n)_{n \in \mathbb{N}}\) converges pointwise to \(f : M \rightarrow \mathbb{K}\) if, and only if, \(\lim_{n \to \infty} f_n(z) = f(z)\) for each \(z \in M\), i.e. if, and only if,

\[ \forall z \in M \forall \epsilon \in \mathbb{R}^+ \exists N \in \mathbb{N} \forall n > N \ |f_n(z) - f(z)| < \epsilon. \]  

(8.1)

So, in general, \(N\) in (8.1) depends on both \(z\) and \(\epsilon\).

(b) We say \((f_n)_{n \in \mathbb{N}}\) converges uniformly to \(f : M \rightarrow \mathbb{K}\) if, and only if,

\[ \forall \epsilon \in \mathbb{R}^+ \exists N \in \mathbb{N} \forall n > N \forall z \in M \ |f_n(z) - f(z)| < \epsilon. \]  

(8.2)

In (8.2), \(N\) is still allowed to depend on \(\epsilon\), but, in contrast to the situation of (8.1), not on \(z\) – in that sense, the convergence is uniform in \(z\).

Remark 8.2. It is immediate from Def. 8.1(a),(b) that uniform convergence implies pointwise convergence, but Ex. 8.3(b) below will show the converse is not true.

Example 8.3. (a) Let \(\emptyset \neq M \subseteq \mathbb{C}\) (for example \(M = [0,1]\) or \(M = B_1(0)\)), and \(f_n : M \rightarrow \mathbb{K}, f_n(z) = 1/n\) for each \(n \in \mathbb{N}\). Then, clearly, \((f_n)_{n \in \mathbb{N}}\) converges uniformly to \(f \equiv 0\).

(b) The sequence \((f_n)_{n \in \mathbb{N}}\), where \(f_n : [0,1] \rightarrow \mathbb{R}, f_n(x) := x^n\), converges pointwise, but not uniformly, to

\[ f : [0,1] \rightarrow \mathbb{R}, \quad f(x) := \begin{cases} 0 & \text{for } 0 \leq x < 1, \\ 1 & \text{for } x = 1. \end{cases} \]  

(8.3)
For $x = 1$, $\lim_{n \to \infty} x^n = \lim_{n \to \infty} 1 = 1$, and, for $0 \leq x < 1$, $\lim_{n \to \infty} x^n = 0$ by Ex. 7.6. To see that the convergence is not uniform, consider $\epsilon := \frac{1}{2}$. Then, for every $n \in \mathbb{N}$, according to the intermediate value Th. 7.57, there exists $\xi_n \in ]0, 1[\$ such that $f_n(\xi_n) = \xi_n^n = \frac{1}{2}$, i.e.

$$\forall n \in \mathbb{N} \quad |f_n(\xi_n) - f(\xi_n)| = \xi_n^n = \frac{1}{2} = \epsilon,$$

(8.4)

proving the convergence is not uniform.

**Theorem 8.4.** Let $(f_n)_{n \in \mathbb{N}}$ be a sequence of functions, $f_n : M \to K$, $\emptyset \neq M \subseteq \mathbb{C}$. If $(f_n)_{n \in \mathbb{N}}$ converges uniformly to $f : M \to K$ and all $f_n$ are continuous at $\zeta \in M$, then $f$ is also continuous at $\zeta$. In particular, if each $f_n$ is continuous, then so is $f$ (uniform limits of continuous functions are continuous).

**Proof.** Let $\epsilon > 0$. Due to the uniform convergence of $(f_n)_{n \in \mathbb{N}}$,

$$\exists m \in \mathbb{N} \quad \forall z \in M \quad |f_m(z) - f(z)| < \frac{\epsilon}{3}.$$  

(8.5)

Due to the continuity of $f_m$ in $\zeta$,

$$\exists \delta > 0 \quad \forall z \in M \cap B_\delta(\zeta) \quad |f_m(z) - f_m(\zeta)| < \frac{\epsilon}{3}.$$  

(8.6)

Thus,

$$\forall z \in M \cap B_\delta(\zeta) \quad |f(z) - f(\zeta)| \leq |f(z) - f_m(z)| + |f_m(z) - f_m(\zeta)| + |f_m(\zeta) - f(\zeta)| < 3 \cdot \frac{\epsilon}{3} = \epsilon,$$

(8.7)

proving continuity of $f$ in $\zeta$.  

\[ \blacksquare \]

### 8.2 Power Series

**Definition 8.5.** (a) In Def. 7.77, it was mentioned that series can be formed from each sequence in a set $A$, where an addition is defined. Letting $\emptyset \neq M \subseteq \mathbb{C}$, we now consider $A := \mathcal{F}(M, \mathbb{K})$, i.e. the set of functions from $M$ into $\mathbb{K}$. Then the addition on $A$ is defined according to (6.1a) and, given a sequence of functions $(f_n)_{n \in \mathbb{N}}$ in $A$, the series

$$\sum_{j=1}^{\infty} f_j := (s_n)_{n \in \mathbb{N}}$$

(8.8)

is defined as the sequence of partial sums $s_n := \sum_{j=1}^{n} f_j$.

(b) Given a sequence of functions $(f_n)_{n \in \mathbb{N}_0}$, where $f_n : \mathbb{K} \to \mathbb{K}$, $f_n(z) = a_n z^n$ with $a_n \in \mathbb{K}$, the series

$$\sum_{j=0}^{\infty} a_j z^j := \sum_{j=0}^{\infty} f_j$$

(8.9)
is called a power series and the $a_j$ are called the coefficients of the power series. Note: The notation $\sum_{j=0}^{\infty} a_j z^j$ introduced in (8.9) is very common, but not entirely correct, since one writes $a_j z^j = f_j(z)$ for the summands, even though one actually means $f_j$. Moreover, one uses the same notation if one actually does mean the series $\sum_{j=0}^{\infty} f_j(z)$ in $\mathbb{K}$, so one has to see from the context if $\sum_{j=0}^{\infty} a_j z^j$ means a series of $\mathbb{K}$-valued functions or a series of numbers.

Definition 8.6. Consider a series of $\mathbb{K}$-valued functions $\sum_{j=1}^{\infty} f_j$ as in Def. 8.5(a), in particular, $s_n := \sum_{j=1}^{n} f_j$ for each $n \in \mathbb{N}$.

(a) The series converges pointwise to $f : M \rightarrow \mathbb{K}$ if, and only if, it (i.e. $(s_n)_{n \in \mathbb{N}}$) converges pointwise in the sense of Def. 8.1(a). In that case, we use the notation

$$f = \sum_{j=1}^{\infty} f_j. \quad (8.10)$$

If (8.10) holds, then the series is sometimes called a series expansion of $f$, in particular, a power series expansion if the series happens to be a power series.

Analogous to the situation of series in $\mathbb{K}$, the notation $\sum_{j=1}^{\infty} f_j$ is also used with two different meanings – it can mean the sequence of partial sums as in (8.8) or, in the case of convergent series, the limit function as in (8.10) (cf. Caveat 7.79).

(b) The series converges uniformly to $f : M \rightarrow \mathbb{K}$ if, and only if, it converges uniformly in the sense of Def. 8.1(b).

Corollary 8.7. Consider a function series $\sum_{j=1}^{\infty} f_j$ with $f_j : M \rightarrow \mathbb{K}$, $\emptyset \neq M \subseteq \mathbb{C}$.

(a) The series converges uniformly to some $f : M \rightarrow \mathbb{K}$ if, and only if, for each $n \in \mathbb{N}$ and each $z \in M$, the remainder series $\sum_{j=n+1}^{\infty} f_j(z)$ in $\mathbb{K}$ converges to some $r_n(z) \in \mathbb{K}$ such that

$$\forall \epsilon \in \mathbb{R}^+ \exists N \in \mathbb{N} \forall n > N \forall z \in M \ |r_n(z)| < \epsilon. \quad (8.11)$$

(b) If $\sum_{j=1}^{\infty} a_j$ is a convergent series in $\mathbb{R}^+$, then the condition

$$\forall z \in M \forall j \in \mathbb{N} \ |f_j(z)| \leq a_j \quad (8.12)$$

implies uniform convergence of $\sum_{j=1}^{\infty} f_j$.

(c) If each $f_j$ is continuous in $\zeta \in M$ and the series converges uniformly to $f : M \rightarrow \mathbb{K}$, then $f$ is continuous in $\zeta$. In particular, if each $f_j$ is continuous, then $f$ is continuous.

Proof. (a): If $\sum_{j=1}^{\infty} f_j$ converges uniformly to $f$, then $f(z) = \sum_{j=1}^{\infty} f_j(z)$ holds for each $z \in M$, $r_n(z) = f(z) - s_n(z)$ for each $n \in \mathbb{N}$, $z \in M$ according to (7.76), where $s_n(z) := \sum_{j=1}^{n} f_j(z)$. Then (8.11) is just (8.2), where the $s_n$ now play the role of the $f_n$ in (8.2). Conversely, if the remainder series converge for each $z \in M$, then we can define
8 CONVERGENCE OF $\mathbb{K}$-VALUED FUNCTIONS

$f : M \rightarrow \mathbb{K}$, $f(z) := f_1(z) + r_1(z) = \sum_{j=1}^{\infty} f_j(z)$. Then, once again, $r_n(z) = f(z) - s_n(z)$ for each $n \in \mathbb{N}$, $z \in M$, and (8.11) is just (8.2), yielding the uniform convergence of the series.

(b): First, (8.12) implies each remainder series $\sum_{j=n+1}^{\infty} f_j(z)$ converges absolutely. Thus, with $r_n(z)$ as in (a),

$$\forall z \in \mathbb{M} \ |r_n(z)| \leq \sum_{j=n+1}^{\infty} |f_j(z)| \leq \sum_{j=n+1}^{\infty} a_j \rightarrow 0 \quad \text{for } n \rightarrow \infty,$$

such that (a) yields uniform convergence.

(c) is immediate from Th. 8.4. ■

Remark 8.8. Given a function series $\sum_{j=1}^{\infty} f_j$ with $f_j : M \rightarrow \mathbb{K}$, $\emptyset \neq M \subseteq \mathbb{C}$; for each $z \in M$, $\sum_{j=1}^{\infty} f_j(z)$ constitutes a series in $\mathbb{K}$. Typically, one will only have convergence of $\sum_{j=1}^{\infty} f_j(z)$ in $\mathbb{K}$ on a subset $C \subseteq M$. The series then converges pointwise in the sense of Def. 8.6(a) if all $f_j$ are restricted to $C$. It can be very difficult to determine if $\sum_{j=1}^{\infty} f_j(z)$ converges or diverges for some $z \in M$, and such investigations are often of particular interest in the context of function series. Even for power series, studying convergence can still be difficult, but the availability of the following Th. 8.9 does help to (at least partially) settle the question in many cases.

Theorem 8.9. For each power series $\sum_{j=0}^{\infty} a_j z^j$, $a_j \in \mathbb{K}$, there exists a number $r \in [0, \infty] := \mathbb{R}_0^+ \cup \{\infty\}$, called the radius of convergence of the power series, such that

$$(z \in \mathbb{K} \land |z| < r) \quad \Rightarrow \quad \sum_{j=0}^{\infty} a_j z^j \quad \text{converges absolutely in } \mathbb{K}, \quad (8.13a)$$

$$(z \in \mathbb{K} \land |z| > r) \quad \Rightarrow \quad \sum_{j=0}^{\infty} a_j z^j \quad \text{diverges in } \mathbb{K} \quad (8.13b)$$

(for $r = \infty$, (8.13a) claims absolute convergence for each $z \in \mathbb{K}$). In particular, $\sum_{j=0}^{\infty} a_j z^j$ converges pointwise in the sense of Def. 8.6(a) for each $z \in B_r(0)$ (cf. Def. 7.1(a)). Moreover,

$$\forall 0 < r_0 < r \left( \sum_{j=0}^{\infty} a_j z^j \quad \text{converges uniformly on } \overline{B}_{r_0}(0) \quad \text{(cf. Ex. 7.47(a))} \right). \quad (8.14)$$

For the radius of convergence, one has the formula

$$r = \frac{1}{L}, \quad \text{where} \quad L := \limsup_{n \rightarrow \infty} \sqrt[n]{|a_n|}. \quad (8.15)$$

In (8.15), $\limsup$ denotes the so-called limit superior, which is defined as the largest cluster point of the sequence $(\sqrt[n]{|a_n|})_{n \in \mathbb{N}}$ if the sequence is bounded (cf. Th. 7.27) and $\infty$ if the sequence is unbounded. As the limit superior can be 0 or $\infty$, we also define $1/0 := \infty$ and $1/\infty := 0$ in (8.15).
One has the simpler formula
\[ r = \lim_{n \to \infty} \left| \frac{a_n}{a_{n+1}} \right|, \tag{8.16} \]
provided all \( a_n \) are nonzero and provided the limit in (8.16) either exists in \( \mathbb{R}^+ \) or is \( \infty \).

**Proof.** For the proof of (8.15), we apply the root test from Th. 7.89(b). Here, for the root test, we have to consider the sequence \((\sqrt[n]{|a_n|z^n})_{n \in \mathbb{N}}\). As a consequence of (7.11a) and Prop. 7.26, \( \limsup_{n \to \infty}(\lambda x_n) = \lambda \limsup_{n \to \infty} x_n \) for each \( \lambda > 0 \) and each sequence \((x_n)_{n \in \mathbb{N}}\) in \( \mathbb{R} \) (with \( \lambda \cdot \infty = \infty \), this also holds if the limit superior is infinite). Thus,
\[ \limsup_{n \to \infty} \sqrt[n]{|a_n|z^n} = |z| \limsup_{n \to \infty} \sqrt[n]{|a_n|} = |z| L. \]
If \( |z| > 1/L \), then \( |z| L > 1 \) and (7.82b) applies, i.e. (8.13b) holds for \( r = 1/L \). If \( |z| < 1/L \), then \( |z| L < 1 \), and, recalling the Bolzano-Weierstrass Th. 7.27, one sees that (7.82a) applies, i.e. (8.13a) holds for \( r = 1/L \).

Next, if \( 0 < r_0 < r \), then \( \sum_{j=0}^{\infty} |a_j r_0^j| \) converges according to (8.13a). Since, for each \( z \in B_{r_0}(0) \) and each \( j \in \mathbb{N} \), we have \( |a_j z^j| \leq |a_j r_0^j| \), (8.14) is a consequence of Cor. 8.7(b).

The validity of (8.16) follows from the ratio test of Th. 7.89(c): If all \( a_n \neq 0 \) and \( z \neq 0 \), then
\[ \lim_{n \to \infty} \left| \frac{a_{n+1}z^{n+1}}{a_n z^n} \right| = |z| \lim_{n \to \infty} \left| \frac{a_{n+1}}{a_n} \right| = \frac{|z|}{\lim_{n \to \infty} \left| \frac{a_n}{a_{n+1}} \right|}. \]
If \( |z| < l := \lim_{n \to \infty} \left| \frac{a_n}{a_{n+1}} \right| \), then \( |z|/l < 1 \), i.e. (7.83a) applies, proving (8.13a) for \( r = l \).
If \( |z| > l \), then \( |z|/l > 1 \), i.e. (7.83b) applies, proving (8.13b) for \( r = l \). \( \blacksquare \)

**Corollary 8.10.** If \( \sum_{j=0}^{\infty} a_j z^j \), \( a_j \in \mathbb{K} \), is a power series with radius of convergence \( r \in ]0, \infty[ \), then the function
\[ f : B_r(0) \to \mathbb{K}, \quad f(z) := \sum_{j=0}^{\infty} a_j z^j, \tag{8.17} \]
is continuous. In particular, if \( r = \infty \), then \( f \) is continuous on \( \mathbb{K} \).

**Proof.** Each partial sum \( z \mapsto \sum_{j=0}^{n} a_j z^j \) is a polynomial, i.e. continuous on \( \mathbb{K} \). Moreover, if \( \zeta \in B_r(0) \), then the power series converges uniformly on \( M := B_{|\zeta|}(0) \) by (8.14), i.e. it is continuous at \( \zeta \in M \) by Th. 8.4. \( \blacksquare \)

**Example 8.11.** (a) For each \( \alpha \in \mathbb{R} \), the radius of convergence of \( \sum_{n=1}^{\infty} n^\alpha z^n \) is \( r = 1 \), since
\[ \limsup_{n \to \infty} \sqrt[n]{|a_n|} = \lim_{n \to \infty} \sqrt[n]{n^\alpha} = 1, \tag{8.18} \]
which, for each \( \alpha \in \mathbb{Z} \), follows from (7.46) and Th. 7.13(a), and, then, for all \( \alpha \in \mathbb{R} \) from the Sandwich Th. 7.16.
Let us investigate what can happen for \(|z| = r = 1\) for some cases: The series \(\sum_{n=1}^{\infty} z^n\) \((\alpha = 0)\) is divergent for each \(z \in \mathbb{C}\) with \(z = 1\) by the observation that \((z^n)_{n \in \mathbb{N}}\) does not converge to 0 for \(n \to \infty\) (as \(|z^n| = 1\) for each \(n \in \mathbb{N}\)); the series \(\sum_{n=1}^{\infty} n^{-1} z^n\) \((\alpha = -1)\) is the harmonic series, i.e. divergent, for \(z = 1\), but convergent for \(z = -1\) according to Ex. 7.86(a).

(b) The radius of convergence of both \(\sum_{n=0}^{\infty} \frac{z^n}{n!}\) and \(\sum_{n=0}^{\infty} \frac{z^n}{n^m}\) is \(r = \infty\) by (8.16) and (8.15), respectively, since

\[
\lim_{n \to \infty} \left| \frac{a_n}{a_{n+1}} \right| = \lim_{n \to \infty} \frac{(n+1)!}{n!} = \lim_{n \to \infty} (n+1) = \infty, \tag{8.19a}
\]

\[
\limsup_{n \to \infty} \sqrt[n]{|a_n|} = \lim_{n \to \infty} \sqrt[n]{\frac{1}{n^n}} = \lim_{n \to \infty} \frac{1}{n} = 0. \tag{8.19b}
\]

(c) The radius of convergence of \(\sum_{n=0}^{\infty} n! z^n\) is \(r = 0\) by (8.16), since

\[
\lim_{n \to \infty} \left| \frac{a_n}{a_{n+1}} \right| = \lim_{n \to \infty} \frac{n!}{(n+1)!} = \lim_{n \to \infty} \frac{1}{n+1} = 0. \tag{8.20}
\]

**Caveat 8.12.** Theorem 8.9 does not claim the uniform convergence of \(\sum_{j=0}^{\infty} a_j z^j\) on \(B_r(0)\), which is usually not true (e.g., it is an exercise to show that \(\sum_{j=0}^{\infty} z^j\) does not converge uniformly on \(B_1(0)\)). Theorem 8.9 also claims nothing about the convergence or divergence of \(\sum_{j=0}^{\infty} a_j z^j\) for \(|z| = r\), which has to be determined case by case (cf. Ex. 8.11(a)).

**Definition and Remark 8.13.** Given two power series \(p := \sum_{j=0}^{\infty} a_j z^j\) and \(q := \sum_{j=0}^{\infty} b_j z^j\) in \(\mathbb{K}\), we define their **Cauchy product**

\[
p \ast q := \sum_{j=0}^{\infty} c_j z^j, \quad \text{where} \quad c_j := \sum_{k=0}^{j} a_k b_{j-k} = a_0 b_j + a_1 b_{j-1} + \cdots + a_j b_0. \tag{8.21}
\]

Note that we have not assumed any convergence of the series so far, i.e. \(p, q,\) and \(p \ast q\) are not \(\mathbb{K}\)-valued functions, but sequences of \(\mathbb{K}\)-valued functions according to Def. 8.5 (sequences of polynomials, actually). Sometimes one also calls the Cauchy product \(p \ast q\) the **convolution** of \(p\) and \(q\).

Now, if we do assume \(p\) and \(q\) to have some nonzero radii of convergence, say \(r_p, r_q \in ]0, \infty]\), respectively, then, by (8.13a), both series are absolutely convergent for each \(z \in B_r(0)\), where \(r := \min\{r_p, r_q\}\). Thus, the functions

\[
f : B_r(0) \to \mathbb{K}, \quad f(z) := \sum_{j=0}^{\infty} a_j z^j, \quad g : B_r(0) \to \mathbb{K}, \quad g(z) := \sum_{j=0}^{\infty} b_j z^j, \tag{8.22}
\]

are well-defined, and (7.94) implies

\[
\forall_{z \in B_r(0)} f(z) g(z) = \sum_{j=0}^{\infty} c_j z^j \quad \text{with} \quad c_j \text{ as in (8.21)}. \tag{8.23}
\]
8.3 Exponential Functions

The notion of power series allows us to extend the definition of exponential functions to complex arguments:

**Definition and Remark 8.14.** We define the *exponential function*

$$\exp : \mathbb{C} \to \mathbb{C}, \quad \exp(z) := \sum_{n=0}^{\infty} \frac{z^n}{n!} = 1 + z + \frac{z^2}{2!} + \frac{z^3}{3!} + \ldots \quad (8.24)$$

From Ex. 8.11(b), we already know the radius of convergence of the power series in (8.24) is $\infty$, such that the function in (8.24) is well-defined.

For the time being, we also redefine Euler’s number as $e := \exp(1) > 1 > 0$ and, for each $x \in \mathbb{R}^+$, $\ln x := \log_{\exp(1)}(x)$. This, as well as calling the function of (8.24) exponential function, will be justified as soon as we will have proved

$$\lim_{n \to \infty} \left( 1 + \frac{1}{n} \right)^n = \sum_{n=0}^{\infty} \frac{1}{n!} = 1 + \frac{1}{1!} + \frac{1}{2!} + \frac{1}{3!} + \ldots \quad (8.25)$$

and

$$\forall x \in \mathbb{R} \quad e^x = \sum_{n=0}^{\infty} \frac{x^n}{n!} = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \ldots \quad (8.26)$$

in (8.36) of Th. 8.18 and in Th. 8.16(c) below, respectively.

**Proposition 8.15.** If a continuous function $E : \mathbb{R} \to \mathbb{R}$ satisfies

$$a := E(1) > 0 \quad \text{and} \quad \forall x,y\in\mathbb{R} \quad E(x+y) = E(x)E(y), \quad (8.27a)$$

then $f$ is an exponential function – more precisely

$$\forall x \in \mathbb{R} \quad E(x) = a^x. \quad (8.28)$$

**Proof.** First, $a = E(1) = E(0 + 1) = E(0)E(1) = E(0) a$ and $a > 0$ shows $E(0) = 1$. Then, for each $x \in \mathbb{R}$, $1 = E(0) = E(x-x) = E(x)E(-x)$, i.e. $E(-x) = (E(x))^{-1}$, showing $E(x) \neq 0$ for each $x \in \mathbb{R}$. Thus, $E(1) > 0$, the continuity of $E$, and the intermediate value Th. 7.57 imply $E(x) > 0$ for each $x \in \mathbb{R}$. Next, an induction shows

$$\forall x \in \mathbb{R} \quad \forall n\in\mathbb{N} \quad E(n \cdot x) = (E(x))^n \quad (8.29)$$

The base case is trivially true and the induction step is

$$E((n+1)x) = E(nx)E(x) \overset{\text{ind. hyp.}}{=} (E(x))^n E(x) = (E(x))^{n+1}.$$
Applying (8.29) with \( x = 1 \) shows \( E(n) = a^n \) for each \( n \in \mathbb{N} \). Applying (8.29) with \( x = 1/n, n \in \mathbb{N} \), shows \( a = E(1) = (E(1/n))^n \), i.e. \( E(1/n) = a^{1/n} \) since \( E(1/n) > 0 \). Next,
\[
\forall \ n,k \in \mathbb{N} \quad E(k/n) = (E(1/n))^k = (a^{1/n})^k = a^k,
\]
showing (8.28) holds for each \( x \in \mathbb{Q}^+ \). Then (8.28) also holds for each \( x \in \mathbb{R}^+ \), since, if \( (q_n)_{n \in \mathbb{N}} \) is a sequence in \( \mathbb{Q}^+ \) with \( \lim_{n \to \infty} q_n = x \), then the continuity of \( E \) implies
\[
a^x = \lim_{n \to \infty} a^{q_n} = \lim_{n \to \infty} E(q_n) = E(x).
\]
Finally, if \( x \in \mathbb{R}^- \), then
\[
a^x = (a^{-x})^{-1} = (E(-x))^{-1} = E(x),
\]
completing the proof that (8.28) holds for each \( x \in \mathbb{R} \).

**Theorem 8.16.** We consider the exponential function \( \exp \) as defined in (8.24). The following holds:

(a) \( \exp \) is continuous on \( \mathbb{C} \).

(b) \( \exp(z + w) = \exp(z) \exp(w) \) is valid for all \( z, w \in \mathbb{C} \).

(c) With \( e := \exp(1) \) (cf. Def. and Rem. 8.14), it is
\[
\forall \ x \in \mathbb{R} \quad e^x = \exp(x) = \sum_{n=0}^{\infty} \frac{x^n}{n!}.
\]

**Proof.** (a) holds by Cor. 8.10; for (b), we compute (using (7.94)),
\[
\forall \ z, w \in \mathbb{C} \quad \left( \exp(z) \exp(w) = \sum_{n=0}^{\infty} c_n, \right.
\]
where
\[
c_n = \sum_{j=0}^{n} \frac{z^j w^{n-j}}{j! (n-j)!} = \frac{1}{n!} \sum_{j=0}^{n} \binom{n}{j} z^j w^{n-j} = \frac{(z + w)^n}{n!},
\]
and then (c) is an immediate consequence of (a), (b), and Prop. 8.15.

**Definition 8.17.** Let \( M \subseteq \mathbb{C} \). If \( \zeta \in \mathbb{C} \) is a cluster point of \( M \), then a function \( f : M \to \mathbb{K} \) is said to tend to \( \eta \in \mathbb{K} \) (or to have the limit \( \eta \in \mathbb{K} \) for \( z \to \zeta \) (denoted by \( \lim_{z \to \zeta} f(z) = \eta \)) if, and only if, for each sequence \( (z_k)_{k \in \mathbb{N}} \) in \( M \setminus \{\zeta\} \) with \( \lim_{k \to \infty} z_k = \zeta \), the sequence \( (f(z_k))_{k \in \mathbb{N}} \) converges to \( \eta \in \mathbb{K} \), i.e.
\[
\lim_{z \to \zeta} f(z) = \eta \iff \forall \ (z_k)_{k \in \mathbb{N}} \text{ in } M \setminus \{\zeta\} \left( \lim_{k \to \infty} z_k = \zeta \Rightarrow \lim_{k \to \infty} f(z_k) = \eta \right). \tag{8.31}
\]
Theorem 8.18. We consider the exponential function $\exp$ as defined in (8.24). With $e^z := \exp(z)$ for each $z \in \mathbb{C}$ and $\ln x := \log_{\exp(1)}(x)$ for each $x \in \mathbb{R}^+$ (cf. Th. 8.16(c) and Def. and Rem. 8.14), we have the following limits:

\[
\lim_{z \to 0} \frac{e^z - 1}{z} = 1 \quad (z \in M := \mathbb{C} \setminus \{0\}), \quad (8.32)
\]

\[
\lim_{x \to 0} \frac{\ln(1 + x)}{x} = 1 \quad (x \in M := ] - 1, \infty[ \setminus \{0\}), \quad (8.33)
\]

\[
\lim_{x \to 0} \ln(1 + \xi x)^{1/2} = \xi \quad (x \in M := \{x \in \mathbb{R} : 1 + \xi x > 0\} \setminus \{0\}), \quad (8.34)
\]

\[
\lim_{x \to 0} (1 + \xi x)^{1/2} = e^\xi \quad (x \in M := \{x \in \mathbb{R} : 1 + \xi x > 0\} \setminus \{0\}), \quad (8.35)
\]

\[
\lim_{n \to \infty} \left(1 + \frac{x}{n}\right)^n = e^x = \sum_{n=0}^{\infty} \frac{x^n}{n!}, \quad (8.36)
\]

Proof. (8.32): From (8.24) and $e^z = \exp(z)$, we obtain

\[
\forall z \neq 0 \quad \frac{e^z - 1}{z} = \sum_{n=0}^{\infty} \frac{z^n}{(n+1)!} = 1 + \frac{z}{2!} + \frac{z^2}{3!} + \ldots,
\]

which, since $z \mapsto \sum_{n=0}^{\infty} \frac{z^n}{(n+1)!}$ is continuous on $\mathbb{C}$ by Cor. 8.10, implies (8.32).

(8.33): Consider the auxiliary function $f : ] - 1, \infty[ \to \mathbb{R}$, $f(x) := \ln(x + 1)$, with $f^{-1}(x) = e^x - 1$. Now, given a sequence $(x_k)_{k \in \mathbb{N}}$ in $] - 1, \infty[ \setminus \{0\}$ with $\lim_{k \to \infty} x_k = 0$, one obtains

\[
\lim_{k \to \infty} \frac{\ln(1 + x_k)}{x_k} = \lim_{k \to \infty} \frac{\ln(1 + f^{-1}(f(x_k)))}{f^{-1}(f(x_k))} = \lim_{k \to \infty} \frac{\ln(1 + e^{f(x_k)} - 1)}{e^{f(x_k)} - 1} = 1,
\]

where, in the last step, it was used that $\lim_{k \to \infty} x_k = 0$ and the continuity of $f$ implies $\lim_{k \to \infty} f(x_k) = \ln 1 = 0$.

Similarly, but simpler, one obtains (8.34) and (8.35) (exercise). Finally, for the sequence $(x_n)_{n \in \mathbb{N}}$ with $x_n := 1/n$, (8.35) implies (8.36). \[\blacksquare\]

Definition 8.19 (Exponentiation with Complex Exponents). For each $(a, z) \in \mathbb{R}^+ \times \mathbb{C}$, we define

\[
a^z := \exp(z \ln a), \quad (8.37)
\]

where $\exp$ is the function defined in (8.24). For $a = e$, (8.37) yields $e^z = \exp(z)$, i.e. (8.37) is consistent with (8.26).

Theorem 8.20. (a) The first two exponentiation rules of (7.54) still hold for each $a, b > 0$ and each $z, w \in \mathbb{C}$:

\[
a^{z+w} = a^z a^w, \quad (8.38a)
\]

\[
a^z b^w = (ab)^{z+w}. \quad (8.38b)
\]
(b) For each $a \in \mathbb{R}^+$, the exponential function
\[ f : \mathbb{C} \rightarrow \mathbb{C}, \quad f(z) := a^z, \quad (8.39a) \]
is continuous, and, for each $\zeta \in \mathbb{C}$, the power function
\[ g : \mathbb{R}^+ \rightarrow \mathbb{C}, \quad g(x) := x^{\zeta}, \quad (8.39b) \]
is continuous.

(c) The limit in (8.36) extends to complex numbers:
\[ \forall z \in \mathbb{C} \lim_{n \to \infty} \left( 1 + \frac{z}{n} \right)^n = e^z = \sum_{n=0}^{\infty} \frac{z^n}{n!}. \quad (8.40) \]

Proof. (a): We compute
\[ a^{z+w} \overset{(8.37)}{=} \exp((z + w) \ln a) = \exp(z \ln a + w \ln a) \]
Th. 8.16(b)
\[ = \exp(z \ln a) \exp(w \ln a) \overset{(8.37)}{=} a^z a^w, \]
proving (8.38a), and
\[ a^{z} b^{z} \overset{(8.37)}{=} \exp(z \ln a) \exp(z \ln b) \overset{\text{Th. 8.16(b)}}{=} \exp(z \ln a + z \ln b) \]
\[ \overset{(7.65e)}{=} \exp(z \ln(ab)) \overset{(8.37)}{=} (ab)^z, \]
proving (8.38b).

(b): The continuity of both functions follows from the continuity of $\exp$ (according to Th. 8.16(a)) and from the fact that continuity is preserved by compositions (according to Th. 7.41): The exponential function $f$, given by $f(z) = e^{z \ln a}$, is the composition of the continuous functions $z \mapsto z \ln a$ and $w \mapsto e^w$, whereas (analogous to Ex. 7.76(a)), the power function $g$, given by $g(x) = e^{\zeta \ln x}$, is the composition $g = \exp \circ (\zeta \ln)$, where $\ln$ is continuous by Cor. 7.74.

(c): Exercise. ■

8.4 Trigonometric Functions

The first “definition” of the trigonometric functions sine and cosine is the one based on geometric visualization usually given in high school: $\cos x$ and $\sin x$ are the coordinates of the point $p = (p_1, p_2) \in \mathbb{R}^2$ on the unit circle, such that $x$ is the angle measured in radian between the line segment between $(0, 0)$ and $(1, 0)$ and the line segment between $(0, 0)$ and $p$.

While this “definition” allows to obtain many important properties of sine and cosine using geometric arguments, it is not mathematically rigorous, and, for example, provides
no clue how to compute values like \( \sin 1 \). The problem is related to the fact that the angle measured in radian between the line segment between \((0, 0)\) and \((1, 0)\) and the line segment between \((0, 0)\) and \(p\) is supposed to be the length of the segment of the unit circle between \((1, 0)\) and \(p\) (taken in the counter-clockwise direction).

In the following Def. and Rem. 8.21, we will provide a mathematically rigorous definition of sine and cosine using power series, and we will then verify that the functions have the familiar properties one learns in high school. However, as the computation of lengths of curved paths is actually beyond the scope of this lecture, we will not be able to see that our sine and cosine functions are precisely the same we visualized in high school (the interested reader is referred to Ex. 1 in Sec. 5.14 of [Wal02] and to [Phi17, Ex. 3.13(b)]).

**Definition and Remark 8.21.** We define the **sine function**, denoted \( \sin \), and the **cosine function**, denoted \( \cos \) by

\[
\sin : \mathbb{C} \rightarrow \mathbb{C}, \quad \sin z := \sum_{n=0}^{\infty} \frac{(-1)^n z^{2n+1}}{(2n+1)!} = z - \frac{z^3}{3!} + \frac{z^5}{5!} - + \ldots, \tag{8.41a}
\]

\[
\cos : \mathbb{C} \rightarrow \mathbb{C}, \quad \cos z := \sum_{n=0}^{\infty} \frac{(-1)^n z^{2n}}{(2n)!} = 1 - \frac{z^2}{2!} + \frac{z^4}{4!} - + \ldots. \tag{8.41b}
\]

(a) \( \sin \) and \( \cos \) are well-defined and continuous: For both series and each \( z \in \mathbb{C} \), we can estimate the absolute value of the \( n \)th summand by the \( n \)th summand of the series for the exponential function \( e^{\|z\|} \) (cf. (8.36)), which we know to be convergent from Ex. 8.11(b). Thus, by Th. 8.9, both series in (8.41) have radius of convergence \( \infty \) and are continuous by Cor. 8.10.

(b) \( \cos : \mathbb{R} \rightarrow \mathbb{R} \) (i.e. \( \cos \big|_{\mathbb{R}} \)) has a smallest positive zero \( \alpha \in \mathbb{R}^+ \). We *define* \( \pi := 2\alpha \). One can show \( \pi \) is an irrational number (see Appendix H.2) and its first digits are \( \pi = 3.14159 \ldots \)

To see \( \cos \) has a smallest positive zero and to obtain a first (very coarse) estimate, note

\[
\forall \ x \in \mathbb{R}^+ \ \forall \ k \in \mathbb{N} \quad \left( \frac{x^k}{k!} > \frac{x^{k+1}}{(k+1)!} \iff 1 > \frac{x}{k+1} \iff k+1 > x \right),
\]

showing \( \frac{x^k}{k!} > \frac{x^{k+1}}{(k+1)!} \) holds for each \( k \geq 2 \) and each \( x \in ]0, 3[ \). In particular, the summands of the series in (8.41) converge monotonically to 0 (for \( k \geq 2 \)) and, since the series are alternating for \( x \neq 0 \), Th. 7.85 applies and (7.79) yields

\[
\forall \ 0 < x < 3 \left( f(x) := 1 - \frac{x^2}{2} < \cos x < 1 - \frac{x^2}{2} + \frac{x^4}{24} =: g(x), \quad x - \frac{x^3}{6} < \sin x < x - \frac{x^3}{6} + \frac{x^5}{120}. \right) \tag{8.42}
\]

The zeros of \( x \mapsto f(x) \) are \(-\sqrt{2}, \sqrt{2} \), i.e. \( \sqrt{2} \) is its smallest positive zero; the zeros of \( x \mapsto g(x) \) are \(-\sqrt{6 - 2\sqrt{3}}, -\sqrt{6 + 2\sqrt{3}}, \sqrt{6 - 2\sqrt{3}}, \sqrt{6 + 2\sqrt{3}}, \) i.e. \( \sqrt{6 + 2\sqrt{3}} \)
is its smallest positive zero. Thus, as \( f(0) = g(0) = 1 \), the intermediate value Th.
7.57 implies \( \cos \) has a smallest positive zero \( \alpha \) and

\[
1.4 < \sqrt{2} < \frac{\pi}{2} = \alpha < \sqrt{6 - 2\sqrt{3}} < 1.6
\]

(8.43)

**Theorem 8.22.** We have the following identities:

\[
\begin{align*}
\forall z \in \mathbb{C} & \quad \sin 0 = 0, \quad \cos 0 = 1, \quad (8.44a) \\
\forall z, w \in \mathbb{C} & \quad \sin(z + w) = \sin z \cos w + \cos z \sin w, \quad (8.44c) \\
\forall z, w \in \mathbb{C} & \quad \cos(z + w) = \cos z \cos w - \sin z \sin w, \quad (8.44d) \\
\forall z \in \mathbb{C} & \quad (\sin z)^2 + (\cos z)^2 = 1, \quad (8.44e) \\
\forall z \in \mathbb{C} & \quad \sin \left( \frac{\pi}{2} \right) = 0, \quad \sin \left( \frac{\pi}{2} \right) = 1, \quad (8.44f) \\
\forall z \in \mathbb{C} & \quad \sin \left( z + \frac{\pi}{2} \right) = \cos z, \quad (8.44g) \\
\forall z \in \mathbb{C} & \quad \sin(z + \pi) = -\sin z, \quad (8.44h) \\
\forall z \in \mathbb{C} & \quad \sin(z + 2\pi) = \sin z, \quad \cos(z + 2\pi) = \cos z, \quad (8.44i) \\
\lim_{z \to 0} \frac{\sin z}{z} = 1, \quad \lim_{z \to 0} \frac{\cos z - 1}{z^2} = -\frac{1}{2}. \quad (8.44j)
\end{align*}
\]

Identities (8.44i) can be restated as sine and cosine being periodic functions with period 2\( \pi \).

**Proof.** (8.44a) is immediate from (8.41) since, for \( z = 0 \), all summands of the sine series are 0 and all summands of the cosine series are 0, except the first one, which is \( \frac{(-1)^0 0^0}{0!} = 1 \).

(8.44b) is also immediate from (8.41), since \( (-z)^{2n+1} = (-1)^{2n+1} z^{2n+1} = -z^{2n+1} \) and \( (-z)^{2n} = (-1)^{2n} z^{2n} = z^{2n} \).

(8.44c) and (8.44d) can be verified using the Cauchy product: According to (7.94),

\[
\begin{align*}
\forall z, w \in \mathbb{C} & \quad \left( \sin z \cos w = \sum_{n=0}^{\infty} c_n, \quad \cos z \sin w = \sum_{n=0}^{\infty} d_n, \right) \\
\text{where} & \quad c_n = \sum_{j=0}^{n} \frac{(-1)^j z^{2j+1} (-1)^{n-j} w^{2(n-j)}}{(2j+1)! (2(n-j))!}, \\
& \quad d_n = \sum_{j=0}^{n} \frac{(-1)^j z^{2j} (-1)^{n-j} w^{2(n-j)+1}}{(2j)! (2(n-j)+1)!},
\end{align*}
\]
that means, for each \( z, w \in \mathbb{C} \),

\[
c_n + d_n = \sum_{j=0}^{n} \left( \frac{(-1)^n z^{2j+1}}{(2j+1)!} \frac{w^{2(n-j)}}{(2(n-j))!} \right) + \sum_{j=0}^{n} \left( \frac{(-1)^n z^{2j}}{(2j)!} \frac{w^{2(n-j)+1}}{(2(n-j)+1)!} \right)
\]

\[
= \sum_{j=0}^{n} \left( \frac{(-1)^n z^{2j+1}}{(2j+1)!} \frac{w^{2n+1-(2j+1)}}{(2n+1-(2j+1))!} \right) + \sum_{j=0}^{n} \left( \frac{(-1)^n z^{2j}}{(2j)!} \frac{w^{2n+1-2j}}{(2n+1-2j)!} \right)
\]

\[
= (-1)^n \sum_{j=0}^{2n+1} \frac{z^j w^{2n+1-j}}{j!(2n+1-j)!} = (-1)^n \sum_{j=0}^{2n+1} \left( \frac{2n+1}{j} \right) z^j w^{2n+1-j}
\]

proving (8.44c). Similarly, according to (7.94),

\[
\forall \ z, w \in \mathbb{C} \quad \begin{cases} 
\cos z \cos w = \sum_{n=0}^{\infty} c_n, \\
\sin z \sin w = \sum_{n=0}^{\infty} d_n, 
\end{cases}
\]

where

\[
c_n = \sum_{j=0}^{n} \left( \frac{(-1)^j z^{2j}}{(2j)!} \frac{(-1)^{n-j} w^{2(n-j)}}{(2(n-j))!} \right),
\]

\[
d_n = \sum_{j=0}^{n} \left( \frac{(-1)^j z^{2j+1}}{(2j+1)!} \frac{(-1)^{n-j} w^{2(n-j)+1}}{(2(n-j)+1)!} \right),
\]

that means, for each \( z, w \in \mathbb{C} \),

\[
c_0 = 1 \quad \text{and}
\]

\[
\forall n \in \mathbb{N} \quad c_n - d_{n-1} = \sum_{j=0}^{n} \left( \frac{(-1)^n z^{2j}}{(2j)!} \frac{w^{2(n-j)}}{(2(n-j))!} \right) - \sum_{j=0}^{n-1} \left( \frac{(-1)^{n-1} z^{2j+1}}{(2j+1)!} \frac{w^{2(n-1-j)+1}}{(2(n-1-j)+1)!} \right)
\]

\[
= \sum_{j=0}^{n} \left( \frac{(-1)^n z^{2j}}{(2j)!} \frac{w^{2n-2j}}{(2n-2j)!} \right) + \sum_{j=0}^{n-1} \left( \frac{(-1)^n z^{2j+1}}{(2j+1)!} \frac{w^{2n-(2j+1)}}{(2n-(2j+1))!} \right)
\]

\[
= (-1)^n \sum_{j=0}^{2n} \frac{z^j w^{2n-j}}{j!(2n-j)!} = (-1)^n \sum_{j=0}^{2n} \left( \frac{2n}{j} \right) z^j w^{2n-j}
\]

proving (8.44d).

(8.44e): One computes for each \( z \in \mathbb{C} \):

\[
(sin z)^2 + (cos z)^2 = cos z cos(-z) - sin z sin(-z) \quad \text{(8.44d)}
\]

\[
= \cos(z - z) = \cos 0 = 1.
\]

(8.44f): \( \cos \frac{\pi}{2} = 0 \) and \( \cos x > 0 \) for \( 0 \leq x < \frac{\pi}{2} \) hold according to the definition of \( \pi \) in Def. and Rem. 8.21(b). Then

\[
(sin \frac{\pi}{2})^2 = 1 - (cos \frac{\pi}{2})^2 = 1 \quad \text{and} \quad \sin \frac{\pi}{2} \Rightarrow \frac{\pi}{2} \geq \frac{(\pi/2)^3}{6} \quad \text{(8.43)}
\]

\[
> 1.4 - \frac{(1.6)^3}{6} > 0.7 > 0.
\]
(8.44g) is immediate from (8.44c), (8.44d), and (8.44f).

(8.44h): One obtains
\[
\sin \pi = \sin \left( \frac{\pi}{2} + \frac{\pi}{2} \right) = 1 \cdot 0 + 0 \cdot 1 = 0,
\]
\[
\cos \pi = \cos \left( \frac{\pi}{2} + \frac{\pi}{2} \right) = 0 \cdot 0 - 1 \cdot 1 = -1,
\]
\[
\forall z \in \mathbb{C}, \quad \sin(z + \pi) \overset{(8.44c)}{=} -\sin z + 0 = -\sin z,
\]
\[
\forall z \in \mathbb{C}, \quad \cos(z + \pi) \overset{(8.44d)}{=} -\cos z + 0 = -\cos z.
\]

(8.44i): One obtains
\[
\sin(2\pi) = \sin(\pi + \pi) \overset{(8.44c)}{=} 0 + 0 = 0,
\]
\[
\cos(2\pi) = \cos(\pi + \pi) \overset{(8.44d)}{=} (-1)(-1) - 0 = 1,
\]
\[
\forall z \in \mathbb{C}, \quad \sin(z + 2\pi) \overset{(8.44c)}{=} \sin z + 0 = \sin z,
\]
\[
\forall z \in \mathbb{C}, \quad \cos(z + 2\pi) \overset{(8.44d)}{=} \cos z - 0 = \cos z.
\]

(8.44j): One obtains
\[
\forall z \in \mathbb{C}\backslash\{0\}, \quad \begin{pmatrix}
\frac{\sin z}{z} = \sum_{n=0}^{\infty} \frac{(-1)^n z^{2n}}{(2n+1)!} = 1 - \frac{z^2}{3!} + \frac{z^4}{5!} - + \ldots, \\
\frac{\cos z - 1}{z^2} = \sum_{n=0}^{\infty} \frac{(-1)^{n+1} z^{2n}}{(2(n+1))!} = -\frac{1}{2!} + \frac{z^2}{4!} - \frac{z^4}{6!} + \ldots
\end{pmatrix}.
\]

For both series on the right-hand side and each \( z \in \mathbb{C} \), we can estimate the absolute value of each summand by the corresponding summand of the exponential series for \( e^{iz} \) (cf. (8.36)), showing they have radius of convergence \( \infty \) and are continuous by Cor. 8.10. In particular, their continuity in \( z = 0 \) proves (8.44j).

\[\blacksquare\]

**Theorem 8.23.** One has \( \sin(\mathbb{R}) = \cos(\mathbb{R}) = [-1, 1] \), i.e. the image of both sine and cosine is \([-1, 1]\). Moreover, for each \( k \in \mathbb{Z} \):

\[
\sin \text{ is strictly increasing on } \left[ -\frac{\pi}{2} + 2k\pi, \frac{\pi}{2} + 2k\pi \right], \quad (8.45a)
\]
\[
\sin \text{ is strictly decreasing on } \left[ \frac{\pi}{2} + 2k\pi, \frac{3\pi}{2} + 2k\pi \right], \quad (8.45b)
\]
\[
\cos \text{ is strictly increasing on } \left[ (2k - 1)\pi, 2k\pi \right], \quad (8.45c)
\]
\[
\cos \text{ is strictly decreasing on } \left[ 2k\pi, (2k + 1)\pi \right], \quad (8.45d)
\]

which, due to (8.44e), can be summarized (and visualized) by saying that, if \( x \) runs from \( 2k\pi \) to \( 2(k+1)\pi \), then \((\cos x, \sin x)\) runs once counterclockwise through the unit circle, starting at \((1, 0)\).
Proof. From (8.44e), we know \( \sin (\mathbb{R}) \subseteq [-1, 1] \) and \( \cos (\mathbb{R}) \subseteq [-1, 1] \). As
\[
\sin \frac{\pi}{2} (8.44f) = 1, \quad \sin \left(-\frac{\pi}{2}\right) (8.44b) = -1, \quad \cos 0 (8.44a) = 1, \quad \cos \pi (8.44h) = -1 = -\cos 0 = -1,
\]
the continuity of sine and cosine together with the intermediate value Th. 7.57 implies \( \sin (\mathbb{R}) = \cos (\mathbb{R}) = [-1, 1] \).

From (8.42), we know \( 0 < x - \frac{x^3}{6} < \sin x \) and \( \cos x < 1 - \frac{x^2}{2} + \frac{x^4}{24} < 1 \) for each \( x \in [0, \frac{\pi}{2}] \), implying
\[
\forall 0 \leq x < x + y \leq \frac{\pi}{2}, \quad \cos (x + y) = \cos x \cos y - \sin x \sin y \leq \cos x \cos y < \cos x,
\]
showing \( \cos \) is strictly decreasing on \( [0, \frac{\pi}{2}] \). Then \( \cos \) is strictly increasing on \( [-\frac{\pi}{2}, 0] \) by (8.44b), \( \sin \) is strictly increasing on \( [0, \frac{\pi}{2}] \) and strictly decreasing on \( [\frac{\pi}{2}, \pi] \) by (8.44g), implying \( \sin \) is strictly increasing on \( [-\frac{\pi}{2}, 0] \) and strictly decreasing on \( [-\pi, -\frac{\pi}{2}] \) by (8.44b), i.e. \( \sin \) is strictly increasing on \( [\frac{3\pi}{2}, 2\pi] \) and strictly decreasing on \( [\pi, \frac{3\pi}{2}] \) by (8.44i), implying \( \cos \) is strictly decreasing on \( [\frac{\pi}{2}, \pi] \) and strictly increasing on \( [-\pi, -\frac{\pi}{2}] \) by (8.44g). Since this fixes the monotonicity properties of both sine and cosine over more than one period, the general statements in (8.45) are provided by (8.44i).

We now come to important complex number relations between sine, cosine, and the exponential function.

**Theorem 8.24.** One has the following formulas, relating the (complex) sine, cosine, and exponential function:
\[
\forall z \in \mathbb{C} \quad e^{iz} = \cos z + i \sin z \quad \text{(Euler formula),} \quad (8.46a)
\]
\[
\forall z \in \mathbb{C} \quad \cos z = \frac{e^{iz} + e^{-iz}}{2}, \quad (8.46b)
\]
\[
\forall z \in \mathbb{C} \quad \sin z = \frac{e^{iz} - e^{-iz}}{2i}. \quad (8.46c)
\]

**Proof.** Let \( z \in \mathbb{C} \). For (8.46a), one computes
\[
e^{iz} (8.37), (8.24) = \sum_{n=0}^{\infty} \frac{(iz)^n}{n!} = \sum_{n=0}^{\infty} \frac{(-1)^n z^{2n}}{(2n)!} + i \frac{(-1)^n z^{2n+1}}{(2n+1)!} = \cos z + i \sin z.
\]

Then
\[
e^{iz} + e^{-iz} = \cos z + i \sin z + \cos(-z) + i \sin(-z) = 2 \cos z \quad \text{proves (8.46b), and}
\]
\[
e^{iz} - e^{-iz} = \cos z + i \sin z - \cos(-z) - i \sin(-z) = 2i \sin z \quad \text{proves (8.46c).} \]
As a first application of (8.46), we can now determine all solutions to the equation $e^z = 1$ and all zeros (if any) of exp, sin, and cos:

Theorem 8.25. The set of (complex) solutions to the equation $e^z = 1$ consists precisely of all integer multiples of $2\pi i$, the exponential function has no zeros (neither in $\mathbb{R}$ nor in $\mathbb{C}$), and the set of all (real or complex) zeros of sine and cosine consists of a discrete set of real numbers. More precisely:

\[
\exp^{-1}\{1\} = \{2k\pi i : k \in \mathbb{Z}\}, \quad (8.47a)
\exp^{-1}\{0\} = \emptyset, \quad (8.47b)
\sin^{-1}\{0\} = \{k\pi : k \in \mathbb{Z}\}, \quad (8.47c)
\cos^{-1}\{0\} = \{(2k+1)\frac{\pi}{2} : k \in \mathbb{Z}\}. \quad (8.47d)
\]

Proof. Exercise.■

Definition and Remark 8.26. We define tangent and cotangent by

\[
\tan : \mathbb{C} \setminus \cos^{-1}\{0\} \rightarrow \mathbb{C}, \quad \tan z := \frac{\sin z}{\cos z}, \quad (8.48a)
\]

\[
\cot : \mathbb{C} \setminus \sin^{-1}\{0\} \rightarrow \mathbb{C}, \quad \cot z := \frac{\cos z}{\sin z}, \quad (8.48b)
\]

respectively. Since sine and cosine are both continuous, tangent and cotangent are also both continuous on their respective domains. Both functions have period $\pi$, since, for each $z$ in the respective domains,

\[
\tan(z + \pi) = \frac{\sin(z + \pi)}{\cos(z + \pi)} = \frac{-\sin z}{-\cos z} = \tan z, \quad \cot(z + \pi) = \frac{-\cos z}{-\sin z} = \cot z. \quad (8.49)
\]

Since

\[
\lim_{n \to \infty} \sin \left(\frac{\pi}{2} - \frac{1}{n}\right) = \sin \frac{\pi}{2} = 1 \quad \land \quad \lim_{n \to \infty} \cos \left(\frac{\pi}{2} - \frac{1}{n}\right) = \cos \frac{\pi}{2} = 0
\]

\[
\land \quad \cos \left(\frac{\pi}{2} - \frac{1}{n}\right) > 0 \quad \Rightarrow \quad \lim_{n \to \infty} \tan \left(\frac{\pi}{2} - \frac{1}{n}\right) = \infty,
\]

\[
\lim_{n \to \infty} \sin \left(-\frac{\pi}{2} + \frac{1}{n}\right) = \sin \left(-\frac{\pi}{2}\right) = -1 \quad \land \quad \lim_{n \to \infty} \cos \left(-\frac{\pi}{2} + \frac{1}{n}\right) = \cos \left(-\frac{\pi}{2}\right) = 0
\]

\[
\land \quad \cos \left(-\frac{\pi}{2} + \frac{1}{n}\right) > 0 \quad \Rightarrow \quad \lim_{n \to \infty} \tan \left(-\frac{\pi}{2} + \frac{1}{n}\right) = -\infty,
\]

\[
\lim_{n \to \infty} \sin \frac{1}{n} = \sin 0 = 0 \quad \land \quad \lim_{n \to \infty} \cos \frac{1}{n} = \cos 0 = 1 \quad \land \quad \sin \frac{1}{n} > 0
\]

\[
\Rightarrow \quad \lim_{n \to \infty} \cot \frac{1}{n} = \infty,
\]
we obtain \( \tan(\mathbb{R} \setminus \cos^{-1}\{0\}) = \cot(\mathbb{R} \setminus \sin^{-1}\{0\}) = \mathbb{R} \).

For each \( k \in \mathbb{Z} \),

\[
\tan \text{ is strictly increasing on } \left[ -\frac{\pi}{2} + k\pi, \frac{\pi}{2} + k\pi \right], \tag{8.50a}
\]
\[
\cot \text{ is strictly decreasing on } \left[ k\pi, (k+1)\pi \right]. \tag{8.50b}
\]

On \( ]0, \frac{\pi}{2}[ \), sin is strictly increasing and cos is strictly decreasing, i.e. tan is strictly increasing and cot is strictly decreasing. Since \( \tan(-x) = \sin(-x)/\cos(-x) = -\tan(x) \), on \( ] -\frac{\pi}{2}, 0[ \), tan is strictly increasing and cot is strictly decreasing. Taking into account the signs of tan and cot on the respective intervals and their \( \pi \)-periodicity according to (8.49) proves (8.50).

**Definition and Remark 8.27.** Since we have seen sin to be strictly increasing on \( [-\frac{\pi}{2}, \frac{\pi}{2}] \) with image \([-1, 1]\), cos to be strictly decreasing on \([0, \pi]\) with image \([-1, 1]\), tan to be strictly increasing on \( ] -\frac{\pi}{2}, \frac{\pi}{2} [\) with image \( \mathbb{R} \), and cot to be strictly decreasing on \( ]0, \pi[ \) with image \( \mathbb{R} \); and since all four functions are continuous, Th. 7.60 implies the existence of inverse functions, denoted by

\[
\text{arcsin} : [-1, 1] \rightarrow [-\pi/2, \pi/2], \tag{8.51a}
\]
\[
\text{arccos} : [-1, 1] \rightarrow [0, \pi], \tag{8.51b}
\]
\[
\text{arctan} : \mathbb{R} \rightarrow ]-\pi/2, \pi/2[, \tag{8.51c}
\]
\[
\text{arccot} : \mathbb{R} \rightarrow ]0, \pi[, \tag{8.51d}
\]

respectively, where all four inverse functions are continuous, arcsin is strictly increasing, arccos is strictly decreasing, arctan is strictly increasing, and arccot is strictly decreasing.

Of course, using (8.45) and (8.50), respectively, one can also obtain the inverse functions on different intervals, and, in the literature, such inverse functions are, indeed, considered as well. Somewhat confusingly, it is common to denote all these different functions by the same symbols, namely the ones introduced in (8.51). Here, we will not need to pursue this any further, i.e. we will only consider the inverse functions precisely as defined in (8.51), which are also known as the *principle* inverse functions of sin, cos, tan, and cot, respectively.

### 8.5 Polar Form of Complex Numbers, Fundamental Theorem of Algebra

**Theorem 8.28.** For each complex number \( z \in \mathbb{C} \), there exist real numbers \( r \geq 0 \) and \( \varphi \in \mathbb{R} \) such that

\[
z = r e^{i\varphi}. \tag{8.52}
\]
Moreover, if \( (8.52) \) holds with \( r \geq 0 \) and \( \varphi \in \mathbb{R} \), then \( r \) is the modulus of \( z \) and, for \( z \neq 0 \), \( \varphi \) is uniquely determined up to addition of an integer multiple of \( 2\pi \), i.e.

\[
\forall z \in \mathbb{C} \setminus \{0\} \quad \left( z = re^{i\varphi_1} = re^{i\varphi_2} \land r \geq 0 \Rightarrow r = |z| \land \exists k \in \mathbb{Z} \quad \varphi_1 - \varphi_2 = 2\pi k \right). \quad (8.53)
\]

**Proof.** For \( z = 0 \), there is nothing to prove, so we assume \( z \neq 0 \) and set \( r := |z| \). We write \( z = x + iy \) with \( x, y \in \mathbb{R} \), first assuming \( y \geq 0 \). Then

\[
\frac{z}{r} = \xi + i\eta, \quad \text{where} \quad \xi = \frac{x}{r}, \quad \eta = \frac{y}{r} \geq 0, \quad \xi^2 + \eta^2 = 1. \quad (8.54)
\]

In particular, \( -1 \leq \xi \leq 1 \). Thus, letting

\[ \varphi := \arccos \xi, \]

we obtain \( \varphi \in [0, \pi] \), \( \xi = \cos \varphi \), and \( \sin \varphi \geq 0 \), yielding

\[ \sin \varphi = \sqrt{1 - (\cos \varphi)^2} = \sqrt{1 - \frac{\xi^2}{\xi^2}} = \eta. \]

In consequence,

\[ \frac{z}{r} = \xi + i\eta = \cos \varphi + i \sin \varphi = e^{i\varphi}, \]

as desired. If \( y \leq 0 \), then the above shows the existence of \( \psi \in \mathbb{R} \) such that \( z = x - iy = re^{i\psi} = r \cos \psi + ir \sin \psi \). Letting \( \varphi := -\psi \), we, once again, have \( z = r \cos \psi - ir \sin \psi = re^{-i\psi} = re^{i\varphi} \), as desired, completing the existence proof for the representation \((8.52)\).

Now assume \((8.52)\) holds with \( r \geq 0 \). Then

\[ |z| = r|e^{i\varphi}| = r\sqrt{(\sin \varphi)^2 + (\cos \varphi)^2} = r. \]

Finally, if \( r e^{i\varphi_1} = r e^{i\varphi_2} \) with \( r > 0 \), then \( e^{i(\varphi_1 - \varphi_2)} = 1 \), i.e. \( i(\varphi_1 - \varphi_2) \in \{2k\pi i : k \in \mathbb{Z}\} \) by \((8.47a)\).

**Definition and Remark 8.29.** The representation of \( z \in \mathbb{C} \) given by \((8.52)\) is called its **polar form**, where \((r, \varphi)\) are also called **polar coordinates** of \( z \), \( \varphi \) is called an **argument** of \( z \). For \( z \neq 0 \), one can fix the argument uniquely by the additional requirement \( \varphi \in [0, 2\pi[ \) (but one also finds other choices, for example \( \varphi \in ]-\pi, \pi] \), in the literature). The above terminology is consistent with the common use of calling \((r, \varphi)\) **polar coordinates** of the vector \( z = (x, y) \in \mathbb{R}^2(= \mathbb{C}) \) (in contrast to the **Cartesian coordiantes** \((x, y)\)), where \( r \) constitutes the distance of the point \( z = (x, y) \) from the origin \((0, 0)\) and \( \varphi \) is the angle between the vector \( z = (x, y) \) and the \( x \)-axis (cf. the three introductory paragraphs of the previous Sec. 8.4). As promised, we can now better understand the geometric interpretation of complex multiplication already described in Rem. 5.10: If \( z_1 = r_1 e^{i\varphi_1} \) and \( z_2 = r_2 e^{i\varphi_2} \), then \( z_1 z_2 = r_1 r_2 e^{i(\varphi_1 + \varphi_2)} \), i.e. complex multiplication, indeed, means multiplying absolute values and adding arguments.

**Corollary 8.30.** If \( z \in \mathbb{C} \), then \(|z| = 1\) holds if, and only if, there exists \( \varphi \in \mathbb{R} \) such that \( z = e^{i\varphi} \) — in other words, the map

\[
 f : \mathbb{R} \longrightarrow \{ z \in \mathbb{C} : |z| = 1 \}, \quad f(\varphi) := e^{i\varphi}, \quad (8.55)
\]

is surjective. Moreover \( f(\varphi_1) = f(\varphi_2) \) holds if, and only if, \( \varphi_1 - \varphi_2 = 2\pi k \) for some \( k \in \mathbb{Z} \).
Proof. Everything is immediate from Th. 8.28. \hfill \blacksquare

**Corollary 8.31** (Roots of Unity). For each \( n \in \mathbb{N} \), the equation \( z^n = 1 \) has precisely \( n \) distinct solutions \( \zeta_1, \ldots, \zeta_n \in \mathbb{C} \), where

\[
\forall_{k=1,\ldots,n} \quad \zeta_k := e^{\frac{2k\pi i}{n}} = \cos \frac{k2\pi}{n} + i \sin \frac{k2\pi}{n} = \zeta^k_1. \quad (8.56)
\]

The numbers \( \zeta_1, \ldots, \zeta_n \) defined in (8.56) are called the \( n \)th roots of unity.

**Proof.** It is \( \zeta^n_k = e^{2k\pi i} = 1 \) for each \( k \in \{1, \ldots, n\} \) and the \( \zeta_1, \ldots, \zeta_n \) are all distinct by Cor. 8.30, since, for \( k, l \in \{1, \ldots, n\} \) with \( k \neq l \), \( (k-l)/n \notin \mathbb{Z} \). As \( \zeta_1, \ldots, \zeta_n \) are \( n \) distinct zeros of the polynomial \( P : \mathbb{C} \rightarrow \mathbb{C}, P(z) := z^n - 1 \), and \( P \) has at most \( n \) zeros by Th. 6.6(a), \( \zeta_1, \ldots, \zeta_n \) constitute all solutions to \( z^n = 1 \). \hfill \blacksquare

We are now in a position to prove one of the central results of analysis and algebra, namely the **fundamental theorem of algebra**. The following proof does not need any tools beyond the ones provided by this class – it is actually mainly founded on continuous functions attaining a min and a max on compact sets according to Th. 7.54 and the existence of \( n \)th roots of unity according to Cor. 8.31.

**Theorem 8.32** (Fundamental Theorem of Algebra). *Every polynomial \( P : \mathbb{C} \rightarrow \mathbb{C}, P(z) := \sum_{j=0}^n a_j z^j \), of degree \( n \geq 1 \) (i.e. \( a_0, \ldots, a_n \in \mathbb{C} \) with \( a_n \neq 0 \)) has at least one zero \( z_0 \in \mathbb{C} \).*

**Proof.** Dividing the equation \( P(z) = 0 \) by \( a_n \neq 0 \), it suffices to consider the case \( a_n = 1 \). We therefore assume

\[
\forall_{z \in \mathbb{C}} \quad P(z) = z^n + a_{n-1} z^{n-1} + \cdots + a_1 z + a_0.
\]

**Claim 1.** The function \( |P| \) attains its global min on \( \mathbb{C} \), i.e. there exists \( z_0 \in \mathbb{C} \) such that \(|P(x)|\) is minimal in \( x = z_0 \).

**Proof.** We first note

\[
\forall_{z \neq 0} \quad P(z) = z^n (1 + r(z)), \quad \text{where} \quad r(z) := \frac{a_{n-1}}{z} + \cdots + \frac{a_0}{z^n}.
\]

Set \( M := |a_0| + \cdots + |a_{n-1}| \) and \( R := \max\{1, 2M\} \). Then

\[
\forall_{|z| \geq R} \quad |r(z)| \leq \frac{|z|^{n-1} M}{|z|} \leq \frac{|z|^{2M}}{2} \leq \frac{1}{2}.
\]

and, thus,

\[
\forall_{|z| \geq R} \quad |P(z)| = |z|^n |1 + r(z)| \geq \frac{|z|^n}{2} \geq M.
\]

This estimate together with \(|P(0)| = |a_0| \leq M\) shows that the min of \(|P|\) on the compact disk \( \overline{B}_R(0) \) (see Ex. 7.47(a)) (such a min \( z_0 \in \overline{B}_R(0) \) exists due to Th. 7.54) must be the global min of \(|P|\) on \( \mathbb{C} \). \hfill \blacksquare
Claim 2. If $|P|$ has a min in $z_0 \in \mathbb{C}$, then $P(z_0) = 0$.

Proof. Proceeding by contraposition, we assume $P(z_0) \neq 0$ and show that $|P|$ does not have a min in $z_0$. We need to construct $z_1 \in \mathbb{C}$ such that $|P(z_1)| < |P(z_0)|$. To this end, define

$$p : \mathbb{C} \rightarrow \mathbb{C}, \quad p(z) := \frac{P(z_0 + z)}{P(z_0)}.$$ 

Then $p$ is still a polynomial of degree $n$. Since $p(0) = 1$,

$$\exists_{k \in \{1, \ldots, n\}} \exists_{b_k, \ldots, b_n \in \mathbb{C}} \forall_{z \in \mathbb{C}} p(z) = 1 + \sum_{j=k}^{n} b_j z^j, \quad b_k \neq 0.$$ 

Write $-b_k^{-1}$ in polar form, i.e. $-b_k^{-1} = re^{i\varphi}$ with $r \in \mathbb{R}^+$ and $\varphi \in \mathbb{R}$. Define

$$\beta := \sqrt[r]{r} e^{i\varphi/k} \quad (\text{i.e. } \beta^k = re^{i\varphi} = -b_k^{-1})$$

and

$$q : \mathbb{C} \rightarrow \mathbb{C}, \quad q(z) := p(\beta z) = 1 + b_k \beta^k z^k + \sum_{j=k+1}^{n} b_j \beta^j z^j = 1 - z^k + z^{k+1} S(z),$$

where $S$ is the polynomial

$$S : \mathbb{C} \rightarrow \mathbb{C}, \quad S(z) := \sum_{j=0}^{n-k-1} b_{k+1+j} \beta^{k+1+j} z^j \quad (S \equiv 0 \text{ in case } k = n).$$

Then, according to Th. 7.54,

$$\exists_{C \in \mathbb{R}^+} \forall_{z \in \mathbb{B}_1(0)} |S(z)| \leq C.$$ 

Letting

$$c := \min\{1, C^{-1}\},$$

one obtains

$$\forall_{0 < |z| < c} |z^{k+1} S(z)| \leq C |z|^{k+1} < |z|^k$$

and, thus,

$$\forall_{x \in [0, c]} |q(x)| \leq 1 - x^k + |x^{k+1} S(x)| < 1 - x^k + x^k = 1.$$ 

Thus, finally,

$$\forall_{x \in [0, c]} \frac{|P(z_0 + \beta x)|}{|P(z_0)|} = |p(\beta x)| = |q(x)| < 1,$$

showing $|P|$ does not have a min in $z_0$. \hfill \Box

Combining Claims 1 and 2 completes the proof of the theorem. \hfill \blacksquare
Corollary 8.33. (a) For every polynomial \( P : \mathbb{C} \rightarrow \mathbb{C} \) of degree \( n \geq 1 \), there exist numbers \( c, \zeta_1, \ldots, \zeta_n \in \mathbb{C} \) such that

\[
P(z) = c \prod_{j=1}^{n} (z - \zeta_j) = c(z - \zeta_1)(z - \zeta_2) \cdots (z - \zeta_n)
\]  
(8.57)

(the \( \zeta_1, \ldots, \zeta_n \) are precisely all the zeros of \( P \), some or all of which might be identical).

(b) For every polynomial \( P : \mathbb{R} \rightarrow \mathbb{R} \) of degree \( n \geq 1 \), there exist numbers \( n_1, n_2 \in \mathbb{N}_0 \) and \( c, \xi_1, \ldots, \xi_{n_1}, \alpha_1, \ldots, \alpha_{n_2}, \beta_1, \ldots, \beta_{n_2} \in \mathbb{R} \) such that

\[
n = n_1 + 2n_2,
\]  
(8.58a)

and

\[
P(x) = c \prod_{j=1}^{n_1} (x - \xi_j) \prod_{j=1}^{n_2} (x^2 + \alpha_j x + \beta_j).
\]  
(8.58b)

Proof. For (a), one just combines Th. 8.32 with Rem. 6.7.

(b): If \( P \) has only real coefficients, then we can take complex conjugates to obtain

\[
P(\zeta) = 0 \Rightarrow \overline{P(\zeta)} = P(\overline{\zeta}) = 0,
\]  
(8.59)

showing that the nonreal zeros of \( P \) (if any) must occur in conjugate pairs. Moreover,

\[
(x - \zeta)(x - \overline{\zeta}) = x^2 - (\zeta + \overline{\zeta})x + \zeta\overline{\zeta} = x^2 - 2x \text{ Re } \zeta + |\zeta|^2,
\]  
(8.60)

showing that (8.57) implies (8.58). \( \blacksquare \)

9 Differential Calculus

9.1 Definition of Differentiability and Rules

The basic idea of differential calculus is to locally approximate nonlinear functions \( f \) by linear functions. In our case, \( f \) will be defined on a subset \( M \) of \( \mathbb{R} \) and, given \( \xi \in M \) and \( \mathbb{R} \)-valued \( f \), we will investigate the question if we can define a number \( f'(\xi) \in \mathbb{R} \) that represents the slope of the graph of \( f \) at \( \xi \) such that the line through \( \xi \) with slope \( f'(\xi) \) (called the tangent of \( f \) in \( \xi \)) can be considered as a local approximation of the graph of \( f \).

If such a local approximation of \( f \) in \( \xi \) is at all reasonable, then, for \( x \neq \xi \),

\[
f(x) - f(\xi) \overline{x - \xi}
\]
should provide “good” approximations of \( f' (\xi) \) if \( x \) tends to \( \xi \). This leads to the following Def. 9.1, where we also allow \( \mathbb{C} \)-valued functions (while the above-described geometric interpretation only works for \( \mathbb{R} \)-valued functions, it can be applied to both the real and the imaginary parts of a \( \mathbb{C} \)-valued function, cf. Rem. 9.2 below); but note that we do not consider differentiability of functions \( f : \mathbb{C} \rightarrow \mathbb{C} \), which would lead to the notion of complex differentiability or holomorphicity, which is studied in the field of Complex Analysis and is beyond the scope of this class.

**Definition 9.1.** Let \( a < b, f : ]a, b[ \rightarrow \mathbb{K} \ (a = -\infty, b = \infty \text{ is admissible}), \) and \( \xi \in ]a, b[ \). Then \( f \) is said to be **differentiable** at \( \xi \) if, and only if, the following limit in (9.1) exists in the sense of Def. 8.17 (where \( x \mapsto \frac{f(x) - f(\xi)}{x - \xi} \) plays the role of \( x \mapsto f(x) \) in Def. 8.17).

The limit is then called the **derivative** of \( f \) in \( \xi \). Many symbols are used in the literature to denote derivatives, the following provides a selection:

\[
\frac{df}{d\xi} := \frac{df(\xi)}{d\xi} := \lim_{x \to \xi} \frac{f(x) - f(\xi)}{x - \xi} = \lim_{h \to 0} \frac{f(\xi + h) - f(\xi)}{h}. \tag{9.1}
\]

Note both limits occurring in (9.1) are, indeed, identical, since the sequence \((x_k)_{k \in \mathbb{N}}\) in \( ]a, b[ \) converges to \( \xi \) if, and only if, the sequence \((h_k)_{k \to \infty}\) with \( h_k := x_k - \xi \) converges to 0. The number in (9.1) (if it exists) is also called a **differential quotient**, whereas \( \frac{f(x) - f(\xi)}{x - \xi} \) is known as a **difference quotient**.

\( f \) is called **differentiable** if, and only if, it is differentiable at each \( \xi \in ]a, b[ \). In that case, one calls the function

\[
f' : ]a, b[ \rightarrow \mathbb{K}, \quad x \mapsto f'(x), \tag{9.2}\]

the **derivative** of \( f \).

**Remark 9.2.** In the situation of Def. 9.1, the complex-valued function \( f : ]a, b[ \rightarrow \mathbb{C} \) is differentiable at \( \xi \in ]a, b[ \) if, and only if, both functions \( \text{Re } f, \text{Im } f : ]a, b[ \rightarrow \mathbb{R} \) are differentiable, and, in that case

\[
f' (\xi) = (\text{Re } f)' (\xi) + i \text{ (Im } f)' (\xi). \tag{9.3}\]

Indeed, we merely have to note

\[
\forall \underbrace{x, \xi \in ]a, b[}_{x \neq \xi}, \quad \frac{f(x) - f(\xi)}{x - \xi} = \frac{\text{Re } f(x) - \text{Re } f(\xi)}{x - \xi} + \frac{\text{Im } f(x) - \text{Im } f(\xi)}{x - \xi}. \tag{9.4}\]

and that, by (7.2) a sequence \((z_n)_{n \in \mathbb{N}}\) in \( \mathbb{C} \) converges to \( \zeta \in \mathbb{C} \) if, and only if, both \( \lim_{n \to \infty} \text{Re } z_n = \text{Re } \zeta \) and \( \lim_{n \to \infty} \text{Im } z_n = \text{Im } \zeta \) hold.

**Definition 9.3.** If \( f : ]a, b[ \rightarrow \mathbb{R} \) as in Def. 9.1 is differentiable at \( \xi \in ]a, b[ \), then the graph of the affine function

\[
L : \mathbb{R} \rightarrow \mathbb{R}, \quad L(x) := f(\xi) + f'(\xi)(x - \xi), \tag{9.5}\]

i.e. the line through \((\xi, f(\xi))\) with slope \( f'(\xi) \) is called the **tangent** to the graph of \( f \) at \( \xi \).
Theorem 9.4. If \( f : [a, b] \rightarrow \mathbb{K} \) as in Def. 9.1 is differentiable at \( \xi \in [a, b] \), then it is continuous at \( \xi \). In particular, if \( f \) is everywhere differentiable, then it is everywhere continuous.

Proof. Let \((x_k)_{k \in \mathbb{N}}\) be a sequence in \([a, b]\setminus\{\xi\}\) such that \(\lim_{k \to \infty} x_k = \xi\). Then

\[
\lim_{k \to \infty} \left( f(x_k) - f(\xi) \right) = \lim_{k \to \infty} \frac{(x_k - \xi)(f(x_k) - f(\xi))}{x_k - \xi} = 0 \cdot f'(\xi) = 0, \tag{9.6}
\]

proving the continuity of \( f \) in \( \xi \).

Example 9.5. (a) For each \( a, b \in \mathbb{K} \), the affine function \( f : \mathbb{R} \rightarrow \mathbb{K}, f(x) := ax + b \), is differentiable with \( f'(x) = a \) for each \( x \in \mathbb{R} \); if \( x \in \mathbb{R} \) and \((h_k)_{k \in \mathbb{N}}\) is a sequence with \( h_k \neq 0 \) such that \(\lim_{k \to \infty} h_k = 0\), then

\[
\lim_{k \to \infty} \frac{f(x + h_k) - f(x)}{h_k} = \lim_{k \to \infty} \frac{a(x + h_k) + b - ax - b}{h_k} = \lim_{k \to \infty} \frac{ah_k}{h_k} = a. \tag{9.7}
\]

In particular, each constant function \( f \equiv b \) has derivative \( f' \equiv 0 \).

(b) For each \( c \in \mathbb{K} \), the function \( f : \mathbb{R} \rightarrow \mathbb{K}, f(x) := e^{cx} \), is differentiable with \( f'(x) = ce^{cx} \) for each \( x \in \mathbb{R} \) (in particular, \( c = 1 \) yields \( f'(x) = e^x \) for \( f(x) = e^x \), and \( c = \ln a \) yields \( f'(x) = (\ln a) a^x \) for \( f(x) = a^x = e^{\ln a} a \in \mathbb{R}^+ \)); the case \( c = 0 \) was treated in (a). Thus, let \( c \neq 0 \). If \( x \in \mathbb{R} \) and \((h_k)_{k \in \mathbb{N}}\) is a sequence with \( h_k \neq 0 \) such that \(\lim_{k \to \infty} h_k = 0\), then

\[
\lim_{k \to \infty} \frac{f(x + h_k) - f(x)}{h_k} = \lim_{k \to \infty} \frac{e^{cx+ch_k} - e^{cx}}{h_k} = e^{cx} \lim_{k \to \infty} \frac{e^{ch_k} - 1}{ch_k} = ce^{cx}. \tag{9.8}
\]

(c) The sine and the cosine function \( f, g : \mathbb{R} \rightarrow \mathbb{K}, f(x) := \sin x, g(x) := \cos x \), are differentiable with \( f'(x) = \cos x \) and \( g'(x) = -\sin x \) for each \( x \in \mathbb{R} \); if \( x \in \mathbb{R} \) and \((h_k)_{k \in \mathbb{N}}\) is a sequence with \( h_k \neq 0 \) such that \(\lim_{k \to \infty} h_k = 0\), then

\[
\lim_{k \to \infty} \frac{f(x + h_k) - f(x)}{h_k} = \lim_{k \to \infty} \frac{\sin(x + h_k) - \sin x}{h_k} \equiv (8.44c)
\]

\[
= \lim_{k \to \infty} \frac{\sin x \cos h_k + \cos x \sin h_k - \sin x}{h_k}
\]

\[
= \sin x \lim_{k \to \infty} \frac{h_k(\cos h_k - 1)}{h_k^2} + \cos x \lim_{k \to \infty} \frac{\sin h_k}{h_k}
\]

\[
= (8.44i) \quad (\sin x) \cdot 0 \cdot \left(-\frac{1}{2}\right) + (\cos x) \cdot 1 = \cos x. \tag{9.9}
\]

The proof of \( g'(x) = -\sin x \) is left as an exercise.

(d) The absolute value function \( f : \mathbb{R} \rightarrow \mathbb{R}, f(x) := |x| \), is not differentiable at \( \xi = 0 \):

\[
\lim_{n \to \infty} \frac{f(0 + \frac{1}{n}) - f(0)}{\frac{1}{n}} = \lim_{n \to \infty} 1 = 1, \tag{9.10a}
\]

\[
\lim_{n \to \infty} \frac{f(0 - \frac{1}{n}) - f(0)}{-\frac{1}{n}} = \lim_{n \to \infty} \frac{1}{-\frac{1}{n}} = -1, \tag{9.10b}
\]
showing that \( \frac{f(0+h) - f(0)}{h} \) does not have a limit for \( h \to 0 \).

**Remark 9.6.** As just seen in Ex. 9.5(d), the absolute value function shows continuous functions do not need to be differentiable. Somewhat surprisingly, with a bit more effort, one can even construct continuous functions \( f : \mathbb{R} \to \mathbb{R} \) that are not differentiable at any \( x \in \mathbb{R} \) (see Appendix J.1).

**Theorem 9.7.** Let \( a < b, f, g :]a, b[ \to K \) (\( a = -\infty, b = \infty \) is admissible), and \( \xi \in ]a, b[ \). Assume \( f \) and \( g \) are differentiable at \( \xi \).

(a) For each \( \lambda \in K \), \( \lambda f \) is differentiable at \( \xi \) and (\( \lambda f \)'(\( \xi \)) = \( \lambda f'(\xi) \)).

(b) \( f + g \) is differentiable at \( \xi \) and (\( f + g \)'(\( \xi \)) = \( f'(\xi) + g'(\xi) \)).

(c) **Product Rule:** \( fg \) is differentiable at \( \xi \) and (\( fg \)'(\( \xi \)) = \( f'(\xi)g(\xi) + f(\xi)g'(\xi) \)).

(d) **Quotient Rule:** If \( g(\xi) \neq 0 \), then \( f/g \) is differentiable at \( \xi \) and

\[
(f/g)'(\xi) = \frac{f'(\xi)g(\xi) - f(\xi)g'(\xi)}{(g(\xi))^2}, \quad \text{in particular} \quad (1/g)'(\xi) = -\frac{g'(\xi)}{(g(\xi))^2}.
\]

**Proof.** Let \( (h_k)_{k \in \mathbb{N}} \) be a sequence with \( h_k \neq 0 \) such that \( \lim_{k \to \infty} h_k = 0 \).

For (a), one computes

\[
\lim_{k \to \infty} \frac{(\lambda f)(\xi + h_k) - (\lambda f)(\xi)}{h_k} = \lim_{k \to \infty} \frac{\lambda f(\xi + h_k) - \lambda f(\xi)}{h_k} = \lambda \lim_{k \to \infty} \frac{f(\xi + h_k) - f(\xi)}{h_k} = \lambda f'(\xi).
\]

For (b), one computes

\[
\lim_{k \to \infty} \frac{(f + g)(\xi + h_k) - (f + g)(\xi)}{h_k} = \lim_{k \to \infty} \frac{f(\xi + h_k) - f(\xi) + g(\xi + h_k) - g(\xi)}{h_k} = f'(\xi) + g'(\xi).
\]

For (c), one computes

\[
\lim_{k \to \infty} \frac{(fg)(\xi + h_k) - (fg)(\xi)}{h_k} = \lim_{k \to \infty} \frac{f(\xi + h_k)g(\xi + h_k) - f(\xi)g(\xi + h_k) + f(\xi)g(\xi + h_k) - f(\xi)g(\xi)}{h_k} = f'(\xi)g(\xi) + f(\xi)g'(\xi),
\]

where, in the last equality, we used the continuity of \( g \) in \( \xi \) according to Th. 9.4.
For (d), one first proves the special case \( f \equiv 1 \) by

\[
\lim_{k \to \infty} \frac{(1/g)(\xi + h_k) - (1/g)(\xi)}{h_k} = \lim_{k \to \infty} \frac{g(\xi) - g(\xi + h_k)}{g(\xi + h_k)g(\xi)h_k} = -\frac{g'(\xi)}{(g(\xi))^2},
\]

which implies the general case using (c):

\[
(f/g)'(\xi) = \left( f \cdot \frac{1}{g} \right)'(\xi) = \frac{f'(\xi)}{g(\xi)} - \frac{f(\xi)g'(\xi)}{(g(\xi))^2} = \frac{f'(\xi)g(\xi) - f(\xi)g'(\xi)}{(g(\xi))^2},
\]

completing the proof.

\[\blacksquare\]

**Example 9.8. (a)** Each polynomial is differentiable and the derivative is, again, a polynomial. More precisely,

\[
P : \mathbb{R} \to \mathbb{K}, \quad P(x) = \sum_{j=0}^{n} a_j x^j, \quad a_j \in \mathbb{K}
\]

implies

\[
P' : \mathbb{R} \to \mathbb{K}, \quad P'(x) = \sum_{j=1}^{n} j a_j x^{j-1}:
\]

The cases \( n = 0, 1 \) are provided by Ex. 9.5(a). To complete the induction proof of (9.11), we carry out the induction step for each \( n \in \mathbb{N} \): Writing \( P(x) = \sum_{j=0}^{n} a_j x^j + a_{n+1} x^{n+1} \) and applying the induction hypothesis as well as the rules of Th. 9.7 yields

\[
P'(x) = \sum_{j=1}^{n} j a_j x^{j-1} + a_{n+1}(1 \cdot x^n + x \cdot n \cdot x^n) = \sum_{j=1}^{n+1} j a_j x^{j-1},
\]

which establishes the case.

(b) Clearly, the derivatives of rational functions \( P/Q \) with polynomials \( P \) and \( Q \) can be computed from (9.11) and the quotient rule of Th. 9.7(d).

(c) The functions \( \tan \) and \( \cot \) as defined in (8.48) and restricted to \( \mathbb{R} \setminus \cos^{-1}\{0\} \) and \( \mathbb{R} \setminus \sin^{-1}\{0\} \), respectively, are differentiable and one obtains

\[
\tan' : \underbrace{\mathbb{R} \setminus \cos^{-1}\{0\}}_{\mathbb{R} \setminus \{2k+1\pi/2, k \in \mathbb{Z}\}} \to \mathbb{R}, \quad \tan' x = \frac{1}{(\cos x)^2} = 1 + (\tan x)^2, \quad (9.12a)
\]

\[
\cot' : \underbrace{\mathbb{R} \setminus \sin^{-1}\{0\}}_{\mathbb{R} \setminus \{k \pi : k \in \mathbb{Z}\}} \to \mathbb{R}, \quad \cot' x = -\frac{1}{(\sin x)^2} = -(1 + (\cot x)^2) : \quad (9.12b)
\]

One merely needs the derivatives of \( \sin \) and \( \cos \) from Ex. 9.5(c) and the quotient rule of Th. 9.7(d):

\[
\tan' x = \frac{\cos x \cos x - \sin x(-\sin x)}{(\cos x)^2} = \frac{1}{(\cos x)^2} = 1 + (\tan x)^2, \quad (8.44e)
\]

\[
\cot' x = -\frac{\sin x \sin x - \cos x \cos x}{(\sin x)^2} = -\frac{1}{(\sin x)^2} = -(1 + (\cot x)^2), \quad (8.44e)
\]
Theorem 9.9 (Derivative of Inverse Functions). Let \( a < b, I := ]a, b[ \) (\( a = -\infty, b = \infty \) is admissible). If \( f : I \longrightarrow \mathbb{R} \) is differentiable and strictly increasing (resp. decreasing), then \( f \) has a continuous, strictly increasing (resp. decreasing) inverse function \( f^{-1} \) defined on the interval \( J := f(I) \), i.e. \( f^{-1} : J \longrightarrow I \), and, for each \( \xi \in I \) with \( f'(\xi) \neq 0 \), \( f^{-1} \) is differentiable at \( \eta := f(\xi) \) with

\[
(f^{-1})'(\eta) = \frac{1}{f'(\xi)} = \frac{1}{f'(f^{-1}(\eta))}.
\]

(9.13)

Proof. As a differentiable function, \( f \) is continuous by Th. 9.4, i.e. Th. 7.60 provides all the present assertions, except differentiability at \( \eta \) and (9.13). Let \( (y_k)_{k \in \mathbb{N}} \) be a sequence in \( J \setminus \{\eta\} \) such that \( \lim_{k \to \infty} y_k = \eta \). Then, as \( f^{-1} \) is bijective and continuous, \( (f^{-1}(y_k))_{k \in \mathbb{N}} \) is a sequence in \( I \setminus \{\xi\} \) such that \( \lim_{k \to \infty} f^{-1}(y_k) = \xi \), and one obtains

\[
\lim_{k \to \infty} \frac{f^{-1}(y_k) - f^{-1}(\eta)}{y_k - \eta} = \lim_{k \to \infty} \frac{f^{-1}(y_k) - f^{-1}(\eta)}{f(f^{-1}(y_k)) - f(f^{-1}(\eta))} = \frac{1}{f'(f^{-1}(\eta))},
\]

(9.14)

establishing the case. \( \square \)

Example 9.10. (a) The function \( \ln : \mathbb{R}^+ \longrightarrow \mathbb{R} \) is differentiable and, for each \( x \in \mathbb{R}^+ \), \( \ln' x = 1/x \): If \( f(x) = e^x \), then \( f'(x) = e^x \neq 0 \) for each \( x \in \mathbb{R} \), \( \ln x = f^{-1}(x) \), and (9.13) yields

\[
\ln' x = \frac{1}{f'(\ln x)} = \frac{1}{e^{\ln x}} = \frac{1}{x}.
\]

(b) The function \( \arcsin ]-1,1[ \longrightarrow ]-\pi/2, \pi/2[ \) is differentiable and, for each \( x \in ]-1,1[ \), \( \arcsin' x = 1/\sqrt{1-x^2} \): If \( f(x) = \sin x \), then \( f'(x) = \cos x \neq 0 \) for each \( x \in ]-\pi/2, \pi/2[ \), \( \arcsin x = f^{-1}(x) \), and (9.13) yields

\[
\arcsin' x = \frac{1}{f'(\arcsin x)} = \frac{1}{\cos \arcsin x} = \frac{1}{\sqrt{1-(\sin \arcsin x)^2}} = \frac{1}{\sqrt{1-x^2}},
\]

where, at (*), it was used that \( \cos^2 x = 1 - \sin^2 x \) and \( \cos t > 0 \) for each \( t \in ]-\pi/2, \pi/2[ \).

(c) The function \( \arccos ]-1,1[ \longrightarrow ]0, \pi[ \) is differentiable and, for each \( x \in ]-1,1[ \), \( \arccos' x = -1/\sqrt{1-x^2} \): If \( f(x) = \cos x \), then \( f'(x) = -\sin x \neq 0 \) for each \( x \in ]0, \pi[ \), \( \arccos x = f^{-1}(x) \), and (9.13) yields

\[
\arccos' x = \frac{1}{f'(\arccos x)} = \frac{1}{-\sin \arccos x} = \frac{1}{\sqrt{1-(\cos \arccos x)^2}} = \frac{1}{\sqrt{1-x^2}},
\]

where, at (*), it was used that \( \sin^2 x = 1 - \cos^2 x \) and \( \sin t > 0 \) for each \( t \in ]0, \pi[ \).

(d) The function \( \arctan : \mathbb{R} \longrightarrow ]-\pi/2, \pi/2[ \) is differentiable and, for each \( x \in \mathbb{R} \), \( \arctan' x = 1/(1+x^2) \): Apply Th. 9.9 with \( f(x) = \tan x \) as an exercise.

(e) The function \( \arccot : \mathbb{R} \longrightarrow ]0, \pi[ \) is differentiable and, for each \( x \in \mathbb{R} \), \( \arccot' x = -1/(1+x^2) \): Apply Th. 9.9 with \( f(x) = \cot x \) as an exercise.
Theorem 9.11 (Chain Rule). Let \( a < b, \ c < d, \ f : ]a, b[ \rightarrow \mathbb{R}, \ g : ]c, d[ \rightarrow \mathbb{K}, \ f \circ g \) be differentiable in \( ]a, b[ \subseteq ]c, d[ \) \((a, c = -\infty; \ b, d = \infty \) is admissible). If \( f \) is differentiable in \( \xi \in ]a, b[ \) and \( g \) is differentiable in \( f(\xi) \in ]c, d[ \), then \( g \circ f : ]a, b[ \rightarrow \mathbb{K} \) is differentiable in \( \xi \) and

\[
(g \circ f)'(\xi) = f'(\xi)g'(f(\xi)).
\]

Proof. Let \( \eta := f(\xi) \) and define the auxiliary function

\[
\tilde{g} : ]c, d[ \rightarrow \mathbb{K}, \quad \tilde{g}(x) := \begin{cases} \frac{g(x)-g(\eta)}{x-\eta} & \text{for } x \neq \eta, \\ g'(x) & \text{for } x = \eta. \end{cases}
\]

Then

\[
\lim_{x \to \xi} \frac{g(f(x)) - g(\xi)}{x - \xi} = \lim_{k \to \infty} \frac{\tilde{g}(f(x_k))(f(x_k) - f(\xi))}{x_k - \xi} = \lim_{k \to \infty} \tilde{g}(f(x_k)) \lim_{k \to \infty} \frac{f(x_k) - f(\xi)}{x_k - \xi} = f'(\xi)g'(f(\xi)),
\]

establishing the case. \hspace{1cm} \blacksquare

Example 9.12. (a) According to the chain rule of Th. 9.11, the function \( h : \mathbb{R} \rightarrow \mathbb{R}, \ h(x) := \sin(-x^3) \) is differentiable and, for each \( x \in \mathbb{R}, \ h'(x) = -3x^2\cos(-x^3) \).

(b) According to the chain rule of Th. 9.11, each power function \( h : \mathbb{R}^+ \rightarrow \mathbb{K}, \ h(x) := x^\alpha = e^{\alpha \ln x}, \ \alpha \in \mathbb{K}, \) is differentiable and, for each \( x \in \mathbb{R}^+, \ h'(x) = \frac{\alpha}{x} e^{\alpha \ln x} = \alpha x^{\alpha - 1} \). Indeed, \( h = g \circ f \), where \( f : \mathbb{R}^+ \rightarrow \mathbb{R}, \ f(x) := \ln x \) with \( f' : \mathbb{R}^+ \rightarrow \mathbb{R}^+ \), \( f'(x) := \frac{1}{x} \), according to Ex. 9.5(b), and \( g : \mathbb{R} \rightarrow \mathbb{K}, \ g(x) := e^{\alpha x} \) with \( g' : \mathbb{R} \rightarrow \mathbb{K}, \ g'(x) := \alpha e^{\alpha x} \) according to Ex. 9.10(a).

9.2 Higher Order Derivatives and the Sets \( C^k \)

Definition 9.13. Let \( a < b, \ I := ]a, b[, \ f : I \rightarrow \mathbb{K} \ (a = -\infty, \ b = \infty \) is admissible). If \( f \) is differentiable, then \( f' \) might or might not itself be differentiable. If \( f' \) is differentiable, then its derivative is denoted by \( f'' \) and is called the second derivative of \( f \). Clearly, this process can be iterated, leading to the following general recursive definition of higher-order derivatives:

Let \( f^{(0)} := f \). For \( k \in \mathbb{N}_0 \) assume the \( k \)th derivative of \( f \), denoted by \( f^{(k)} \) exists on \( I \). Then \( f \) is said to have a derivative of order \( k + 1 \) at \( \xi \in I \) if, and only if, \( f^{(k)} \) is differentiable at \( \xi \). In that case, define

\[
f^{(k+1)}(\xi) := (f^{(k)})'(\xi).
\]
A simple induction shows, for each polynomial $f$ local min or max in $\xi$.

Proof. Theorem 9.15. 9.3 Mean Value Theorem, Monotonicity, and Extrema

Remark 9.16. For $f : R \rightarrow R$, $f(x) := x^3$, it is $f'(0) = 0$, but $f$ does not have a local min or max at 0, showing that, while being necessary for an differentiable function $f$ to have a local extremum at $\xi$, $f'(\xi) = 0$ is not a sufficient condition for such an extremum at $\xi$. Points $\xi$ with $f'(\xi) = 0$ are sometimes called stationary or critical points of $f$. 

If $f^{(k+1)}(\xi)$ exists for all $\xi \in I$, then $f$ is said to be $(k+1)$-times differentiable and the function $f^{(k+1)} : I \rightarrow K$, $x \mapsto f^{(k+1)}(\xi)$, is called the $(k+1)$st derivative of $f$. It is common to write $f' := f^{(1)}$, $f'' := f^{(2)}$, $f''' := f^{(3)}$, but $f^{(k)}$ if $k \geq 4$.

If $f^{(k)}$ exists, it might or might not be continuous (cf. Ex. 9.14(c) below). One defines

$$\forall k \in N_0 \quad C^k(I, K) := \left\{ f \in F(I, K) : f^{(k)} \text{ exists and is continuous on } I \right\}, \quad (9.20)$$

$$C^\infty(I, K) := \bigcap_{k \in N_0} C^k(I, K) \quad (9.21)$$

(note $C^0(I, K) = C(I, K)$ and $C(I, K) \supseteq C^1(I, K) \supseteq C^2(I, K) \supseteq \ldots$). Finally, we define the notation $C^k(I) := C^k(I, R)$ for $k \in N_0 \cup \{\infty\}$.

Example 9.14. (a) One has $\sin \in C^\infty(R)$ with $\sin' = \cos$, $\sin'' = -\sin$, $\sin''' = -\cos$, $\sin^{(4)} = \sin$, \ldots

(b) A simple induction shows, for each polynomial $P : R \rightarrow K$, $P(x) = \sum_{j=0}^n a_j x^j$, $a_j \in K$, $n \in N_0$, that $P^{(n)}(x) = n! a_n$. In particular, $P \in C^\infty(R, K)$.

(c) It is an exercise to show the following function $f$ is differentiable, but $f'$ is not continuous, i.e. $f \notin C^1(R)$:

$$f : R \rightarrow R, \quad f(x) := \begin{cases} x^2 \cos \left(\frac{1}{x}\right) & \text{for } x \neq 0, \\ 0 & \text{for } x = 0. \end{cases}$$

9.3 Mean Value Theorem, Monotonicity, and Extrema

Theorem 9.15. Let $a < b$. If $f : ]a, b[ \rightarrow R$ is differentiable in $\xi \in ]a, b[\text{ and } f$ has a local min or max in $\xi$, then $f'(\xi) = 0$.

Proof. Suppose $f$ has a local max at $\xi$. Then there exists $\varepsilon > 0$ such that $|h| < \varepsilon$ implies $f(\xi + h) - f(\xi) \leq 0$. Now let $(h_k)_{k \in N}$ be a sequence in $]0, \epsilon[\text{ with } \lim_{k \to \infty} h_k = 0$. Then $f(\xi + h_k) - f(\xi) \leq 0$ for all $k \in N$ implies

$$f'(\xi) = \lim_{k \to \infty} \frac{f(\xi + h_k) - f(\xi)}{h_k} \leq 0, \quad f'(\xi) = \lim_{k \to \infty} \frac{f(\xi - h_k) - f(\xi)}{-h_k} \geq 0, \quad (9.22)$$

showing $f'(\xi) = 0$. Now, if $f$ has a local min at $\xi$, then $-f$ has a local max at $\xi$, and $f'(\xi) = -(-f)'(\xi) = 0$ establishes the case.

Remark 9.16. For $f : R \rightarrow R$, $f(x) := x^3$, it is $f'(0) = 0$, but $f$ does not have a local min or max at 0, showing that, while being necessary for an differentiable function $f$ to have a local extremum at $\xi$, $f'(\xi) = 0$ is not a sufficient condition for such an extremum at $\xi$. Points $\xi$ with $f'(\xi) = 0$ are sometimes called stationary or critical points of $f$. 

---
Now, we first prove an important special case of the mean value theorem:

**Theorem 9.17** (Rolle’s Theorem). Let \( a < b \). If \( f : [a, b] \to \mathbb{R} \) is continuous on the compact interval \([a, b]\), differentiable on the open interval \([a, b]\), and \( f(a) = f(b) \), then there exists \( \xi \in ]a, b[ \) such that \( f'(\xi) = 0 \).

Proof. If \( f \) is constant, then \( f'(\xi) = 0 \) holds for each \( \xi \in ]a, b[ \). If \( f \) is nonconstant, then there exists \( x \in ]a, b[ \) with \( f(x) \neq f(a) \). If \( f(x) > f(a) \), then Th. 7.54 implies the existence of \( \xi \in ]a, b[ \) such that \( f \) attains its (global and, thus, local) max in \( \xi \). Then Th. 9.15 yields \( f'(\xi) = 0 \). The case \( f(x) < f(a) \) is treated analogously. ■

**Theorem 9.18** (Mean Value Theorem). Let \( a < b \). If \( f, g : [a, b] \to \mathbb{R} \) are continuous on the compact interval \([a, b]\), differentiable on the open interval \([a, b]\), and \( g'(x) \neq 0 \) for each \( x \in ]a, b[ \), then there exists \( \xi \in ]a, b[ \) such that

\[
\frac{f(b) - f(a)}{g(b) - g(a)} = \frac{f'(\xi)}{g'(\xi)}.
\] (9.23a)

In the special case that \( g : [a, b] \to \mathbb{R}, g(x) = x \), one obtains the standard form

\[
\frac{f(b) - f(a)}{b - a} = f'(\xi).
\] (9.23b)

Proof. First note that Rolle’s Th. 9.17 and \( g' \neq 0 \) imply \( g(b) - g(a) \neq 0 \). Next, one applies Rolle’s Th. 9.17 to the auxiliary function

\[
h : [a, b] \to \mathbb{R}, \quad h(x) := f(x) - (g(x) - g(a)) \frac{f(b) - f(a)}{g(b) - g(a)}.
\] (9.24)

Since \( f \) and \( g \) are continuous on \([a, b]\) and differentiable on \([a, b]\), so is \( h \). Moreover, \( h(a) = f(a) = h(b) \), i.e. Rolle’s Th. 9.17 applies and yields \( \xi \in ]a, b[ \) satisfying \( h'(\xi) = 0 \). However, (9.24) implies \( h'(\xi) = 0 \) is equivalent to (9.23a). ■

**Corollary 9.19.** Let \( c < d \) and \( f : [c, d[ \to \mathbb{R} \) be differentiable (\( c = -\infty \), \( d = \infty \) is admissible).

(a) If \( f' \geq 0 \) (resp. \( f' \leq 0 \)), then \( f \) is increasing (resp. decreasing). Moreover, if the inequalities are strict, then the monotonicity of \( f \) is strict as well.

(b) If \( f' \equiv 0 \), then \( f \) is constant.

Proof. If \( c < a < b < d \) and \( f' \geq 0 \) (resp. \( f' \leq 0 \), resp. \( f' \equiv 0 \)), then (9.23b) implies \( f(b) \geq f(a) \) (resp. \( f(b) \leq f(a) \), resp. \( f(b) = f(a) \)). Moreover, strict inequalities for \( f' \) yield strict inequality between \( f(b) \) and \( f(a) \). ■

**Lemma 9.20.** Let \( a < b \), \( f : ]a, b[ \to \mathbb{R}, \xi \in ]a, b[, \) and assume \( f \) is differentiable at \( \xi \). If \( f'(\xi) > 0 \) (resp. \( f'(\xi) < 0 \)), then there exists \( \epsilon > 0 \) such that \( ]\xi - \epsilon, \xi + \epsilon[ \subseteq ]a, b[ \) and

\[
\forall a_1 \in ]\xi - \epsilon, \xi[ \quad \forall b_1 \in ]\xi, \xi + \epsilon[ \quad f(a_1) < f(\xi) < f(b_1) \quad (\text{resp.} \quad f(a_1) > f(\xi) > f(b_1)).
\]
Proof. If there does not exist $\epsilon > 0$ such that $f(a_1) < f(\xi) < f(b_1)$ for each $a_1 \in ]\xi - \epsilon, \xi[ \: \text{and each } b_1 \in ]\xi, \xi + \epsilon[,$ then there exists a sequence $(x_k)_{k \in \mathbb{N}}$ in $]a, b[ \setminus \{\xi\}$ such that $\lim_{k \to \infty} x_k = \xi$ and

$$\forall k \in \mathbb{N} \quad \frac{f(x_k) - f(\xi)}{x_k - \xi} \leq 0,$$

showing $f'(\xi) \leq 0$. Analogously, one obtains that $f'(\xi) \geq 0$ provided there does not exist $\epsilon > 0$ such that $f(a_1) > f(\xi) > f(b_1)$ for each $a_1 \in ]\xi - \epsilon, \xi[ \: \text{and each } b_1 \in ]\xi, \xi + \epsilon[.$

Caveat 9.21. The hypotheses of Lem. 9.20 are not sufficient for $f$ to be increasing or decreasing in any neighborhood of $\xi$: It is an exercise to find a counterexample.

**Theorem 9.22** (Sufficient Conditions for Extrema). Let $c < d$, let $f : ]c, d[ \longrightarrow \mathbb{R}$ be differentiable, and assume $f'(\xi) = 0$ for some $\xi \in ]c, d[.$

(a) If $f'(x) > 0$ for each $x \in ]c, \xi[$ and $f'(x) < 0$ for each $x \in ]\xi, d[,$ then $f$ has a strict max at $\xi.$ Likewise, if $f''(\xi)$ exists and is negative, then $f$ has a strict max at $\xi.$

(b) If $f'(x) < 0$ for each $x \in ]c, \xi[$ and $f'(x) > 0$ for each $x \in ]\xi, d[,$ then $f$ has a strict min at $\xi.$ Likewise, if $f''(\xi)$ exists and is positive, then $f$ has a strict min at $\xi.$

Proof. We just present the proof for (a); (b) is proved analogously. If $f'(x) > 0$ for each $x \in ]c, \xi[,$ then (9.23b) shows $f(\xi) - f(a) > 0$ for each $c < a < \xi;$ analogously, if $f'(x) < 0$ for each $x \in ]\xi, d[,$ then (9.23b) shows $f(\xi) - f(b) > 0$ for each $\xi < b < d.$ Altogether, we have shown $f$ to have a strict max at $\xi.$ If $f''(\xi)$ exists and is negative, then Lem. 9.20 yields the existence of $\epsilon > 0$ such that $f'$ is positive on $]\xi - \epsilon, \xi[ \: \text{and negative on } ]\xi, \xi + \epsilon[.$ Applying what we have already proved with $c := \xi - \epsilon$ and $d := \xi + \epsilon$ establishes the case.

**Example 9.23.** One obtains

$$f : \mathbb{R} \longrightarrow \mathbb{R}, \quad f(x) := x e^x,$$

$$f' : \mathbb{R} \longrightarrow \mathbb{R}, \quad f'(x) = e^x + x e^x = (1 + x) e^x,$$

$$f'' : \mathbb{R} \longrightarrow \mathbb{R}, \quad f''(x) = 2 e^x + x e^x = (2 + x) e^x.$$

From Th. 9.15, we know that $f$ can have at most one extremum, namely at $\xi = -1,$ where $f'(\xi) = 0.$ Since $f''(\xi) = e^{-x} > 0,$ Th. 9.22(b) implies that $f$ has a strict min at $-1.$

The following Th. 9.24, the intermediate value theorem for derivatives, is another application of Lem. 9.20. Even though Ex. 9.14(c) shows that derivatives do not have to be continuous, not every function can occur as a derivative: The following Th. 9.24 shows that derivatives always satisfy an intermediate value property, even if they are not continuous (that also means that discontinuities in derivatives are always due to oscillations rather than jumps).
**Theorem 9.24** (Intermediate Value Theorem for Derivatives). Let $a, b, c, d \in \mathbb{R}$ with $a < c < d < b$. If $f : ]a, b[ \rightarrow \mathbb{R}$ is differentiable, then $f'$ assumes every value between $f'(c)$ and $f'(d)$, i.e.

$$\left[ \min\{f'(c), f'(d)\}, \max\{f'(c), f'(d)\} \right] \subseteq f'([c, d]). \quad (9.26)$$

**Proof.** Exercise (hint: use a suitable auxiliary function and apply Lem. 9.20 together with Th. 7.54 and Th. 9.15).

### 9.4 L'Hôpital's Rule

L'Hôpital's rule is a result that can help to determine (function) limits (cf. Def. 8.17).

**Theorem 9.25** (L'Hôpital’s Rule). Let $\xi \in \mathbb{R}$ and either $I = ]a, \xi[ \, \text{with} \, a < \xi$ or $I := ]\xi, b[ \, \text{with} \, \xi < b$. Moreover, assume $f, g : I \rightarrow \mathbb{R}$ are differentiable, $g'(x) \neq 0$ for each $x \in I$, and one of the following two conditions (a), (b) is satisfied:

(a) $\lim_{x \to \xi} f(x) = \lim_{x \to \xi} g(x) = 0$.

(b) $\lim_{x \to \xi} g(x) = \infty$ or $\lim_{x \to \xi} g(x) = -\infty$, where Def. 8.17 is extended to the case $\eta \in \{-\infty, \infty\}$ in the obvious way.

Then

$$\lim_{x \to \xi} \frac{f'(x)}{g'(x)} = \eta \quad \Rightarrow \quad \lim_{x \to \xi} \frac{f(x)}{g(x)} = \eta. \quad (9.27)$$

The above statement also holds for $\xi \in \{-\infty, \infty\}$ and/or $\eta \in \{-\infty, \infty\}$ if, as in (b), one extends Def. 8.17 to these cases in the obvious way.

**Proof.** First, we assume (a). Consider the case $\xi \in \mathbb{R}$. Since $f$ and $g$ are continuous, (a) implies $f$ and $g$ remain continuous, if we extend them to $\xi$ by letting $f(\xi) := g(\xi) = 0$. This extension will now allow us to apply Th. 9.18 to $f$ and $g$. To prove (9.27), let $(x_k)_{k \in \mathbb{N}}$ be a sequence in $I$ with $\lim_{k \to \infty} x_k = \xi$. Then (9.23a) yields, for each $k \in \mathbb{N}$, some $\xi_k \in ]x_k, \xi[$ if $x_k < \xi$ and some $\xi_k \in ]\xi, x_k[ \, \text{if} \, \xi < x_k$, satisfying

$$\frac{f(x_k)}{g(x_k)} = \frac{f(x_k) - f(\xi)}{g(x_k) - g(\xi)} = \frac{f'(\xi_k)}{g'(\xi_k)}. \quad (9.28)$$

From the Sandwich Th. 7.16, we obtain $\lim_{k \to \infty} \xi_k = \xi$, i.e. (9.28) and $\lim_{x \to \xi} \frac{f(x)}{g(x)} = \eta$ imply $\lim_{x \to \xi} \frac{f(x)}{g(x)} = \eta$ (also for $\eta \in \{-\infty, \infty\}$). Now consider the case $\xi \in \{-\infty, \infty\}$ and let $(x_k)_{k \in \mathbb{N}}$ be as before. If $\xi = \infty$, then choose $1 \leq c \in I$ and set $\tilde{I} := ]0, c^{-1}[; \, \text{if} \, \xi = -\infty$, then choose $-1 \geq c \in I$ and set $\tilde{I} := ]c^{-1}, 0[. \, \text{We apply what we have already proved above to the auxiliary functions}$

$$\tilde{f} : \tilde{I} \rightarrow \mathbb{R}, \quad \tilde{f}(x) := f(1/x), \quad \tilde{g} : \tilde{I} \rightarrow \mathbb{R}, \quad \tilde{g}(x) := g(1/x)$$
at $\xi := 0$. From the chain rule (9.15), we know $f'(x) = -\frac{f'(1/x)}{x^2}$ and $g'(x) = -\frac{g'(1/x)}{x^2}$ for each $x \in \tilde{I}$. Thus, $\lim_{x \to \xi} \frac{f'(x)}{g'(x)} = \eta$ implies,

$$\eta = \lim_{k \to \infty} \frac{f'(x_k)}{g'(x_k)} = \lim_{k \to \infty} -\frac{x_k^2 f'(x_k)}{g'(x_k)} = \lim_{k \to \infty} \frac{f'(1/x_k)}{g'(1/x_k)} = \lim_{k \to \infty} \frac{f(1/x_k)}{g(1/x_k)} = \lim_{k \to \infty} f(x_k),$$

proving $\lim_{x \to \xi} \frac{f(x)}{g(x)} = \eta$.

We now assume (b), still letting $(x_k)_{k \in \mathbb{N}}$ be as before. Note that $g' \neq 0$ implies $g$ is either strictly increasing or strictly decreasing. We proceed with the proof for the case $I = ]a, \xi[$, the proof for $I = ]\xi, b]$ can be done completely analogous. We first consider the case where $g$ is strictly increasing, i.e. $\lim_{x \to \xi} g(x) = \infty$. Assume $\eta \in \mathbb{R}$ and $\epsilon > 0$. Then $\lim_{x \to \xi} \frac{f(x)}{g(x)} = \eta$ and $\lim_{x \to \xi} g(x) = \infty$ imply

$$\exists c \in ]a, \xi[ \quad \forall x \in ]c, \xi[ \quad \left( g(x) > 0 \land \eta - \frac{\epsilon}{2} < \frac{f(x)}{g(x)} < \eta + \frac{\epsilon}{2} \right).$$

Since $\lim_{k \to \infty} x_k = \xi$, there exists $N_0 \in \mathbb{N}$ such that, for each $k > N_0$, $c < x_k < \xi$. Next, according to Th. 9.18,

$$\forall k > N_0 \quad \exists \xi_k \in ]c, x_k[ \quad \eta - \frac{\epsilon}{2} < \frac{f(x_k) - f(c)}{g(x_k) - g(c)} = \frac{f'(\xi_k)}{g'(\xi_k)} < \eta + \frac{\epsilon}{2}.$$

In consequence, using $g(x_k) > g(c)$, as $g$ is strictly increasing,

$$\forall k > N_0 \quad \left( \eta - \frac{\epsilon}{2} \right) (g(x_k) - g(c)) < f(x_k) - f(c) < \left( \eta + \frac{\epsilon}{2} \right) (g(x_k) - g(c))$$

and

$$\forall k > N_0 \quad \left( \eta - \frac{\epsilon}{2} \right) + \frac{f(c) - \left( \eta - \frac{\epsilon}{2} \right) g(c)}{g(x_k)} < \frac{f(x_k)}{g(x_k)} < \left( \eta + \frac{\epsilon}{2} \right) + \frac{f(c) - \left( \eta + \frac{\epsilon}{2} \right) g(c)}{g(x_k)}.$$

Since $\lim_{k \to \infty} g(x_k) = \infty$,

$$\exists N \geq N_0 \quad \forall k > N \quad \left( \left| \frac{f(c) - \left( \eta - \frac{\epsilon}{2} \right) g(c)}{g(x_k)} \right| < \frac{\epsilon}{2} \land \left| \frac{f(c) - \left( \eta + \frac{\epsilon}{2} \right) g(c)}{g(x_k)} \right| < \frac{\epsilon}{2} \right),$$

that means

$$\forall k > N \quad \eta - \epsilon < \frac{f(x_k)}{g(x_k)} < \eta + \epsilon,$$

proving $\lim_{x \to \xi} \frac{f(x)}{g(x)} = \eta$. For $\eta = \infty$ and given $n \in \mathbb{N}$, the argument is similar: $\lim_{x \to \xi} \frac{f'(x)}{g'(x)} = \eta$ and $\lim_{x \to \xi} g(x) = \infty$ imply

$$\exists c \in ]a, \xi[ \quad \forall x \in ]c, \xi[ \quad \left( g(x) > 0 \land n < \frac{f'(x)}{g'(x)} \right).$$
As before, since \( \lim_{k \to \infty} x_k = \xi \), there exists \( N_0 \in \mathbb{N} \) such that, for each \( k > N_0 \), \( c < x_k < \xi \). Again, according to Th. 9.18,

\[
\forall \ k > N_0 \quad \exists \ \xi \in (c, x_k] \quad n < \frac{f(x_k) - f(c)}{g(x_k) - g(c)} = \frac{f'(\xi_k)}{g'(\xi_k)}.
\]

In consequence, using \( g(x_k) > g(c) \), as \( g \) is strictly increasing,

\[
\forall \ k > N_0 \quad n (g(x_k) - g(c)) < f(x_k) - f(c)
\]

and

\[
\forall \ k > N_0 \quad n + \frac{f(c) - n g(c)}{g(x_k)} < f(x_k) \frac{g(x_k)}{g(x_k)}.
\]

Since \( \lim_{k \to \infty} g(x_k) = \infty \),

\[
\exists \ N \geq N_0 \quad \forall \ k > N \quad \left| \frac{f(c) - n g(c)}{g(x_k)} \right| < 1,
\]

that means

\[
\forall \ k > N \quad n - 1 < \frac{f(x_k)}{g(x_k)}.
\]

proving \( \lim_{x \to \xi} \frac{f(x)}{g(x)} = \eta \). If \( \eta = -\infty \), then, using what we have already shown,

\[
\lim_{x \to \xi} = \frac{f'(x)}{g'(x)} = \eta \Rightarrow \lim_{x \to \xi} - \frac{f'(x)}{g'(x)} = \infty = \lim_{x \to \xi} - \frac{f(x)}{g(x)} \Rightarrow \lim_{x \to \xi} \frac{f(x)}{g(x)} = \eta.
\]

Finally, if \( g \) strictly decreasing, then \(-g\) is strictly increasing and we obtain

\[
\lim_{x \to \xi} = \frac{f'(x)}{g'(x)} = \eta \Rightarrow \lim_{x \to \xi} - \frac{f'(x)}{g'(x)} = -\eta = \lim_{x \to \xi} - \frac{f(x)}{g(x)} \Rightarrow \lim_{x \to \xi} \frac{f(x)}{g(x)} = \eta,
\]

concluding the proof. 

**Example 9.26.** (a) Applying L'Hôpital's rule to \( f : ] - \pi/2, \pi/2[ \to \mathbb{R}, f(x) := \tan x, \ g : ] - \pi/2, \pi/2[ \to \mathbb{R}, g(x) := e^x - 1 \), with \( \xi = 0 \) yields

\[
\lim_{x \to 0} \frac{\tan x}{e^x - 1} = \lim_{x \to 0} \frac{1 + \tan^2 x}{e^x} = \frac{1}{1} = 1 \quad (9.29)
\]

(note \( g'(x) = e^x \neq 0 \) for each \( x \in ] - \pi/2, \pi/2[ \)).

(b) It can happen that a single application of L'Hôpital's rule does not, yet, yield a useful result, but that a repeated application does. An example is provided by considering \( \alpha > 0, n \in \mathbb{N} \), and \( f : \mathbb{R}^+ \to \mathbb{R}, f(x) := e^{\alpha x}, g : \mathbb{R}^+ \to \mathbb{R}, g(x) := x^n, \xi := \infty \). Applying L'Hôpital's rule \( n \) times yields

\[
\forall \alpha \in \mathbb{R}^+ \quad \forall \ n \in \mathbb{N} \quad \lim_{x \to \infty} \frac{e^{\alpha x}}{x^n} = \lim_{x \to \infty} \frac{\alpha^n e^{\alpha x}}{n!} = \infty \quad (9.30)
\]

(note \( g^{(k)}(x) = n(n-1) \cdots (n-k+1)x^{n-k} \neq 0 \) for each \( k \in \{1, \ldots, n\} \) and each \( x \in \mathbb{R}^+ \)).
(c) It can also happen that even repeated applications of L'Hôpital's rule do not help at all, even though \( \lim_{x \to \xi} \frac{f(x)}{g(x)} \) does exist and the hypotheses of Th. 9.25 are all satisfied. A simple example is given by \( f : \mathbb{R} \to \mathbb{R}, f(x) := e^x \), \( g : \mathbb{R} \to \mathbb{R}, g(x) := 2e^x \), and \( \xi = -\infty \). Even though \( \lim_{x \to -\infty} \frac{f(x)}{g(x)} = \frac{1}{2} \), one has \( \lim_{x \to -\infty} f^{(n)}(x) = \lim_{x \to -\infty} g^{(n)}(x) = 0 \) for every \( n \in \mathbb{N} \).

## 9.5 Convex Functions

In the present section, we provide an introduction to (one-dimensional) convex functions. They have many important applications, some of which we will need in Analysis II, when studying so-called norms on \( \mathbb{K}^n \).

The idea is to call a function \( f : I \to \mathbb{R} \) (where \( I \subseteq \mathbb{R} \) is an interval) convex if, and only if, each line segment connecting two points on the graph of \( f \) lies above this graph, and to call \( f \) concave if, and only if, each such line segment lies below the graph of \( f \).

Noting that, for \( x_1 < x_2 \), the line through the two points \((x_1, f(x_1))\) and \((x_2, f(x_2))\) is represented by the equation

\[
L(x) = \frac{x_2 - x}{x_2 - x_1} f(x_1) + \frac{x - x_1}{x_2 - x_1} f(x_2),
\]

this leads to the following definition:

**Definition 9.27.** Let \( I \subseteq \mathbb{R} \) be an interval (\( I \) can be open, closed, or half-open, it can be for finite or of infinite length) and \( f : I \to \mathbb{R} \). Then \( f \) is called convex if, and only if, for each \( x_1, x, x_2 \in I \) such that \( x_1 < x < x_2 \), one has

\[
f(x) \leq \frac{x_2 - x}{x_2 - x_1} f(x_1) + \frac{x - x_1}{x_2 - x_1} f(x_2);
\]

\[\tag{9.32a}
\]

\( f \) is called concave if, and only if, for each \( x_1, x, x_2 \in I \) such that \( x_1 < x < x_2 \), one has

\[
f(x) \geq \frac{x_2 - x}{x_2 - x_1} f(x_1) + \frac{x - x_1}{x_2 - x_1} f(x_2).
\]

\[\tag{9.32b}
\]

Moreover, \( f \) is called strictly convex (resp. strictly concave) if, and only if, \( (9.32a) \) (resp. \( (9.32b) \)) always holds with strict inequality.

**Lemma 9.28.** Let \( I \subseteq \mathbb{R} \) be an interval. Then \( f : I \to \mathbb{R} \) is (strictly) convex if, and only if, \( -f \) is (strictly) concave.

**Proof.** Merely multiply \((9.32b)\) by \((-1)\) and compare with \((9.32a)\). 

The following Prop. 9.29 provides equivalences for convexity. One can easily obtain the corresponding equivalences for concavity by combining Prop. 9.29 with Lem. 9.28.

**Proposition 9.29.** Let \( I \subseteq \mathbb{R} \) be an interval and \( f : I \to \mathbb{R} \). Then the following statements are equivalent:
\( f \) is (strictly) convex.

(ii) For each \( a, b \in I \) such that \( a \neq b \) and each \( \lambda \in \) \( ]0, 1[ \), the following estimate holds (with strict inequality):

\[
f(\lambda a + (1 - \lambda)b) \leq \lambda f(a) + (1 - \lambda) f(b).
\] (9.33)

(iii) For each \( x_1, x, x_2 \in I \) such that \( x_1 < x < x_2 \), one has (with strict inequality)

\[
\frac{f(x) - f(x_1)}{x - x_1} \leq \frac{f(x_2) - f(x)}{x_2 - x}.
\] (9.34)

(iv) For each \( x_1, x, x_2 \in I \) such that \( x_1 < x < x_2 \), one has (with strict inequality)

\[
\frac{f(x) - f(x_1)}{x - x_1} \leq \frac{f(x_2) - f(x_1)}{x_2 - x_1} \leq \frac{f(x_2) - f(x)}{x_2 - x}.
\] (9.35)

Proof. We leave the proof as an exercise. Suggestion: Establish the following implications: (i) \( \iff \) (ii), (i) \( \iff \) (iii), (iv) \( \Rightarrow \) (iii), (i) \( \Rightarrow \) (iv).

Example 9.30. Since \( |x + (1 - \lambda)y| \leq \lambda|x| + (1 - \lambda)|y| \) for each \( 0 < \lambda < 1 \) and each \( x, y \in \mathbb{R} \), the absolute value function is convex. This example also shows that a convex function does not need to be differentiable.

For differentiable functions, one can formulate convexity criteria in terms of the derivative:

Proposition 9.31. Let \( a < b \), and suppose that \( f : [a, b] \rightarrow \mathbb{R} \) is continuous on \([a, b]\) and differentiable on \( ]a, b[ \). Then \( f \) is (strictly) convex (resp. (strictly) concave) on \([a, b]\) if, and only if, the derivative \( f' \) is (strictly) increasing (resp. (strictly) decreasing) on \([a, b]\).

Proof. Since \((-f)' = -f'\) and \(-f'\) is (strictly) increasing if, and only if, \( f' \) is (strictly) decreasing, it suffices to consider the (strictly) convex case. So assume that \( f \) is (strictly) convex. Then for each \( x_1, x, x_0, y, x_2 \in ]a, b[ \) such that \( x_1 < x < x_0 < y < x_2 \), applying Prop. 9.29(iv), one has (with strict inequalities),

\[
\frac{f(x) - f(x_1)}{x - x_1} \leq \frac{f(x_0) - f(x_1)}{x_0 - x_1} \leq \frac{f(x_2) - f(x_1)}{x_2 - x_1} \leq \frac{f(x_2) - f(y)}{x_2 - y}.
\] (9.36)

Thus,

\[
f'(x_1) = \lim_{x \downarrow x_1} \frac{f(x) - f(x_1)}{x - x_1} \leq \frac{f(x_0) - f(x_1)}{x_0 - x_1} \leq \frac{f(x_2) - f(x_1)}{x_2 - x_1} \leq \lim_{y \uparrow x_2} \frac{f(x_2) - f(y)}{x_2 - y} = f'(x_2)
\] (9.37)
(where the inequality at (\#) is strict if it is strict in (9.36)), showing that \( f' \) is (strictly) increasing on \([a, b] \). On the other hand, if \( f' \) is (strictly) increasing on \([a, b] \), then for each \( x_1, x_2 \in [a, b] \) such that \( x_1 < x < x_2 \), Th. 9.18 yields \( \xi_1 \in ]x_1, x[ \) and \( \xi_2 \in ]x, x_2[ \) such that

\[
\frac{f(x) - f(x_1)}{x - x_1} = f'(\xi_1) \quad \text{and} \quad \frac{f(x_2) - f(x)}{x_2 - x} = f'(\xi_2). \tag{9.38}
\]

As \( \xi_1 < \xi_2 \) and \( f' \) is (strictly) increasing, (9.38) implies (9.34) (with strict inequality) and, thus, the (strict) convexity of \( f \).

\[\begin{array}{c}
\text{Proposition 9.32.} \quad \text{Let} \ a < b, \ \text{and suppose that} \ f : [a, b] \rightarrow \mathbb{R} \ \text{is continuous on} \ [a, b] \\
\text{and twice differentiable on} \ [a, b]. \\
\text{(a) } f \text{ is convex (resp. concave) on} \ [a, b] \ \text{if, and only if,} \ f'' \geq 0 \ (\text{resp.} \ f'' \leq 0) \ \text{on} \ ]a, b]. \\
\text{(b) If} \ f'' > 0 \ (\text{resp.} \ f'' < 0) \ \text{on} \ ]a, b], \ \text{then} \ f \text{ is strictly convex (resp. strictly concave)} \\
\text{(as a caveat we remark that, here, the converse does not hold – for example} \ x \mapsto x^4 \text{is strictly convex, but its second derivative} \ x \mapsto 12x^2 \text{is 0 at} \ x = 0). \\
\end{array}\]

\[\begin{array}{c}
\text{Proof.} \quad \text{Since} \ -f'' \geq 0 \ \text{if, and only if} \ f'' \leq 0; \ \text{and} \ -f'' > 0 \ \text{if, and only if} \ f'' < 0, \ \text{if}
\text{suffices to consider the convex cases. Moreover, for (a), one merely has to combine Prop.} \\
\text{9.31 with the fact that} \ f' \ \text{is increasing on} \ ]a, b[ \ \text{if, and only if,} \ f'' \geq 0 \ \text{on} \ ]a, b[. \ \text{The}
\text{proof of (b) is left as an exercise.} \\
\end{array}\]

\[\begin{array}{c}
\text{Example 9.33.} \ (a) \ \text{Since for} \ f : \mathbb{R} \rightarrow \mathbb{R}, \ f(x) = e^x, \ \text{it is} \ f''(x) = e^x > 0, \ \text{the}
\text{exponential function is strictly convex on} \ \mathbb{R}. \\
\text{(b) Since for} \ f : \mathbb{R}^+ \rightarrow \mathbb{R}, \ f(x) = \ln x, \ \text{it is} \ f''(x) = -1/x^2 < 0, \ \text{the natural logarithm}
\text{is strictly concave on} \ \mathbb{R}^+. \\
\end{array}\]

\[\begin{array}{c}
\text{Theorem 9.34 (Jensen’s inequality).} \ \text{Let} \ I \subseteq \mathbb{R} \ \text{be an interval and let} \ f : I \rightarrow \mathbb{R} \ \text{be}
\text{convex. If} \ n \in \mathbb{N} \ \text{and} \ \lambda_1, \ldots, \lambda_n > 0 \ \text{such that} \ \lambda_1 + \cdots + \lambda_n = 1, \ \text{then}
\forall \ x_1, \ldots, x_n \in I \quad f(\lambda_1 x_1 + \cdots + \lambda_n x_n) \leq \lambda_1 f(x_1) + \cdots + \lambda_n f(x_n). \tag{9.39a}
\end{array}\]

\[\begin{array}{c}
\text{If} \ f \ \text{is concave, then}
\forall \ x_1, \ldots, x_n \in I \quad f(\lambda_1 x_1 + \cdots + \lambda_n x_n) \geq \lambda_1 f(x_1) + \cdots + \lambda_n f(x_n). \tag{9.39b}
\end{array}\]

\[\begin{array}{c}
\text{If} \ f \ \text{is strictly convex or strictly concave, then equality in the above inequalities can only}
\text{hold if} \ x_1 = \cdots = x_n. \\
\end{array}\]

\[\begin{array}{c}
\text{Proof.} \ \text{If one lets} \ a := \min \{x_1, \ldots, x_n\}, \ b := \max \{x_1, \ldots, x_n\}, \ \text{and} \ \bar{x} := \lambda_1 x_1 + \cdots + \lambda_n x_n, \ \text{then}
\end{array}\]

\[
a = \sum_{j=1}^n \lambda_j a \leq \bar{x} \leq \sum_{j=1}^n \lambda_j b = b \quad \Rightarrow \quad \bar{x} \in I. \tag{9.40}
\]
Since \( f \) is (strictly) concave if, and only if, \(-f\) is (strictly) convex, it suffices to consider the cases where \( f \) is convex and where \( f \) is strictly convex. Thus, we assume that \( f \) is convex and prove (9.39a) by induction. For \( n = 1 \), one has \( \lambda_1 = 1 \) and there is nothing to prove. For \( n = 2 \), (9.39a) reduces to (9.33), which holds due to the convexity of \( f \).

Finally, let \( n > 2 \) and assume that (9.39a) already holds for each \( 1 \leq l \leq n - 1 \). Set

\[
\lambda := \lambda_1 + \cdots + \lambda_{n-1}, \quad x := \frac{\lambda_1}{\lambda} x_1 + \cdots + \frac{\lambda_{n-1}}{\lambda} x_{n-1}.
\]

Then \( x \in I \) follows as in (9.40). One computes

\[
f(\lambda_1 x_1 + \cdots + \lambda_n x_n) = f \left( \sum_{j=1}^{n-1} \lambda_j x_j + \lambda_n x_n \right) = f(\lambda x + \lambda_n x_n)
\]

\[
\leq \lambda f(x) + \lambda_n f(x_n) \leq \lambda \sum_{j=1}^{n-1} \frac{\lambda_j}{\lambda} f(x_j) + \lambda_n f(x_n)
\]

\[
= \lambda_1 f(x_1) + \cdots + \lambda_n f(x_n),
\]

thereby completing the induction, and, thus, the proof of (9.39a). If \( f \) is strictly convex and (9.39a) holds with equality, then one can also proceed by induction to prove the equality of the \( x_j \). Again, if \( n = 1 \), then there is nothing to prove. If \( n = 2 \), and \( x_1 \neq x_2 \), then strict convexity requires (9.39a) to hold with strict inequality. Thus \( x_1 = x_2 \). Now let \( n > 2 \). It is noted that (9.42) still holds. By hypothesis, the first and last term in (9.42) are now equal, implying that all terms in (9.42) must be equal. Using the induction hypothesis for \( l = 2 \) and the corresponding equality in (9.42), we conclude that \( x = x_n \). Using the induction hypothesis for \( l = n - 1 \) and the corresponding equality in (9.42), we conclude that \( x_1 = \cdots = x_{n-1} \). Finally, \( x = x_n \) and \( x_1 = \cdots = x_{n-1} \) are combined using (9.41) to get \( x_1 = x_n \), finishing the proof of the theorem.

**Theorem 9.35** (Inequality Between the Weighted Arithmetic Mean and the Weighted Geometric Mean). If \( n \in \mathbb{N}, x_1, \ldots, x_n \geq 0 \) and \( \lambda_1, \ldots, \lambda_n > 0 \) such that \( \lambda_1 + \cdots + \lambda_n = 1 \), then

\[
x_1^{\lambda_1} \cdots x_n^{\lambda_n} \leq \lambda_1 x_1 + \cdots + \lambda_n x_n,
\]

where equality occurs if, and only if, \( x_1 = \cdots = x_n \). In particular, for \( \lambda_1 = \cdots = \lambda_n = \frac{1}{n} \), one recovers the inequality between the arithmetic and the geometric mean without weights, known from Th. 7.63.

**Proof.** If at least one of the \( x_j \) is 0, then (9.43) becomes the true statement \( 0 \leq \sum_{j=1}^{n} \lambda_j x_j \) with strict inequality if, and only if, at least one \( x_j > 0 \). Thus, it remains to consider the case \( x_1, \ldots, x_n > 0 \). As we noted in Ex. 9.33(b), the natural logarithm \( \ln : \mathbb{R}^+ \rightarrow \mathbb{R} \) is concave and even strictly concave. Employing Jensen’s inequality (9.39b) yields

\[
\ln(\lambda_1 x_1 + \cdots + \lambda_n x_n) \geq \lambda_1 \ln x_1 + \cdots + \lambda_n \ln x_n = \ln(x_1^{\lambda_1} \cdots x_n^{\lambda_n}).
\]

Applying the exponential function to both sides of (9.44), one obtains (9.43). Since (9.44) is equivalent to (9.43), the strict concavity of \( \ln \) yields that equality in (9.44) implies \( x_1 = \cdots = x_n \). \( \blacksquare \)
10 The Riemann Integral on Intervals in $\mathbb{R}$

10.1 Definition and Simple Properties

Given a nonnegative function $f : M \rightarrow \mathbb{R}_0^+$, $M \subseteq \mathbb{R}$, we aim to compute the area $\int_M f$ of the set “under the graph” of $f$, i.e., of the set

$$\{(x, y) \in \mathbb{R}^2 : x \in M \text{ and } 0 \leq y \leq f(x)\}.$$  \hspace{1cm} (10.1)

This area $\int_M f$ (if it exists) will be called the integral of $f$ over $M$. Moreover, for functions $f : M \rightarrow \mathbb{R}$ that are not necessarily nonnegative, we would like to count areas of sets of the form (10.1) (which are below the graph of $f$ and above the set $M \approx \{(x, 0) \in \mathbb{R}^2 : x \in M \} \subseteq \mathbb{R}^2$) with a positive sign, and whereas we would like to count areas of sets above the graph of $f$ and below the set $M$ with a negative sign. In other words, making use of the positive and negative parts $f^+$ and $f^-$ of $f=f^+-f^-$ as defined in (6.1i) and (6.1j), respectively, we would like our integral to satisfy

$$\int_M f = \int_M f^+ - \int_M f^-.$$ \hspace{1cm} (10.2)

Difficulties arise from the fact that both the function $f$ and the set $M$ can be extremely complicated. To avoid dealing with complicated sets $M$, we restrict ourselves to the situation of integrals over compact intervals, i.e., to integrals over sets of the form $M = [a, b]$. Moreover, we will also restrict ourselves to bounded functions $f$, which we now define:

**Definition 10.1.** Let $\emptyset \neq M \subseteq \mathbb{R}$ and $f : M \rightarrow \mathbb{R}$. Then $f$ is called bounded if, and only if, the set $\{|f(x)| : x \in M\} \subseteq \mathbb{R}_0^+$ is bounded, i.e., if, and only if,

$$\|f\|_{\text{sup}} := \sup\{|f(x)| : x \in M\} \in \mathbb{R}_0^+.$$ \hspace{1cm} (10.3)

The basic idea for the definition of the Riemann integral $\int_M f$ is rather simple: Decompose the set $M$ into small pieces $I_1, \ldots, I_N$ and approximate $\int_M f$ by the finite sum $\sum_{j=1}^N f(x_j)|I_j|$, where $x_j \in I_j$ and $|I_j|$ denotes the size of the set $I_j$. Define $\int_M f$ as the limit of such sums as the size of the $I_j$ tends to zero (if the limit exists). However, to carry out this idea precisely and rigorously does require some work.

As stated before, we will assume that $M$ is a closed bounded interval, and we will choose the $I_j$ to be closed bounded intervals as well. To emphasize we are dealing with intervals, in the following, we will prefer to use the symbol $I$ instead of $M$.

**Definition 10.2.** If $a, b \in \mathbb{R}$, $a \leq b$, and $I := [a, b]$, then we call

$$|I| := b - a = |a - b|,$$ \hspace{1cm} (10.4)

the length or the (1-dimensional) size, volume, or measure of $I$. 

Definition 10.3. Given a real interval \( I := [a, b] \subset \mathbb{R}, a, b \in \mathbb{R}, a < b, \) the \((N+1)\)-tuple \( \Delta := (x_0, \ldots, x_N) \in \mathbb{R}^{N+1}, N \in \mathbb{N}, \) is called a partition of \( I \) if, and only if, \( a = x_0 < x_1 < \cdots < x_N = b. \) We call \( x_0, \ldots, x_N \) the nodes of \( \Delta, \) and let \( \nu(\Delta) := \{x_0, \ldots, x_N\} \) be the set of all nodes. A tagged partition of \( I \) is a partition together with an \( N\)-tuple \( (t_1, \ldots, t_N) \in \mathbb{R}^N \) such that \( t_j \in [x_{j-1}, x_j] \) for each \( j \in \{1, \ldots, N\}. \) Given a partition \( \Delta \) (with or without tags) of \( I \) as above and letting \( I_j := [x_{j-1}, x_j], \) the number

\[
|\Delta| := \max \{|I_j| : j \in \{1, \ldots, N\}\},
\]

is called the mesh size of \( \Delta. \) It is sometimes convenient, if we extend our definitions to trivial intervals, consisting of just one point: For \( a = b, \) we have \( I = [a, a] = \{a\}. \) We then define \( \Delta = x_0 = a \) to be a partition of \( I, \) \( \nu(\Delta) = \{x_0\}, \) and \( a \) is then the only tag that makes \( \Delta \) into a tagged partition. We also set \( I_0 := I = \{a\}, \) and the mesh size in this case is \( |\Delta| := 0. \)

Definition 10.4. Let \( \Delta \) be a partition of \( I = [a, b] \subset \mathbb{R}, a \leq b, \) as in Def. 10.3. Given a function \( f : I \to \mathbb{R} \) that is bounded according to Def. 10.1, define

\[
m_j := m_j(f) := \inf \{f(x) : x \in I_j\}, \quad M_j := M_j(f) := \sup \{f(x) : x \in I_j\},
\]

and

\[
r(\Delta, f) := \sum_{j=1}^{N} m_j |I_j| = \sum_{j=1}^{N} m_j (x_j - x_{j-1}), \quad (10.7a)
\]

\[
R(\Delta, f) := \sum_{j=1}^{N} M_j |I_j| = \sum_{j=1}^{N} M_j (x_j - x_{j-1}), \quad (10.7b)
\]

where \( r(\Delta, f) \) is called the lower Riemann sum and \( R(\Delta, f) \) is called the upper Riemann sum associated with \( \Delta \) and \( f. \) If \( \Delta \) is tagged by \( \tau := (t_1, \ldots, t_N), \) then we also define the intermediate Riemann sum

\[
\rho(\Delta, f) := \sum_{j=1}^{N} f(t_j) |I_j| = \sum_{j=1}^{N} f(t_j) (x_j - x_{j-1}). \quad (10.7c)
\]

Note that, for \( a = b, \) all the above sums are empty and we have \( r(\Delta, f) = R(\Delta, f) = \rho(\Delta, f) = 0. \)

Definition 10.5. Let \( I := [a, b] \subset \mathbb{R} \) be an interval, \( a \leq b, \) and suppose \( f : I \to \mathbb{R} \) is bounded.

(a) Define

\[
J_*(f, I) := \sup \left\{ r(\Delta, f) : \Delta \text{ is a partition of } I \right\}, \quad (10.8a)
\]

\[
J^*(f, I) := \inf \left\{ R(\Delta, f) : \Delta \text{ is a partition of } I \right\}. \quad (10.8b)
\]

We call \( J_*(f, I) \) the lower Riemann integral of \( f \) over \( I \) and \( J^*(f, I) \) the upper Riemann integral of \( f \) over \( I. \)
(b) The function \( f \) is called \textit{Riemann integrable} over \( I \) if, and only if, \( J_*(f, I) = J^*(f, I) \). If \( f \) is Riemann integrable over \( I \), then
\[
\int_a^b f(x) \, dx := \int_I f(x) \, dx := \int_a^b f := J_*(f, I) = J^*(f, I)
\]
(10.9) is called the \textit{Riemann integral} of \( f \) over \( I \). The set of all functions \( f : I \rightarrow \mathbb{R} \) that are Riemann integrable over \( I \) is denoted by \( \mathcal{R}(I, \mathbb{R}) \) or just by \( \mathcal{R}(I) \).

(c) The function \( g : I \rightarrow \mathbb{C} \) is called \textit{Riemann integrable} over \( I \) if, and only if, both \( \text{Re} \, g \) and \( \text{Im} \, g \) are Riemann integrable. The set of all Riemann integrable functions \( g : I \rightarrow \mathbb{C} \) is denoted by \( \mathcal{R}(I, \mathbb{C}) \). If \( g \in \mathcal{R}(I, \mathbb{C}) \), then
\[
\int_I g := \left( \int_I \text{Re} \, g, \int_I \text{Im} \, g \right) = \int_I \text{Re} \, g + i \int_I \text{Im} \, g \in \mathbb{C}
\]
(10.10) is called the Riemann integral of \( g \) over \( I \).

\textbf{Remark 10.6.} If \( I = [a, b] \subseteq \mathbb{R} \), \( \Delta \) is a partition of \( I \), and \( f : I \rightarrow \mathbb{R} \) is bounded, then (10.6) implies
\[
m_j(f) \overset{(4.6c)}{=} -M_j(-f) \quad \text{and} \quad m_j(-f) \overset{(4.6d)}{=} -M_j(f),
\]
(10.11a) (10.7) implies
\[
r(\Delta, f) = -R(\Delta, -f) \quad \text{and} \quad r(\Delta, -f) = -R(\Delta, f),
\]
(10.11b) and (10.8) implies
\[
J_*(f, I) = -J^*(-f, I) \quad \text{and} \quad J_*(-f, I) = -J^*(f, I).
\]
(10.11c)

\textbf{Example 10.7. (a)} If \( I = [a, b] \subseteq \mathbb{R} \) as before and \( f : I \rightarrow \mathbb{R} \) is constant, i.e. \( f \equiv c \) with \( c \in \mathbb{R} \), then \( f \in \mathcal{R}(I) \) and
\[
\int_a^b f = c(b - a) = c|I| : \quad (10.12)
\]
We have, for each partition \( \Delta \) of \( I \),
\[
r(\Delta, f) = \sum_{j=1}^N m_j |I_j| = c \sum_{j=1}^N |I_j| = c|I| = c(b - a) = \sum_{j=1}^N M_j |I_j| = R(\Delta, f),
\]
(10.13) proving \( J_*(f, I) = c(b - a) = J^*(f, I) \).

(b) An example of a function that is not Riemann integrable for \( a < b \) is given by the \textit{Dirichlet function}
\[
f : [a, b] \rightarrow \mathbb{R}, \quad f(x) := \begin{cases} 0 & \text{for } x \text{ irrational,} \\ 1 & \text{for } x \text{ rational,} \end{cases} \quad a < b.
\]
(10.14)
Since \( r(\Delta, f) = 0 \) and \( R(\Delta, f) = \sum_{j=1}^N |I_j| = b - a \) for every partition \( \Delta \) of \( I \), one obtains \( J_*(f, I) = 0 \neq (b - a) = J^*(f, I) \), showing that \( f \notin \mathcal{R}(I) \).
Definition 10.8. (a) If $\Delta$ is a partition of $[a, b] \subseteq \mathbb{R}$ as in Def. 10.3, then another partition $\Delta'$ of $[a, b]$ is called a refinement of $\Delta$ if, and only if, $\nu(\Delta) \subseteq \nu(\Delta')$, i.e., if, and only if, the nodes of $\Delta'$ include all the nodes of $\Delta$.

(b) If $\Delta$ and $\Delta'$ are partitions of $[a, b] \subseteq \mathbb{R}$, then the superposition of $\Delta$ and $\Delta'$, denoted $\Delta + \Delta'$, is the unique partition of $[a, b]$ having $\nu(\Delta) \cup \nu(\Delta')$ as its set of nodes. Note that the superposition of $\Delta$ and $\Delta'$ is always a common refinement of $\Delta$ and $\Delta'$.

Lemma 10.9. Let $a, b \in \mathbb{R}$, $a < b$, $I := [a, b]$, and suppose $f : I \rightarrow \mathbb{R}$ is bounded with $M := \|f\|_{\text{sup}} \in \mathbb{R}_0^+$. Let $\Delta'$ be a partition of $I$ and assume

$$\alpha := \#\left(\nu(\Delta') \setminus \{a, b\}\right) \geq 1 \quad (10.15)$$

is the number of interior nodes that occur in $\Delta'$. Then, for each partition $\Delta$ of $I$, the following holds:

$$r(\Delta, f) \leq r(\Delta + \Delta', f) \leq r(\Delta, f) + 2\alpha M |\Delta|, \quad (10.16a)$$

$$R(\Delta, f) \geq R(\Delta + \Delta', f) \geq R(\Delta, f) - 2\alpha M |\Delta|. \quad (10.16b)$$

Proof. We carry out the proof of (10.16a) – the proof of (10.16b) can be conducted completely analogous. Consider the case $\alpha = 1$ and let $\xi$ be the single element of $\nu(\Delta') \setminus \{a, b\}$. If $\xi \in \nu(\Delta)$, then $\Delta + \Delta' = \Delta$, and (10.16a) is trivially true. If $\xi \notin \nu(\Delta)$, then $x_{k-1} < \xi < x_k$ for a suitable $k \in \{1, \ldots, N\}$. Define

$$I' := [x_{k-1}, \xi], \quad I'' := [\xi, x_k] \quad (10.17)$$

and

$$m' := \inf\{f(x) : x \in I'\}, \quad m'' := \inf\{f(x) : x \in I''\}. \quad (10.18)$$

Then we obtain

$$r(\Delta + \Delta', f) - r(\Delta, f) = m' |I'| + m'' |I''| - m_k |I_k| = (m' - m_k) |I'| + (m'' - m_k) |I''|. \quad (10.19)$$

Together with the observation

$$0 \leq m' - m_k \leq 2M, \quad 0 \leq m'' - m_k \leq 2M, \quad (10.20)$$

(10.19) implies

$$0 \leq r(\Delta + \Delta', f) - r(\Delta, f) \leq 2M \left(|I'| + |I''|\right) \leq 2M |\Delta|. \quad (10.21)$$

The general form of (10.16a) follows by an induction on $\alpha$.  

Theorem 10.10. Let $a, b \in \mathbb{R}$, $a \leq b$, $I := [a, b]$, and let $f : I \rightarrow \mathbb{R}$ be bounded.

(a) Suppose $\Delta$ and $\Delta'$ are partitions of $I$ such that $\Delta'$ is a refinement of $\Delta$. Then

$$r(\Delta, f) \leq r(\Delta', f), \quad R(\Delta, f) \geq R(\Delta', f). \quad (10.22)$$
For arbitrary partitions $\Delta$ and $\Delta'$, the following holds:
\[
 r(\Delta, f) \leq R(\Delta', f).
\] (10.23)

(c) $J_*(f, I) \leq J^*(f, I)$.

(d) For each sequence of partitions $(\Delta_n)_{n \in \mathbb{N}}$ of $I$ such that $\lim_{n \to \infty} |\Delta_n| = 0$, one has
\[
 \lim_{n \to \infty} r(\Delta_n, f) = J_*(f, I), \quad \lim_{n \to \infty} R(\Delta_n, f) = J^*(f, I).
\] (10.24)

In particular, if $f \in \mathcal{R}(I)$, then
\[
 \lim_{n \to \infty} r(\Delta_n, f) = \lim_{n \to \infty} R(\Delta_n, f) = \int_I f,
\] (10.25a)

and if $f \in \mathcal{R}(I)$ and the $\Delta_n$ are tagged, then also
\[
 \lim_{n \to \infty} \rho(\Delta_n, f) = \int_I f.
\] (10.25b)

Proof. (a): If $\Delta'$ is a refinement of $\Delta$, then $\Delta' = \Delta + \Delta'$. Thus, (10.22) is immediate from (10.16).

(b): This also follows from (10.16):
\[
 r(\Delta, f) \leq^{(10.16a)} r(\Delta + \Delta', f) \leq^{(10.7)} R(\Delta + \Delta', f) \leq^{(10.16b)} R(\Delta', f).
\] (10.26)

(c): One just combines (10.8) with (b).

(d): For $a = b$, there is nothing to show. For $a < b$, let $(\Delta_n)_{n \in \mathbb{N}}$ be a sequence of partitions of $I$ such that $\lim_{n \to \infty} |\Delta_n| = 0$, and let $\Delta'$ be an arbitrary partition of $I$ with numbers $\alpha$ and $M$ defined as in Lem. 10.9. Then, according to (10.16a):
\[
 r(\Delta_n, f) \leq r(\Delta_n + \Delta', f) \leq r(\Delta_n, f) + 2 \alpha M |\Delta_n| \quad \text{for each } n \in \mathbb{N}.
\] (10.27)

From (b), we conclude the sequence $(r(\Delta_n, f))_{n \in \mathbb{N}}$ is bounded. According to the Bolzano-Weierstrass Th. 7.27, if we can show that the sequence has $J_*(f, I)$ as its only cluster point, then the first equality of (10.24) must hold. Thus, according to Prop. 7.26, it suffices to show that every converging subsequence of $(r(\Delta_n, f))_{n \in \mathbb{N}}$ converges to $J_*(f, I)$. To this end, suppose $(r(\Delta_{n_k}, f))_{k \in \mathbb{N}}$ is a converging subsequence of $(r(\Delta_n, f))_{n \in \mathbb{N}}$ with $\beta := \lim_{k \to \infty} r(\Delta_{n_k}, f)$. First note $\beta \leq J_*(f, I)$ due to the definition of $J_*(f, I)$. Moreover, (10.27) implies $\lim_{k \to \infty} r(\Delta_{n_k} + \Delta', f) = \beta$. Since $r(\Delta', f) \leq r(\Delta_{n_k} + \Delta', f)$ and $\Delta'$ is arbitrary, we obtain $J_*(f, I) \leq \beta$, i.e. $J_*(f, I) = \beta$. Thus, we have shown that, indeed, every subsequence of $(r(\Delta_n, f))_{n \in \mathbb{N}}$ converges to $\beta = J_*(f, I)$. In the same manner, one conducts the proof of $J^*(f, I) = \lim_{n \to \infty} R(\Delta_n, f)$. Then (10.25a) is immediate from the definition of Riemann integrability, and (10.25b) follows from (10.25a), since (10.7) implies $r(\Delta, f) \leq \rho(\Delta, f) \leq R(\Delta, f)$ for each tagged partition $\Delta$ of $I$. \[\blacksquare\]
Theorem 10.11. Let \( a, b \in \mathbb{R}, \ a \leq b, \ I := [a, b] \).

(a) The integral is linear: More precisely, if \( f, g \in \mathcal{R}(I, \mathbb{K}) \) and \( \lambda, \mu \in \mathbb{K} \), then \( \lambda f + \mu g \in \mathcal{R}(I, \mathbb{K}) \) and
\[
\int_I (\lambda f + \mu g) = \lambda \int_I f + \mu \int_I g. \tag{10.28}
\]

(b) Let \( \Delta = (y_0, \ldots, y_M), M \in \mathbb{N}, \) be a partition of \( I, J_k := [y_{k-1}, y_k] \). Then \( f \in \mathcal{R}(I, \mathbb{K}) \) if, and only if, \( f \in \mathcal{R}(J_k, \mathbb{K}) \) for each \( k \in \{1, \ldots, M\} \). If \( f \in \mathcal{R}(I, \mathbb{K}) \), then
\[
\int_a^b f = \int_I f = \sum_{k=1}^M \int_{J_k} f = \sum_{k=1}^M \int_{y_{k-1}}^{y_k} f. \tag{10.29}
\]

(c) Monotonicity of the Integral: If \( f, g : I \to \mathbb{R} \) are bounded and \( f \leq g \) (i.e. \( f(x) \leq g(x) \) for each \( x \in I \)), then \( J_*(f, I) \leq J_*(g, I) \) and \( J^*(f, I) \leq J^*(g, I) \). In particular, if \( f, g \in \mathcal{R}(I) \) and \( f \leq g \), then
\[
\int_I f \leq \int_I g. \tag{10.30}
\]

(d) Triangle Inequality: For each \( f \in \mathcal{R}(I, \mathbb{C}) \), one has
\[
\left| \int_I f \right| \leq \int_I |f|. \tag{10.31}
\]

Proof. (a): First, consider \( \mathbb{K} = \mathbb{R} \), i.e. \( f, g : I \to \mathbb{R} \) and \( \lambda, \mu \in \mathbb{R} \). For \( a = b \), there is nothing to prove, so let \( a < b \). Let \( (\Delta_n)_{n \in \mathbb{N}} \) be a sequence of partitions of \( I, \Delta_n = (x_{n,0}, \ldots, x_{n,N_n}), I_{n,j} := [x_{n,j-1}, x_{n,j}], \) satisfying \( \lim_{n \to \infty} |\Delta_n| = 0 \). Note that, for each \( n \in \mathbb{N} \) and each \( j \in \{1, \ldots, N_n\} \),
\[
\begin{align*}
m_{n,j}(f + g) &= \inf \{ f(x) + g(x) : x \in I_{n,j} \} \\
&\geq \inf \{ f(x) : x \in I_{n,j} \} + \inf \{ g(x) : x \in I_{n,j} \} \\
&= m_{n,j}(f) + m_{n,j}(g), \tag{10.32a} \\
M_{n,j}(f + g) &= \sup \{ f(x) + g(x) : x \in I_{n,j} \} \\
&\leq \sup \{ f(x) : x \in I_{n,j} \} + \sup \{ g(x) : x \in I_{n,j} \} \\
&= M_{n,j}(f) + M_{n,j}(g), \tag{10.32b}
\end{align*}
\]
\[
\forall \lambda \in \mathbb{R} \quad m_{n,j}(\lambda f) = \inf \{ \lambda f(x) : x \in I_{n,j} \} = \lambda m_{n,j}(f) \quad \text{for } \lambda \geq 0, \tag{10.32c}
\]
\[
\forall \lambda \in \mathbb{R} \quad M_{n,j}(\lambda f) = \sup \{ \lambda f(x) : x \in I_{n,j} \} = \lambda M_{n,j}(f) \quad \text{for } \lambda < 0, \tag{10.32d}
\]

\[
\begin{align*}
(\lambda f)(x) &= \begin{cases} 
\lambda \inf \{ f(x) : x \in I_{n,j} \} = \lambda m_{n,j}(f) & \text{for } \lambda \geq 0, \\
\lambda \sup \{ f(x) : x \in I_{n,j} \} = \lambda M_{n,j}(f) & \text{for } \lambda < 0,
\end{cases} \\
(\lambda g)(x) &= \begin{cases} 
\lambda \inf \{ g(x) : x \in I_{n,j} \} = \lambda m_{n,j}(g) & \text{for } \lambda \geq 0, \\
\lambda \sup \{ g(x) : x \in I_{n,j} \} = \lambda M_{n,j}(g) & \text{for } \lambda < 0.
\end{cases}
\end{align*}
\]
Thus,

\[ J_*(f + g, I) \quad (10.24) \overset{\text{(10.32a)}}{=} \lim_{n \to \infty} r(\Delta_n, f + g) \overset{\text{(10.7a)}}{=} \lim_{n \to \infty} \sum_{j=1}^{N_n} m_{n,j}(f + g) |I_{n,j}| \]

\[ \geq \lim_{n \to \infty} (r(\Delta_n, f) + r(\Delta_n, g)) = J_*(f, I) + J_*(g, I), \quad (10.33a) \]

\[ J^*(f + g, I) \quad (10.24) \overset{\text{(10.32b)}}{=} \lim_{n \to \infty} R(\Delta_n, f + g) \overset{\text{(10.7b)}}{=} \lim_{n \to \infty} \sum_{j=1}^{N_n} M_{n,j}(f + g) |I_{n,j}| \]

\[ \leq \lim_{n \to \infty} (R(\Delta_n, f) + R(\Delta_n, g)) = J^*(f, I) + J^*(g, I), (10.33b) \]

\[ \forall \lambda \in \mathbb{R} \quad J_*(\lambda f, I) \quad (10.24) \overset{\text{(10.32c)}}{=} \lim_{n \to \infty} r(\Delta_n, \lambda f) \overset{\text{(10.7a)}}{=} \lim_{n \to \infty} \sum_{j=1}^{N_n} m_{n,j}(\lambda f) |I_{n,j}| \]

\[ = \begin{cases} 
\lambda \lim_{n \to \infty} r(\Delta_n, f) = \lambda J_*(f, I) & \text{for } \lambda \geq 0, \\
\lambda \lim_{n \to \infty} R(\Delta_n, f) = \lambda J^*(f, I) & \text{for } \lambda < 0, 
\end{cases} \quad (10.33c) \]

\[ \forall \lambda \in \mathbb{R} \quad J^*(\lambda f, I) \quad (10.24) \overset{\text{(10.32d)}}{=} \lim_{n \to \infty} R(\Delta_n, \lambda f) \overset{\text{(10.7b)}}{=} \lim_{n \to \infty} \sum_{j=1}^{N_n} M_{n,j}(\lambda f) |I_{n,j}| \]

\[ = \begin{cases} 
\lambda \lim_{n \to \infty} R(\Delta_n, f) = \lambda J^*(f, I) & \text{for } \lambda \geq 0, \\
\lambda \lim_{n \to \infty} r(\Delta_n, f) = \lambda J_*(f, I) & \text{for } \lambda < 0. 
\end{cases} \quad (10.33d) \]

Thus, if \( f \) and \( g \) are both Riemann integrable over \( I \), then we obtain \( J_*(f + g, I) \geq J_*(f, I) + J_*(g, I) = J^*(f, I) + J^*(g, I) \geq J^*(f + g, I) \), i.e., by Th. 10.10(c), \((f + g) \in \mathcal{R}(I)\); and \( J_*(\lambda f, I) = \lambda J_*(f, I) = \lambda J^*(f, I) \) for \( \lambda \geq 0 \), \( J_*(\lambda f, I) = \lambda J^*(f, I) \) for \( \lambda < 0 \), i.e. \((\lambda f) \in \mathcal{R}(I)\) in each case. In particular, for each \( \lambda, \mu \in \mathbb{R} \),

\[ \int_I (\lambda f + \mu g) = J_*(\lambda f + \mu g, I) = \lambda J_*(f, I) + \mu J_*(g, I) = \lambda \int_I f + \mu \int_I g, \]

proving (10.28) for \( \mathbb{K} = \mathbb{R} \). It remains to consider \( f, g \in \mathcal{R}(I, \mathbb{C}) \) and \( \lambda, \mu \in \mathbb{C} \). One computes, using the real-valued case,

\[ \int_I (\lambda f) = \left( \int_I (\text{Re } \lambda \text{ Re } f - \text{Im } \lambda \text{ Im } f), \int_I (\text{Re } \lambda \text{ Im } f + \text{Im } \lambda \text{ Re } f) \right) \]

\[ = \left( \text{Re } \lambda \int_I \text{ Re } f - \text{Im } \lambda \int_I \text{ Im } f, \text{Re } \lambda \int_I \text{ Im } f + \text{Im } \lambda \int_I \text{ Re } f \right) \]

\[ = \lambda \int_I f \]

and

\[ \int_I (f + g) = \left( \int_I \text{Re } (f + g), \int_I \text{Im } (f + g) \right) = \left( \int_I \text{Re } f + \int_I \text{Re } g, \int_I \text{Im } g + \int_I \text{Im } g \right) \]

\[ = \left( \int_I \text{Re } f, \int_I \text{Im } f \right) + \left( \int_I \text{Re } g, \int_I \text{Im } g \right) = \int_I f + \int_I g. \]
(b): Once again, consider \( K = \mathbb{R} \) first. For \( a = b \), there is nothing to prove, so let \( a < b \). For \( M = 1 \), there is still nothing to prove. For \( M = 2 \), we have \( a = y_0 < y_1 < y_2 = b \). Consider a sequence \( (\Delta_n)_{n \in \mathbb{N}} \) of partitions of \( I \), \( \Delta_n = (x_{n,0}, \ldots, x_{n,N_n}) \), such that \( \lim_{n \to \infty} |\Delta_n| = 0 \) and \( y_1 \in \nu(\Delta_n) \) for each \( n \in \mathbb{N} \). Define \( \Delta'_n := (x_{n,0}, \ldots, y_1) \), \( \Delta''_n := (y_1, \ldots, x_{n,N_n}) \). Then \( \Delta'_n \) and \( \Delta''_n \) are partitions of \( J_1 \) and \( J_2 \), respectively, and \( \lim_{n \to \infty} |\Delta'_n| = \lim_{n \to \infty} |\Delta''_n| = 0 \). Moreover,

\[
\forall_{n \in \mathbb{N}} \quad \left( r(\Delta_n, f) = r(\Delta'_n, f) + r(\Delta''_n, f), \quad R(\Delta_n, f) = R(\Delta'_n, f) + R(\Delta''_n, f) \right),
\]

implying \( J_*(f, I) = J_*(f, J_1) + J_*(f, J_2) \) and \( J^*(f, I) = J^*(f, J_1) + J^*(f, J_2) \). This proves \( \int_I f = \int_{J_1} f + \int_{J_2} f \) provided \( f \in \mathcal{R}(I) \cap \mathcal{R}(J_1) \cap \mathcal{R}(J_2) \). So it just remains to show the claimed equivalence between \( f \in \mathcal{R}(I) \) and \( f \in \mathcal{R}(J_1) \cap \mathcal{R}(J_2) \). If \( f \in \mathcal{R}(J_1) \cap \mathcal{R}(J_2) \), then \( J_*(f, I) = J_*(f, J_1) + J_*(f, J_2) = J^*(f, J_1) + J^*(f, J_2) = J^*(f, I) \), showing \( f \in \mathcal{R}(I) \).

Conversely, \( J_*(f, I) = J^*(f, I) \) implies \( J_*(f, J_1) = J^*(f, J_1) + J^*(f, J_2) - J_*(f, J_2) \geq J^*(f, J_1) \), showing \( J_*(f, J_1) = J^*(f, J_1) \) and \( f \in \mathcal{R}(J_1) \cap \mathcal{R}(J_2) \) follows completely analogous. The general case now follows by induction on \( M \). If, \( f \in \mathcal{R}(I, \mathbb{C}) \), then one computes, using the real-valued case,

\[
\int_I f = \left( \int_I \text{Re} f, \int_I \text{Im} f \right) = \left( \sum_{k=1}^M \int_{J_k} \text{Re} f, \sum_{k=1}^M \int_{J_k} \text{Im} f \right) = \sum_{k=1}^M \int_{J_k} f.
\]

(c): If \( f, g : I \to \mathbb{R} \) are bounded and \( f \leq g \), then, for each partition \( \Delta \) of \( I \), \( r(\Delta, f) \leq r(\Delta, g) \) and \( R(\Delta, f) \leq R(\Delta, g) \) are immediate from (10.7). As these inequalities are preserved when taking the sup and the inf, respectively, all claims of (c) are established.

(d): We will see in Th. 10.18(c) below, that \( f \in \mathcal{R}(I, \mathbb{K}) \) implies \( |f| \in \mathcal{R}(I) \). Let \( \Delta \) be an arbitrary partition of \( I \), tagged by \( (t_1, \ldots, t_N) \). Then, using the same notation as in Def. 10.3 and Def. 10.4,

\[
\left| \rho(\Delta, \text{Re} f), \rho(\Delta, \text{Im} f) \right| := \left( \sum_{j=1}^N \text{Re} f(t_j) |I_j|, \sum_{j=1}^N \text{Im} f(t_j) |I_j| \right) \leq \sum_{j=1}^N \left| \left( \text{Re} f(t_j), \text{Im} f(t_j) \right) \right| |I_j| = \sum_{j=1}^N |f(t_j)||I_j| =: \rho(\Delta, |f|).
\]

Since the intermediate Riemann sums in (10.34) converge to the respective integrals by (10.25b), one obtains

\[
\left| \int_I f \right| = \lim_{|\Delta| \to 0} \left| \rho(\Delta, \text{Re} f), \rho(\Delta, \text{Im} f) \right| \leq \lim_{|\Delta| \to 0} \rho(\Delta, |f|) = \int_I |f|,
\]

proving (10.31). \( \square \)
Theorem 10.12 (Mean Value Theorem). Let \( a, b \in \mathbb{R}, \ a \leq b, \ I := [a, b] \). If \( f, p \in \mathcal{R}(I) \) and \( p \geq 0 \), then, for each \( m, M \in \mathbb{R} \) with \( m \leq f \leq M \):

\[
m \int_\mathcal{I} p \leq \int_\mathcal{I} fp \leq M \int_\mathcal{I} p.
\]

(10.35a)

In particular, if \( f \) is continuous, then

\[
\exists \xi \in \mathcal{I} \int_\mathcal{I} fp = f(\xi) \int_\mathcal{I} p.
\]

(10.35b)

Returning to a general \( f \in \mathcal{R}(I) \), if \( p \equiv 1 \), then we obtain the theorem’s classical form:

\[
m(b - a) = m|\mathcal{I}| \leq \int_a^b f = \int_\mathcal{I} f \leq M|\mathcal{I}| = M(b - a).
\]

(10.35c)

The theorem’s name comes from the fact that, for \( a < b \), \(|\mathcal{I}|^{-1} \int_\mathcal{I} f\) is sometimes referred to as the mean value of \( f \) on \( I \).

Proof. Since \( mp \leq fp \leqMp \), we compute

\[
m \int_\mathcal{I} p \overset{\text{Th. 10.11(c)}}{\leq} \int_\mathcal{I} fp \overset{\text{Th. 10.11(c)}}{\leq} M \int_\mathcal{I} p.
\]

(10.36)

If \( \int_\mathcal{I} p = 0 \), then (10.35a) implies \( \int_\mathcal{I} fp = 0 \) and (10.35b) is immediate. It remains to consider that \( \int_\mathcal{I} p > 0 \). If \( f \) is continuous, then we can let \( m, M \) be such that \( f(I) = [m, M] \) by Th. 7.54 and the intermediate value Th. 7.57. Then (10.35a) shows there is \( \xi \in \mathcal{I} \) such that \( f(\xi) = \frac{\int_\mathcal{I} fp}{\int_\mathcal{I} p} \), i.e. (10.35b) holds.

\[\blacksquare\]

Theorem 10.13 (Riemann’s Integrability Criterion). Let \( I = [a, b] \subseteq \mathbb{R} \) and suppose \( f : I \rightarrow \mathbb{R} \) is bounded. Then \( f \) is Riemann integrable over \( I \) if, and only if, for each \( \epsilon > 0 \), there exists a partition \( \Delta \) of \( I \) such that

\[
R(\Delta, f) - r(\Delta, f) < \epsilon.
\]

(10.37)

Proof. Suppose, for each \( \epsilon > 0 \), there exists a partition \( \Delta \) of \( I \) such that (10.37) is satisfied. Then

\[
J^*(f, I) - J_*(f, I) \leq R(\Delta, f) - r(\Delta, f) < \epsilon,
\]

(10.38)

showing \( J^*(f, I) \leq J_*(f, I) \). As the opposite inequality always holds, we have \( J^*(f, I) = J_*(f, I) \), i.e. \( f \in \mathcal{R}(I) \) as claimed. Conversely, if \( f \in \mathcal{R}(I) \) and \((\Delta_n)_{n \in \mathbb{N}}\) is a sequence of partitions of \( I \) with \( \lim_{n \to \infty} |\Delta_n| = 0 \), then (10.25a) implies that, for each \( \epsilon > 0 \), there is \( N \in \mathbb{N} \) such that \( R(\Delta_n, f) - r(\Delta_n, f) < \epsilon \) for each \( n > N \).

\[\blacksquare\]

The previous theorem will allow us to prove that every continuous function on \([a, b]\) is Riemann integrable. However, we will also need to make use of the following result:
Proposition 10.14. Let \( I = [a, b] \subseteq \mathbb{R}, a \leq b, f : I \rightarrow \mathbb{R} \). If \( f \) is continuous, then \( f \) is even uniformly continuous, i.e.
\[
\forall \epsilon \in \mathbb{R}^+ \exists \delta \in \mathbb{R}^+ \forall x, y \in I \ (|x - y| < \delta \Rightarrow |f(x) - f(y)| < \epsilon).
\] (10.39)

Proof. Arguing by contraposition, we assume \( f \) not to be uniformly continuous on \( I \). Then the negation of (10.39) must hold, i.e.
\[
\exists \epsilon_0 \in \mathbb{R}^+ \forall \delta \in \mathbb{R}^+ \exists x, y \in I \ (|x - y| < \delta \land |f(x) - f(y)| \geq \epsilon_0).
\] (10.40)

In particular, for each \( n \in \mathbb{N} \), there exist \( x_n, y_n \in I \) such that
\[
|x_n - y_n| < \delta_n := 1/n
\] (10.41) and \( |f(x_n) - f(y_n)| \geq \epsilon_0 \). Then the sequence \((x_n)_{n \in \mathbb{N}}\) is bounded and the Bolzano-Weierstrass Th. 7.27 provides a convergent subsequence \((x_{\phi(n)})_{n \in \mathbb{N}}\), i.e. there is \( \xi \in \mathbb{R} \) with \( \lim_{n \rightarrow \infty} x_{\phi(n)} = \xi \). Clearly, \( \xi \in [a, b] \) and (10.41) implies \( \lim_{n \rightarrow \infty} y_{\phi(n)} = \xi \) as well. However, due to \( |f(x_{\phi(n)}) - f(y_{\phi(n)})| \geq \epsilon_0 > 0 \), the sequences \((f(x_{\phi(n)}))_{n \in \mathbb{N}}\) and \((f(y_{\phi(n)}))_{n \in \mathbb{N}}\) can not both converge to \( f(\xi) \), showing that \( f \) can not be continuous. \( \blacksquare \)

Caveat 10.15. It is important in Prop. 10.14 that \( f \) is defined on a compact interval \( I \). The examples \( f : [0, 1] \rightarrow \mathbb{R}, f(x) := 1/x \), and \( f : \mathbb{R} \rightarrow \mathbb{R}, f(x) := x^2 \) are examples of continuous functions that are not uniformly continuous.

Theorem 10.16. Let \( I = [a, b] \subseteq \mathbb{R}, a \leq b \).

(a) If \( f : I \rightarrow \mathbb{C} \) is continuous, then \( f \) is Riemann integrable over \( I \).

(b) If \( f : I \rightarrow \mathbb{R} \) is increasing or decreasing, then \( f \) is Riemann integrable over \( I \).

Proof. (a): As \( f \) is continuous if, and only if, \( \text{Re} \ f \) and \( \text{Im} \ f \) are both continuous, it suffices to consider a real-valued continuous \( f \). For \( a = b \), there is nothing to prove, so let \( a < b \). First note that, if \( f \) is continuous on \( I = [a, b] \), then \( f \) is bounded by Th. 7.54. Moreover, \( f \) is uniformly continuous due to Prop. 10.14. Thus, given \( \epsilon > 0 \), there is \( \delta > 0 \) such that \( |x - y| < \delta \) implies \( |f(x) - f(y)| < \epsilon/|I| \) for each \( x, y \in I \). Then, for each partition \( \Delta \) of \( I \) satisfying \( |\Delta| < \delta \), we obtain
\[
R(\Delta, f) - r(\Delta, f) = \sum_{j=1}^{N} (M_j - m_j)|I_j| \leq \epsilon/|I| \sum_{j=1}^{N} |I_j| = \epsilon,
\] (10.42)
as \( |\Delta| < \delta \) implies \( |x - y| < \delta \) for each \( x, y \in I_j \) and each \( j \in \{1, \ldots, N\} \). Finally, (10.42) implies \( f \in \mathcal{R}(I) \) due to Riemann’s integrability criterion of Th. 10.13.

(b): Suppose \( f : [a, b] \rightarrow \mathbb{R} \) is increasing. Then \( f \) is bounded, as \( f(a) \leq f(x) \leq f(b) \) for each \( x \in [a, b] \). If \( f(a) = f(b) \), then \( f \) is constant. Thus, assume \( f(a) < f(b) \). Moreover, if \( \Delta = (x_0, \ldots, x_N) \) is a partition of \( I \) as in Def. 10.3, then
\[
R(\Delta, f) - r(\Delta, f) = \sum_{j=1}^{N} (M_j - m_j)|I_j| = \sum_{j=1}^{N} (f(x_j) - f(x_{j-1}))*|I_j| \leq |\Delta|(f(b) - f(a)).
\] (10.43)
Thus, given $\epsilon > 0$, we have $R(\Delta, f) - r(\Delta, f) < \epsilon$ for each partition $\Delta$ of $I$ satisfying $|\Delta| < \epsilon/(f(b) - f(a))$. In consequence, $f \in \mathcal{R}(I)$, once again due to Riemann’s integrability criterion of Th. 10.13. If $f$ is decreasing, then $-f$ is increasing, and Th. 10.11(a) establishes the case.

**Definition and Remark 10.17.** Let $M \subseteq \mathbb{C}$. A function $f : M \to \mathbb{C}$ is called **Lipschitz continuous** in $M$ with Lipschitz constant $L$ if, and only if,

$$
\exists L \in \mathbb{R}^+ \quad \forall x, y \in M \quad |f(x) - f(y)| \leq L|x - y|.
$$

(10.44)

Every Lipschitz continuous function is, indeed, continuous, since, if $\xi \in M$ and $(y_n)_{n \in \mathbb{N}}$ is a sequence in $M$ with $\lim_{n \to \infty} y_n = \xi$, then (10.44) implies

$$
\forall n \in \mathbb{N} \quad |f(\xi) - f(y_n)| \leq L|\xi - y_n|,
$$

(10.45)

proving $\lim_{n \to \infty} f(y_n) = f(\xi)$. Moreover, it is not too much harder to prove Lipschitz continuous functions are even uniformly continuous, but we will not pursue this right now. On the other hand, $f : \mathbb{R}_0^+ \to \mathbb{R}$, $f(x) := \sqrt{x}$, is an example of a continuous function (actually, even uniformly continuous) that is not Lipschitz continuous.

**Theorem 10.18.** Let $a, b \in \mathbb{R}$, $a \leq b$, $I := [a, b]$.

(a) If $f \in \mathcal{R}(I, \mathbb{R})$ and $\phi : f(I) \to \mathbb{C}$ is Lipschitz continuous, then $\phi \circ f \in \mathcal{R}(I, \mathbb{C})$.

(b) If $f \in \mathcal{R}(I, \mathbb{C})$ and $\phi : f(I) \to \mathbb{R}$ is Lipschitz continuous, then $\phi \circ f \in \mathcal{R}(I, \mathbb{R})$.

(c) If $f \in \mathcal{R}(I)$, then $|f|, f^2, f^+, f^- \in \mathcal{R}(I)$. In particular, we, indeed, have (10.2) from the introduction (with $M$ replaced by $I$). If, in addition, there exists $\delta > 0$ such that $f(x) \geq \delta$ for each $x \in I$, then $1/f \in \mathcal{R}(I)$. Moreover, $|f| \in \mathcal{R}(I)$ also holds for $f \in \mathcal{R}(I, \mathbb{C})$.

(d) If $f, g \in \mathcal{R}(I)$, then $\max(f, g), \min(f, g) \in \mathcal{R}(I)$. If $f, g \in \mathcal{R}(I, \mathbb{K})$, then $\tilde{f}, \tilde{g} \in \mathcal{R}(I, \mathbb{K})$. If, in addition, there exists $\delta > 0$ such that $|g(x)| \geq \delta$ for each $x \in I$, then $f/g \in \mathcal{R}(I, \mathbb{K})$.

**Proof.** (a), (b): We just carry out the proof for the case $f \in \mathcal{R}(I)$, $\phi : f(I) \to \mathbb{R}$, and leave the extension to the case that $f$ or $\phi$ is $\mathbb{C}$-valued as an exercise. Thus, let $f \in \mathcal{R}(I)$ and let $\phi : f(I) \to \mathbb{R}$ be Lipschitz continuous. Then there exists $L \geq 0$ such that

$$
|\phi(x) - \phi(y)| \leq L|x - y|
$$

(10.46)

for each $x, y \in f(I)$. As $f \in \mathcal{R}(I)$, given $\epsilon > 0$, Th. 10.13 provides a partition $\Delta$ of $I$ such that $R(\Delta, f) - r(\Delta, f) < \epsilon/L$, and we obtain

$$
R(\Delta, \phi \circ f) - r(\Delta, \phi \circ f) = \sum_{j=1}^{N} (M_j(\phi \circ f) - m_j(\phi \circ f))|I_j|
$$

$$
\leq \sum_{j=1}^{N} L(M_j(f) - m_j(f))|I_j|
$$

$$
= L \left( R(\Delta, f) - r(\Delta, f) \right) < \epsilon.
$$

(10.47)
Thus, $\phi \circ f \in \mathcal{R}(I)$ by another application of Th. 10.13.

(c): $|f|, f^2, f^+, f^- \in \mathcal{R}(I)$ follows from (b) (for $|f|$ also for $f \in \mathcal{R}(I, \mathbb{C})$), since each of the maps $x \mapsto |x|, x \mapsto x^2, x \mapsto \max\{x, 0\}, x \mapsto -\min\{x, 0\}$ is Lipschitz continuous on the bounded set $f(I)$ (recall that $f \in \mathcal{R}(I)$ implies that $f$ is bounded). Since $f = f^+ - f^-$, (10.2) is implied by (10.28). Finally, if $f(x) \geq \delta > 0$, then $x \mapsto x^{-1}$ is Lipschitz continuous on the bounded set $f(I)$, and $f^{-1} \in \mathcal{R}(I)$ follows from (b).

(d): For $f, g \in \mathcal{R}(I)$, we note that, due to

\begin{align*}
fg &= \frac{1}{4}(f + g)^2 - (f - g)^2, \quad (10.48a) \\
\max(f, g) &= f + (g - f)^+,
\min(f, g) &= g - (f - g)^-,
\end{align*}

everything is a consequence of (c). For $f, g \in \mathcal{R}(I, \mathbb{C})$, due to

\begin{align*}
\bar{f} &= (\text{Re}f, -\text{Im}f), \quad (10.48d) \\
fg &= (\text{Re}f \text{Re}g - \text{Im}f \text{Im}g, \text{Re}f \text{Im}g + \text{Im}f \text{Re}g), \\
1/g &= (\text{Re}g/|g|^2, -\text{Im}g/|g|^2),
\end{align*}

everything follows from the real-valued case together with (c) and Th. 10.11(a), where $|g| \geq \delta > 0$ guarantees $|g|^2 \geq \delta^2 > 0$.

**10.2 Important Theorems**

This section compiles a number of important theorems on Riemann integrals, which, in particular, provide powerful tools to actually evaluate such integrals.

**10.2.1 Fundamental Theorem of Calculus**

We provide two variants of the fundamental theorem with slightly different flavors: In the first variant, Th. 10.20(a), we start with a function $f$, obtain another function $F$ by means of integrating $f$, and recover $f$ by taking the derivative of $F$. In the second variant, Th. 10.20(b), one first differentiates the given function $F$, obtaining $f := F'$, followed by integrating $f$, recovering $F$ up to an additive constant.

**Notation 10.19.** If $a, b \in \mathbb{R}, a \leq b$, $I := [a, b], f : I \rightarrow \mathbb{C}$, then denote

\begin{align*}
\int_a^b f := \int_I f, \quad &\int_a^b f := -\int_b^a f, \quad (10.49a) \\
[f(t)]_a^b := [f]_a^b := f(b) - f(a), \quad &[f(t)]_b^a := [f]_b^a := f(a) - f(b), \quad (10.49b)
\end{align*}

where $f \in \mathcal{R}(I, \mathbb{C})$ for (10.49a).

**Theorem 10.20.** Let $a, b \in \mathbb{R}, a < b$, $I := [a, b]$. 

(a) If \( f \in \mathcal{R}(I, \mathbb{K}) \) is continuous in \( \xi \in I \), then, for each \( c \in I \), the function
\[
F_c : I \rightarrow \mathbb{K}, \quad F_c(x) := \int_c^x f(t) \, dt,
\]
is differentiable in \( \xi \) with \( F'_c(\xi) = f(\xi) \). In particular, if \( f \in C(I, \mathbb{K}) \), then \( F_c \in C^1(I, \mathbb{K}) \) and \( F'_c(x) = f(x) \) for each \( x \in I \).

(b) If \( F \in C^1(I, \mathbb{K}) \) or, alternatively, \( F : I \rightarrow \mathbb{K} \) is differentiable with integrable derivative \( F' \in \mathcal{R}(I, \mathbb{K}) \), then
\[
F(b) - F(a) = [F(t)]_a^b = \int_a^b F'(t) \, dt,
\]and
\[
F(x) = F(c) + \int_c^x F'(t) \, dt \quad \text{for each } c, x \in I.
\]

Proof. It suffices to prove the case \( \mathbb{K} = \mathbb{R} \), since the case \( \mathbb{K} = \mathbb{C} \) then follows by applying the case \( \mathbb{K} = \mathbb{R} \) to \( \text{Re} \, F_c \) and \( \text{Im} \, F_c \) (for (a)) and to \( \text{Re} \, F \) and \( \text{Im} \, F \) (for (b)). Thus, for the rest of the proof, we assume \( \mathbb{K} = \mathbb{R} \).

(a): We need to show that
\[
\lim_{h \rightarrow 0} A(h) = 0, \quad \text{where } A(h) := \frac{F_c(\xi + h) - F_c(\xi)}{h} - f(\xi).
\]

One computes
\[
A(h) = \frac{1}{h} \int_\xi^{\xi+h} f(t) \, dt - \frac{1}{h} f(\xi) \int_\xi^{\xi+h} dt = \frac{1}{h} \int_\xi^{\xi+h} (f(t) - f(\xi)) \, dt.
\]

Now, given \( \epsilon > 0 \), the continuity of \( f \) in \( \xi \) allows us to find \( \delta > 0 \) such that \( |(f(t) - f(\xi))| < \epsilon/2 \) for each \( t \in I \) with \( |t - \xi| < \delta \). Thus, for each \( h \) with \( |h| < \delta \), we obtain
\[
|A(h)| \leq \frac{1}{h} \int_\xi^{\xi+h} |f(t) - f(\xi)| \, dt \leq \frac{\epsilon h}{2h} < \epsilon,
\]
thereby proving \( \lim_{h \rightarrow 0} A(h) = 0 \), i.e. \( f(\xi) = F'_c(\xi) \).

(b): First assume \( F \in C^1(I) \). Then \( F' \) is continuous on \( I \), and we can apply (a) to the function
\[
G : I \rightarrow \mathbb{R}, \quad G(x) := \int_a^x F'(t) \, dt,
\]
to obtain \( G' = F' \). Thus, for \( H := F - G \), we obtain \( H' = 0 \), showing that \( H \) must be constant on \( I \), i.e. \( H(x) = H(a) = F(a) - G(a) = F(a) \) for each \( x \in I \). Evaluating at \( x = b \) yields
\[
F(a) = H(b) = F(b) - \int_a^b F'(t) \, dt,
\]
thereby establishing the case.

Now we consider the remaining case of a differentiable $F$ with integrable derivative $F' \in \mathcal{R}(I)$. Consider a partition $\Delta = (x_0, \ldots, x_N)$ of $I$ as in Def. 10.3. Then, for each $j \in \{1, \ldots, N\}$, the mean value theorem provides $\xi_j \in ]x_{j-1}, x_j[$ such that $F(x_j) - F(x_{j-1}) = (x_j - x_{j-1})F'(\xi_j)$. Thus,

$$F(b) - F(a) = \sum_{j=1}^{N} (F(x_j) - F(x_{j-1})) = \sum_{j=1}^{N} (x_j - x_{j-1}) F'(\xi_j) = \rho(\Delta, F'). \quad (10.57)$$

If we choose a sequence of partitions $\Delta$ of $I$ such that $|\Delta| \to 0$, then the integrability of $f$ implies that the right-hand side of (10.57) converges to $\int_a^b F'$, once again establishing the case. ■

**Definition 10.21.** If $I \subseteq \mathbb{R}$, $f : I \rightarrow \mathbb{K}$, and $F : I \rightarrow \mathbb{K}$ is a differentiable function with $F' = f$, then $F$ is called a primitive or antiderivative of $f$.

**Example 10.22.** (a) Due to the fundamental theorem, if we know a function’s antiderivative, we can easily compute its integral over a given interval. Here are three simple examples:

$$\int_0^1 (x^5 - 3x) \, dx = \left[ \frac{x^6}{6} - \frac{3x^2}{2} \right]_0^1 = \frac{1}{6} - \frac{3}{2} = -\frac{4}{3}, \quad (10.58a)$$

$$\int_1^e \frac{1}{x} \, dx = [\ln x]_1^e = \ln e - \ln 1 = 1, \quad (10.58b)$$

$$\int_0^\pi \sin x \, dx = [-\cos x]_0^\pi = 2. \quad (10.58c)$$

(b) As a more complicated example, consider a rational function $x \mapsto r(x) = p(x)/q(x)$ with polynomials $p$ and $q$. To find an antiderivative of $r$, one first writes $r$ in the form $r = s + \tilde{p}/\tilde{q}$ with polynomials $s, \tilde{p}, \tilde{q}$ such that $\deg(\tilde{p}) < \deg(\tilde{q})$ (this can be done using so-called polynomial long division). One then applies the partial fraction decomposition of Sec. G of the Appendix to $\tilde{p}/\tilde{q}$: As a concrete example, consider

$$r : \mathbb{R} \setminus \{-1, 2\} \rightarrow \mathbb{R}, \quad r(x) := \frac{3x^8 - 9x^6 - 6x^5 + x^3 + 5x^2 - 1}{x^3 - 3x - 2}.$$

One then finds

$$r(x) = 3x^5 + 1 + \frac{5x^2 + 3x + 1}{(x + 1)^2(x - 2)},$$

and the partial fraction decomposition according to (G.5) is

$$r(x) = 3x^5 + 1 + \frac{2}{x + 1} - \frac{1}{(x + 1)^2} + \frac{3}{x - 2}.$$

One can now easily provide an antiderivative for each summand, and putting everything together yields the antiderivative

$$R : \mathbb{R} \setminus \{-1, 2\} \rightarrow \mathbb{R}, \quad R(x) := \frac{1}{2}x^6 + x + 2 \ln |x + 1| + \frac{1}{x + 1} + 3 \ln |x - 2|,$$
for \( r \). We note that, to apply partial fraction decomposition, one always needs the zeros of the denominator (i.e. of \( \tilde{q} \)). For polynomials of high degree, it will usually not be possible to determine these zeros exactly, but merely approximately.

10.2.2 Integration by Parts Formula

**Theorem 10.23.** Let \( a, b \in \mathbb{R}, a < b, I := [a, b] \). If \( f, g \in C^1(I, \mathbb{C}) \), then the following integration by parts formula holds:

\[
\int_a^b fg' = [fg]_a^b - \int_a^b f'g. \tag{10.59}
\]

**Proof.** If \( f, g \in C^1(I, \mathbb{C}) \), then, according to the product rule, \( fg \in C^1(I, \mathbb{C}) \) with \( (fg)' = f'g + fg' \). Applying (10.51a), we obtain

\[
[fg]_a^b = \int_a^b (fg)' = \int_a^b f'g + \int_a^b fg', \tag{10.60}
\]

which is precisely (10.59).

**Example 10.24.** We compute the integral \( \int_0^{2\pi} \sin^2 t \, dt \):

\[
\int_0^{2\pi} \sin^2 t \, dt = [-\sin t \cos t]_0^{2\pi} + \int_0^{2\pi} \cos^2 t \, dt = \int_0^{2\pi} \cos^2 t \, dt. \tag{10.61}
\]

Adding \( \int_0^{2\pi} \sin^2 t \, dt \) on both sides of (10.61) and using \( \sin^2 + \cos^2 \equiv 1 \) yields

\[
2 \int_0^{2\pi} \sin^2 t \, dt = \int_0^{2\pi} 1 \, dt = 2\pi, \tag{10.62}
\]

i.e. \( \int_0^{2\pi} \sin^2 t \, dt = \pi. \)

10.2.3 Change of Variables

**Theorem 10.25.** Let \( I, J \subseteq \mathbb{R} \) be intervals, \( \phi \in C^1(I) \) and \( f \in C(J, \mathbb{C}) \). If \( \phi(I) \subseteq J \), then the following change of variables formula holds for each \( a, b \in I \):

\[
\int_{\phi(a)}^{\phi(b)} f = \int_{\phi(a)}^{\phi(b)} f(x) \, dx = \int_a^b f(\phi(t)) \phi'(t) \, dt = \int_a^b (f \circ \phi) \phi'. \tag{10.63}
\]

**Proof.** Let

\[
F : J \rightarrow \mathbb{C}, \quad F(x) := \int_{\phi(a)}^{x} f(t) \, dt. \tag{10.64}
\]
According to Th. 10.20(a) and the chain rule of Th. 9.11, we obtain
\[(F \circ \phi)' : I \rightarrow \mathbb{C}, \quad (F \circ \phi)'(x) = \phi'(x)f(\phi(x)).\] (10.65)

Thus, we can apply (10.51a), which yields
\[
\int_{\phi(a)}^{\phi(b)} f = F(\phi(b)) - F(\phi(a)) = \int_a^b (f \circ \phi) \phi',
\] (10.66)
proving (10.63).

\[\square\]

**Example 10.26.** We compute the integral \[\int_0^1 t^2 \sqrt{1-t} \, dt\] using the change of variables \(x := \phi(t) := 1 - t, \phi'(t) = -1:\)

\[
\int_0^1 t^2 \sqrt{1-t} \, dt = - \int_1^0 (1-x)^{\frac{3}{2}} \sqrt{x} \, dx = \int_0^1 (\sqrt{x} - 2x\sqrt{x} + x^2\sqrt{x}) \, dx
\]
\[= \left[ \frac{2x^{\frac{7}{2}}}{3} - \frac{4x^{\frac{5}{2}}}{5} + \frac{2x^{\frac{7}{2}}}{7} \right]_0^1 = \frac{16}{105}.
\] (10.67)

### 10.3 Application: Taylor’s Theorem

**Theorem 10.27** (Taylor’s Theorem). Let \(I \subseteq \mathbb{R}\) be an open interval and \(a, x \in I, x \neq a\). If \(m \in \mathbb{N}_0\) and \(f \in C^{m+1}(I, \mathbb{K})\), then

\[f(x) = T_m(x, a) + R_m(x, a),\] (10.68)

where

\[T_m(x, a) := \sum_{k=0}^{m} \frac{f^{(k)}(a)}{k!}(x-a)^k = f(a) + f'(a)(x-a) + \frac{f''(a)}{2!}(x-a)^2 + \cdots + \frac{f^{(m)}(a)}{m!}(x-a)^m\] (10.69)

is the \(m\)th Taylor polynomial and

\[R_m(x, a) := \int_a^x \frac{(x-t)^m}{m!} f^{(m+1)}(t) \, dt\] (10.70)

is the integral form of the remainder term. For \(\mathbb{K} = \mathbb{R}\), one can also write the remainder term in Lagrange form:

\[R_m(x, a) = \frac{f^{(m+1)}(\theta)}{(m+1)!}(x-a)^{m+1} \quad \text{with some suitable } \theta \in ]x, a[.\] (10.71)

**Proof.** The integral form (10.70) of the remainder term we prove by using induction on \(m\): For \(m = 0\), the assertion is

\[f(x) = f(a) + \int_a^x f'(t) \, dt,\] (10.72)
which holds according to the fundamental theorem of calculus in the form Th. 10.20(b).

For the induction step, we assume (10.68) holds for fixed $m \in \mathbb{N}_0$ with $R_m(x, a)$ in integral form (10.70) and consider $f \in C^{m+2}(I, \mathbb{K})$. For fixed $x \in I$, we define the function

$$g : I \to \mathbb{K}, \quad g(t) := \frac{(x - t)^{m+1}}{(m+1)!} f^{(m+1)}(t).$$

(10.73)

Using the product rule, its derivative is

$$g' : I \to \mathbb{K}, \quad g'(t) = \frac{(x - t)^{m+1}}{(m+1)!} f^{(m+2)}(t) - \frac{(x - t)^m}{m!} f^{(m+1)}(t).$$

(10.74)

Applying the fundamental theorem to $g$ then yields

$$-g(a) = g(x) - g(a) = \int_a^x g'(t) \, dt \overset{(10.74)}{=} \frac{f^{(m+1)}(a)}{(m+1)!} (x - a)^{m+1} + R_m(x, a) - g(a),$$

(10.75)

with $R_m(x, a)$ and $R_{m+1}(x, a)$ defined according to (10.70). Thus,

$$T_{m+1}(x, a) + R_{m+1}(x, a) \overset{(10.75)}{=} T_m(x, a) + \frac{f^{(m+1)}(a)}{(m+1)!} (x - a)^{m+1} + R_m(x, a) - g(a)$$

$$= T_m(x, a) + R_m(x, a) \overset{\text{ind. hyp.}}{=} f(x),$$

(10.76)

thereby completing the induction and the proof of (10.70).

It remains to prove the Lagrange form (10.71) of the remainder term for $\mathbb{K} = \mathbb{R}$. Since $(x - t)^m > 0$ for $x > t$ or $m = 0$, and $(x - t)^m < 0$ for $x < t$ and $m > 0$, (10.70) and (10.36b) imply the existence of some $\theta \in [x, a]$ such that

$$R_m(x, a) = f^{(m+1)}(\theta) \int_a^x \frac{(x - t)^m}{m!} \, dt = \frac{f^{(m+1)}(\theta)}{(m+1)!} (x - a)^{m+1},$$

(10.77)

proving the Lagrange form, where it remains an exercise to show one can always choose $\theta \in ]x, a[$.

\[\square\]

**Remark 10.28.** The importance of Taylor’s Th. 10.27 does not lie in the decomposition $f = T_m + R_m$, which can be accomplished simply by defining $R_m := f - T_m$. The importance lies rather in the specific formulas for the remainder term.

**Example 10.29.** Depending on $f \in C^\infty(I)$ and $x, a \in I$, the Taylor series

$$(T_m(x, a))_{m \in \mathbb{N}} = \sum_{k=0}^\infty \frac{f^{(k)}(a)}{k!} (x - a)^k$$

can converge to $f(x)$ (see (a) below), diverge (see (b) below), or converge to $\eta \neq f(x)$ (see (c) below):

(a) For $f : \mathbb{R} \to \mathbb{R}$, $f(x) = e^x$, $x \in \mathbb{R}$, $a = 0$, we have $f^{(k)}(a) = e^0 = 1$ for each $k \in \mathbb{N}_0$, recovering the power series for the exponential function, which we already know to converge from Def. and Rem. 8.14:

$$\lim_{m \to \infty} T_m(x, 0) = \lim_{m \to \infty} \sum_{k=0}^m \frac{1}{k!} x^k = e^x.$$
(b) For \( f: \mathbb{R}^+ \to \mathbb{R} \), \( f(x) = \ln x \), we have \( f^{(m)}(x) = (-1)^{m-1}(m-1)!x^{-m} \) for each \( m \in \mathbb{N}_0 \) and each \( x \in \mathbb{R}^+ \). Thus, for \( x = \frac{3}{2} \) and \( a = \frac{1}{2} \), using \( f^{(k)}(a) = (-1)^{k-1}(k-1)!2^k \), we have

\[
T_m(x,a) = \sum_{k=0}^{m} \frac{f^{(k)}(a)}{k!}(x-a)^k = \sum_{k=0}^{m} \frac{(-1)^{k-1}2^k}{k},
\]

which diverges for \( m \to \infty \), since the summands do not converge to 0.

(c) For \( f: \mathbb{R} \to \mathbb{R} \), \( f(x) := \begin{cases} e^{-1/x^2} & \text{for } x \neq 0, \\ 0 & \text{for } x = 0, \end{cases} \)

it is an exercise to show \( f \in C^\infty(\mathbb{R}) \) with \( f^{(k)}(0) = 0 \) for each \( k \in \mathbb{N}_0 \) (hint: for \( x \neq 0 \), one obtains \( f^{(k)}(x) = P_k(x^{-1})e^{-1/x^2} \), where \( P_k \) is a polynomial of degree \( 3k \)). Thus, \( T_m(x,0) = 0 \) for each \( x \in \mathbb{R} \) and each \( m \in \mathbb{N}_0 \), implying

\[
\lim_{m \to \infty} T_m(x,0) = 0 \neq f(x) \quad \text{for each } x \in \mathbb{R} \setminus \{0\}.
\]

10.4 Improper Integrals

For our definition of the Riemann integral in Def. 10.5, it was important that we considered bounded functions on compact intervals (where the boundedness of the intervals was more important than the closedness) – for unbounded functions and/or unbounded intervals, even Def. 10.4 of lower and upper Riemann sums no longer makes sense.

Still, for sufficiently benign functions, it is possible to extend the notion of a definite Riemann integral to both unbounded intervals and unbounded functions, and in such situations we will speak of improper integrals (cf. Def. 10.33 below).

**Definition 10.30.** Let \( \emptyset \neq I \subseteq \mathbb{R} \) be an interval. We call \( f: I \to \mathbb{R} \) to be locally Riemann integrable if, and only if, \( f \in \mathcal{R}(J) \) for each compact interval \( J \subseteq I \). Let \( \mathcal{R}_{\text{loc}}(I) \) denote the set of all locally Riemann integrable functions on \( I \).

**Remark 10.31.** In particular, locally Riemann integrable functions are bounded on every compact interval. Moreover, \( \mathcal{R}_{\text{loc}}(I) = \mathcal{R}(I) \) if, and only if, \( I \) is a compact interval. For example, for each \( a, b \in \mathbb{R} \) with \( a < b \), the function given by the assignment rule

\[
f(x) := \frac{1}{(x-a)(x-b)}
\]

is clearly locally Riemann integrable, but not bounded on each of the intervals \( ]-\infty, a[, ]a, b[, \) and \( ]b, \infty[ \).

Before we can define improper Riemann integrals, we define, in partial extension of Def. 8.17:
Definition 10.32. Let $M \subseteq \mathbb{R}$. If $M$ is unbounded from above (resp. below, then $f : M \to \mathbb{K}$ is said to tend to $\eta \in \mathbb{K}$ (or to have the limit $\eta \in \mathbb{K}$) for $x \to \infty$ (resp., for $x \to -\infty$) (denoted by $\lim_{x \to \pm \infty} f(x) = \eta$) if, and only if, for each sequence $(\xi_k)_{k \in \mathbb{N}}$ in $M$ with $\lim_{k \to \infty} \xi_k = \infty$ (resp. with $\lim_{k \to \infty} \xi_k = -\infty$), the sequence $(f(\xi_k))_{k \in \mathbb{N}}$ converges to $\eta \in \mathbb{K}$, i.e.

$$\lim_{x \to \pm \infty} f(x) = \eta \iff \forall (\xi_k)_{k \in \mathbb{N}} \text{ in } M \left( \lim_{k \to \infty} \xi_k = \pm \infty \implies \lim_{k \to \infty} f(\xi_k) = \eta \right).$$ (10.78)

Definition 10.33. Let $0 < c < b$ (or $a = -\infty$, $b = \infty$ is admissible).

(a) Let $I := [c, b]$, $f \in \mathcal{R}_\text{loc}(I)$, and assume $b = \infty$ or $f$ is unbounded. Consider the function

$$F : I \to \mathbb{R}, \quad F(x) := \int_c^x f.$$

If the limit

$$\lim_{x \to b} F(x) = \lim_{x \to b} \int_c^x f$$

exists in $\mathbb{R}$, then we define

$$\int_I f := \int_c^b f(t) \, dt := \int_c^b f := \lim_{x \to b} \int_c^x f.$$

(b) Let $I := ]a, c]$, $f \in \mathcal{R}_\text{loc}(I)$, and assume $a = -\infty$ or $f$ is unbounded. Consider the function

$$F : I \to \mathbb{R}, \quad F(x) := \int_x^c f.$$

If the limit

$$\lim_{x \to a} F(x) = \lim_{x \to a} \int_x^c f$$

exists in $\mathbb{R}$, then we define

$$\int_I f := \int_a^c f(t) \, dt := \int_a^c f := \lim_{x \to a} \int_x^c f.$$

(c) Let $I := ]a, b]$, $f \in \mathcal{R}_\text{loc}(I)$. If the conditions of both (a) and (b) hold, i.e. (i) – (iv), where

(i) $b = \infty$ or $f$ is unbounded on $[c, b[$,
(ii) $\lim_{x \to b} \int_c^x f$ exists in $\mathbb{R}$,
(iii) $a = -\infty$ or $f$ is unbounded on $]a, c]$,
(iv) $\lim_{x \to a} \int_x^c f$ exists in $\mathbb{R}$,

then we define

$$\int_I f := \int_a^b f(t) \, dt := \int_a^b f := \int_a^c f + \int_c^b f.$$
All the above limits of Riemann integrals (if they exist) are called *improper* Riemann integrals. In each case, if the limit exists, we call \( f \) *improperly Riemann integrable* and write \( f \in \mathcal{R}(I) \).

**Remark 10.34. (a)** The definitions in Def. 10.33 are consistent with what occurs if the limits are *proper* Riemann integrals: Let \( a, c, b \in \mathbb{R} \), \( a < c < b \), and \( f \in \mathcal{R}[a, b] \). Then

\[
\lim_{x \to b} \int_c^x f = \int_c^b f \quad \text{and} \quad \lim_{x \to a} \int_x^c f = \int_a^c f.
\]

Indeed, since \( f \in \mathcal{R}[a, b] \), \(|f| \) is bounded by some \( M \in \mathbb{R}^+ \); and if \((x_k)_{k \in \mathbb{N}} \) is a sequence in \([a, b]\) such that \( \lim_{k \to \infty} x_k = b \), then

\[
\left| \int_{x_k}^b f \right| \leq \int_{x_k}^b |f| \leq M (b - x_k) \to 0 \quad \text{for} \quad k \to \infty,
\]

implying

\[
\lim_{k \to \infty} \int_{c}^{x_k} f \overset{\text{Th. 10.11(b)}}{=} \lim_{k \to \infty} \left( \int_{c}^{b} f - \int_{x_k}^{b} f \right) = \int_{c}^{b} f - 0 = \int_{c}^{b} f.
\]

An analogous argument shows the remaining equality in (10.81).

**Remark 10.34. (b)** In Def. 10.33(c), it can occur that \( \int_{-\infty}^{\infty} f \) does *not* exist, even though the limit \( \lim_{x \to \infty} \int_{-x}^{x} f \) exists: For example, if \( f : \mathbb{R} \to \mathbb{R} \), \( f(x) = x \), then \( f \in \mathcal{R}_{\text{loc}}(\mathbb{R}) \), and, for each sequence \((x_k)_{k \in \mathbb{N}} \) in \( \mathbb{R} \) such that \( \lim_{k \to \infty} x_k = \infty \) and each \( c \in \mathbb{R} \), one has

\[
\lim_{k \to \infty} \int_{-x_k}^{x_k} t \, dt = \lim_{k \to \infty} \left[ \frac{t^2}{2} \right]_{-x_k}^{x_k} = \lim_{k \to \infty} \frac{x_k^2 - x_k^2}{2} = 0,
\]

\[
\lim_{k \to \infty} \int_{c}^{x_k} t \, dt = \lim_{k \to \infty} \left[ \frac{t^2}{2} \right]_{c}^{x_k} = \lim_{k \to \infty} \frac{x_k^2 - c^2}{2} = \infty,
\]

\[
\lim_{k \to \infty} \int_{-x_k}^{c} t \, dt = \lim_{k \to \infty} \left[ \frac{t^2}{2} \right]_{-x_k}^{c} = \lim_{k \to \infty} \frac{c^2 - x_k^2}{2} = -\infty,
\]

i.e. \( \lim_{x \to \infty} \int_{-x}^{x} t \, dt = 0 \), but neither \( \lim_{x \to \infty} \int_{c}^{x} t \, dt \) nor \( \lim_{x \to -\infty} \int_{x}^{c} t \, dt \) exists in \( \mathbb{R} \).

**Remark 10.34. (c)** Let \( a < c_1 < c_2 < b \) (\( a = -\infty \), \( b = \infty \) is admissible). If \( I := [c_1, b] \), \( f \in \mathcal{R}_{\text{loc}}(I) \), and \( b = \infty \) or \( f \) is unbounded, then \( \int_{c_1}^{b} f \) exists if, and only if, \( \int_{c_2}^{b} f \) exists. Moreover, if the integrals exist, then

\[
\int_{c_1}^{b} f = \int_{c_1}^{c_2} f + \int_{c_2}^{b} f.
\]

Indeed, if \((x_k)_{k \to \infty} \) is a sequence in \([c_1, b]\) such that \( \lim_{k \to \infty} x_k = b \) and if \( \int_{c_1}^{b} f \) exists, then

\[
\lim_{k \to \infty} \int_{c_2}^{x_k} f \overset{\text{Th. 10.11(b)}}{=} \lim_{k \to \infty} \left( \int_{c_2}^{x_k} f - \int_{c_2}^{c_2} f \right) = \int_{c_1}^{b} f - \int_{c_1}^{c_2} f,
\]
proving \( \int_{c_2}^{b} f \) exists and (10.82a) holds. Conversely, if \( \int_{c_2}^{b} f \) exists, then

\[
\lim_{k \to \infty} \int_{c_1}^{x_k} f = \lim_{k \to \infty} \left( \int_{c_2}^{x_k} f + \int_{c_2}^{c_1} f \right) = \int_{c_2}^{b} f + \int_{c_1}^{c_2} f,
\]

proving \( \int_{c_1}^{b} f \) exists and (10.82a) holds. Analogously, one shows that if \( I := [a, c_2] \), \( f \in \mathcal{R}_{\text{loc}}(I) \), and \( a = -\infty \) or \( f \) is unbounded, then \( \int_{c_1}^{c_2} f \) exists if, and only if, \( \int_{a}^{c_2} f \) exists, where, if the integrals exist, then

\[
\int_{a}^{c_2} f = \int_{c_1}^{c_2} f + \int_{a}^{c_1} f. \tag{10.82b}
\]

In particular, we see that neither the existence nor the value of the improper integral in Def. 10.33(c) depends on the choice of \( c \).

**Example 10.35.** (a) Let \( 0 < \alpha < 1 \). We claim that

\[
\int_{0}^{1} \frac{1}{t^\alpha} \, dt = \frac{1}{1 - \alpha} \quad \left( \alpha = \frac{1}{2} \text{ yields } \int_{0}^{1} \frac{1}{\sqrt{t}} \, dt = 2 \right). \tag{10.83}
\]

Indeed, if \((x_k)_{k \in \mathbb{N}}\) is a sequence in \([0,1]\) such that \( \lim_{k \to \infty} x_k = 0 \), then

\[
\lim_{k \to \infty} \int_{x_k}^{1} \frac{1}{t^\alpha} \, dt = \lim_{k \to \infty} \left[ t^{1-\alpha} \right]_{x_k}^{1} = \lim_{k \to \infty} \frac{1 - x_k^{1-\alpha}}{1 - \alpha} = \frac{1}{1 - \alpha}.
\]

(b) If \((x_k)_{k \in \mathbb{N}}\) is a sequence in \([0,1]\) such that \( \lim_{k \to \infty} x_k = 0 \), then

\[
\lim_{k \to \infty} \int_{x_k}^{1} \frac{1}{t} \, dt = \lim_{k \to \infty} \left[ \ln t \right]_{x_k}^{1} = \lim_{k \to \infty} \left( 0 - \ln x_k \right) = \infty,
\]

showing the limit does not exist in \( \mathbb{R} \), but diverges to \( \infty \). Sometimes, this is stated in the form

\[
\int_{0}^{1} \frac{1}{t} \, dt = \infty. \tag{10.84}
\]

(c) We claim that

\[
\int_{-\infty}^{0} e^t \, dt = 1. \tag{10.85}
\]

Indeed, if \((x_k)_{k \in \mathbb{N}}\) is a sequence in \( \mathbb{R}_0^- \) such that \( \lim_{k \to \infty} x_k = -\infty \), then

\[
\lim_{k \to \infty} \int_{x_k}^{0} e^t \, dt = \lim_{k \to \infty} \left[ e^t \right]_{x_k}^{0} = \lim_{k \to \infty} \left( 1 - e^{x_k} \right) = 1.
\]

(d) Consider the function

\[
f : \mathbb{R}_0^+ \to \mathbb{R}, \quad f(t) := \begin{cases} 
n \quad \text{for } n \leq t \leq n + \frac{1}{n^2}, \ n \in \mathbb{N}, \\
0 \quad \text{otherwise}.
\end{cases}
\]

Then \( \lim_{t \to \infty} f(t) \) does not exist and \( f \) is not even bounded. However \( f \in \mathcal{R}(\mathbb{R}_0^+) \) and

\[
\int_{0}^{\infty} f = \sum_{n=1}^{\infty} \int_{n}^{n+1/(n^2)} n \, dt = \sum_{n=1}^{\infty} 2^{-n} = \frac{1}{1 - \frac{1}{2}} - 1 = 1.
\]
Lemma 10.36. Let $a < c < b$ ($a = -\infty$, $b = \infty$ is admissible). Let $I \subseteq ]a, b[$ be one of the three kinds of intervals occurring in Def. 10.33 (i.e. $I = [c, b[$, $I = ]a, c]$, or $I = ]a, b[$), and assume $f, g : I \rightarrow \mathbb{R}$ to be improperly Riemann integrable over $I$.

(a) Linearity: For each $\lambda, \mu \in \mathbb{R}$, $\lambda f + \mu g$ is improperly Riemann integrable over $I$ and
\[
\int_I (\lambda f + \mu g) = \lambda \int_I f + \mu \int_I g.
\]

(b) Monotonicity: If $f \leq g$, then
\[
\int_I f \leq \int_I g.
\]

Proof. We conduct the proof for the case $I = [c, b[$ – the case $I = ]a, c]$ can be shown analogously, and the case $I = ]a, b[$ then also follows. Let $(x_k)_{k \in \mathbb{N}}$ be a sequence in $I$ such that $\lim_{k \to \infty} x_k = b$.

(a): One computes
\[
\lim_{k \to \infty} \int_c^{x_k} (\lambda f + \mu g) = \lim_{k \to \infty} \left( \lambda \int_c^{x_k} f + \mu \int_c^{x_k} g \right) = \lambda \int_c^b f + \mu \int_c^b g,
\]
showing $(\lambda f + \mu g) \in \mathcal{R}(I)$ and proving (a).

(b): One estimates
\[
\int_c^b f = \lim_{k \to \infty} \int_c^{x_k} f \leq \lim_{k \to \infty} \int_c^{x_k} g = \int_c^b g,
\]
proving (b). \[\Box\]

Definition 10.37. Let $a < c < b$ ($a = -\infty$, $b = \infty$ is admissible). Let $I \subseteq ]a, b[$ be one of the three kinds of intervals occurring in Def. 10.33 (i.e. $I = [c, b[$, $I = ]a, c]$, or $I = ]a, b[$), and assume $f \in \mathcal{R}_{\text{loc}}(I)$. Then, by Th. 10.18(c), $|f| \in \mathcal{R}_{\text{loc}}(I)$.

Before we can proceed to Prop. 10.39 about convergence criteria for improper integrals, we need to prove the analogon of Th. 7.19 for limits of functions.

**Proposition 10.38.** Let $\emptyset \neq M \subseteq \mathbb{R}$, $a \in \mathbb{R} \cup \{-\infty\}$, $b \in \mathbb{R} \cup \{\infty\}$, and assume
\[
a = \begin{cases} 
\inf(M \setminus \{a\}) & \text{if } M \text{ is bounded from below,} \\
-\infty & \text{if } M \text{ is unbounded from below,}
\end{cases}
\]
\[
b = \begin{cases} 
\sup(M \setminus \{a\}) & \text{if } M \text{ is bounded from above,} \\
\infty & \text{if } M \text{ is unbounded from above.}
\end{cases}
\]
Let \( f : M \rightarrow \mathbb{R} \) be monotone (increasing or decreasing). Defining \( A := f(M) = \{ f(x) : x \in M \} \), the following holds:

\[
\lim_{x \to b} f(x) = \begin{cases} 
\sup A & \text{if } f \text{ is increasing and } A \text{ is bounded from above}, \\
\infty & \text{if } f \text{ is increasing and } A \text{ is not bounded from above}, \\
\inf A & \text{if } f \text{ is decreasing and } A \text{ is bounded from below}, \\
-\infty & \text{if } f \text{ is decreasing and } A \text{ is not bounded from below},
\end{cases} \tag{10.87a}
\]

\[
\lim_{x \to a} f(x) = \begin{cases} 
\sup A & \text{if } f \text{ is decreasing and } A \text{ is bounded from above}, \\
\infty & \text{if } f \text{ is decreasing and } A \text{ is not bounded from above}, \\
\inf A & \text{if } f \text{ is increasing and } A \text{ is bounded from below}, \\
-\infty & \text{if } f \text{ is increasing and } A \text{ is not bounded from below}.
\end{cases} \tag{10.87b}
\]

**Proof.** We prove (10.87a) for the case where \( f \) is increasing – the remaining case of (10.87a) as well as (10.87b) can be proved completely analogous. Let \((x_k)_{k \in \mathbb{N}}\) be a sequence in \( M \setminus \{b\} \) such that \( \lim_{k \to \infty} x_k = b \). We have to show that \( \lim_{k \to \infty} f(x_k) = \eta \), where \( \eta := \sup A \) for \( A \) bounded from above and \( \eta := \infty \) for \( A \) not bounded from above. Seeking a contradiction, assume \( \lim_{k \to \infty} f(x_k) = \eta \) does not hold. Due to the choice of \( b \), there then must be \( \epsilon > 0 \) and a subsequence \((y_k)_{k \in \mathbb{N}}\) of \((x_k)_{k \in \mathbb{N}}\) such that \((y_k)_{k \in \mathbb{N}}\) is strictly increasing and

\[
\forall k \in \mathbb{N} \quad f(y_k) \leq \begin{cases} 
\eta - \epsilon & \text{if } \eta = \sup A, \\
\epsilon & \text{if } \eta = \infty.
\end{cases}
\]

Since \( \lim_{k \to \infty} y_k = b \) and \( f \) is increasing, this means \( \sup A \leq \eta - \epsilon \) or \( \sup A = \epsilon \), which means a contradiction in each case. Thus, \( \lim_{k \to \infty} f(x_k) = \eta \) must hold and the proof is complete.

**Proposition 10.39.** Let \( a < c < b \) (\( a = -\infty \), \( b = \infty \) is admissible). Let \( I \subseteq ]a, b[ \) be one of the three kinds of intervals occurring in Def. 10.33 (i.e. \( I = [c, b[ \), \( I = ]a, c] \), or \( I = ]a, b[ \)), and assume \( f \in \mathcal{R}_{\text{loc}}(I) \).

(a) If \( g \in \mathcal{R}_{\text{loc}}(I) \), \( 0 \leq f \leq g \), and \( \int_I g \) exists, then \( \int_I f \) exists as well. Conversely, if \( 0 \leq g \leq f \) and \( \int_I g \) diverges, then \( \int_I f \) diverges as well.

(b) If \( \int_I f \) is an improper integral that is absolutely convergent, then it is also convergent.

**Proof.** (a): We consider the case \( I = [c, b[ \) – the proof for the case \( I = ]a, c] \) is completely analogous, and the case \( I = ]a, b[ \) then also follows. First, suppose \( 0 \leq f \leq g \), and \( \int_I g \) exists. Since \( 0 \leq f \), the function

\[
F : [c, b[ \rightarrow \mathbb{R}_+^*, \quad F(x) := \int_c^x f,
\]

satisfies...
is increasing. Due to
\[ F(x) = \int_c^x f \leq \int_c^x g \leq \int_c^b g \in \mathbb{R}_0^+ , \]
\( F \) is also bounded from above (in the sense that \( \{ F(x) : x \in [c,b] \} \) is bounded from above), i.e. Prop. 10.38 yields that \( \lim_{x \to b} F(x) = \lim_{x \to b} \int_c^x f \) exists in \( \mathbb{R} \) as claimed.
Now suppose \( 0 \leq g \leq f \) and \( \int_I g \) diverges. As the function \( F \) above, the function
\[ G : [c,b] \to \mathbb{R}_0^+, \quad G(x) := \int_c^x g , \]
is increasing. Since we assume that \( \lim_{x \to b} G(x) \) does not exist in \( \mathbb{R} \), Prop. 10.38 implies \( \lim_{x \to b} G(x) = \infty \). As a consequence, if \( (x_k)_{k \in \mathbb{N}} \) is a sequence in \( [c,b] \) such that \( \lim_{k \to \infty} x_k = b \), then
\[ \lim_{k \to \infty} F(x_k) = \lim_{k \to \infty} \int_c^{x_k} f = \lim_{k \to \infty} \int_c^{x_k} g = \infty , \]
showing that \( \int_I f \) diverges as well.

(b): We assume \( \int_I f \) to converge absolutely, i.e. \( \int_I |f| \) must exist in \( \mathbb{R} \). Since \( 0 \leq f^+ \leq |f| \) and \( 0 \leq f^- \leq |f| \), (a) then implies the existence of \( \int_I f^+ + \int_I f^- \) and of \( \int_I f^- \). Thus, according to Lem. 10.36(a), \( \int_I f = \int_I f^+ - \int_I f^- \) must also exist.

**Example 10.40.** (a) We will use Prop. 10.39(a) to show that the improper integral
\[ \int_0^\infty e^{-t^2} dt \]
exists. Indeed,
\[ \forall t \in \mathbb{R} \quad ((t - 1)^2 = t^2 - 2t + 1 \geq 0 \implies -t^2 \leq -2t + 1 \implies 0 \leq e^{-t^2} \leq e^{-2t+1}) , \]
and, since
\[ \int_0^\infty e^{-2t+1} dt = \lim_{x \to \infty} \int_0^x e^{-2t+1} dt = \lim_{x \to \infty} \left[ -\frac{e^{-2t+1}}{2} \right]_0^x = \lim_{x \to \infty} \frac{e - e^{-2x+1}}{2} = \frac{e}{2} , \]
Prop. 10.39(a) implies that \( \int_0^\infty e^{-t^2} dt \) exists in \( \mathbb{R} \).

(b) We will use Prop. 10.39(a) to show that
\[ \int_0^\infty e^{t^2} dt \]
diverges. Indeed,
\[ \forall t \in \mathbb{R} \quad (t^2 \geq 0 \implies e^{t^2} \geq 1) , \]
and, since
\[ \lim_{x \to \infty} \int_0^x 1 dt = \lim_{x \to \infty} x = \infty , \]
Prop. 10.39(a) implies that \( \int_0^\infty e^{t^2} dt = \infty \).
We provide an example that shows an improper integral can converge without converging absolutely: Consider the function

$$f : [0, \infty[ \rightarrow \mathbb{R}, \quad f(t) := \begin{cases} (-1)^{n+1} & \text{for } n \leq t \leq n + \frac{1}{n}, \ n \in \mathbb{N}, \\ 0 & \text{otherwise.} \end{cases}$$

Then

$$\int_{0}^{\infty} |f| = \lim_{k \to \infty} \sum_{k=1}^{n} \int_{k}^{k+\frac{1}{k}} 1 \, dt = \lim_{k \to \infty} \sum_{k=1}^{n} \frac{1}{k} = \infty,$$

showing $\int_{0}^{\infty} f$ does not converge absolutely. However, we will show

$$\int_{0}^{\infty} f = \sum_{j=1}^{\infty} \frac{(-1)^{j+1}}{j} =: \alpha > 0.$$  \hfill (10.90)

We know $\alpha > 0$ from Ex. 7.86(a) and Th. 7.85. Let $(x_k)_{k \in \mathbb{N}}$ be a sequence in $\mathbb{R}^+_{>0}$ such that $\lim_{k \to \infty} x_k = \infty$. Given $\epsilon > 0$, choose $K \in \mathbb{N}$ such that $\frac{1}{K} < \frac{\epsilon}{2}$ and $N \in \mathbb{N}$ such that

$$\forall_{k > N} \quad x_k > K.$$  \hfill (10.91)

Then, for each $k > N$, there exists $K_1 \in \mathbb{N}$ such that $K < K_1 \leq x_k < K_1 + 1$. Thus

$$\int_{0}^{x_k} f(t) \, dt = \min \left\{ x_k - K_1, \frac{1}{K_1} \right\} + \sum_{j=1}^{K_1-1} \frac{(-1)^{j+1}}{j}$$

and

$$\left| \alpha - \int_{0}^{x_k} f(t) \, dt \right| = \left| \sum_{j=K_1}^{\infty} \frac{(-1)^{j+1}}{j} - \min \left\{ x_k - K_1, \frac{1}{K_1} \right\} \right| < \frac{1}{K_1} + \frac{1}{K_1}$$

$$< \frac{2}{K} < 2 \cdot \frac{\epsilon}{2} = \epsilon,$$

thereby proving (10.90).

## A Axiomatic Set Theory

### A.1 Motivation, Russell’s Antinomy

As it turns out, _naive set theory_, founded on the definition of a set according to Cantor (as stated at the beginning of Sec. 1.3) is not suitable to be used in the foundation of mathematics. The problem lies in the possibility of obtaining contradictions such as _Russell’s antinomy_, after Bertrand Russell, who described it in 1901.

Russell’s antinomy is obtained when considering the set $X$ of all sets that do not contain themselves as an element: When asking the question if $X \in X$, one obtains the contradiction that $X \in X \iff X \not\in X$:
Suppose $X \in X$. Then $X$ is a set that contains itself. But $X$ was defined to contain only sets that do not contain themselves, i.e. $X \notin X$.

So suppose $X \notin X$. Then $X$ is a set that does not contain itself. Thus, by the definition of $X$, $X \in X$.

Perhaps you think Russell's construction is rather academic, but it is easily translated into a practical situation. Consider a library. The catalog $C$ of the library should contain all the library's books. Since the catalog itself is a book of the library, it should occur as an entry in the catalog. So there can be catalogs such as $C$ that have themselves as an entry and there can be other catalogs that do not have themselves as an entry. Now one might want to have a catalog $X$ of all catalogs that do not have themselves as an entry. As in Russell's antinomy, one is led to the contradiction that the catalog $X$ must have itself as an entry if, and only if, it does not have itself as an entry.

One can construct arbitrarily many versions, which we will not do. Just one more: Consider a small town with a barber, who, each day, shaves all inhabitants, who do not shave themselves. The poor barber now faces a terrible dilemma: He will have to shave himself if, and only if, he does not shave himself.

To avoid contradictions such as Russell's antinomy, *axiomatic set theory* restricts the construction of sets via so-called axioms, as we will see below.

### A.2 Set-Theoretic Formulas

The contradiction of Russell’s antinomy is related to Cantor’s sets not being hierarchical. Another source of contradictions in naive set theory is the imprecise nature of informal languages such as English. In (1.6), we said that

$$ A := \{ x \in B : P(x) \} $$

defines a subset of $B$ if $P(x)$ is a statement about an element $x$ of $B$. Now take $B := \mathbb{N} := \{ 1, 2, \ldots \}$ to be the set of the natural numbers and let

$$ P(x) := \text{"The number } x \text{ can be defined by fifty English words or less"}. \quad (A.1) $$

Then $A$ is a finite subset of $\mathbb{N}$, since there are only finitely many English words (if you think there might be infinitely many English words, just restrict yourself to the words contained in some concrete dictionary). Then there is a smallest natural number $n$ that is not in $A$. But then $n$ is the smallest natural number that can not be defined by fifty English words or less, which, actually, defines $n$ by less than fifty English words, in contradiction to $n \notin A$.

To avoid contradictions of this type, we require $P(x)$ to be a so-called *set-theoretic formula*.

**Definition A.1. (a)** The language of set theory consists precisely of the following symbols: $\land, \neg, \exists, (,), \in, =, v_j$, where $j = 1, 2, \ldots$. 

(b) A set-theoretic formula is a finite string of symbols from the above language of set theory that can be built using the following recursive rules:

(i) \( v_i \in v_j \) is a set-theoretic formula for \( i, j = 1, 2, \ldots \).

(ii) \( v_i = v_j \) is a set-theoretic formula for \( i, j = 1, 2, \ldots \).

(iii) If \( \phi \) and \( \psi \) are set-theoretic formulas, then \( (\phi) \land (\psi) \) is a set-theoretic formula.

(iv) If \( \phi \) is a set-theoretic formulas, then \( \neg(\phi) \) is a set-theoretic formula.

(v) If \( \phi \) is a set-theoretic formulas, then \( \exists v_j(\phi) \) is a set-theoretic formula for \( j = 1, 2, \ldots \).

Example A.2. Examples of set-theoretic formulas are \((v_3 \in v_5) \land (\neg(v_2 = v_3)), \exists v_1(\neg(v_1 = v_1))\); examples of symbol strings that are not set-theoretic formulas are \( v_1 \in v_2 \in v_3 \), \( \exists \exists \neg \), and \( \in v_3 \exists \).

Remark A.3. It is noted that, for a given finite string of symbols, a computer can, in principle, check in finitely many steps, if the string constitutes a set-theoretic formula or not. The symbols that can occur in a set-theoretic formula are to be interpreted as follows: The variables \( v_1, v_2, \ldots \) are variables for sets. The symbols \( \land \) and \( \neg \) are to be interpreted as the logical operators of conjunction and negation as described in Sec. 1.2.2. Similarly, \( \exists \) stands for an existential quantifier as in Sec. 1.4: The statement \( \exists v_j(\phi) \) means “there exists a set \( v_j \) that has the property \( \phi \)”. Parentheses ( and ) are used to make clear the scope of the logical symbols \( \exists, \land, \neg \). Where the symbol \( \in \) occurs, it is interpreted to mean that the set to the left of \( \in \) is contained as an element in the set to the right of \( \in \). Similarly, \( = \) is interpreted to mean that the sets occurring to the left and to the right of \( = \) are equal.

Remark A.4. A disadvantage of set-theoretic formulas as defined in Def. A.1 is that they quickly become lengthy and unreadable (at least to the human eye). To make formulas more readable and concise, one introduces additional symbols and notation. Formally, additional symbols and notation are always to be interpreted as abbreviations or transcriptions of actual set-theoretic formulas. For example, we use the rules of Th. 1.11 to define the additional logical symbols \( \lor, \Rightarrow, \Leftrightarrow \) as abbreviations:

\[
(\phi) \lor (\psi) \quad \text{is short for} \quad \neg((\neg(\phi)) \land (\neg(\psi))) \quad \text{(cf. Th. 1.11(j))}, \tag{A.2a}
\]

\[
(\phi) \Rightarrow (\psi) \quad \text{is short for} \quad (\neg(\phi)) \lor (\psi) \quad \text{(cf. Th. 1.11(a))}, \tag{A.2b}
\]

\[
(\phi) \Leftrightarrow (\psi) \quad \text{is short for} \quad ((\phi) \Rightarrow (\psi)) \land ((\psi) \Rightarrow (\phi)) \quad \text{(cf. Th. 1.11(b))}. \tag{A.2c}
\]

Similarly, we use (1.17a) to define the universal quantifier:

\[
\forall v_j(\phi) \quad \text{is short for} \quad \neg(\exists v_j(\neg(\phi))). \tag{A.2d}
\]

Further abbreviations and transcriptions are obtained from omitting parentheses if it is clear from the context and/or from Convention 1.10 where to put them in, by writing variables bound by quantifiers under the respective quantifiers (as in Sec. 1.4), and by using other symbols than \( v_j \) for set variables. For example,

\[
\forall_x (\phi \Rightarrow \psi) \quad \text{transcribes} \quad \neg(\exists v_1(\neg((\neg(\phi)) \lor (\psi))))).
\]
Moreover,

\[ v_i \neq v_j \text{ is short for } \neg (v_i = v_j); \quad v_i \notin v_j \text{ is short for } \neg (v_i \in v_j). \quad (A.2e) \]

**Remark A.5.** Even though axiomatic set theory requires the use of set-theoretic formulas as described above, the systematic study of formal symbolic languages is the subject of the field of mathematical logic and is beyond the scope of this class (see, e.g., [EFT07]). In Def. and Rem. 1.15, we defined a proof of statement \( B \) from statement \( A_i \) as a finite sequence of statements \( A_1, A_2, \ldots, A_n \) such that, for \( 1 \leq i < n \), \( A_i \) implies \( A_{i+1} \), and \( A_n \) implies \( B \). In the field of proof theory, also beyond the scope of this class, such proofs are formalized via a finite set of rules that can be applied to (set-theoretic) formulas (see, e.g., [EFT07, Sec. IV], [Kun12, Sec. II]). Once proofs have been formalized in this way, one can, in principle, mechanically check if a given sequence of symbols does, indeed, constitute a valid proof (without even having to understand the actual meaning of the statements). Indeed, several different computer programs have been devised that can be used for automatic proof checking, for example Coq [Wik15a], HOL Light [Wik15b], and Isabelle [Wik15c] to name just a few.

### A.3 The Axioms of Zermelo-Fraenkel Set Theory

Axiomatic set theory seems to provide a solid and consistent foundation for conducting mathematics, and most mathematicians have accepted it as the basis of their everyday work. However, there do remain some deep, difficult, and subtle philosophical issues regarding the foundation of logic and mathematics (see, e.g., [Kun12, Sec. 0, Sec. III]).

**Definition and Remark A.6.** An axiom is a statement that is assumed to be true without any formal logical justification. The most basic axioms (for example, the standard axioms of set theory) are taken to be justified by common sense or some underlying philosophy. However, on a less fundamental (and less philosophical) level, it is a common mathematical strategy to state a number of axioms (for example, the axioms defining the mathematical structure called a group), and then to study the logical consequences of these axioms (for example, group theory studies the statements that are true for all groups as a consequence of the group axioms). For a given system of axioms, the question if there exists an object satisfying all the axioms in the system (i.e. if the system of axioms is consistent, i.e. free of contradictions) can be extremely difficult to answer.

We are now in a position to formulate and discuss the axioms of axiomatic set theory. More precisely, we will present the axioms of Zermelo-Fraenkel set theory, usually abbreviated as ZF, which are Axiom 0 – Axiom 8 below. While there exist various set theories in the literature, each set theory defined by some collection of axioms, the axioms of ZFC, consisting of the axioms of ZF plus the axiom of choice (Axiom 9, see Sec. A.4 below), are used as the foundation of mathematics currently accepted by most mathematicians.
A.3.1 Existence, Extensionality, Comprehension

**Axiom 0** *Existence:*

\[ \exists X (X = X) . \]

Recall that this is just meant to be a more readable transcription of the set-theoretic formula \( \exists v_1 (v_1 = v_1) \). The axiom of existence states that there exists (at least one) set \( X \).

In Def. 1.18 two sets are defined to be equal if, and only if, they contain precisely the same elements. In axiomatic set theory, this is guaranteed by the axiom of extensionality:

**Axiom 1** *Extensionality:*

\[ \forall X \forall Y \left( \forall z (z \in X \iff z \in Y) \Rightarrow X = Y \right) . \]

Following [Kun12], we assume that the substitution property of equality is part of the underlying logic, i.e. if \( X = Y \), then \( X \) can be substituted for \( Y \) and vice versa without changing the truth value of a (set-theoretic) formula. In particular, this yields the converse to extensionality:

\[ \forall X \forall Y \left( X = Y \Rightarrow \forall z (z \in X \iff z \in Y) \right) . \]

Before we discuss further consequences of extensionality, we would like to have the existence of the empty set. However, Axioms 0 and 1 do not suffice to prove the existence of an empty set (see [Kun12, I.6.3]). This, rather, needs the additional axiom of comprehension. More precisely, in the case of comprehension, we do not have a single axiom, but a scheme of infinitely many axioms, one for each set-theoretic formula. Its formulation makes use of the following definition:

**Definition A.7.** One obtains the *universal closure* of a set-theoretic formula \( \phi \), by writing \( \forall v_j \) in front of \( \phi \) for each variable \( v_j \) that occurs as a free variable in \( \phi \) (recall from Def. 1.31 that \( v_j \) is free in \( \phi \) if, and only if, it is not bound by a quantifier in \( \phi \)).

**Axiom 2** *Comprehension Scheme:* For each set-theoretic formula \( \phi \), not containing \( Y \) as a free variable, the universal closure of

\[ \exists Y \forall x \left( x \in Y \iff (x \in X \land \phi) \right) \]

is an axiom. Thus, the comprehension scheme states that, given the set \( X \), there exists (at least one) set \( Y \), containing precisely the elements of \( X \) that have the property \( \phi \).
Remark A.8. Comprehension does not provide uniqueness. However, if both $Y$ and $Y'$ are sets containing precisely the elements of $X$ that have the property $\phi$, then

$$\forall x \left( x \in Y \iff (x \in X \land \phi) \iff x \in Y' \right),$$

and, then, extensionality implies $Y = Y'$. Thus, due to extensionality, the set $Y$ given by comprehension is unique, justifying the notation

$$\{ x : x \in X \land \phi \} := \{ x \in X : \phi \} := Y$$

(A.3)

(this is the axiomatic justification for (1.6)).

Theorem A.9. There exists a unique empty set (which we denote by $\emptyset$ or by 0 – it is common to identify the empty set with the number zero in axiomatic set theory).

Proof. Axiom 0 provides the existence of a set $X$. Then comprehension allows us to define the empty set by

$$0 := \emptyset := \{ x \in X : x \neq x \},$$

where, as explained in Rem. A.8, extensionality guarantees uniqueness. ■

Remark A.10. In Rem. A.4 we said that every formula with additional symbols and notation is to be regarded as an abbreviation or transcription of a set-theoretic formula as defined in Def. A.1(b). Thus, formulas containing symbols for defined sets (e.g. 0 or $\emptyset$ for the empty set) are to be regarded as abbreviations for formulas without such symbols. Some logical subtleties arise from the fact that there is some ambiguity in the way such abbreviations can be resolved: For example, $0 \in X$ can abbreviate either

$$\psi : \exists_y (\phi(y) \land y \in X) \quad \text{or} \quad \chi : \forall_y (\phi(y) \Rightarrow y \in X),$$

where $\phi(y)$ stands for $\forall_v (v \notin y)$. Then $\psi$ and $\chi$ are equivalent if $\exists! \phi(y)$ is true (e.g., if Axioms 0 – 2 hold), but they can be nonequivalent, otherwise (see discussion between Lem. 2.9 and Lem. 2.10 in [Kun80]).

At first glance, the role played by the free variables in $\phi$, which are allowed to occur in Axiom 2, might seem a bit obscure. So let us consider examples to illustrate that allowing free variables (i.e. set parameters) in comprehension is quite natural:

Example A.11. (a) Suppose $\phi$ in comprehension is the formula $x \in Z$ (having $Z$ as a free variable), then the set given by the resulting axiom is merely the intersection of $X$ and $Z$:

$$X \cap Z := \{ x \in X : \phi \} = \{ x \in X : x \in Z \}.$$

(b) Note that it is even allowed for $\phi$ in comprehension to have $X$ as a free variable, so one can let $\phi$ be the formula $\exists_u (x \in u \land u \in X)$ to define the set

$$X^* := \left\{ x \in X : \exists_u (x \in u \land u \in X) \right\}.$$
Then, if \(0 := \emptyset\), \(1 := \{0\}\), \(2 := \{0, 1\}\), we obtain
\[
2^* = \{0\} = 1.
\]

It is a consequence of extensionality that the mathematical universe consists of sets and only of sets: Suppose there were other objects in the mathematical universe, for example a cow \(C\) and a monkey \(M\) (or any other object without elements, other than the empty set) – this would be equivalent to allowing a cow or a monkey (or any other object without elements, other than the empty set) to be considered a set, which would mean that our set-theoretic variables \(v_j\) were allowed to be a cow or a monkey as well. However, extensionality then implies the false statement \(C = M = \emptyset\), thereby excluding cows and monkeys from the mathematical universe.

Similarly, \(\{C\}\) and \(\{M\}\) (or any other object that contains a non-set), can not be inside the mathematical universe. Indeed, otherwise we had
\[
\forall \ x \ (x \in \{C\} \iff x \in \{M\})
\]
(as \(C\) and \(M\) are non-sets) and, by extensionality, \(\{C\} = \{M\}\) were true, in contradiction to a set with a cow inside not being the same as a set with a monkey inside. Thus, we see that all objects of the mathematical universe must be so-called hereditary sets, i.e. sets all of whose elements (thinking of the elements as being the children of the sets) are also sets.

### A.3.2 Classes

As we need to avoid contradictions such as Russell’s antinomy, we must not require the existence of a set \(\{x : \phi\}\) for each set-theoretic formula \(\phi\). However, it can still be useful to think of a “collection” of all sets having the property \(\phi\). Such collections are commonly called classes:

**Definition A.12.**

(a) If \(\phi\) is a set-theoretic formula, then we call \(\{x : \phi\}\) a class, namely the class of all sets that have the property \(\phi\) (typically, \(\phi\) will have \(x\) as a free variable).

(b) If \(\phi\) is a set-theoretic formula, then we say the class \(\{x : \phi\}\) exists (as a set) if, and only if
\[
\exists X \left( \forall x \left( x \in X \iff \phi \right) \right)
\]
(A.4)
is true. Then \(X\) is actually unique by extensionality and we identify \(X\) with the class \(\{x : \phi\}\). If (A.4) is false, then \(\{x : \phi\}\) is called a proper class (and the usual interpretation is that the class is in some sense “too large” to be a set).

**Example A.13.**

(a) Due to Russell’s antinomy of Sec. A.1, we know that \(R := \{x : x \notin x\}\) forms a proper class.
(b) The \textit{universal class} of all sets, \( V := \{ x : x = x \} \), is a proper class. Once again, this is related to Russell’s antinomy: If \( V \) were a set, then
\[
R = \{ x : x \notin x \} = \{ x : x = x \land x \notin x \} = \{ x : x \in V \land x \notin x \}
\]
would also be a set by comprehension. However, this is in contradiction to \( R \) being a proper class by (a).

**Remark A.14.** From the perspective of formal logic, statements involving proper classes are to be regarded as abbreviations for statements without proper classes. For example, it turns out that the class \( G \) of all sets forming a group is a proper class. But we might write \( G \in G \) as an abbreviation for the statement “The set \( G \) is a group.”

### A.3.3 Pairing, Union, Replacement

Axioms 0 – 2 are still consistent with the empty set being the only set in existence (see [Kun12, I.6.13]). The next axiom provides the existence of nonempty sets:

**Axiom 3** \textit{Pairing}:
\[
\forall x \forall y \exists Z (x \in Z \land y \in Z).
\]

Thus, the pairing axiom states that, for all sets \( x \) and \( y \), there exists a set \( Z \) that contains \( x \) and \( y \) as elements.

In consequence of the pairing axiom, the sets
\[
0 := \emptyset, \quad 1 := \{ 0 \}, \quad 2 := \{ 0, 1 \}
\]
all exist. More generally, we may define:

**Definition A.15.** If \( x, y \) are sets and \( Z \) is given by the pairing axiom, then we call

(a) \( \{ x, y \} := \{ u \in Z : u = x \lor u = y \} \) the \textit{unordered pair} given by \( x \) and \( y \),

(b) \( \{ x \} := \{ x, x \} \) the \textit{singleton set} given by \( x \),

(c) \( (x, y) := \{ \{ x \}, \{ x, y \} \} \) the \textit{ordered pair} given by \( x \) and \( y \) (cf. Def. 2.1).

We can now show that ordered pairs behave as expected:

**Lemma A.16.** The following holds true:
\[
\forall x,y,x',y' \left( (x, y) = (x', y') \iff (x = x') \land (y = y') \right).
\]
Proof. “⇐” is merely
\[(x, y) = \{\{x\}, \{x, y\}\} \xrightarrow{x=x', y=y'} \{\{x'\}, \{x', y'\}\} = (x', y').\]

“⇒” is done by distinguishing two cases: If \(x = y\), then
\[\{\{x\}\} = (x, y) = (x', y') = \{\{x'\}, \{x', y'\}\}.\]
Next, by extensionality, we first get \(\{x\} = \{x'\} = \{x', y'\}\), followed by \(x = x' = y'\), establishing the case. If \(x \neq y\), then
\[\{\{x\}, \{x, y\}\} = (x, y) = (x', y') = \{\{x'\}, \{x', y'\}\},\]
where, by extensionality \(\{x\} \neq \{x, y\} \neq \{x'\}\). Thus, using extensionality again, \(\{x\} = \{x'\}\) and \(x = x'\). Next, we conclude
\[\{x, y\} = \{x', y'\} = \{x, y'\}\]
and a last application of extensionality yields \(y = y'\). 

While we now have the existence of the infinitely many different sets 0, \(\{0\}\), \(\{\{0\}\}\), . . . , we are not, yet, able to form sets containing more than two elements. This is remedied by the following axiom:

**Axiom 4 Union:**
\[\forall X \exists Y \forall x \forall M \left( (x \in X \land X \in M) \Rightarrow x \in Y \right).\]

Thus, the union axiom states that, for each set of sets \(M\), there exists a set \(Y\) containing all elements of elements of \(M\).

**Definition A.17.** (a) If \(M\) is a set and \(Y\) is given by the union axiom, then define
\[\bigcup M := \bigcup_{X \in M} X := \left\{ x \in Y : \exists x \in \bigcup \{X, Y\} \right\}.\]
(b) If \(X \) and \(Y\) are sets, then define
\[X \cup Y := \bigcup \{X, Y\}.\]
(c) If \(x, y, z\) are sets, then define
\[\{x, y, z\} := \{x, y\} \cup \{z\}.\]

**Remark A.18.** (a) The definition of set-theoretic unions as
\[\bigcup_{i \in I} A_i := \left\{ x : \exists x \in \bigcup_{i \in I} A_i \right\}\]
in (1.25b) will be equivalent to the definition in Def. A.17(a) if we are allowed to form the set
\[M := \{A_i : i \in I\}.\]
If \(I\) is a set and \(A_i\) is a set for each \(i \in I\), then \(M\) as above will be a set by Axiom 5 below (the axiom of replacement).
(b) In contrast to unions, intersections can be obtained directly from comprehension without the introduction of an additional axiom: For example

\[ X \cap Y := \{ x \in X : x \in Y \}, \]

\[ \bigcap_{i \in I} A_i := \left\{ x \in A_{i_0} : \forall_{i \in I} x \in A_i \right\}, \]

where \( i_0 \in I \neq \emptyset \) is an arbitrary fixed element of \( I \).

(c) The union

\[ \bigcup \emptyset = \bigcup_{X \in \emptyset} X = \bigcup_{i \in \emptyset} A_i = \emptyset \]

is the empty set – in particular, a set. However,

\[ \bigcap \emptyset = \left\{ x : \forall_{X \in \emptyset} x \in X \right\} = V = \left\{ x : \forall_{i \in \emptyset} x \in A_i \right\} = \bigcap_{i \in \emptyset} A_i, \]

i.e. the intersection over the empty set is the class of all sets – in particular, a proper class and not a set.

**Definition A.19.** We define the successor function

\[ x \mapsto S(x) := x \cup \{ x \} \quad \text{(for each set } x). \]

Thus, recalling (A.5), we have \( 1 = S(0), 2 = S(1) \); and we can define \( 3 := S(2), \ldots \) In general, we call the set \( S(x) \) the successor of the set \( x \).

In Def. 2.3 and Def. 2.19, respectively, we define functions and relations in the usual manner, making use of the Cartesian product \( A \times B \) of two sets \( A \) and \( B \), which, according to (2.2) consists of all ordered pairs \( (x, y) \), where \( x \in A \) and \( y \in B \). However, Axioms 0 – 4 are not sufficient to justify the existence of Cartesian products. To obtain Cartesian products, we employ the axiom of replacement. Analogous to the axiom of comprehension, the following axiom of replacement actually consists of a scheme of infinitely many axioms, one for each set-theoretic formula:

**Axiom 5 Replacement Scheme:** For each set-theoretic formula, not containing \( Y \) as a free variable, the universal closure of

\[ \left( \forall_{x \in X} \exists! \phi \right) \Rightarrow \left( \exists_{Y} \forall_{x \in X} \exists_{y \in Y} \phi \right) \]

is an axiom. Thus, the replacement scheme states that if, for each \( x \in X \), there exists a unique \( y \) having the property \( \phi \) (where, in general, \( \phi \) will depend on \( x \)), then there exists a set \( Y \) that, for each \( x \in X \), contains this \( y \) with property \( \phi \). One can view this as obtaining \( Y \) by replacing each \( x \in X \) by the corresponding \( y = y(x) \).
Theorem A.20. If $A$ and $B$ are sets, then the Cartesian product of $A$ and $B$, i.e. the class

$$A \times B := \left\{ x : \exists a \in A \exists b \in B \ x = (a, b) \right\}$$

exists as a set.

Proof. For each $a \in A$, we can use replacement with $X := B$ and $\phi := \phi_a$ being the formula $y = (a, x)$ to obtain the existence of the set

$$\{a\} \times B := \{(a, x) : x \in B\} \tag{A.6a}$$

(in the usual way, comprehension and extensionality were used as well). Analogously, using replacement again with $X := A$ and $\phi$ being the formula $y = \{x\} \times B$, we obtain the existence of the set

$$\mathcal{M} := \{\{x\} \times B : x \in A\}. \tag{A.6b}$$

In a final step, the union axiom now shows

$$\bigcup \mathcal{M} = \bigcup_{a \in A} \{a\} \times B = A \times B \tag{A.6c}$$

to be a set as well. $\blacksquare$

A.3.4 Infinity, Ordinals, Natural Numbers

The following axiom of infinity guarantees the existence of infinite sets (e.g., it will allow us to define the set of natural numbers $\mathbb{N}$, which is infinite by Th. A.46 below).

Axiom 6 Infinity:

$$\exists X \left( 0 \in X \land \forall x \in X \ (x \cup \{x\} \in X) \right).$$

Thus, the infinity axiom states the existence of a set $X$ containing $\emptyset$ (identified with the number 0), and, for each of its elements $x$, its successor $S(x) = x \cup \{x\}$.

In preparation for our official definition of $\mathbb{N}$ in Def. A.41 below, we will study so-called ordinals, which are special sets also of further interest to the field of set theory (the natural numbers will turn out to be precisely the finite ordinals). We also need some notions from the theory of relations, in particular, order relations (cf. Def. 2.19 and Def. 2.25).

Definition A.21. Let $R$ be a relation on a set $X$.

(a) $R$ is called asymmetric if, and only if,

$$\forall_{x,y \in X} \ (xRy \Rightarrow \neg(yRx)), \tag{A.7}$$

i.e. if $x$ is related to $y$ only if $y$ is not related to $x$. 

(b) $R$ is called a **strict partial order** if, and only if, $R$ is asymmetric and transitive. It is noted that this is consistent with Not. 2.26, since, recalling the notation $\Delta(X) := \{(x,x) : x \in X\}$, $R$ is a partial order on $X$ if, and only if, $R \setminus \Delta(X)$ is a strict partial order on $X$. We extend the notions lower/upper bound, min, max, inf, sup of Def. 2.27 to strict partial orders $R$ by applying them to $R \cup \Delta(X)$: We call $x \in X$ a lower bound of $Y \subseteq X$ with respect to $R$, and only if, $x$ is a lower bound of $Y$ with respect to $R \cup \Delta(X)$, and analogous for the other notions.

(c) A strict partial order $R$ is called a **strict total order** or a **strict linear order** if, and only if, for each $x,y \in X$, one has $x = y$ or $xRy$ or $yRx$.

(d) $R$ is called a (strict) **well-order** if, and only if, $R$ is a (strict) total order and every nonempty subset of $X$ has a min with respect to $R$ (for example, the usual $\leq$ constitutes a well-order on $\mathbb{N}$ (see Th. D.5 below), but not on $\mathbb{R}$ (e.g., $\mathbb{R}^+$ does not have a min)).

(e) If $Y \subseteq X$, then the relation on $Y$ defined by

$$xSy \iff xRy$$

is called the **restriction** of $R$ to $Y$, denoted $S = R|_Y$ (usually, one still writes $R$ for the restriction).

**Lemma A.22.** Let $R$ be a relation on a set $X$ and $Y \subseteq X$.

(a) If $R$ is transitive, then $R|_Y$ is transitive.

(b) If $R$ is reflexive, then $R|_Y$ is reflexive.

(c) If $R$ is antisymmetric, then $R|_Y$ is antisymmetric.

(d) If $R$ is asymmetric, then $R|_Y$ is asymmetric.

(e) If $R$ is a (strict) partial order, then $R|_Y$ is a (strict) partial order.

(f) If $R$ is a (strict) total order, then $R|_Y$ is a (strict) total order.

(g) If $R$ is a (strict) well-order, then $R|_Y$ is a (strict) well-order.

**Proof.** (a): If $a, b, c \in Y$ with $aRb$ and $bRc$, then $aRc$, since $a, b, c \in X$ and $R$ is transitive on $X$.

(b): If $a \in Y$, then $a \in X$ and $aRa$, since $R$ is reflexive on $X$.

(c): If $a, b \in Y$ with $aRb$ and $bRa$, then $a = b$, since $a, b \in X$ and $R$ is antisymmetric on $X$.

(d): If $a, b \in Y$ with $aRb$, then $\neg bRa$, since $a, b \in X$ and $R$ is asymmetric on $X$.

(e) follows by combining (a) – (d).
(f): If \( a, b \in Y \) with \( a = b \) and \( \neg aRb \), then \( bRa \), since \( a, b \in X \) and \( R \) is total on \( X \). Combining this with (e) yields (f).

(g): Due to (f), it merely remains to show that every nonempty subset \( Z \subseteq Y \) has a min. However, since \( Z \subseteq X \) and \( R \) is a well-order on \( X \), there is \( m \in Z \) such that \( m \) is a min for \( R \) on \( X \), implying \( m \) to be a min for \( R \) on \( Y \) as well. ■

Remark A.23. Since the universal class \( V \) is not a set, \( \in \) is not a relation in the sense of Def. 2.19. It can be considered as a “class relation”, i.e. a subclass of \( V \times V \), but it is a proper class. However, \( \in \) does constitute a relation in the sense of Def. 2.19 on each set \( X \) (recalling that each element of \( X \) must be a set as well). More precisely, if \( X \) is a set, then so is

\[
R_\in := \{(x, y) \in X \times X : x \in y\}. \tag{A.8a}
\]

Then

\[
\forall_{x,y\in X} (x, y) \in R_\in \iff x \in y. \tag{A.8b}
\]

Definition A.24. A set \( X \) is called transitive if, and only if, every element of \( X \) is also a subset of \( X \):

\[
\forall_{x\in X} x \subseteq X. \tag{A.9a}
\]

Clearly, (A.9a) is equivalent to

\[
\forall_{x,y} \left( x \in y \land y \in X \implies x \in X \right). \tag{A.9b}
\]

Lemma A.25. If \( X, Y \) are transitive sets, then \( X \cap Y \) is a transitive set.

Proof. If \( x \in X \cap Y \) and \( y \in x \), then \( y \in X \) (since \( X \) is transitive) and \( y \in Y \) (since \( Y \) is transitive). Thus \( y \in X \cap Y \), showing \( X \cap Y \) is transitive. ■

Definition A.26. (a) A set \( \alpha \) is called an ordinal number or just an ordinal if, and only if, \( \alpha \) is transitive and \( \in \) constitutes a strict well-order on \( \alpha \). An ordinal \( \alpha \) is called a successor ordinal if, and only if, there exists an ordinal \( \beta \) such that \( \alpha = S(\beta) \), where \( S \) is the successor function of Def. A.19. An ordinal \( \alpha \neq 0 \) is called a limit ordinal if, and only if, it is not a successor ordinal. We denote the class of all ordinals by \( \text{ON} \) (it is a proper class by Cor. A.33 below).

(b) We define

\[
\forall_{\alpha, \beta \in \text{ON}} (\alpha < \beta :\iff \alpha \in \beta), \tag{A.10a}
\]

\[
\forall_{\alpha, \beta \in \text{ON}} (\alpha \leq \beta :\iff \alpha < \beta \lor \alpha = \beta). \tag{A.10b}
\]

Example A.27. Using (A.5), \( 0 = \emptyset \) is an ordinal, and \( 1 = S(0), 2 = S(1) \) are both successor ordinals (in Prop. A.43, we will identify \( \mathbb{N}_0 \) as the smallest limit ordinal). Even though \( X := \{1\} \) and \( Y := \{0, 2\} \) are well-ordered by \( \in \), they are not ordinals, since they are not transitive sets: \( 1 \in X \), but \( 1 \not\in X \) (since \( 0 \in 1 \), but \( 0 \not\in X \)); similarly, \( 1 \in 2 \in Y \), but \( 1 \not\in Y \).
Lemma A.28. No ordinal contains itself, i.e.
\[ \forall \alpha \in \text{ON} \quad \alpha \notin \alpha. \]

Proof. If \( \alpha \) is an ordinal, then \( \in \) is a strict order on \( \alpha \). Due to asymmetry of strict orders, \( x \in x \) can not be true for any element of \( \alpha \), implying that \( \alpha \in \alpha \) can not be true. \( \blacksquare \)

Proposition A.29. Every element of an ordinal is an ordinal, i.e.
\[ \forall \alpha \in \text{ON} \quad (X \in \alpha \Rightarrow X \in \text{ON}) \]
(in other words, ON is a transitive class).

Proof. Let \( \alpha \in \text{ON} \) and \( X \in \alpha \). Since \( \alpha \) is transitive, we have \( X \subseteq \alpha \). As \( \in \) is a strict well-order on \( \alpha \), it must also be a strict well-order on \( X \) by Lem. A.22(g). In consequence, it only remains to prove that \( X \) is transitive as well. To this end, let \( x \in X \). Then \( x \in \alpha \), as \( \alpha \) is transitive. If \( y \in x \), then, using transitivity of \( \alpha \) again, \( y \in \alpha \). Now \( y \in X \), as \( \in \) is transitive on \( \alpha \), proving \( x \subseteq X \), i.e. \( X \) is transitive. \( \blacksquare \)

Proposition A.30. If \( \alpha, \beta \in \text{ON} \), then \( X := \alpha \cap \beta \in \text{ON} \) (we will see in Th. A.35(a) below that, actually, \( \alpha \cap \beta = \min \{ \alpha, \beta \} \)).

Proof. \( X \) is transitive by Lem. A.25, and, since \( X \subseteq \alpha \), \( \in \) is a strict well-order on \( X \) by Lem. A.22(g). \( \blacksquare \)

Proposition A.31. On the class \( \text{ON} \), the relation \( \leq \) (as defined in (A.10)) is the same as the relation \( \subseteq \), i.e.
\[ \forall_{\alpha, \beta \in \text{ON}} \quad (\alpha \leq \beta \Leftrightarrow \alpha \subseteq \beta \Leftrightarrow (\alpha \in \beta \lor \alpha = \beta)). \] (A.11)

Proof. Let \( \alpha, \beta \in \text{ON} \).

Assume \( \alpha \leq \beta \). If \( \alpha = \beta \), then \( \alpha \subseteq \beta \). If \( \alpha \in \beta \), then \( \alpha \subseteq \beta \), since \( \beta \) is transitive.

Conversely, assume \( \alpha \subseteq \beta \) and \( \alpha \neq \beta \). We have to show \( \alpha \in \beta \). To this end, we set \( X := \beta \setminus \alpha \). Then \( X \neq \emptyset \) and, as \( \in \) well-orders \( \beta \), we can let \( m := \min X \). We will show \( m = \alpha \) (note that this will complete the proof, due to \( \alpha = m \in X \subseteq \beta \)). If \( m \in m \), then \( \mu \in \beta \) (since \( m \in \beta \) and \( \beta \) is transitive) and \( \mu \notin X \) (since \( m = \min X \)), implying \( \mu \in \alpha \) (since \( X = \beta \setminus \alpha \)) and, thus, \( m \subseteq \alpha \). Seeking a contradiction, assume \( m \neq \alpha \). Then there must be some \( \gamma \in \alpha \setminus m \subseteq \alpha \subseteq \beta \). In consequence \( \gamma, m \in \beta \). As \( \gamma \notin m \) and \( \in \) is a total order on \( \beta \), we must have either \( m = \gamma \) or \( m \in \gamma \). However, \( m \neq \gamma \), since \( \gamma \in \alpha \) and \( m \notin \alpha \) (as \( m \in X \)). So it must be \( m \in \gamma \in \alpha \), implying \( m \in \alpha \), as \( \beta \) is transitive. This contradiction proves \( m = \alpha \) and establishes the proposition. \( \blacksquare \)

Theorem A.32. The class \( \text{ON} \) is well-ordered by \( \in \), i.e.
Axiomatic Set Theory

(i) $\in$ is transitive on $\text{ON}$:

$$\forall_{\alpha, \beta, \gamma \in \text{ON}} \left( \alpha < \beta \land \beta < \gamma \Rightarrow \alpha < \gamma \right).$$

(ii) $\in$ is asymmetric on $\text{ON}$:

$$\forall_{\alpha, \beta \in \text{ON}} \left( \alpha < \beta \Rightarrow \neg (\beta < \alpha) \right).$$

(iii) Ordinals are always comparable:

$$\forall_{\alpha, \beta \in \text{ON}} \left( \alpha < \beta \lor \beta < \alpha \lor \alpha = \beta \right).$$

(iv) Every nonempty set of ordinals has a min.

Proof. (i) is clear, as $\gamma$ is a transitive set.

(ii): If $\alpha, \beta \in \text{ON}$, then $\alpha \in \beta \in \alpha$ implies $\alpha \in \alpha$ by (i), which is a contradiction to Lem. A.28.

(iii): Let $\gamma := \alpha \cap \beta$. Then $\gamma \in \text{ON}$ by Prop. A.30. Thus

$$\gamma \subseteq \alpha \land \gamma \subseteq \beta \Rightarrow (\gamma \in \alpha \lor \gamma = \alpha) \land (\gamma \in \beta \lor \gamma = \beta).$$

(A.12)

If $\gamma \in \alpha$ and $\gamma \in \beta$, then $\gamma \in \alpha \cap \beta = \gamma$, in contradiction to Lem. A.28. Thus, by (A.12), $\gamma = \alpha$ or $\gamma = \beta$. If $\gamma = \alpha$, then $\alpha \subseteq \beta$. If $\gamma = \beta$, then $\beta \subseteq \alpha$, completing the proof of (iii).

(iv): Let $X$ be a nonempty set of ordinals and consider $\alpha \in X$. If $\alpha = \min X$, then we are already done. Otherwise, $Y := \alpha \cap X = \beta \in X : \beta \in \alpha \neq \emptyset$. Since $\alpha$ is well-ordered by $\in$, there is $m := \min Y$. If $\beta \in X$, then either $\beta < \alpha$ or $\alpha \leq \beta$ by (iii). If $\beta < \alpha$, then $\beta \in Y$ and $m \leq \beta$. If $\alpha \leq \beta$, then $m < \alpha \leq \beta$. Thus, $m = \min X$, proving (iv).

Corollary A.33. $\text{ON}$ is a proper class (i.e. there is no set containing all the ordinals).

Proof. If there is a set $X$ containing all ordinals, then, by comprehension, $\beta := \text{ON} = \{ \alpha \in X : \alpha \text{ is an ordinal} \}$ must be a set as well. But then Prop. A.29 says that the set $\beta$ is transitive and Th. A.32 yields that the set $\beta$ is well-ordered by $\in$, implying $\beta$ to be an ordinal, i.e. $\beta \in \beta$ in contradiction to Lem. A.28.

Corollary A.34. For each set $X$ of ordinals, we have:

(a) $X$ is well-ordered by $\in$.

(b) $X$ is an ordinal if, and only if, $X$ is transitive. Note: A transitive set of ordinals $X$ is sometimes called an initial segment of $\text{ON}$, since, here, transitivity can be restated in the form

$$\forall_{\alpha \in \text{ON}} \forall_{\beta \in X} (\alpha < \beta \Rightarrow \alpha \in X).$$

(A.13)
Proof. (a) is a simple consequence of Th. A.32(i)-(iv).
(b) is immediate from (a).

Theorem A.35. Let $X$ be a nonempty set of ordinals.

(a) Then $\gamma := \bigcap X$ is an ordinal, namely $\gamma = \min X$. In particular, if $\alpha, \beta \in \text{ON}$, then $\min\{\alpha, \beta\} = \alpha \cap \beta$.

(b) Then $\delta := \bigcup X$ is an ordinal, namely $\delta = \sup X$. In particular, if $\alpha, \beta \in \text{ON}$, then $\max\{\alpha, \beta\} = \alpha \cup \beta$.

Proof. (a): Let $m := \min X$. Then $\gamma \subseteq m$, since $m \in X$. Conversely, if $\alpha \in X$, then $m \leq \alpha$ implies $m \subseteq \alpha$ by Prop. A.31, i.e. $m \subseteq \gamma$. Thus, $m = \gamma$, proving (a).

(b): To show $\delta \in \text{ON}$, we need to show $\delta$ is transitive (then $\delta$ is an ordinal by Cor. A.34(b)). If $\alpha \in \delta$, then there is $\beta \in X$ such that $\alpha \in \beta$. Thus, if $\gamma \in \beta$, then $\gamma \in \delta$, i.e. $\alpha \leq \delta$, showing $\delta$ to be an upper bound for $X$. Now let $u \in \text{ON}$ be an arbitrary upper bound for $X$, i.e.

$$\forall_{\alpha \in X} \quad \alpha \subseteq u.$$ 

Thus, $\delta \subseteq u$, i.e. $\delta \leq u$, proving $\delta = \sup X$. ■

Next, we obtain some results regarding the successor function of Def. A.19 in the context of ordinals.

Lemma A.36. We have

$$\forall_{\alpha \in \text{ON}} \quad \left( x, y \in S(\alpha) \land x \in y \Rightarrow x \neq \alpha \right).$$

Proof. Seeking a contradiction, we reason as follows:

$$x = \alpha \quad \Rightarrow \quad y \neq \alpha \quad \Rightarrow \quad y \in S(\alpha) \quad \Rightarrow \quad y \in \alpha \quad \text{transitive} \quad \Rightarrow \quad y \subseteq \alpha \quad \Rightarrow \quad x \notin \alpha \quad \Rightarrow \quad \alpha \in \alpha.$$

This contradiction to $\alpha \notin \alpha$ yields $x \neq \alpha$, concluding the proof. ■

Proposition A.37. For each $\alpha \in \text{ON}$, the following holds:

(a) $S(\alpha) \in \text{ON}$.

(b) $\alpha < S(\alpha)$.

(c) For each ordinal $\beta$, $\beta < S(\alpha)$ holds if, and only if, $\beta \leq \alpha$.

(d) For each ordinal $\beta$, if $\beta < \alpha$, then $S(\beta) < S(\alpha)$.

(e) For each ordinal $\beta$, if $S(\beta) < S(\alpha)$, then $\beta < \alpha$. 
Proof. (a): Due to Prop. A.29, \( S(\alpha) \) is a set of ordinals. Thus, by Cor. A.34(b), it merely remains to prove that \( S(\alpha) \) is transitive. Let \( x \in S(\alpha) \). If \( x = \alpha \), then \( x = \alpha \subseteq \alpha \cup \{ \alpha \} = S(\alpha) \). If \( x \neq \alpha \), then \( x \in \alpha \) and, since \( \alpha \) is transitive, this implies \( x \subseteq \alpha \subseteq S(\alpha) \), showing \( S(\alpha) \) to be transitive, thereby completing the proof of (a).

(b) holds, as \( \alpha \in S(\alpha) \) holds by the definition of \( S(\alpha) \).

(c) is clear, since, for each ordinal \( \beta \),
\[
\beta < S(\alpha) \iff \beta \in S(\alpha) \iff \beta \in \alpha \lor \beta = \alpha \iff \beta \leq \alpha.
\]

(d): If \( \beta < \alpha \), then \( S(\beta) = \beta \cup \{ \beta \} \subseteq \alpha \), i.e. \( S(\beta) \leq \alpha < S(\alpha) \).

(e) follows from (d) using contraposition: If \( \neg(\beta < \alpha) \), then \( \beta = \alpha \) or \( \alpha < \beta \), implying \( S(\beta) = S(\alpha) \) or \( S(\alpha) < S(\beta) \), i.e. \( \neg(S(\beta) < S(\alpha)) \).

We now proceed to define the natural numbers:

**Definition A.38.** An ordinal \( n \) is called a natural number if, and only if,
\[
n \neq 0 \land \forall m \in \text{ON} \ (m \leq n \Rightarrow m = 0 \lor m \text{ is successor ordinal}).
\]

**Proposition A.39.** If \( n = 0 \) or \( n \) is a natural number, then \( S(n) \) is a natural number and every element of \( n \) is a natural number or 0.

Proof. Suppose \( n \) is 0 or a natural number. If \( m \in n \), then \( m \) is an ordinal by Prop. A.29. Suppose \( m \neq 0 \) and \( k \in m \). Then \( k \in n \), since \( n \) is transitive. Since \( n \) is a natural number, \( k = 0 \) or \( k \) is a successor ordinal. Thus, \( m \) is a natural number. It remains to show that \( S(n) \) is a natural number. By definition, \( S(n) = n \cup \{ n \} \neq 0 \). Moreover, \( S(n) \in \text{ON} \) by Prop. A.37(a), and, thus, \( S(n) \) is a successor ordinal. If \( m \in S(n) \), then \( m \leq n \), implying \( m = 0 \) or \( m \) is a successor ordinal, completing the proof that \( S(n) \) is a natural number.

**Theorem A.40** (Principle of Induction). If \( X \) is a set satisfying
\[
0 \in X \land \forall x \in X \ S(x) \in X,
\]
then \( X \) contains 0 and all natural numbers.

Proof. Let \( X \) be a set satisfying (A.14). Then 0 \( \in X \) is immediate. Let \( n \) be a natural number and, seeking a contradiction, assume \( n \notin X \). Consider \( N := S(n) \setminus X \). According to Prop. A.39, \( S(n) \) is a natural number and all nonzero elements of \( S(n) \) are natural numbers. Since \( N \subseteq S(n) \) and 0 \( \in X \), 0 \( \notin N \) and all elements of \( N \) must be natural numbers. As \( n \in N \), \( N \neq 0 \). Since \( S(n) \) is well-ordered by \( \in \) and \( 0 \neq N \subseteq S(n) \), \( N \) must have a \( \min m \in N \), \( 0 \neq m \leq n \). Since \( m \) is a natural number, there must be \( k \) such that \( m = S(k) \). Then \( k < m \), implying \( k \notin N \). On the other hand
\[
k < m \land m \leq n \Rightarrow k \leq n \Rightarrow k \in S(n).
\]
Thus, \( k \in X \), implying \( m = S(k) \in X \), in contradiction to \( m \in N \). This contradiction proves \( n \in X \), thereby establishing the case.
**Definition A.41.** If the set $X$ is given by the axiom of infinity, then we use comprehension to define the set

$$
\mathbb{N}_0 := \{ n \in X : n = 0 \lor n \text{ is a natural number} \}
$$

and note $\mathbb{N}_0$ to be unique by extensionality. We also denote $\mathbb{N} := \mathbb{N}_0 \setminus \{0\}$. In set theory, it is also very common to use the symbol $\omega$ for the set $\mathbb{N}_0$.

**Corollary A.42.** $\mathbb{N}_0$ is the set of all natural numbers and 0, i.e.

$$
\forall n \quad (n \in \mathbb{N}_0 \iff n = 0 \lor n \text{ is a natural number}).
$$

**Proof.** “$\Rightarrow$” is clear from Def. A.41 and “$\Leftarrow$” is due to Th. A.40.

**Proposition A.43.** $\omega = \mathbb{N}_0$ is the smallest limit ordinal.

**Proof.** Since $\omega$ is a set of ordinals and $\omega$ is transitive by Prop. A.39, $\omega$ is an ordinal by Cor. A.34(b). Moreover $\omega \neq 0$, since $0 \in \omega$; and $\omega$ is not a successor ordinal (if $\omega = S(n) = n \cup \{n\}$, then $n \in \omega$ and $S(n) \in \omega$ by Prop. A.39, in contradiction to $\omega = S(n)$), implying it is a limit ordinal. To see that $\omega$ is the smallest limit ordinal, let $\alpha \in \text{ON}$, $\alpha < \omega$. Then $\alpha \in \omega$, that means $\alpha = 0$ or $\alpha$ is a natural number (in particular, a successor ordinal).

In the following Th. A.44, we will prove that $\mathbb{N}$ satisfies the Peano axioms P1 – P3 of Sec. 3.1 (if one prefers, one can show the same for $\mathbb{N}_0$, where 0 takes over the role of 1).

**Theorem A.44.** The set of natural numbers $\mathbb{N}$ satisfies the Peano axioms P1 – P3 of Sec. 3.1.

**Proof.** For P1 and P2, we have to show that, for each $n \in \mathbb{N}$, one has $S(n) \in \mathbb{N} \setminus \{1\}$ and that $S(m) \neq S(n)$ for each $m, n \in \mathbb{N}, m \neq n$. Let $n \in \mathbb{N}$. Then $S(n) \in \mathbb{N}$ by Prop. A.39. If $S(n) = 1$, then $n \neq S(n) = 1$ by Prop. A.37(b), i.e. $n = 0$, in contradiction to $n \in \mathbb{N}$. If $m, n \in \mathbb{N}$ with $m \neq n$, then $S(m) \neq S(n)$ is due to Prop. A.37(d). To prove P3, suppose $A \subseteq \mathbb{N}$ has the property that $1 \in A$ and $S(n) \in A$ for each $n \in A$. We need to show $A = \mathbb{N}$ (i.e. $\mathbb{N} \subseteq A$, as $A \subseteq \mathbb{N}$ is assumed). Let $X := A \cup \{0\}$. Then $X$ satisfies (A.14) and Th. A.40 yields $\mathbb{N}_0 \subseteq X$. Thus, if $n \in \mathbb{N}$, then $n \in X \setminus \{0\} = A$, showing $\mathbb{N} \subseteq A$.

**Notation A.45.** For each $n \in \mathbb{N}_0$, we introduce the notation $n + 1 := S(n)$ (more generally, one also defines $\alpha + 1 := S(\alpha)$ for each ordinal $\alpha$).

**Theorem A.46.** Let $n \in \mathbb{N}_0$. Then $A := \mathbb{N}_0 \setminus n$ is infinite (see Def. 3.12(b)). In particular, $\mathbb{N}_0$ and $\mathbb{N} = \mathbb{N}_0 \setminus \{0\} = \mathbb{N}_0 \setminus \{0\}$ are infinite.

**Proof.** Since $n \notin n$, we have $n \in A \neq \emptyset$. Thus, if $A$ were finite, then there were a bijection $f : A \rightarrow A_m := \{1, \ldots, m\} = \{k \in \mathbb{N} : k \leq m\}$ for some $m \in \mathbb{N}$. However, we will show by induction on $m \in \mathbb{N}$ that there is no injective map $f : A \rightarrow A_m$. Since
Axiomatic Set Theory

$S(n) \notin n$, we have $S(n) \in A$. Thus, if $f : A \to A_1 = \{1\}$, then $f(n) = f(S(n))$, showing that $f$ is not injective and proving the cases $m = 1$. For the induction step, we proceed by contraposition and show that the existence of an injective map $f : A \to A_{m+1}$, $m \in \mathbb{N}$, (cf. Not. A.45) implies the existence of an injective map $g : A \to A_m$. To this end, let $m \in \mathbb{N}$ and $f : A \to A_{m+1}$ be injective. If $m + 1 \notin f(A)$, then $f$ itself is an injective map into $A_m$. If $m + 1 \in f(A)$, then there is a unique $a \in A$ such that $f(a) = m + 1$. Define

$$g : A \to A_m, \quad g(k) := \begin{cases} f(k) & \text{for } k < a, \\ f(k + 1) & \text{for } a \leq k. \end{cases} \quad (A.15)$$

Then $g$ is well-defined: If $k \in A$ and $a \leq k$, then $k + 1 \in A \setminus \{a\}$, and, since $f$ is injective, $g$ does, indeed, map into $A_m$. We verify $g$ to be injective: If $k, l \in A$, $k < l$, then also $k < l + 1$ and $k + 1 \neq l + 1$ (by Peano axiom P2 – $k + 1 < l + 1$ then also follows, but we do not make use of that here). In each case, $g(k) \neq g(l)$, proving $g$ to be injective. ■

For more basic information regarding ordinals see, e.g., [Kun12, Sec. I.8].

A.3.5 Power Set

There is one more basic construction principle for sets that is not covered by Axioms 0 – 6, namely the formation of power sets. This needs another axiom:

Axiom 7 Power Set:

$$\forall X \exists M \forall Y \left( Y \subseteq X \Rightarrow Y \in M \right).$$

Thus, the power set axiom states that, for each set $X$, there exists a set $M$ that contains all subsets $Y$ of $X$ as elements.

Definition A.47. If $X$ is a set and $M$ is given by the power set axiom, then we call

$$\mathcal{P}(X) := \{Y \in M : Y \subseteq X\}$$

the power set of $X$. Another common notation for $\mathcal{P}(X)$ is $2^X$ (cf. Prop. 2.18).

A.3.6 Foundation

Foundation is, perhaps, the least important of the axioms in ZF. It basically cleanses the mathematical universe of unnecessary “clutter”, i.e. of certain pathological sets that are of no importance to standard mathematics anyway.

Axiom 8 Foundation:

$$\forall X \left( \exists x \in X \Rightarrow \exists y \exists z \left( z \in x \land z \in X \right) \right).$$

Thus, the foundation axiom states that every nonempty set $X$ contains an element $x$ that is disjoint to $X$. 
Theorem A.48. Due to the foundation axiom, the ∈ relation can have no cycles, i.e. there do not exist sets \( x_1, x_2, \ldots, x_n, n \in \mathbb{N} \), such that

\[
x_1 \in x_2 \in \cdots \in x_n \in x_1.
\]  

(A.16a)

In particular, sets can not be members of themselves:

\[
\neg \exists x \in x.
\]  

(A.16b)

Proof. If there were sets \( x_1, x_2, \ldots, x_n, n \in \mathbb{N} \), such that (A.16a) were true, then, by using the pairing axiom and the union axiom, we could form the set

\[
X := \{x_1, \ldots, x_n\}.
\]

Then, in contradiction to the foundation axiom, \( X \cap x_i \neq \emptyset \), for each \( i = 1, \ldots, n \):

Indeed, \( x_n \in X \cap x_1 \), and \( x_{i-1} \in X \cap x_i \) for each \( i = 2, \ldots, n \).  

For a detailed explanation, why “sets” forbidden by foundation do not occur in standard mathematics, anyway, see, e.g., [Kun12, Sec. I.14].

A.4 The Axiom of Choice

In addition to the axioms of ZF discussed in the previous section, there is one more axiom, namely the axiom of choice (AC) that, together with ZF, makes up ZFC, the axiom system at the basis of current standard mathematics. Even though AC is used and accepted by most mathematicians, it does have the reputation of being somewhat less “natural”. Thus, many mathematicians try to avoid the use of AC, where possible, and it is often pointed out explicitly, if a result depends on the use of AC (but this practise is by no means consistent, neither in the literature nor in this class, and one might sometimes be surprised, which seemingly harmless result does actually depend on AC in some subtle nonobvious way). We will now state the axiom:

Axiom 9 Axiom of Choice (AC):

\[
\forall \mathcal{M} \left( \emptyset \notin \mathcal{M} \Rightarrow \exists f: \mathcal{M} \rightarrow \bigcup_{N \in \mathcal{M}} N \left( \forall M \in \mathcal{M} \ f(M) \in M \right) \right).
\]

Thus, the axiom of choice postulates, for each nonempty set \( \mathcal{M} \), whose elements are all nonempty sets, the existence of a choice function, that means a function that assigns, to each \( M \in \mathcal{M} \), an element \( m \in M \).

Example A.49. For example, the axiom of choice postulates, for each nonempty set \( A \), the existence of a choice function on \( \mathcal{P}(A) \setminus \{\emptyset\} \) that assigns each subset of \( A \) one of its elements.
The axiom of choice is remarkable since, at first glance, it seems so natural that one can hardly believe it is not provable from the axioms in ZF. However, one can actually show that it is neither provable nor disprovable from ZF (see, e.g., [Jec73, Th. 3.5, Th. 5.16] – such a result is called an independence proof, see [Kun80] for further material). If you want to convince yourself that the existence of choice functions is, indeed, a tricky matter, try to define a choice function on $\mathcal{P}(\mathbb{R}) \setminus \{\emptyset\}$ without AC (but do not spend too much time on it – one can show this is actually impossible to accomplish).

Theorem A.52 below provides several important equivalences of AC. Its statement and proof needs some preparation. We start by introducing some more relevant notions from the theory of partial orders:

**Definition A.50.** Let $X$ be a set and let $\leq$ be a partial order on $X$.

(a) An element $m \in X$ is called maximal (with respect to $\leq$) if, and only if, there exists no $x \in X$ such that $m < x$ (note that a maximal element does not have to be a max and that a maximal element is not necessarily unique).

(b) A nonempty subset $C$ of $X$ is called a chain if, and only if, $C$ is totally ordered by $\leq$. Moreover, a chain $C$ is called maximal if, and only if, no strict superset $Y$ of $C$ (i.e. no $Y \subseteq X$ such that $C \subsetneq Y$) is a chain.

The following lemma is a bit technical and will be used to prove the implication $\text{AC} \Rightarrow (\text{ii})$ in Th. A.52 (other proofs in the literature often make use of so-called transfinite recursion, but that would mean further developing the theory of ordinals, and we will not pursue this route in this class).

**Lemma A.51.** Let $X$ be a set and let $\emptyset \neq \mathcal{M} \subseteq \mathcal{P}(X)$ be a nonempty set of subsets of $X$. We let $\mathcal{M}$ be partially ordered by inclusion, i.e. setting $A \leq B :\iff A \subseteq B$ for each $A, B \in \mathcal{M}$. Moreover, define

$$\forall_{S \subseteq \mathcal{M}} \bigcup S := \bigcup_{S \in S} S \quad (A.17)$$

and assume

$$\forall_{C \subseteq \mathcal{M}} \big(C \text{ is a chain } \Rightarrow \bigcup C \in \mathcal{M}\big). \quad (A.18)$$

If the function $g : \mathcal{M} \longrightarrow \mathcal{M}$ has the property that

$$\forall_{M \in \mathcal{M}} \big(M \subseteq g(M) \land \#(g(M) \setminus M) \leq 1\big), \quad (A.19)$$

then $g$ has a fixed point, i.e.

$$\exists_{M \in \mathcal{M}} g(M) = M. \quad (A.20)$$

**Proof.** Fix some arbitrary $M_0 \in \mathcal{M}$. We call $\mathcal{T} \subseteq \mathcal{M}$ an $M_0$-tower if, and only if, $\mathcal{T}$ satisfies the following three properties
(i) $M_0 \in \mathcal{T}$.

(ii) If $C \subseteq \mathcal{T}$ is a chain, then $\bigcup C \in \mathcal{T}$.

(iii) Next. For this purpose, fix $N \in \mathcal{T}$.

Case 2: Then, clearly, $\mathcal{T}_1$ is an $M_0$-tower and, in particular, $\mathcal{T} \neq \emptyset$. Next, we note that the intersection of all $M_0$-towers, i.e. $\mathcal{T}_0 := \bigcap_{T \in \mathcal{T}} T$, is also an $M_0$-tower. Clearly, no strict subset of $\mathcal{T}_0$ can be an $M_0$-tower and

$$M \in \mathcal{T}_0 \Rightarrow M \in \mathcal{T}_1 \Rightarrow M_0 \subseteq M.$$  \hspace{1cm} (A.21)

The main work of the rest of the proof consists of showing that $\mathcal{T}_0$ is a chain. To show $\mathcal{T}_0$ to be a chain, define

$$\Gamma := \left\{ M \in \mathcal{T}_0 : \forall N \in \mathcal{T}_0 \left( M \subseteq N \lor N \subseteq M \right) \right\}.  \hspace{1cm} (A.22)$$

We intend to show that $\Gamma = \mathcal{T}_0$ by verifying that $\Gamma$ is an $M_0$-tower. As an intermediate step, we define

$$\forall M \in \Gamma \Phi(M) := \left\{ N \in \mathcal{T}_0 : N \subseteq M \lor g(M) \subseteq N \right\}$$

and also show each $\Phi(M)$ to be an $M_0$-tower. Actually, $\Gamma$ and each $\Phi(M)$ satisfy (i) due to (A.21). To verify $\Gamma$ satisfies (ii), let $C \subseteq \Gamma$ be a chain and $U := \bigcup C$. Then $U \in \mathcal{T}_0$, since $\mathcal{T}_0$ satisfies (ii). If $N \in \mathcal{T}_0$, and $C \subseteq N$ for each $C \in C$, then $U \subseteq N$. If $N \in \mathcal{T}_0$, and there is $C \in C$ such that $C \not\subseteq N$, then $N \subseteq C$ (since $C \in \Gamma$), i.e. $N \subseteq U$, showing $U \in \Gamma$ and $\Gamma$ satisfying (ii). Now, let $M \in \Gamma$. To verify $\Phi(M)$ satisfies (ii), let $C \subseteq \Phi(M)$ be a chain and $U := \bigcup C$. Then $U \in \mathcal{T}_0$, since $\mathcal{T}_0$ satisfies (ii). If $U \subseteq M$, then $U \in \Phi(M)$ as desired. If $U \not\subseteq M$, then there is $x \in U$ such that $x \notin M$. Thus, there is $C \in C$ such that $x \in C$ and $g(M) \subseteq C$ (since $C \in \Phi(M)$), i.e. $g(M) \subseteq U$, showing $U \in \Phi(M)$ also in this case, and $\Phi(M)$ satisfies (ii). We will verify that $\Phi(M)$ satisfies (iii) next. For this purpose, fix $N \in \Phi(M)$. We need to show $g(N) \in \Phi(M)$. We already know $g(N) \in \mathcal{T}_0$, as $\mathcal{T}_0$ satisfies (iii). As $N \in \Phi(M)$, we can now distinguish three cases.

Case 1: $N \not\subseteq M$. In this case, we cannot have $M \subseteq g(N)$ (otherwise, $\#(g(N) \setminus N) \geq 2$ in contradiction to (A.19)). Thus, $g(N) \not\subseteq M$ (since $M \in \Gamma$), showing $g(N) \not\in \Phi(M)$.

Case 2: $N = M$. Then $g(N) = g(M) \in \Phi(M)$ (since $g(M) \in \mathcal{T}_0$ and $g(M) \subseteq g(M)$).

Case 3: $g(M) \subseteq N$. Then $g(M) \subseteq g(N)$ by (A.19), again showing $g(N) \in \Phi(M)$. Thus, we have verified that $\Phi(M)$ satisfies (iii) and, therefore, is an $M_0$-tower. Then, by the definition of $\mathcal{T}_0$, we have $\mathcal{T}_0 \subseteq \Phi(M)$. As we also have $\Phi(M) \subseteq \mathcal{T}_0$ (from the definition of $\Phi(M)$), we have shown

$$\forall M \in \Gamma \Phi(M) = \mathcal{T}_0.$$  \hspace{1cm} (A.23)

As a consequence, if $N \in \mathcal{T}_0$ and $M \in \Gamma$, then $N \in \Phi(M)$ and this means $N \subseteq M \subseteq g(M)$ or $g(M) \subseteq N$, i.e. each $N \in \mathcal{T}_0$ is comparable to $g(M)$, showing $g(M) \in \Gamma$ and $\Gamma$ satisfying (iii), completing the proof that $\Gamma$ is an $M_0$-tower. As with the $\Phi(M)$ above,
we conclude \( \Gamma = \mathcal{T}_0 \), as desired. To conclude the proof of the lemma, we note \( \Gamma = \mathcal{T}_0 \) implies \( \mathcal{T}_0 \) is a chain. We claim that

\[ M := \bigcup \mathcal{T}_0 \]

satisfies (A.20): Indeed, \( M \in \mathcal{T}_0 \), since \( \mathcal{T}_0 \) satisfies (ii). Then \( g(M) \in \mathcal{T}_0 \), since \( \mathcal{T}_0 \) satisfies (iii). We then conclude \( g(M) \subseteq M \) from the definition of \( M \). As we always have \( M \subseteq g(M) \) by (A.19), we have established \( g(M) = M \) and proved the lemma. ■

**Theorem A.52** (Equivalences to the Axiom of Choice). The following statements (i) – (v) are equivalent to the axiom of choice (as stated as Axiom 9 above).

(i) Every Cartesian product \( \prod_{i \in I} A_i \) of nonempty sets \( A_i \), where \( I \) is a nonempty index set, is nonempty (cf. Def. 2.15(c)).

(ii) Hausdorff's Maximal Principle: Every nonempty partially ordered set \( X \) contains a maximal chain (i.e. a maximal totally ordered subset).

(iii) Zorn's Lemma: Let \( X \) be a nonempty partially ordered set. If every chain \( C \subseteq X \) (i.e. every nonempty totally ordered subset of \( X \)) has an upper bound in \( X \) (such chains with upper bounds are sometimes called inductive), then \( X \) contains a maximal element (cf. Def. A.50(a)).

(iv) Zermelo's Well-Ordering Theorem: Every set can be well-ordered (recall the definition of a well-order from Def. A.21(d)).

(v) Every vector space \( V \) over a field \( F \) has a basis \( B \subseteq V \).

**Proof.** "(i) \( \Leftrightarrow \) AC": Assume (i). Given a nonempty set of nonempty sets \( \mathcal{M} \), let \( I := \mathcal{M} \) and, for each \( M \in \mathcal{M} \), let \( A_M := M \). If \( f \in \prod_{M \in I} A_M \), then, according to Def. 2.15(c), for each \( M \in I = \mathcal{M} \), one has \( f(M) \in A_M = M \), proving AC holds. Conversely, assume AC. Consider a family \((A_i)_{i \in I}\) such that \( I \neq \emptyset \) and each \( A_i \neq \emptyset \). Let \( \mathcal{M} := \{ A_i : i \in I \} \). Then, by AC, there is a map \( g : \mathcal{M} \rightarrow \bigcup_{N \in \mathcal{M}} N = \bigcup_{j \in I} A_j \) such that \( g(M) \in M \) for each \( M \in \mathcal{M} \). Then we can define

\[ f : I \rightarrow \bigcup_{j \in I} A_j, \quad f(i) := g(A_i) \in A_i, \]

to prove (i).

Next, we will show AC \( \Rightarrow \) (ii) \( \Rightarrow \) (iii) \( \Rightarrow \) (iv) \( \Rightarrow \) AC.

"AC \( \Rightarrow \) (ii)": Assume AC and let \( X \) be a nonempty partially ordered set. Let \( \mathcal{M} \) be the set of all chains in \( X \) (i.e. the set of all nonempty totally ordered subsets of \( X \)). Then \( \emptyset \notin \mathcal{M} \) and \( \mathcal{M} \neq \emptyset \) (since \( X \neq \emptyset \) and \( \{x\} \in \mathcal{M} \) for each \( x \in X \)). Moreover, \( \mathcal{M} \) satisfies the hypothesis of Lem. A.51, since, if \( \mathcal{C} \subseteq \mathcal{M} \) is a chain of totally ordered subsets of \( X \), then \( \bigcup \mathcal{C} \) is a totally ordered subset of \( X \), i.e. in \( \mathcal{M} \) (here we have used the notation of (A.17); also note that we are dealing with two different types of chains here, namely
those with respect to the order on $X$ and those with respect to the order given by $\subseteq$ on $\mathcal{M}$). Let $f : \mathcal{P}(X) \setminus \{\emptyset\} \to X$ be a choice function given by AC, i.e. such that

$$\forall_{Y \in \mathcal{P}(X) \setminus \{\emptyset\}} f(Y) \in Y.$$ 

As an auxiliary notation, we set

$$\forall_{M \in \mathcal{M}} M^* := \{x \in X \setminus M : M \cup \{x\} \in \mathcal{M}\}.$$ 

With the intention of applying Lem. A.51, we define

$$g : \mathcal{M} \to \mathcal{M}, \quad g(M) := \begin{cases} M \cup \{f(M^*)\} & \text{if } M^* \neq \emptyset, \\ M & \text{if } M^* = \emptyset. \end{cases}$$

Since $g$ clearly satisfies (A.19), Lem. A.51 applies, providing an $M \in \mathcal{M}$ such that $g(M) = M$. Thus, $M^* = \emptyset$, i.e. $M$ is a maximal chain, proving (ii).

“(ii) $\Rightarrow$ (iii)”: Assume (ii). To prove Zorn’s lemma, let $X$ be a nonempty set, partially ordered by $\leq$, such that every chain $C \subseteq X$ has an upper bound. Due to Hausdorff’s maximality principle, we can assume $C \subseteq X$ to be a maximal chain. Let $m \in X$ be an upper bound for the maximal chain $C$. We claim that $m$ is a maximal element: Indeed, if there were $x \in X$ such that $m < x$, then $x \notin C$ (since $m$ is upper bound for $C$) and $C \cup \{x\}$ would constitute a strict superset of $C$ that is also a chain, contradicting the maximality of $C$.

“(iii) $\Rightarrow$ (iv)”: Assume (iii) and let $X$ be a nonempty set. We need to construct a well-order on $X$. Let $\mathcal{W}$ be the set of all well-orders on subsets of $X$, i.e.

$$\mathcal{W} := \{(Y, W) : Y \subseteq X \land W \subseteq Y \times Y \subseteq X \times X \text{ is a well-order on } Y\}.$$ 

We define a partial order $\leq$ on $\mathcal{W}$ by setting

$$\forall_{(Y, W), (Y', W') \in \mathcal{W}} \left( (Y, W) \leq (Y', W') \iff Y \subseteq Y' \land W = W'|_Y \land (y \in Y, y' \in Y', y'W'y \Rightarrow y' \in Y) \right)$$

(recall the definition of the restriction of a relation from Def. A.21(e)). To apply Zorn’s lemma to $(\mathcal{W}, \leq)$, we need to check that every chain $\mathcal{C} \subseteq \mathcal{W}$ has an upper bound. To this end, if $\mathcal{C} \subseteq \mathcal{W}$ is a chain, let

$$U_\mathcal{C} := (Y_\mathcal{C}, W_\mathcal{C}), \quad \text{where } Y_\mathcal{C} := \bigcup_{(Y, W) \in \mathcal{C}} Y, \quad W_\mathcal{C} := \bigcup_{(Y, W) \in \mathcal{C}} W.$$ 

We need to verify $U_\mathcal{C} \in \mathcal{W}$: If $aW_\mathcal{C}b$, then there is $(Y, C) \in \mathcal{C}$ such that $aWb$. In particular, $(a, b) \in Y \times Y \subseteq Y_\mathcal{C} \times Y_\mathcal{C}$, showing $W_\mathcal{C}$ to be a relation on $Y_\mathcal{C}$. Clearly, $W_\mathcal{C}$ is a total order on $Y_\mathcal{C}$ (one just uses that, if $a, b \in Y_\mathcal{C}$, then, as $\mathcal{C}$ is a chain, there is $(Y, W) \in \mathcal{C}$ such that $a, b \in Y$ and $W = W|_Y$ is a total order on $Y$). To see that $W_\mathcal{C}$ is a well-order on $Y_\mathcal{C}$, let $\emptyset \neq A \subseteq Y_\mathcal{C}$. If $a \in A$, then there is $(Y, W) \in \mathcal{C}$ such that
exists a bijective map \( m \) that 

\[
(iv) \quad A \text{ transitive. According to Def. 3.12(a), } M \text{ constitutes an equivalence relation on } \mathbb{U}_m. \text{ If } a \in A, \text{ then } M \text{ is reflexive. If } b \in Y \cap A \text{ and } mWb. \text{ If } Y \subseteq B, \text{ then } m,b \in B. \text{ If } mUb, \text{ then we are done. If } bUm, \text{ then } b \in Y (\text{since } (Y,W) \leq (B,U)), \text{ i.e., again, } b \in Y \cap A \text{ and } mWb (\text{actually } m = b \text{ in this case}), \text{ proving } m = \min A. \text{ This completes the proof that } W_C \text{ is a well-order on } Y_C \text{ and, thus, shows } U_C \in W. \text{ Next, we check } U_C \text{ to be an upper bound for } C: \text{ If } (Y,W) \in C, \text{ then } Y \subseteq Y_C \text{ and } W = W_C|_Y \text{ are immediate. If } y \in Y, \text{ then } y' \in Y_C, \text{ and } y'Wy, \text{ then } y' \in Y (\text{otherwise, } y' \in A \text{ with } (A,U) \in C, \text{ (Y,W) } \leq (A,U), \text{ y'Wy, in contradiction to } y' \notin Y). \text{ Thus, } (Y,W) \leq U_C, \text{ showing } U_C \text{ to be an upper bound for } C. \text{ By Zorn’s lemma, we conclude that } W \text{ contains a maximal element } (M,W_M). \text{ But then } M = X \text{ and } W_M \text{ is the desired well-order on } X: \text{ Indeed, if there is } x \in X \setminus M, \text{ then we can let } Y := M \cup \{x\} \text{ and,}

\[
\forall a,b \in Y \quad (aWb :\Leftrightarrow (a,b \in M \land aW_Mb) \lor b = x).
\]

Then \((Y,W) \in W \text{ with } (M,W_M) < (Y,W) \text{ in contradiction to the maximality of } (M,W_M).

\[\text{“(iv) } \Rightarrow \text{ AC”}: \text{ Assume (iv). Given a nonempty set of nonempty sets } M, \text{ let } X := \bigcup_{M \in M} M. \text{ By (iv), there exists a well-order } R \text{ on } X. \text{ Then every nonempty } Y \subseteq X \text{ has a unique min. As every } M \in M \text{ is a nonempty subset of } X, \text{ we can define a choice function}
\]

\[f : M \rightarrow X, \quad f(M) := \min M \in M,
\]

proving AC.

\[\text{“(v) } \Leftrightarrow \text{ AC”}: \text{ That every vector space has a basis is proved in } [\Phi19, \text{ Th. 5.23}] \text{ by use of Zorn’s lemma. That, conversely, (v) implies AC was first shown in } [\text{Bla84}], \text{ but the proof needs more algebraic tools than we have available in this class.} \]

\section{A.5 Cardinality}

\subsection{A.5.1 Relations to Injective, Surjective, and Bijective Maps; Schröder-Bernstein Theorem}

\textbf{Theorem A.53.} \textit{Let } \mathcal{M} \textit{ be a set of sets. Then the relation } \sim \textit{ on } \mathcal{M}, \textit{ defined by}

\[A \sim B :\Leftrightarrow A \text{ and } B \text{ have the same cardinality,}\]

\textit{constitutes an equivalence relation on } \mathcal{M}.

\textit{Proof.} According to Def. 2.23, we have to prove that \( \sim \) is reflexive, symmetric, and transitive. According to Def. 3.12(a), \( A \sim B \) holds for \( A, B \in \mathcal{M} \) if, and only if, there exists a bijective map \( f : A \rightarrow B \). Thus, since the identity \( \text{Id} : A \rightarrow A \) is bijective, \( A \sim A \), showing \( \sim \) is reflexive. If \( A \sim B \), then there exists a bijective map \( f : A \rightarrow B \), and \( f^{-1} \) is a bijective map \( f^{-1} : B \rightarrow A \), showing \( B \sim A \) and that \( \sim \) is symmetric. If \( A \sim B \) and \( B \sim C \), then there are bijective maps \( f : A \rightarrow B \) and \( g : B \rightarrow C \).
Then, according to Th. 2.14, the composition \((g \circ f) : A \rightarrow C\) is also bijective, proving \(A \sim C\) and that \(\sim\) is transitive. \(\blacksquare\)

The next theorem provides two interesting, and sometimes useful, characterizations of infinite sets:

**Theorem A.54.** Let \(A\) be a set. Using the axiom of choice (AC) of Sec. A.4, the following statements

(i) \[\text{Theorem A.54.}\]

Theorem A.54. Let \(A\) be a set. Using the axiom of choice (AC) of Sec. A.4, the following statements (i) – (iii) are equivalent. More precisely, (ii) and (iii) are equivalent even without AC (a set \(A\) is sometimes called Dedekind-infinite if, and only if, it satisfies (iii)), (iii) implies (i) without AC, but AC is needed to show (i) implies (ii), (iii).

(i) \(A\) is infinite.

(ii) There exists \(M \subseteq A\) and a bijective map \(f : M \rightarrow \mathbb{N}\).

(iii) There exists a strict subset \(B \subset A\) and a bijective map \(g : A \rightarrow B\).

One sometimes expresses the equivalence between (i) and (ii) by saying that a set is infinite if, and only if, it contains a copy of the natural numbers. The property stated in (iii) might seem strange at first, but infinite sets are, indeed, precisely those identical in size to some of their strict subsets (as an example think of the natural bijection \(n \mapsto 2n\) between all natural numbers and the even numbers).

Proof. We first prove, without AC, the equivalence between (ii) and (iii).

“(ii) ⇒ (iii)”: Let \(E\) denote the even numbers. Then \(E \subseteq \mathbb{N}\) and \(h : \mathbb{N} \rightarrow E\), \(h(n) := 2n\), is a bijection, showing that (iii) holds for the natural numbers. According to (ii), there exists \(M \subseteq A\) and a bijective map \(f : M \rightarrow \mathbb{N}\). Define \(B := (A \setminus M) \cup f^{-1}(E)\) and

\[
h : A \rightarrow B, \quad h(x) := \begin{cases} x & \text{for } x \in A \setminus M, \\ f^{-1} \circ h \circ f(x) & \text{for } x \in M. \end{cases} \tag{A.24}\]

Then \(B \subset A\) since \(B\) does not contain the elements of \(M\) that are mapped to odd numbers under \(f\). Still, \(h\) is bijective, since \(h|_{A \setminus M} = \text{Id}_{A \setminus M}\) and \(h|_{M} = f^{-1} \circ h \circ f\) is the composition of the bijective maps \(f, h,\) and \(f^{-1}|_{E} : E \rightarrow f^{-1}(E)\).

“(iii) ⇒ (ii)”: As (iii) is assumed, there exist \(B \subseteq A, a \in A \setminus B\), and a bijective map \(g : A \rightarrow B\). Set

\[M := \{a_n := g^n(a) : n \in \mathbb{N}\}.\]

We show that \(a_n \neq a_m\) for each \(m, n \in \mathbb{N}\) with \(m \neq n\): Indeed, suppose \(m, n \in \mathbb{N}\) with \(n > m\) and \(a_n = a_m\). Then, since \(g\) is bijective, we can apply \(g^{-1}\) \(m\) times to \(a_n = a_m\) to obtain

\[a = (g^{-1})^m(a_m) = (g^{-1})^m(a_n) = g^{n-m}(a).\]

Since \(l := n - m \geq 1\), we have \(a = g(g^{l-1}(a))\), in contradiction to \(a \in A \setminus B\). Thus, all the \(a_n \in M\) are distinct and we can define \(f : M \rightarrow \mathbb{N}, f(a_n) := n\), which is clearly bijective, proving (ii).
“(iii) \Rightarrow (i)”: The proof is conducted by contraposition, i.e. we assume \( A \) to be finite and proof that (iii) does not hold. If \( A = \emptyset \), then there is nothing to prove. If \( \emptyset \neq A \) is finite, then, by Def. 3.12(b), there exists \( n \in \mathbb{N} \) and a bijective map \( f : A \rightarrow \{1, \ldots, n\} \). If \( B \subsetneq A \), then, according to Th. A.63(a), there exists \( m \in \mathbb{N}_0, m < n \), and a bijective map \( h : B \rightarrow \{1, \ldots, m\} \). If there were a bijective map \( g : A \rightarrow B \), then \( h \circ g \circ f^{-1} \) were a bijective map from \( \{1, \ldots, n\} \) onto \( \{1, \ldots, m\} \) with \( m < n \) in contradiction to Th. A.61.

“(i) \Rightarrow (ii)” : Inductively, we construct a strictly increasing sequence \( M_1 \subseteq M_2 \subseteq \ldots \) of subsets \( M_n \) of \( A \) \( n \in \mathbb{N} \), and a sequence of functions \( f_n : M_n \rightarrow \{1, \ldots, n\} \) satisfying

\[
\forall n \in \mathbb{N} \quad f_n \text{ is bijective,} \tag{A.25a}
\]

\[
\forall m, n \in \mathbb{N} \quad (m \leq n \Rightarrow f_n|M_m = f_m) : \tag{A.25b}
\]

Since \( A \neq \emptyset \), there exists \( m_1 \in A \). Set \( M_1 := \{m_1\} \) and \( f_1 : M_1 \rightarrow \{1\}, f_1(m_1) := 1 \). Then \( M_1 \subseteq A \) and \( f_1 \) bijective are trivially clear. Now let \( n \in \mathbb{N} \) and suppose \( M_1, \ldots, M_n \) and \( f_1, \ldots, f_n \) satisfying (A.25) have already been constructed. Since \( A \) is infinite, there must be \( m_{n+1} \in A \setminus M_n \) (otherwise \( M_n = A \) and the bijectivity of \( f_n : M_n \rightarrow \{1, \ldots, n\} \) shows \( A \) is finite with \( \#A = n \); AC is used to select the \( m_{n+1} \in A \setminus M_n \). Set \( M_{n+1} := M_n \cup \{m_{n+1}\} \) and

\[
f_{n+1} : M_{n+1} \rightarrow \{1, \ldots, n+1\}, \quad f_{n+1}(x) := \begin{cases} f_n(x) & \text{for } x \in M_n, \\ n+1 & \text{for } x = m_{n+1}. \end{cases} \tag{A.26}
\]

Then the bijectivity of \( f_n \) implies the bijectivity of \( f_{n+1} \), and, since \( f_{n+1}|M_n = f_n \) holds by definition of \( f_{n+1} \), the implication

\[
m \leq n+1 \Rightarrow f_{n+1}|M_n = f_n
\]

holds true as well. An induction also shows \( M_n = \{m_1, \ldots, m_n\} \) and \( f_n(m_n) = n \) for each \( n \in \mathbb{N} \). We now define

\[
M := \bigcup_{n \in \mathbb{N}} M_n = \{m_n : n \in \mathbb{N}\}, \quad f : M \rightarrow \mathbb{N}, \quad f(m_n) := f_n(m_n) = n. \tag{A.27}
\]

Clearly, \( M \subseteq A \), and \( f \) is bijective with \( f^{-1} : \mathbb{N} \rightarrow M, f^{-1}(n) = m_n \).

\[ \textbf{Theorem A.55} \] (Schröder-Bernstein). \textit{Let} \( A, B \) \textit{be sets. The following statements are equivalent (even without assuming the axiom of choice):}

\[(i) \text{ The sets } A \text{ and } B \text{ have the same cardinality (i.e. there exists a bijective map } \phi : A \rightarrow B\).

\[(ii) \text{ There exist an injective map } f : A \rightarrow B \text{ and an injective map } g : B \rightarrow A.\]

We will give two proofs of the Schröder-Bernstein theorem. The first proof is rather elegant, but also quite abstract. The second proof is longer, but less abstract. Even though it is still nonconstructive in the general situation, in many concrete cases, it does provide a method for actually constructing a bijective map from two injective maps. The first proof is based on the following lemma:
Lemma A.56. Let \( A \) be a set. Consider \( \mathcal{P}(A) \) to be endowed with the partial order given by set inclusion, i.e., for each \( X,Y \in \mathcal{P}(A) \), \( X \leq Y \) if, and only if, \( X \subseteq Y \). If \( F : \mathcal{P}(A) \longrightarrow \mathcal{P}(A) \) is isotone with respect to that order, then \( F \) has a fixed point, i.e. \( F(X_0) = X_0 \) for some \( X_0 \in \mathcal{P}(A) \).

Proof. Define

\[
\mathcal{A} := \{ X \in \mathcal{P}(A) : F(X) \subseteq X \}, \quad X_0 := \bigcap_{X \in \mathcal{A}} X \quad (A.28)
\]

(\( X_0 \) is well-defined, since \( F(A) \subseteq A \)). Suppose \( X \in \mathcal{A} \), i.e. \( F(X) \subseteq X \) and \( X_0 \subseteq X \). Then \( F(X_0) \subseteq F(X) \subseteq X \) due to the isotonicity of \( F \). Thus, \( F(X_0) \subseteq X \) for every \( X \in \mathcal{A} \), i.e. \( F(X_0) \subseteq X_0 \). Using the isotonicity of \( F \) again shows \( F(F(X_0)) \subseteq F(X_0) \), implying \( F(X_0) \in \mathcal{A} \) and \( X_0 \subseteq F(X_0) \), i.e. \( F(X_0) = X_0 \) as desired. \( \blacksquare \)

First Proof of Th. A.55. (i) trivially implies (ii), as one can simply set \( f := \phi \) and \( g := \phi^{-1} \). It remains to show (ii) implies (i). Thus, let \( f : A \longrightarrow B \) and \( g : B \longrightarrow A \) be injective. To apply Lem. A.56, define

\[
F : \mathcal{P}(A) \longrightarrow \mathcal{P}(A), \quad F(X) := A \setminus g(B \setminus f(X)),
\]

and note

\[
X \subseteq Y \subseteq A \implies f(X) \subseteq f(Y) \implies B \setminus f(Y) \subseteq B \setminus f(X) \implies g(B \setminus f(Y)) \subseteq g(B \setminus f(X)) \implies F(X) \subseteq F(Y).
\]

Thus, by Lem. A.56, \( F \) has a fixed point \( X_0 \). We claim that a bijection is obtained via setting

\[
\phi : A \longrightarrow B, \quad \phi(x) := \begin{cases} f(x) & \text{for } x \in X_0, \\ g^{-1}(x) & \text{for } x \notin X_0. \end{cases}
\]

First, \( \phi \) is well-defined, since \( x \notin X_0 = F(X_0) \) implies \( x \in g(B \setminus f(X_0)) \). To verify that \( \phi \) is injective, let \( x,y \in A \), \( x \neq y \). If \( x,y \in X_0 \), then \( \phi(x) \neq \phi(y) \), as \( f \) is injective. If \( x,y \in A \setminus X_0 \), then \( \phi(x) \neq \phi(y) \), as \( g^{-1} \) is well-defined. If \( x \in X_0 \) and \( y \notin X_0 \), then \( \phi(x) \in f(X_0) \) and \( \phi(y) \) is arbitrary, implying \( \phi(x) \neq \phi(y) \). If \( y \in X_0 \) and \( x \notin X_0 \), then \( \phi(x) \neq \phi(y) \). Thus, \( \phi \) is injective. To verify surjectivity, if \( b \in f(X_0) \), then \( \phi(f^{-1}(b)) = b \). If \( b \in B \setminus f(X_0) \), then \( g(b) \notin X_0 = F(X_0) \), i.e. \( \phi(g(b)) = b \), showing \( \phi \) to be surjective. \( \blacksquare \)

Second Proof of Th. A.55. As in the first proof, we only need to show (ii) implies (i). We first assume that \( A \) and \( B \) are disjoint. To define \( \phi \), we first construct a suitable partition of \( A \cup B \), where the subsets of the partition are given via sequences defined by using \( f \) and \( g \). The idea is to assign a unique sequence \( \sigma(a) \) to each \( a \in A \) and a unique sequence \( \sigma(b) \) to each \( b \in B \) by alternately applying \( f \) and \( g \) to advance the sequence to the right and by alternately applying \( f^{-1} \) and \( g^{-1} \) to advance the sequence to the left, if possible (for a given \( a \in A \), \( g^{-1}(a) \) might not be defined and, for a given \( b \in B \), \( f^{-1}(a) \) might not be defined). Thus, for \( a \in A \), \( \sigma(a) \) has the form

\[
\ldots, f^{-1}(g^{-1}(a)), g^{-1}(a), a, f(a), g(f(a)), \ldots \quad (A.29)
\]
More precisely, for each \( a \in A \), we define \( \sigma(a) = (\sigma_i(a))_{i \in I_a} \) recursively by

\[
\begin{align*}
\sigma_i(a) &:= a & \text{for } i = 0, \\
\sigma_i(a) &:= f(\sigma_{i-1}(a)) & \text{for } i > 0 \text{ odd}, \\
\sigma_i(a) &:= g(\sigma_{i-1}(a)) & \text{for } i > 0 \text{ even}, \\
\sigma_i(a) &:= g^{-1}(\sigma_{i+1}(a)) & \text{for } i < 0 \text{ odd and } \sigma_{i+1}(a) \in g(B), \\
m_a &:= i + 1, I_a := \{ k \in \mathbb{Z} : m_a \leq k \} & \text{for } i < 0 \text{ odd and } \sigma_{i+1}(a) \notin g(B), \\
\sigma_i(a) &:= f^{-1}(\sigma_{i+1}(a)) & \text{for } i < 0 \text{ even and } \sigma_{i+1}(a) \in f(A), \\
m_a &:= i + 1, I_a := \{ k \in \mathbb{Z} : m_a \leq k \} & \text{for } i < 0 \text{ even and } \sigma_{i+1}(a) \notin f(A),
\end{align*}
\]

where the conditions in (A.30c) and (A.30g) are meant to implicitly require \( \sigma_{i+1}(a) \) to be defined for \( i+1 \). By induction, one shows \( \sigma_{i-1}(a) \in A \) for each \( i > 0 \) odd, \( \sigma_{i-1}(a) \in B \) for each \( i > 0 \) even, \( \sigma_{i+1}(a) \in A \) for each \( m_a \leq i < 0 \) odd, and \( \sigma_{i+1}(a) \in B \) for each \( m_a \leq i < 0 \) even, such that \( \sigma_i(a) \) is well-defined by (A.30) for each \( i \in I_a \) (with \( I_a = \mathbb{Z} \) if (A.30e) and (A.30g) are never satisfied). Analogously, for each \( b \in B \), we define \( \sigma(b) = (\sigma_i(b))_{i \in I_b} \) recursively by

\[
\begin{align*}
\sigma_i(b) &:= b & \text{for } i = 0, \\
\sigma_i(b) &:= g(\sigma_{i-1}(b)) & \text{for } i > 0 \text{ odd}, \\
\sigma_i(b) &:= f(\sigma_{i-1}(b)) & \text{for } i > 0 \text{ even}, \\
\sigma_i(b) &:= f^{-1}(\sigma_{i+1}(b)) & \text{for } i < 0 \text{ odd and } \sigma_{i+1}(b) \in f(A), \\
m_b &:= i + 1, I_b := \{ k \in \mathbb{Z} : m_b \leq k \} & \text{for } i < 0 \text{ odd and } \sigma_{i+1}(b) \notin f(A), \\
\sigma_i(b) &:= g^{-1}(\sigma_{i+1}(b)) & \text{for } i < 0 \text{ even and } \sigma_{i+1}(b) \in g(B), \\
m_b &:= i + 1, I_b := \{ k \in \mathbb{Z} : m_b \leq k \} & \text{for } i < 0 \text{ even and } \sigma_{i+1}(b) \notin g(B),
\end{align*}
\]

where the conditions in (A.31e) and (A.31g) are meant to implicitly require \( \sigma_{i+1}(b) \) to be defined for \( i+1 \). By induction, one shows \( \sigma_{i-1}(b) \in B \) for each \( i > 0 \) odd, \( \sigma_{i-1}(b) \in A \) for each \( i > 0 \) even, \( \sigma_{i+1}(b) \in B \) for each \( m_b \leq i < 0 \) odd, and \( \sigma_{i+1}(b) \in A \) for each \( m_b \leq i < 0 \) even, such that \( \sigma_i(b) \) is well-defined by (A.31) for each \( i \in I_b \) (with \( I_b = \mathbb{Z} \) if (A.31e) and (A.31g) are never satisfied). The \( \sigma(a) \) and \( \sigma(b) \) now allow us to define the sets

\[
\forall_{x \in A \cup B} S_x := \{ \sigma_i(x) : i \in I_x \} \subseteq A \cup B. \tag{A.32}
\]

Moreover, we call \( x \in A \cup B \) an \( A \)-stopper if, and only if, \( \sigma(x) \) terminates to the left with some element in \( A \); a \( B \)-stopper, if, and only if, \( \sigma(x) \) terminates to the left with some element in \( B \); and a non-stopper, if \( \sigma(x) \) does never terminate to the left – thus,

\[
\begin{align*}
x \text{ A-stopper} & \iff \left( I_x \neq \mathbb{Z} \land \left( (x \in A \land m_x \text{ even}) \lor (x \in B \land m_x \text{ odd}) \right) \right), \\
x \text{ B-stopper} & \iff \left( I_x \neq \mathbb{Z} \land \left( (x \in A \land m_x \text{ odd}) \lor (x \in B \land m_x \text{ even}) \right) \right), \\
x \text{ non-stopper} & \iff I_x = \mathbb{Z}. \tag{A.33}
\end{align*}
\]
Next, we prove that the \( S_x \) form a partition of \( A \cup B \). Since, for each \( x \in A \cup B \), \( x = \sigma_0(x) \in S_x \), it only remains to show
\[
\forall \, x, y \in A \cup B \quad \left( S_x = S_y \lor S_x \cap S_y = \emptyset \right). \tag{A.34}
\]
To prove (A.34), it clearly suffices to show
\[
\forall \, x, z \in A \cup B \quad \left( z \in S_x \Rightarrow S_x = S_z \right). \tag{A.35}
\]
To verify (A.35), let \( z \in S_x \). Then there exists \( i \in I_x \) such that \( z = \sigma_0(z) = \sigma_i(x) \) and simple inductions show \( \sigma_k(z) = \sigma_{k+i}(x) \) for each \( k \in I_z \) and \( \sigma_{k-i}(z) = \sigma_k(x) \) for each \( k \in I_x \) (in particular, \( i + I_z = I_x \)), proving \( S_x = S_z \).

We are now in a position to define the desired bijection \( \phi : A \to B \):
\[
\phi : A \to B, \quad \phi(a) := \begin{cases} f(a) & \text{if } a \text{ is an } A\text{-stopper or a non-stopper}, \\ g^{-1}(a) & \text{if } a \text{ is a } B\text{-stopper}. \end{cases} \tag{A.36}
\]
Indeed, \( \phi \) is injective: If \( a_1, a_2 \in \{ a \in A : a \text{ an } A\text{-stopper or non-stopper} \} \) with \( a_1 \neq a_2 \), then \( \phi(a_1) \neq \phi(a_2) \) due to \( f \) being injective; if \( a_1, a_2 \in \{ a \in A : a \text{ a } B\text{-stopper} \} \) with \( a_1 \neq a_2 \), then \( \phi(a_1) \neq \phi(a_2) \) due to \( g^{-1} \) being injective; and \( a_1, a_2 \in A \) with \( a_2 \) a \( B\)-stopper and \( a_1 \) not a \( B\)-stopper, \( S_{a_1} = S_{f(a_1)} \) and \( S_{a_2} = S_{g^{-1}(a_2)} \), i.e. \( \phi(a_2) \) is also a \( B\)-stopper, whereas \( \phi(a_1) \) is not a \( B\)-stopper, in particular, \( \phi(a_1) \neq \phi(a_2) \). Moreover, \( \phi \) is also surjective: If \( b \in B \) is a \( B\)-stopper, then, due to \( S_b = S_{g(b)} \), so is \( g(b) \), and \( b = g^{-1}(g(b)) = \phi(g(b)) \); if \( b \in B \) is not a \( B\)-stopper, then \( f^{-1}(b) \) is defined and in \( S_b \), i.e. \( f^{-1}(b) \) is not a \( B\)-stopper, either, and \( b = f(f^{-1}(b)) = \phi(f^{-1}(b)) \).

To conclude the proof, we consider the case that \( A \) and \( B \) are not necessarily disjoint. Since \( A \times \{0\} \) and \( B \times \{1\} \) are always disjoint with
\[
\tilde{f} : A \times \{0\} \to B \times \{1\}, \quad \tilde{f}(a, 0) := (f(a), 1), \tag{A.37a}
\]
\[
\tilde{g} : B \times \{1\} \to A \times \{0\}, \quad \tilde{g}(b, 0) := (g(b), 0), \tag{A.37b}
\]
still being injective if \( f, g \) are, the first part of the proof yields a bijective function \( \tilde{\phi} : A \times \{0\} \to B \times \{1\} \). Then, using the clearly bijective functions
\[
\alpha : A \to A \times \{0\}, \quad \alpha(a) := (a, 0), \tag{A.38a}
\]
\[
\beta : B \to B \times \{1\}, \quad \beta(b) := (b, 1), \tag{A.38b}
\]
\( \phi := \beta^{-1} \circ \tilde{\phi} \circ \alpha : A \to B \) is also bijective.

**Remark A.57.** In general, the second proof of the Schröder-Bernstein Th. A.55 is still nonconstructive, since one has, in general, no algorithm to determine if a given element is an \( A\)-stopper, a \( B\)-stopper, or a non-stopper. However, as the following Ex. A.58 shows, in particular situations, determining \( A\)-stoppers, \( B\)-stoppers, and non-stoppers does not have to be difficult.
Example A.58. Let $A := \mathbb{N}_0, B := \{n \in \mathbb{N}_0 : n \text{ even}\}$. We consider $A$ and $B$ as being made disjoint (for example, by using the trick employed in the last part of the proof of Th. A.55 above), but, for the sake of readability, we will not reflect this in the used notation. Define the maps

$$f : A \rightarrow B, \quad f(n) := 4n,$$

$$g : B \rightarrow A, \quad g(n) := n,$$  \hspace{1cm} \text{(A.39a)}

both being clearly injective, but not surjective. The goal is to, explicitly, find the bijective map $\phi : A \rightarrow B$, given by (A.36). As an intermediate step, we determine which elements of $A$ are non-stoppers, $A$-stoppers, and $B$-stoppers, and likewise for the elements of $B$. Clearly $0 \in A$ and $0 \in B$ are non-stoppers. We will see that all other elements are either $A$-stoppers or $B$-stoppers. The precise claim is

$$A_1 := \{a \in A : a \text{ is } A\text{-stopper}\} = C_1 := \{a \in A : a = n \cdot 4^k, \text{ n odd, } k \in \mathbb{N}_0\},$$  \hspace{1cm} \text{(A.40a)}

$$A_2 := \{a \in A : a \text{ is } B\text{-stopper}\} = C_2 := A \setminus (C_1 \cup \{0\}),$$  \hspace{1cm} \text{(A.40b)}

$$B_1 := \{b \in B : b \text{ is } A\text{-stopper}\} = D_1 := B \setminus (D_2 \cup \{0\}),$$  \hspace{1cm} \text{(A.40c)}

$$B_2 := \{b \in B : b \text{ is } B\text{-stopper}\} = D_2 := \{b \in B : b = n \cdot 2^k; \text{ n, k odd; } n, k \geq 1\}.$$  \hspace{1cm} \text{(A.40d)}

Indeed, if $c = n \cdot 4^k \in C_1$, then $(f^{-1} \circ g^{-1})^k(c) = n$ is odd, i.e. $n \notin g(B)$, showing $c$ is an $A$-stopper, proving $C_1 \subseteq A_1$. If $d = n \cdot 2^k \in D_2$, then $k - 1 = 2m$ with $m \in \mathbb{N}_0$, i.e. $d = n \cdot 2 \cdot 4^m$ and $(g^{-1} \circ f^{-1})^m(d) = 2n$ is not divisible by $4$, i.e. $2n \notin f(A)$, showing $d$ is a $B$-stopper, proving $D_2 \subseteq B_2$. Clearly, each $a \in \mathbb{N}$ either has the form $a = n \cdot 4^k$ with odd and $k \in \mathbb{N}_0$ (i.e. $a \in C_1$) or $a = 2 \cdot n \cdot 4^k$ with odd and $k \in \mathbb{N}_0$, i.e.

$$C_2 = \{a \in A : a = 2 \cdot n \cdot (2 \cdot 2)^k; \text{ n odd; } k \in \mathbb{N}_0\}$$

$$= \{a \in A : a = n \cdot 2^k; \text{ n, k odd; } n, k \geq 1\} = g(D_2).$$  \hspace{1cm} \text{(A.41)}

Since $D_2 \subseteq B_2$, all elements of $D_2$ are $B$-stoppers, and, thus, so are all elements of $C_2$, proving $C_2 \subseteq A_2$. Since $A = C_1 \cup C_2 \cup \{0\}$, we then also obtain $A_1 = C_1$ and $A_2 = C_2$. Clearly, each even $b \in \mathbb{N}$ either has the form $b = n \cdot 2^k$ with odd $n, k \geq 1$ (i.e. $b \in D_2$) or $b = n \cdot 4^k$ with odd and $k \in \mathbb{N}$, i.e.

$$D_1 = \{b \in B : b = n \cdot 4^k, \text{ n odd, } k \in \mathbb{N}\} = f(C_1).$$  \hspace{1cm} \text{(A.42)}

Since $C_1 = A_1$, all elements of $C_1$ are $A$-stoppers, and, thus, so are all elements of $D_1$. Since $B = D_1 \cup D_2 \cup \{0\}$, we then also obtain $B_1 = D_1$ and $B_2 = D_2$.

Now that we have identified explicit formulas for $A_1$ and $A_2$, we can write the assignment rule for the bijective $\phi : A \rightarrow B$, given by (A.36), in the explicit form

$$\phi(a) := \begin{cases} 0 & \text{if } a = 0, \\ 4a & \text{if } a = n \cdot 4^k \text{ with } \text{n odd and } k \in \mathbb{N}_0, \\ a & \text{if } a = 2 \cdot n \cdot 4^k \text{ with } \text{n odd and } k \in \mathbb{N}_0. \end{cases}$$  \hspace{1cm} \text{(A.43)}
Thus, $\phi$ starts out with the assignments

\[ \begin{array}{cccccccc}
0 & 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 \\
\phi : & \downarrow & \downarrow & \downarrow & \downarrow & \downarrow & \downarrow & \downarrow & \downarrow & \ldots \\
0 & 4 & 2 & 12 & 16 & 20 & 6 & 28 & 8 \\
\end{array} \]  \quad (A.44)

**Theorem A.59.** Let $A, B$ be nonempty sets. Then the following statements are equivalent (where the implication “(ii) $\Rightarrow$ (i)” makes use of the axiom of choice (AC) of Sec. A.4).

(i) There exists an injective map $f : A \rightarrow B$.

(ii) There exists a surjective map $g : B \rightarrow A$.

**Proof.** According to Th. 2.13(b), (i) is equivalent to $f$ having a left inverse $g : B \rightarrow A$ (i.e. $g \circ f = \text{Id}_A$), which is equivalent to $g$ having a right inverse, which, according to Th. 2.13(a), is equivalent to (ii) (AC is used in the proof of Th. 2.13(a) to show each surjective map has a right inverse). \[ \blacksquare \]

**Corollary A.60.** Let $A, B$ be nonempty sets. Using AC, we can expand the two equivalent statements of Th. A.55 to the following list of equivalent statements:

(i) The sets $A$ and $B$ have the same cardinality (i.e. there exists a bijective map $\phi : A \rightarrow B$).

(ii) There exist an injective map $f : A \rightarrow B$ and an injective map $g : B \rightarrow A$.

(iii) There exist a surjective map $f : A \rightarrow B$ and a surjective map $g : B \rightarrow A$.

(iv) There exist an injective map $f_1 : A \rightarrow B$ and a surjective map $f_2 : A \rightarrow B$.

(v) There exist an injective map $g_1 : B \rightarrow A$ and a surjective map $g_2 : B \rightarrow A$.

**Proof.** The equivalences are an immediate consequence of combining Th. A.55 with Th. A.59. \[ \blacksquare \]

### A.5.2 Finite Sets

It is intuitively clear that finite cardinalities are uniquely determined. Still one has to provide a rigorous proof. The key is the following theorem:

**Theorem A.61.** If $m, n \in \mathbb{N}$ and the map $f : \{1, \ldots, m\} \rightarrow \{1, \ldots, n\}$ is bijective, then $m = n$. 
Proof. We conduct the proof via induction on $m$. If $m = 1$, then the surjectivity of $f$ implies $n = 1$. For the induction step, we now consider $m > 1$. From the bijective map $f$, we define the map

$$g : \{1, \ldots, m\} \longrightarrow \{1, \ldots, n\}, \quad g(x) := \begin{cases} n & \text{for } x = m, \\ f(m) & \text{for } x = f^{-1}(n), \\ f(x) & \text{otherwise.} \end{cases} \quad (A.45)$$

Then $g$ is bijective, since it is the composition $g = h \circ f$ of the bijective map $f$ with the bijective map $h : \{f(m), n\} \longrightarrow \{f(m), n\}$,

$$h : \{f(m), n\} \longrightarrow \{f(m), n\}, \quad h(f(m)) := n, \quad h(n) := f(m). \quad (A.46)$$

Thus, the restriction $g \upharpoonright \{1, \ldots, m-1\} : \{1, \ldots, n\} \longrightarrow \{1, \ldots, n\}$ must also be bijective, such that the induction hypothesis yields $m - 1 = n - 1$, which, in turn, implies $m = n$ as desired. ■

Corollary A.62. Let $m, n \in \mathbb{N}$ and let $A$ be a set. If $\#A = m$ and $\#A = n$, then $m = n$.

Proof. If $\#A = m$, then, according to Def. 3.12(b), there exists a bijective map $f : A \longrightarrow \{1, \ldots, m\}$. Analogously, if $\#A = n$, then there exists a bijective map $g : A \longrightarrow \{1, \ldots, n\}$. In consequence, we have the bijective map $(g \circ f^{-1}) : \{1, \ldots, m\} \longrightarrow \{1, \ldots, n\}$, such that Th. A.61 yields $m = n$. ■

Theorem A.63. Let $A \neq \emptyset$ be a finite set.

(a) If $B \subseteq A$ with $A \neq B$, then $B$ is finite with $\#B < \#A$.

(b) If $a \in A$, then $\#(A \setminus \{a\}) = \#A - 1$.

Proof. For $\#A = n \in \mathbb{N}$, we use induction to prove (a) and (b) simultaneously, i.e. we show

$$\forall \ n \in \mathbb{N} \left( \#A = n \Rightarrow \forall \ B \in \mathcal{P}(A) \setminus \{A\} \forall \ a \in A \ #B \in \{0, \ldots, n-1\} \land \#(A \setminus \{a\}) = n - 1 \right) \ .$$

Base Case ($n = 1$): In this case, $A$ has precisely one element, i.e. $B = A \setminus \{a\} = \emptyset$, and $\#\emptyset = 0 = n - 1$ proves $\phi(1)$.

Induction Step: For the induction hypothesis, we assume $\phi(n)$ to be true, i.e. we assume (a) and (b) hold for each $A$ with $\#A = n$. We have to prove $\phi(n + 1)$, i.e., we consider $A$ with $\#A = n + 1$. From $\#A = n + 1$, we conclude the existence of a bijective map $\varphi : A \longrightarrow \{1, \ldots, n + 1\}$. We have to construct a bijective map $\psi : A \setminus \{a\} \longrightarrow \{1, \ldots, n\}$. To this end, set $k := \varphi(a)$ and define the auxiliary function

$$f : \{1, \ldots, n + 1\} \longrightarrow \{1, \ldots, n + 1\}, \quad f(x) := \begin{cases} n + 1 & \text{for } x = k, \\ k & \text{for } x = n + 1, \\ x & \text{for } x \notin \{k, n + 1\}. \end{cases} \quad (A.47)$$
Then $f \circ \varphi : A \to \{1, \ldots, n + 1\}$ is bijective by Th. 2.14, and

$$(f \circ \varphi)(a) = f(\varphi(a)) = f(k) = n + 1.$$  

Thus, the restriction $\psi := (f \circ \varphi) \upharpoonright_{A \setminus \{a\}}$ is the desired bijective map $\psi : A \setminus \{a\} \to \{1, \ldots, n\}$, proving $\#(A \setminus \{a\}) = n$. It remains to consider the strict subset $B$ of $A$. Since $B$ is a strict subset of $A$, there exists $a \in A \setminus B$. Thus, $B \subseteq A \setminus \{a\}$ and, as we have already shown $\#(A \setminus \{a\}) = n$, the induction hypothesis applies and yields $B$ is finite with $\#B \leq \#(A \setminus \{a\}) = n$, i.e. $\#B \in \{0, \ldots, n\}$, proving $\phi(n + 1)$, thereby completing the induction.  

\begin{thm}
For $\#A = \#B = n \in \mathbb{N}$ and $f : A \to B$, the following statements are equivalent:

(i) $f$ is injective.

(ii) $f$ is surjective.

(iii) $f$ is bijective.

\end{thm}

\begin{proof}
It suffices to prove the equivalence of (i) and (ii).

If $f$ is injective, then $f : A \to f(A)$ is bijective. Since $\#A = n$, there exists a bijective map $\varphi : A \to \{1, \ldots, n\}$. Then $(\varphi \circ f^{-1}) : f(A) \to \{1, \ldots, n\}$ is also bijective, showing $\#f(A) = n$, i.e., according to Th. A.63(a), $f(A)$ can not be a strict subset of $B$, i.e. $f(A) = B$, proving $f$ is surjective.

If $f$ is surjective, then $f$ has a right inverse $g : B \to A$: One can obtain this from Th. 2.13(a), but, here, we can actually construct $g$ without the axiom of choice: We let $\varphi : A \to \{1, \ldots, n\}$ be the bijective map from above and, for $b \in B$, we let $g(b)$ be the unique $a \in C := f^{-1}(\{b\})$ such that $\varphi(a) = \min \varphi(C)$. Then, clearly, $f \circ g = \text{Id}_B$. But this also means $f$ is a left inverse for $g$, such that $g$ must be injective by Th. 2.13(b). According to what we have already proved above, $g$ injective implies $g$ surjective, i.e. $g$ must be bijective. From Th. 2.13(c), we then know the left inverse of $g$ is unique, implying $f = g^{-1}$. In particular, $f$ is injective.

\end{proof}

\begin{lem}
For each finite set $A$ (i.e. $\#A = n \in \mathbb{N}_0$) and each $B \subseteq A$, one has $\#(A \setminus B) = \#A - \#B$.

\end{lem}

\begin{proof}
For $B = \emptyset$, the assertion is true since $\#(A \setminus B) = \#A = \#A - 0 = \#A - \#B$.

For $B \neq \emptyset$, the proof is conducted over the size of $B$, i.e. as a finite induction (cf. Cor. 3.6) over the set $\{1, \ldots, n\}$, showing

$$\forall m \in \{1, \ldots, n\} \left(\#B = m \Rightarrow \#(A \setminus B) = \#A - \#B\right).$$

Base Case $(m = 1)$: $\phi(1)$ is precisely the statement provided by Th. A.63(b).
Induction Step: For the induction hypothesis, we assume $\phi(m)$ with $1 \leq m < n$. To prove $\phi(m + 1)$, consider $B \subseteq A$ with $\#B = m + 1$. Fix an element $b \in B$ and set $B_1 := B \setminus \{b\}$. Then $\#B_1 = m$ by Th. A.63(b), $A \setminus B = (A \setminus B_1) \setminus \{b\}$, and we compute
\[
\#(A \setminus B) = \#((A \setminus B_1) \setminus \{b\}) \overset{\text{Th. A.63(b)}}{=} \#(A \setminus B_1) - 1 \overset{(\phi(m))}{=} #A - #B_1 - 1
\]
proving $\phi(m + 1)$ and completing the induction. ■

**Theorem A.66.** If $A, B$ are finite sets, then $\#(A \cup B) = #A + #B - # (A \cap B)$.

**Proof.** The assertion is clearly true if $A$ or $B$ is empty. If $A$ and $B$ are nonempty, then there exist $m, n \in \mathbb{N}$ such that $\#A = m$ and $\#B = n$, i.e. there are bijective maps $f : A \rightarrow \{1, \ldots, m\}$ and $g : B \rightarrow \{1, \ldots, n\}$.

We first consider the case $A \cap B = \emptyset$. We need to construct a bijective map $h : A \cup B \rightarrow \{1, \ldots, m + n\}$. To this end, we define
\[
h : A \cup B \rightarrow \{1, \ldots, m + n\}, \quad h(x) := \begin{cases} f(x) & \text{for } x \in A, \\ g(x) + m & \text{for } x \in B. \end{cases}
\]

The bijectivity of $f$ and $g$ clearly implies the bijectivity of $h$, proving $\#(A \cup B) = m + n = #A + #B$.

Finally, we consider the case of arbitrary $A, B$. Since $A \cup B = A \cup (B \setminus A)$ and $B \setminus A = B \setminus (A \cap B)$, we can compute
\[
\#(A \cup B) = \#(A \cup (B \setminus A)) = #A + #B \setminus A
\]
\[
= #A + #B - (A \cap B), \overset{\text{Lem. A.65}}{=} #A + #B - # (A \cap B),
\]
thereby establishing the case. ■

**Theorem A.67.** If $(A_1, \ldots, A_n)$, $n \in \mathbb{N}$, is a finite sequence of finite sets, then
\[
\# \prod_{i=1}^{n} A_i = \#(A_1 \times \cdots \times A_n) = \prod_{i=1}^{n} \#A_i. \quad \text{(A.47)}
\]

**Proof.** If at least one $A_i$ is empty, then (A.47) is true, since both sides are 0.

The case where all $A_i$ are nonempty is proved by induction over $n$, i.e. we know $k_i := \#A_i \in \mathbb{N}$ for each $i \in \{1, \ldots, n\}$ and show by induction
\[
\forall n \in \mathbb{N}, \quad \prod_{i=1}^{n} A_i = \prod_{i=1}^{n} k_i, \overset{\phi(n)}{=} \prod_{i=1}^{n} k_i.
\]

Base Case ($n = 1$): $\prod_{i=1}^{1} A_i = \#A_1 = k_1 = \prod_{i=1}^{1} k_i$, i.e. $\phi(1)$ holds.
Induction Step: From the induction hypothesis \( \phi(n) \), we obtain a bijective map \( \varphi : A \rightarrow \{1, \ldots, N\} \), where \( A := \prod_{i=1}^n A_i \) and \( N := \prod_{i=1}^n k_i \). To prove \( \phi(n+1) \), we need to construct a bijective map \( h : A \times A_{n+1} \rightarrow \{1, \ldots, N \cdot k_{n+1}\} \). Since \( \#A_{n+1} = k_{n+1} \), there exists a bijective map \( f : A_{n+1} \rightarrow \{1, \ldots, k_{n+1}\} \). We define

\[
h : A \times A_{n+1} \rightarrow \{1, \ldots, N \cdot k_{n+1}\},
\]

\[
h(a_1, \ldots, a_n, a_{n+1}) := (f(a_{n+1}) - 1) \cdot N + \varphi(a_1, \ldots, a_n).
\]

Since \( \varphi \) and \( f \) are bijective, and since every \( m \in \{1, \ldots, N \cdot k_{n+1}\} \) has a unique representation in the form \( m = a \cdot N + r \) with \( a \in \{0, \ldots, k_{n+1} - 1\} \) and \( r \in \{1, \ldots, N\} \) (exercise), \( h \) is also bijective. This proves \( \phi(n+1) \) and completes the induction.

**Theorem A.68.** For each finite set \( A \) (i.e. \( \#A = n \in \mathbb{N}_0 \)), one has \( \#\mathcal{P}(A) = 2^n \).

**Proof.** The proof is conducted by induction by showing

\[
\forall n \in \mathbb{N}_0 \quad \left( \#A = n \Rightarrow \#\mathcal{P}(A) = 2^n \right) \quad \phi(n).
\]

Base Case \( (n = 0) \): For \( n = 0 \), we have \( A = \emptyset \), i.e. \( \mathcal{P}(A) = \{\emptyset\} \). Thus, \( \#\mathcal{P}(A) = 1 = 2^0 \), proving \( \phi(0) \).

Induction Step: Assume \( \phi(n) \) and consider \( A \) with \( \#A = n + 1 \). Then \( A \) contains at least one element \( a \). For \( B := A \setminus \{a\} \), we then know \( \#B = n \) from Th. A.63(b). Moreover, setting \( \mathcal{M} := \{C \cup \{a\} : C \in \mathcal{P}(B)\} \), we have the disjoint decomposition \( \mathcal{P}(A) = \mathcal{P}(B) \cup \mathcal{M} \). As the map \( \varphi : \mathcal{P}(B) \rightarrow \mathcal{M}, \varphi(C) := C \cup \{a\} \), is clearly bijective, \( \mathcal{P}(B) \) and \( \mathcal{M} \) have the same cardinality. Thus,

\[
\#\mathcal{P}(A) \stackrel{\text{Th. A.66}}{=} \#\mathcal{P}(B) + \#\mathcal{M} = \#\mathcal{P}(B) + \#\mathcal{P}(B) \quad \left( \phi(n) \right)^{\phi(n)} = 2 \cdot 2^n = 2^{n+1},
\]

thereby proving \( \phi(n+1) \) and completing the induction.

**A.5.3 Power Sets**

**Theorem A.69.** Let \( A \) be a set. There can never exist a surjective map from \( A \) onto \( \mathcal{P}(A) \) (in this sense, the size of \( \mathcal{P}(A) \) is always strictly bigger than the size of \( A \); in particular, \( A \) and \( \mathcal{P}(A) \) can never have the same size).

**Proof.** If \( A = \emptyset \), then there is nothing to prove. For nonempty \( A \), the idea is to conduct a proof by contradiction. To this end, assume there does exist a surjective map \( f : A \rightarrow \mathcal{P}(A) \) and define

\[
B := \{x \in A : x \notin f(x)\}. \quad (A.48)
\]

Now \( B \) is a subset of \( A \), i.e. \( B \in \mathcal{P}(A) \) and the assumption that \( f \) is surjective implies the existence of \( a \in A \) such that \( f(a) = B \). If \( a \in B \), then \( a \notin f(a) = B \), i.e. \( a \in B \)
implies $a \in B \land \neg(a \in B)$, so that the principle of contradiction tells us $a \notin B$ must be true. However, $a \notin B$ implies $a \in f(a) = B$, i.e., this time, the principle of contradiction tells us $a \in B$ must be true. In conclusion, we have shown our original assumption that there exists a surjective map $f : A \rightarrow \mathcal{P}(A)$ implies $a \in B \land \neg(a \in B)$, i.e., according to the principle of contradiction, no surjective map from $A$ into $\mathcal{P}(A)$ can exist. ■

B General Forms of the Laws of Associativity and Commutativity

B.1 Associativity

In the literature, the general law of associativity is often stated in the form that $a_1a_2 \cdots a_n$ gives the same result “for every admissible way of inserting parentheses into $a_1a_2 \cdots a_n$”, but a completely precise formulation of what that actually means seems to be rare. As a warm-up, we first prove a special case of the general law:

**Proposition B.1.** Let $A$ be a set with a composition $\cdot : A \times A \rightarrow A$ (which we write as a multiplication, but, clearly, this is not essential, and we could also write it as an addition or with some other symbol). If the composition is associative, i.e. if

$$\forall a,b,c \in A \ (ab)c = a(bc), \tag{B.1}$$

then

$$\forall n \in \mathbb{N}, \ a_1, \ldots, a_n \in A \ \forall k \in \{2, \ldots, n\} \quad \left( \prod_{i=k}^{n} a_i \right) \left( \prod_{i=1}^{k-1} a_i \right) = \prod_{i=1}^{n} a_i, \tag{B.2}$$

where the product symbol is defined according to (3.20a).

**Proof.** If $k = n$, then (B.2) is immediate from (3.20a). For $2 \leq k < n$, we prove (B.2) by induction on $n$: For the base case, $n = 2$, there is nothing to prove. For $n > 2$, one computes

$$\left( \prod_{i=k}^{n} a_i \right) \left( \prod_{i=1}^{k-1} a_i \right) = \left( a_n \cdot \prod_{i=k}^{n-1} a_i \right) \left( \prod_{i=1}^{k-1} a_i \right) \stackrel{(3.20a)}{=} \left( \prod_{i=k}^{n-1} a_i \right) \left( \prod_{i=1}^{k-1} a_i \right) \stackrel{(B.1)}{=} a_n \cdot \left( \prod_{i=k}^{n-1} a_i \right) \left( \prod_{i=1}^{k-1} a_i \right) \stackrel{\text{ind.hyp.}}{=} a_n \cdot \prod_{i=1}^{n-1} a_i \stackrel{(3.20a)}{=} \prod_{i=1}^{n} a_i, \tag{B.3}$$

completing the induction and the proof of the proposition. ■

The difficulty in stating the general form of the law of associativity lies in giving a precise definition of what one means by “an admissible way of inserting parentheses into $a_1a_2 \cdots a_n$”. So how does one actually proceed to calculate the value of $a_1a_2 \cdots a_n$, given...
that parentheses have been inserted in an admissible way? The answer is that one does it in \( n - 1 \) steps, where, in each step, one combines two juxtaposed elements, consistent with the inserted parentheses. There can still be some ambiguity: For example, for \((a_1 a_2)(a_3 (a_4 a_5))\), one has the freedom of first combining \(a_1, a_2\), or of first combining \(a_4, a_5\). In consequence, our general law of associativity will show that, for each admissible sequence of \( n - 1 \) directives for combining two juxtaposed elements, the final result is the same (under the hypothesis that (B.1) holds). This still needs some preparatory work.

In the following, one might see it as a slight notational inconvenience that we have defined \( \prod_{i=1}^{n} a_i \) as \( a_n \cdots a_1 \) rather than \( a_1 \cdots a_n \). For this reason, we will enumerate the elements to be combined by composition from right to left rather than from left to right.

**Definition and Remark B.2.** Let \( A \) be a (nonempty) set with a composition \( \cdot : A \times A \to A \), let \( n \in \mathbb{N}, \ n \geq 2 \), and let \( I \) be a totally ordered index set, \( \#I = n \), \( I = \{i_1, \ldots, i_n\} \) with \( i_1 < \cdots < i_n \). Moreover, let \( F := (a_{i_n}, \ldots, a_{i_1}) \) be a family of \( n \) elements of \( A \).

(a) An admissible composition directive (for combining two juxtaposed elements of the family) is an index \( i_k \in I \) with \( 1 \leq k \leq n - 1 \). It transforms the family \( F \) into the family \( G := (a_{i_n}, \ldots, a_{i_{k+1}} a_{i_k}, \ldots, a_{i_1}) \). In other words, \( G = (b_j)_{j \in J} \), where \( J := I \setminus \{i_{k+1}\} \), \( b_j = a_j \) for each \( j \in J \setminus \{i_k\} \), and \( b_{i_k} = a_{i_{k+1}} a_{i_k} \). We can write this transformation as two maps

\[
F \mapsto \delta_{i_k}^{(1)}(F) := G = (a_{i_n}, \ldots, a_{i_{k+1}} a_{i_k}, \ldots, a_{i_1}) = (b_j)_{j \in J},
\]

\[
I \mapsto \delta_{i_k}^{(2)}(I) := J = I \setminus \{i_{k+1}\}.
\]

Thus, an application of an admissible composition directive reduces the length of the family and the number of indices by one.

(b) Recursively, we define (finite) sequences of families, index sets, and indices as follows:

\[
F_n := F, \quad I_n := I,
\]

\[
\forall_{\alpha \in \{2, \ldots, n\}} F_{\alpha-1} := \delta_{i_{\alpha}}^{(1)}(F_{\alpha}), \quad I_{\alpha-1} := \delta_{i_{\alpha}}^{(2)}(I_{\alpha}), \quad \text{where } j_{\alpha} \in I_{\alpha} \setminus \{\max I_{\alpha}\}.
\]

The corresponding sequence of indices \( D := (j_n, \ldots, j_2) \) in \( I \) is called an admissible evaluation directive. Clearly,

\[
\forall_{\alpha \in \{1, \ldots, n\}} \#I_{\alpha} = \alpha, \quad \text{i.e. } F_{\alpha} \text{ has length } \alpha.
\]

In particular, \( I_1 = \{j_2\} = \{i_1\} \) (where the second equality follows from (B.4b)), \( F_1 = (a) \), and we call

\[
D(F) := a
\]

the result of the admissible evaluation directive \( D \) applied to \( F \).
**Theorem B.3** (General Law of Associativity). Let $A$ be a (nonempty) set with a composition $\cdot : A \times A \to A$, let $n \in \mathbb{N}$, $n \geq 2$, and let $I$ be a totally ordered index set, $\#I = n$, $I = \{i_1, \ldots, i_n\}$ with $i_1 < \cdots < i_n$. Moreover, let $F := (a_{i_1}, \ldots, a_{i_n})$ be a family of $n$ elements of $A$. If the composition is associative, i.e. if (B.1) holds, then, for each admissible evaluation directive as defined in Def. and Rem. B.2(b), the result is the same, namely

$$\mathcal{D}(F) = \prod_{k=1}^{n} a_{i_k}. \quad \text{(B.8)}$$

**Proof.** We conduct the proof via induction on $n$. For $n = 3$, there are only two possible directives and (B.1) guarantees that they yield the same result. For the induction step, let $n > 3$. As in Def. and Rem. B.2(b), we write $\mathcal{D} = (j_n, \ldots, j_2)$ and obtain some $I_2 = \{i_1, i_m\}$, $1 < m \leq n$, as the corresponding penultimate index set. Depending on $i_m$, we partition $(j_n, \ldots, j_3)$ as follows: Set

$$J_1 := \{k \in \{3, \ldots, n\} : j_k < i_m\}, \quad J_2 := \{k \in \{3, \ldots, n\} : j_k \geq i_m\}. \quad \text{(B.9)}$$

Then, for $k \in J_1$, $j_k$ is a composition directive to combine two elements to the right of $a_{i_m}$ and, for $k \in J_2$, $j_k$ is a composition directive to combine two elements to the left of $a_{i_m}$. Moreover, $J_1$ and $J_2$ might or might not be the empty set: If $J_1 = \emptyset$, then $j_k \neq i_1$ for each $k \in \{3, \ldots, n\}$, implying $i_m = i_2$; if $J_2 = \emptyset$, then, in each of the $n - 2$ steps to obtain $I_2$, an $i_k$ with $k < m$ was removed from $I$, implying $i_m = i_n$ (in particular, as $n \neq 2$, $J_1$ and $J_2$ cannot both be empty). If $J_1 \neq \emptyset$, then $\mathcal{D}_1 := (j_k)_{k \in J_1}$ is an admissible evaluation directive for $(a_{i_{m-1}}, \ldots, a_{i_1})$ – this follows from

$$j_k \in K \subseteq \{i_1, \ldots, i_{m-1}\} \Rightarrow \delta^{(2)}_{j_k}(K) \subseteq K \subseteq \{i_1, \ldots, i_{m-1}\}. \quad \text{(B.10)}$$

Since $m - 1 < n$, the induction hypothesis applies and yields

$$\mathcal{D}_1(a_{i_{m-1}}, \ldots, a_{i_1}) = \prod_{k=1}^{m-1} a_{i_k}. \quad \text{(B.11)}$$

Analogously, if $J_2 \neq \emptyset$, then $\mathcal{D}_2 := (j_k)_{k \in J_2}$ is an admissible evaluation directive for $(a_{i_n}, \ldots, a_{i_m})$ – this follows from

$$j_k \in K \subseteq \{i_m, \ldots, i_n\} \Rightarrow \delta^{(2)}_{j_k}(K) \subseteq K \subseteq \{i_m, \ldots, i_n\}. \quad \text{(B.12)}$$

Since $m > 1$, the induction hypothesis applies and yields

$$\mathcal{D}_2(a_{i_n}, \ldots, a_{i_m}) = \prod_{k=m}^{n} a_{i_k}. \quad \text{(B.13)}$$

Thus, if $J_1 \neq \emptyset$ and $J_2 \neq \emptyset$, then we obtain

$$\mathcal{D}(F) \overset{j_2=i_1}{=} \mathcal{D}_2(a_{i_n}, \ldots, a_{i_m}) \cdot \mathcal{D}_1(a_{i_{m-1}}, \ldots, a_{i_1}) = \left(\prod_{k=m}^{n} a_{i_k}\right) \left(\prod_{k=1}^{m-1} a_{i_k}\right) \overset{\text{Prop. B.1}}{=} \prod_{k=1}^{n} a_{i_k} \quad \text{(B.14)}$$
as desired. If $J_1 = \emptyset$, then, as explained above, $i_m = i_2$. Thus, in this case,

$$D(F)^{j_2\equiv i_1} D_2(a_{i_1}, \ldots, a_{i_2}) \cdot a_{i_1} = \left( \prod_{k=2}^{n} a_{i_k} \right) \cdot a_{i_1} \overset{\text{Prop. B.1}}{=} \prod_{k=1}^{n} a_{i_k}$$

(B.15)

as needed. Finally, if $J_2 = \emptyset$, then, as explained above, $i_m = i_n$. Thus, in this case,

$$D(F)^{j_2\equiv i_1} a_{i_n} \cdot D_1(a_{i_{n-1}}, \ldots, a_{i_1}) = \prod_{k=1}^{n} a_{i_k},$$

(B.16)

again, as desired, and completing the induction.  

\[ \square \]

\section*{B.2 Commutativity}

In the present section, we will generalize the law of commutativity $ab = ba$ to a finite number of factors, provided the composition is also associative. For this purpose, we introduce the notion of permutation, also useful in many other mathematical contexts.

**Definition and Remark B.4.** Let $n \in \mathbb{N}$. Each bijective map $\pi : \{1, \ldots, n\} \to \{1, \ldots, n\}$ is called a permutation of $\{1, \ldots, n\}$. The set of permutations of $\{1, \ldots, n\}$ forms a group with respect to the composition of maps, the so-called symmetric group $S_n$: Indeed, the composition of maps is associative by Prop. 2.10(a); the neutral element is the identity map $e : \{1, \ldots, n\} \to \{1, \ldots, n\}$, $e(i) = i$; and, for each $\pi \in S_n$, its inverse map $\pi^{-1}$ is also its inverse element in the group $S_n$. Caveat: Simple examples show that $S_n$ is not commutative.

**Theorem B.5** (General Law of Commutativity). Let $A$ be a set with an associative composition $\cdot : A \times A \to A$ (which we write as a multiplication, but, clearly, this is not essential, and we could also write it as an addition or with some other symbol). If the composition is commutative, i.e. if

$$\forall_{a,b \in A} \quad ab = ba,$$

(B.17)

then

$$\forall_{n \in \mathbb{N}} \forall_{\pi \in S_n} \forall_{a_1, \ldots, a_n \in A} \prod_{i=1}^{n} a_i = \prod_{i=1}^{n} a_{\pi(i)}.$$

(B.18)

Before we can carry out the proof, we need to learn a bit more about permutations.

**Definition B.6.** Let $k, n \in \mathbb{N}$, $k \leq n$. A permutation $\pi \in S_n$ is called a $k$-cycle if, and only if, there exist $k$ distinct numbers $i_1, \ldots, i_k \in \{1, \ldots, n\}$ such that

$$\pi(i) = \begin{cases} 
  i_{j+1} & \text{if } i = i_j, j \in \{1, \ldots, k-1\}, \\
  i_1 & \text{if } i = i_k, \\
  i & \text{if } i \notin \{i_1, \ldots, i_k\}.
\end{cases}$$

(B.19)
If $\pi$ is a cycle as in (B.19), then one writes
\[ \pi = (i_1 i_2 \ldots i_k). \] (B.20)

Each 2-cycle is also known as a transposition.

**Theorem B.7.** Let $n \in \mathbb{N}$.

(a) Each permutation can be decomposed into finitely many disjoint cycles: For each $\pi \in S_n$, there exists a decomposition of $\{1, \ldots, n\}$ into disjoint sets $A_1, \ldots, A_N$, $N \in \mathbb{N}$, i.e.
\[ \{1, \ldots, n\} = \bigcup_{i=1}^{N} A_i \quad \text{and} \quad A_i \cap A_j = \emptyset \quad \text{for} \ i \neq j, \] (B.21)

such that $A_i$ consists of the distinct elements $a_{i1}, \ldots, a_{i,N_i}$ and
\[ \pi = (a_{N1} \ldots a_{N,N_N}) \cdots (a_{11} \ldots a_{1,N_1}). \] (B.22)

The decomposition (B.22) is unique up to the order of the cycles.

(b) If $n \geq 2$, then every permutation $\pi \in S_n$ is the composition of finitely many transpositions, where each transposition permutes two juxtaposed elements, i.e.
\[ \forall \pi \in S_n \ \exists N \in \mathbb{N} \ \exists \tau_1, \ldots, \tau_N \in T \quad \pi = \tau_N \circ \cdots \circ \tau_1, \] (B.23)

where $T := \{(i, i+1) : i \in \{1, \ldots, n-1\}\}$.

**Proof.** (a): We prove the statement by induction on $n$. For $n = 1$, there is nothing to prove. Let $n > 1$ and choose $i \in \{1, \ldots, n\}$. We claim that
\[ \exists k \in \mathbb{N} \quad \left( \pi^k(i) = i \land \forall l \in \{1, \ldots, k-1\} \pi^l(i) \neq i \right). \] (B.24)

Indeed, since $\{1, \ldots, n\}$ is finite, there must be a smallest $k \in \mathbb{N}$ such that $\pi^k(i) \in A_1 := \{i, \pi(i), \ldots, \pi^{k-1}(i)\}$. Since $\pi$ is bijective, it must be $\pi^k(i) = i$ and $(i, \pi(i), \ldots, \pi^{k-1}(i))$ is a $k$-cycle. We are already done in case $k = n$. If $k < n$, then consider $B := \{1, \ldots, n\} \setminus A_1$. Then, again using the bijectivity of $\pi$, $\pi|_B$ is a permutation on $B$ with $1 \leq \#B < n$. By induction, there are disjoint sets $A_2, \ldots, A_N$ such that $B = \bigcup_{j=2}^{N} A_j$, $A_j$ consists of the distinct elements $a_{j1}, \ldots, a_{j,N_j}$ and
\[ \pi|_B = (a_{N1} \ldots a_{N,N_N}) \cdots (a_{21} \ldots a_{2,N_2}). \]

Since $\pi = (i, \pi(i), \ldots, \pi^{k-1}(i)) \circ \pi|_B$, this finishes the proof of (B.22). If there were another, different, decomposition of $\pi$ into cycles, say, given by disjoint sets $B_1, \ldots, B_M$, $\{1, \ldots, n\} = \bigcup_{i=1}^{M} B_i$, $M \in \mathbb{N}$, then there were $A_i \neq B_j$ and $k \in A_i \cap B_j$. But then $k$ were in the cycle given by $A_i$ and in the cycle given by $B_j$, implying $A_i = \{\pi^l(k) : l \in \mathbb{N}\} = B_j$, in contradiction to $A_i \neq B_j$. 
(b): We first show that every \( \pi \in S_n \) is a composition of finitely many transpositions (not necessarily transpositions from the set \( T \)): According to (a), it suffices to show that every cycle is a composition of finitely many transpositions. Since each 1-cycle is the identity, it is \((i) = \text{Id} = (1 2) (1 2)\) for each \( i \in \{1, \ldots, n\} \). If \((i_1 \ldots i_k)\) is a \( k \)-cycle, \( k \in \{2, \ldots, n\} \), then

\[
(i_1 \ldots i_k) = (i_1 i_2)(i_2 i_3) \cdots (i_{k-1} i_k) : \tag{B.25}
\]

Indeed,

\[
\forall i \in \{1, \ldots, n\}, \quad (i_1 i_2)(i_2 i_3) \cdots (i_{k-1} i_k)(i) = \begin{cases} i_1 & \text{for } i = i_k, \\ i_{l+1} & \text{for } i = i_l, l \in \{1, \ldots, k-1\}, \\ i & \text{for } i \notin \{i_1, \ldots, i_k\}, \end{cases} \tag{B.26}
\]

proving (B.25). To finish the proof of (b), we observe that every transposition is a composition of finitely many elements of \( T \): If \( i, j \in \{1, \ldots, n\}, i < j \), then

\[
(i j) = (i i + 1) \cdots (j - 2 j - 1)(j - 1 j) \cdots (i + 1 i + 2)(i i + 1) : \tag{B.27}
\]

Indeed,

\[
\forall k \in \{1, \ldots, n\}, \quad (i i + 1) \cdots (j - 2 j - 1)(j - 1 j) \cdots (i + 1 i + 2)(i i + 1)(k) = \begin{cases} j & \text{for } k = i, \\ i & \text{for } k = j, \\ k & \text{for } i < k < j, \\ k & \text{for } k \notin \{i, i + 1, \ldots, j\}, \end{cases} \tag{B.28}
\]

proving (B.27).

**Proof of Th. B.5.** For \( n = 1 \), there is nothing to prove. So let \( n > 1 \). For \( l \in 1, \ldots, n-1 \), let \( \tau_l : \{1, \ldots, n\} \to \{1, \ldots, n\} \) be the transposition that interchanges \( l \) and \( l + 1 \) and leaves all other elements fixed (i.e. \( \tau_l(l) = l + 1, \tau_l(l + 1) = l, \tau_l(\alpha) = \alpha \) for each \( \alpha \in \{1, \ldots, n\} \setminus \{l, l + 1\} \) and let \( T := \{\tau_1, \ldots, \tau_{n-1}\} \). Then (B.17) and Th. B.3 imply that the theorem holds for \( \pi = \tau \) for each \( \tau \in T \). For a general permutation \( \pi \in S_n \), Th. B.7(b) provides a finite sequence \( (\tau^1, \ldots, \tau^N), N \in \mathbb{N} \), of elements of \( T \) such that \( \pi = \tau^N \circ \cdots \circ \tau^1 \). Thus, as we already know that the theorem holds for \( N = 1 \), the case \( N > 1 \) follows by induction.

The following example shows that, if the composition is not associative, then, in general, (B.17) does *not* imply (B.18):

**Example B.8.** Let \( A := \{a, b, c\} \) with \( \#A = 3 \) (i.e. the elements \( a, b, c \) are all distinct). Let the composition \( \cdot \) on \( A \) be defined according to the composition table

<table>
<thead>
<tr>
<th></th>
<th>a</th>
<th>b</th>
<th>c</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>b</td>
<td>b</td>
<td>b</td>
</tr>
<tr>
<td>b</td>
<td>b</td>
<td>b</td>
<td>a</td>
</tr>
<tr>
<td>c</td>
<td>b</td>
<td>a</td>
<td>a</td>
</tr>
</tbody>
</table>
(the table entry at the intersection of the row labeled with factor $x$ and the column labeled with factor $y$ is definition of $x \cdot y$). Then, clearly, $\cdot$ is commutative. However, $\cdot$ is not associative, since, e.g.,

$$ (cb)a = aa = b \neq a = cb = c(ba), \quad (B.29) $$

and (B.18) does not hold, since, e.g.,

$$ a(bc) = aa = b \neq a = cb = c(ba). \quad (B.30) $$

C Algebraic Structures

C.1 Groups

Definition C.1. Let $G$ be a nonempty set with a map

$$ \circ : G \times G \rightarrow G, \quad (x, y) \mapsto x \circ y \quad (C.1) $$

(called a composition on $G$, the examples we have in mind are addition and multiplication on $\mathbb{R}$). Then $(G, \circ)$ (or just $G$, if the composition $\circ$ is understood) is called a group if, and only if, the following three conditions are satisfied:

(i) Associativity: $x \circ (y \circ z) = (x \circ y) \circ z$ holds for all $x, y, z \in G$.

(ii) There exists a neutral element $e \in G$, i.e. an element $e \in G$ such that

$$ \forall x \in G \quad x \circ e = x. \quad (C.2) $$

(iii) For each $x \in G$, there exists an inverse element $\overline{x} \in G$, i.e. an element $\overline{x} \in G$ such that

$$ x \circ \overline{x} = e. $$

$G$ is called a commutative or abelian group if, and only if, it is a group and satisfies the additional condition:

(iv) Commutativity: $x \circ y = y \circ x$ holds for all $x, y \in G$.

Theorem C.2. The following statements and rules are valid in every group $(G, \circ)$:

(a) If (C.2) holds for $e \in G$, then

$$ \forall x \in G \quad e \circ x = x. \quad (C.3) $$

also holds.
The neutral element is unique: If \( e, f \in G \), then
\[
\left( \left( \forall_{x \in G} x \circ e = x \right) \land \left( \forall_{x \in G} x \circ f = x \right) \right) \implies e = f. \tag{C.4}
\]

(c) If \( x, a \in G \) and \( x \circ a = e \) (where \( e \in G \) is the neutral element), then \( a \circ x = e \) as well. Moreover, inverse elements are unique (for each \( x \in G \), the unique inverse is then denoted by \( x^{-1} \)).

(d) \( (x^{-1})^{-1} = x \) holds for each \( x \in G \).

(e) \( y^{-1} \circ x^{-1} = (x \circ y)^{-1} \) holds for each \( x, y \in G \).

(f) \( x \circ a = y \circ a \Rightarrow x = y \) holds for each \( x, y, a \in G \).

Proof. (a): Let \( x \in G \). By Def. C.1(iii), there exists \( y \in G \) such that \( x \circ y = e \) and, in turn, \( z \in G \) such that \( y \circ z = e \). Thus,
\[
e \circ z = (x \circ y) \circ z = x \circ (y \circ z) = x \circ e = x, \tag{C.5}
\]
implying
\[
x = e \circ z = (e \circ e) \circ z = e \circ (e \circ z) = e \circ x \tag{C.6}
\]
as desired.

(b): If \( e, f \) are both neutral elements, then, using (a), \( f = e \circ f = e \).

(c): Assume \( x \circ a = e \). Then there is \( b \) such that \( a \circ b = e \). One computes
\[
e = a \circ b = (a \circ e) \circ b = (a \circ (x \circ a)) \circ b = a \circ ((x \circ a) \circ b) = a \circ (x \circ (a \circ b))
= a \circ (x \circ e) = a \circ x, \tag{C.7}
\]
establishing the case. Now let \( a, b \) be inverses to \( x \). Then \( a = a \circ e = a \circ x \circ b = e \circ b = b \).

(d): \( x^{-1} \circ x = e \) holds according to (c) and shows that \( x \) is the inverse to \( x^{-1} \). Thus, \( (x^{-1})^{-1} = x \) as claimed.

(e) is due to \( y^{-1} \circ x^{-1} \circ x \circ y = y^{-1} \circ e \circ y = e \).

(f): If \( x \circ a = y \circ a \), then \( x = x \circ a \circ a^{-1} = y \circ a \circ a^{-1} = y \) as claimed. \( \blacksquare \)

Definition C.3. Let \((G, \circ)\) and \((H, \circ)\) be groups. A map \( \phi : G \to H \) is called a (group) homomorphism if, and only if,
\[
\forall_{a,b \in G} \phi(a \circ b) = \phi(a) \circ \phi(b). \tag{C.8}
\]

Proposition C.4. Let \((G, \circ)\) and \((H, \circ)\) be groups and let \( \phi : G \to H \) be a group homomorphism. Let \( e, e' \) denote the neutral elements of \( G \) and \( H \), respectively. Then the following holds:

(a) \( \phi(e) = e' \).
(b) \(\phi(a^{-1}) = (\phi(a))^{-1}\) for each \(a \in G\).

(c) If \(\phi\) is bijective, then \(\phi^{-1}\) is also a group homomorphism.

Proof. (a): We compute
\[
\phi(e) \circ e' = \phi(e) = \phi(e \circ e) = \phi(e) \circ \phi(e).
\]
Applying \((\phi(e))^{-1}\) to both sides of the above equality then proves \(\phi(e) = e'\).

(b): We compute
\[
\phi(a^{-1}) \circ \phi(a) = \phi(a^{-1} \circ a) = \phi(e) = e',
\]
proving (b).

(c): Applying \(\phi^{-1}\) to (C.8) yields
\[
\forall a, b \in G, \quad a \circ b = \phi^{-1}(\phi(a) \circ \phi(b)). \tag{C.9}
\]
Thus, for each \(x, y \in H\), we obtain
\[
\phi^{-1}(x \circ y) = \phi^{-1}(\phi(\phi^{-1}(x)) \circ \phi(\phi^{-1}(y))) = (x \circ y) = \phi^{-1}(x) \circ \phi^{-1}(y),
\]
establishing the case and completing the proof of the proposition. \(\blacksquare\)

Notation C.5. Exponentiation with Integer Exponents: Let \(G\) be a nonempty set with a composition \(\cdot : G \times G \rightarrow G\). Assume there exists a (unique) neutral element \(1 \in G\) (satisfying \(x \cdot 1 = x\) for each \(x \in G\)). Define recursively for each \(x \in G\) and each \(n \in \mathbb{N}_0\):
\[
x^0 := 1, \quad \forall n \in \mathbb{N}_0, \quad x^{n+1} := x \cdot x^n. \tag{C.10a}
\]
Moreover, if \((G, \cdot)\) constitutes a group, then also define for each \(x \in G\) and each \(n \in \mathbb{N}\):
\[
x^{-n} := (x^{-1})^n. \tag{C.10b}
\]

Theorem C.6. Exponentiation Rules: Let \(G\) be a nonempty set with a composition \(\cdot : G \times G \rightarrow G\). Assume that the composition satisfies the law of associativity and that there exists a (unique) neutral element \(1 \in G\) (satisfying \(x \cdot 1 = x\) for each \(x \in G\)). Let \(x, y \in G\). Then the following rules hold for each \(m, n \in \mathbb{N}_0\). If \((G, \cdot)\) is a group, then the rules even hold for every \(m, n \in \mathbb{Z}\).

(a) \(x^{m+n} = x^m \cdot x^n\).

(b) \((x^m)^n = x^{mn}\).

(c) If the composition is commutative (i.e. \(xy = yx\) for each \(x, y \in G\)), then it holds that \(x^n y^n = (xy)^n\).
Proof. (a): First, we fix \( n \in \mathbb{N}_0 \) and prove the statement for each \( m \in \mathbb{N}_0 \) by induction: The base case \((m = 0)\) is \( x^n = x^n \), which is true. For the induction step, we compute

\[
x^{m+1+n} (C.10a) = x \cdot x^{m+n} \text{ ind. hyp.} \cdot x \cdot x^m \cdot x^n = x^{m+1}x^n,
\]

completing the induction step. Now assume \( G \) to be a group. Consider \( m \geq 0 \) and \( n < 0 \). If \( m + n \geq 0 \), then, using what we have already shown,

\[
x^{m} x^{n} (C.10b) \equiv x^{m} (x^{-1})^{-n} = x^{m+n} x^{-n} (x^{-1})^{-n} = x^{m+n}.
\]

Similarly, if \( m + n < 0 \), then

\[
x^{m} x^{n} (C.10b) \equiv x^{m} (x^{-1})^{-n} = x^{m} (x^{-1})^{m} (x^{-1})^{-n-m} (C.10b) \equiv x^{m+n}.
\]

The case \( m < 0, n \geq 0 \) is treated completely analogously. It just remains to consider \( m < 0 \) and \( n < 0 \). In this case,

\[
x^{m+n} = x^{(-m-n)} (C.10b) (x^{-1})^{-m-n} = (x^{-1})^{-m} (x^{-1})^{-n} (C.10b) \equiv x^{m+n}.
\]

(b): First, we prove the statement for each \( n \in \mathbb{N}_0 \) by induction (for \( m < 0 \), we assume \( G \) to be a group): The base case \((n = 0)\) is \( (x^m)^0 = 1 = x^0 \), which is true. For the induction step, we compute

\[
(x^m)^{n+1} (C.10a) = x^m \cdot (x^m)^n \text{ ind. hyp.} \cdot x^m \cdot x^m \cdot x^n = x^m \cdot x^{n+m} = x^{m(n+1)},
\]

completing the induction step. Now, let \( G \) be a group and \( n < 0 \). We already know \( (x^m)^{-1} = x^{-m} \). Thus, using what we have already shown,

\[
(x^m)^{n} (C.10b) \equiv ((x^m)^{-1})^{-n} = (x^{-m})^{-n} = (x^{-m})^{-n} (C.10b) \equiv x^{m}.
\]

(c): For \( n \in \mathbb{N}_0 \), the statement is proved by induction: The base case \((n = 0)\) is \( x^0 y^0 = 1 = (xy)^0 \), which is true. For the induction step, we compute

\[
x^{n+1} y^{n+1} (C.10a) = x \cdot x^n \cdot y \cdot y^n \text{ ind. hyp.} \cdot xy \cdot (xy)^n (C.10a) \equiv (xy)^{n+1},
\]

completing the induction step. If \( G \) is a group and \( n < 0 \), then, using what we have already shown,

\[
x^n y^n (C.10b) (x^{-1})^{-n} (y^{-1})^{-n} = (x^{-1}y^{-1})^{-n} \text{ Th. C.2(e)} \equiv ((xy)^{-1})^{-n} (C.10b) \equiv (xy)^n,
\]

which completes the proof. \(\blacksquare\)
C.2 Rings

Definition C.7. Let \( R \) be a nonempty set with two maps
\[
+ : R \times R \rightarrow R, \quad (x, y) \mapsto x + y,
\]
\[
\cdot : R \times R \rightarrow R, \quad (x, y) \mapsto x \cdot y
\]
(C.11)

(\( + \) is called addition and \( \cdot \) is called multiplication; often one writes \( xy \) instead of \( x \cdot y \)). Then \((R, +, \cdot)\) (or just \( R \), if + and \( \cdot \) are understood) is called a ring if, and only if, the following three conditions are satisfied:

(i) \( R \) is a commutative group with respect to +.

(ii) Multiplication is associative.

(iii) Distributivity:
\[
\forall x, y, z \in R \quad x \cdot (y + z) = x \cdot y + x \cdot z, \tag{C.12a}
\]
\[
\forall x, y, z \in R \quad (y + z) \cdot x = y \cdot x + z \cdot x. \tag{C.12b}
\]

A ring \( R \) is called commutative if, and only if, its multiplication is commutative. Moreover, a ring called a ring with unity if, and only if, \( R \) contains a neutral element with respect to multiplication (i.e. there is \( 1 \in R \) such that \( 1 \cdot x = x \cdot 1 = x \) for each \( x \in R \)) – some authors always require a ring to have a neutral element with respect to multiplication.

Theorem C.8. The following statements and rules are valid in every ring \((R, +, \cdot)\) (let 0 denote the additive neutral element and let \( x, y, z \in R \)):

(a) \( x \cdot 0 = 0 = 0 \cdot x \).

(b) \( x(-y) = -(xy) = (-x)y \).

(c) \( (-x)(-y) = xy \).

(d) \( x(y - z) = xy - xz \).

Proof. (a): One computes
\[
x \cdot 0 + x \cdot 0 \overset{(\text{C.12a})}{=} x \cdot (0 + 0) = x \cdot 0 = 0 + x \cdot 0,
\]
i.e. \( x \cdot 0 = 0 \) follows since \((R, +)\) is a group. The second equality follows analogously using (C.12b).

(b): \( xy + x(-y) = x(y - y) = x \cdot 0 = 0 \), where we used (C.12a) and (a). This shows \( x(-y) \) is the additive inverse to \( xy \). The second equality follows analogously using (C.12b).

(c): \( xy = -((xy)) = -(x(-y)) = (-x)(-y) \), where (b) was used twice.

(d): \( x(y - z) = x(y + (-z)) = xy + x(-z) = xy - xz \).
C.3 Fields

Definition C.9. Let \((F, +, \cdot)\) be a ring with unity. Then \((F, +, \cdot)\) (or just \(F\), if + and \(\cdot\) are understood) is called a field if, and only if, \(F \setminus \{0\}\) is a commutative group with respect to \(\cdot\).

Theorem C.10. The following statements and rules are valid in every field \((F, +, \cdot)\):

(a) Inverse elements are unique. For each \(x \in F\), the unique inverse with respect to addition is denoted by \(-x\). Also define \(y - x := y + (-x)\). For each \(x \in F \setminus \{0\}\), the unique inverse with respect to multiplication is denoted by \(x^{-1}\). For \(x \neq 0\), define

\[
\text{the fractions } \frac{y}{x} := \frac{y}{x} := \frac{yx^{-1}}{x} \text{ with numerator } y \text{ and denominator } x.
\]

(b) \(-(-x) = x\) and \((x^{-1})^{-1} = x\) for \(x \neq 0\).

(c) \((-x) + (-y) = -(x + y)\) and \(x^{-1}y^{-1} = (xy)^{-1}\) for \(x, y \neq 0\).

(d) \(x + a = y + a \Rightarrow x = y\) and, for \(a \neq 0\), \(xa = ya \Rightarrow x = y\).

(e) \(x \cdot 0 = 0\).

(f) \(x(-y) = -(xy)\).

(g) \((-x)(-y) = xy\).

(h) \(x(y - z) = xy - xz\).

(i) \(xy = 0 \Rightarrow x = 0 \lor y = 0\).

(j) Rules for Fractions:

\[
\frac{a}{c} + \frac{b}{d} = \frac{ad + bc}{cd}, \quad \frac{a}{c} \cdot \frac{b}{d} = \frac{ab}{cd}, \quad \frac{a}{c/d} = \frac{ad}{bc},
\]

where all denominators are assumed \(\neq 0\).

Proof. (a) follows by applying Th. C.2(c) to the groups \((F, +)\) and \((F \setminus \{0\}, \cdot)\).

(b) follows by applying Th. C.2(d) to the groups \((F, +)\) and \((F \setminus \{0\}, \cdot)\).

(c) follows by applying Th. C.2(e) to the groups \((F, +)\) and \((F \setminus \{0\}, \cdot)\), plus then using commutativity of the groups.

(d) follows by applying Th. C.2(f) to the groups \((F, +)\) and \((F \setminus \{0\}, \cdot)\) (in the latter situation, the case \(x = y = 0\) is also clear).

(e) follows by applying Th. C.8(a) to the ring with unity \((F, +, \cdot)\).

(f) follows by applying Th. C.8(b) to the ring with unity \((F, +, \cdot)\).

(g) follows by applying Th. C.8(c) to the ring with unity \((F, +, \cdot)\).

(h) follows by applying Th. C.8(d) to the ring with unity \((F, +, \cdot)\).
(i): If $xy = 0$ and $x \neq 0$, then $y = 1 \cdot y = x^{-1}xy = x^{-1} \cdot 0 = 0$.

(j): One computes

$$\frac{a}{c} + \frac{b}{d} = ac^{-1} + bd^{-1} = c^{-1}(ac + bd) = \frac{ad + bc}{cd}$$

and

$$\frac{a}{c} \cdot \frac{b}{d} = ac^{-1}bd^{-1} = \frac{ab}{cd}$$

and

$$\frac{a/c}{b/d} = \frac{ac - 1}{bd - 1} = \frac{ab}{bc},$$

completing the proof. ■

D Construction of the Real Numbers

In Th. 4.4, we have defined the set of real numbers $\mathbb{R}$ as a complete totally ordered field and we claimed that such a complete totally ordered field does actually exist. In the following, we will describe how $\mathbb{R}$ can be constructed. We will follow [EHH+95, Chs. 1,2], which contains several different approaches for the construction of $\mathbb{R}$.

D.1 Natural Numbers

In the first step, one starts with the natural numbers $\mathbb{N}$. The set of natural numbers $\mathbb{N}$ was defined in Def. A.41 and it was shown in Th. A.44 that $\mathbb{N}$ satisfies the Peano axioms P1 – P3 of Sec. 3.1. We denote natural numbers using the usual symbols $0 := \emptyset$, $1 := S(0) = \{0\}$, $2 := S(1) = \{0, 1\}$, $3 := S(2) = \{0, 1, 2\}$, ... , $n + 1 := S(n) = n \cup \{n\} = \{0, 1, \ldots, n\}$ (which is consistent with previous definitions in (A.5) and Not. A.45).

Theorem 3.7 allows to define addition and multiplication on $\mathbb{N}_0$ via recursion:

**Definition D.1.** (a) For each $m, n \in \mathbb{N}_0$, $m + n$ is defined recursively by

$$m + 0 := m, \quad m + 1 := S(m), \quad \forall \ n \in \mathbb{N}, \ m + S(n) := S(m + n). \quad (D.1)$$

This fits into the framework of Th. 3.7, using $A := \mathbb{N}_0$, $x_1 := S(m)$, and, for each $n \in \mathbb{N}$, $f_n : A^n \rightarrow A$, $f_n(x_1, \ldots, x_n) := S(x_n)$ (due to the different initializations, one obtains a different recursion for each $m \in \mathbb{N}_0$).

(b) For each $m, n \in \mathbb{N}_0$, $mn := m \cdot n$ is defined recursively by

$$m \cdot 0 := 0, \quad m \cdot 1 := m, \quad \forall \ n \in \mathbb{N}, \ m \cdot (n + 1) := m \cdot n + m. \quad (D.2)$$

This fits into the framework of Th. 3.7, using $A := \mathbb{N}_0$, $x_1 := m$, and, for each $m, n \in \mathbb{N}$, $f_{m,n} : A^n \rightarrow A$, $f_{m,n}(x_1, \ldots, x_n) := x_n + m$. 

**Theorem D.2.** The set $\mathbb{N}_0$ of the natural numbers (including 0) with the maps of addition and multiplication
\[ + : \mathbb{N}_0 \times \mathbb{N}_0 \to \mathbb{N}_0, \quad (x, y) \mapsto x + y, \]
\[ \cdot : \mathbb{N}_0 \times \mathbb{N}_0 \to \mathbb{N}_0, \quad (x, y) \mapsto x \cdot y, \]
as defined in Def. D.1(a) and Def. D.1(b), respectively, satisfies Def. C.1(i),(ii),(iv) for both addition and multiplication, i.e. associativity, commutativity, and the existence of a neutral element. This can be summarized as the statement that $\mathbb{N}_0$ forms a commutative semigroup with respect to both addition and multiplication (however, no group, as the existence of inverse elements is lacking). Moreover, distributivity, i.e. Def. C.7(iii) is also satisfied.

**Proof.**

**Associativity of Addition:** We have to show
\[ \forall k, m, n \in \mathbb{N}_0 \quad (k + m) + n = k + (m + n). \quad \text{(D.3a)} \]
The proof of (D.3a) is carried out by induction on $n$. The base case ($n = 0$) follows from the first definition in (D.1): $(k + m) + 0 = k + m = k + (m + 0)$ for every $k, m \in \mathbb{N}_0$. For the induction step, one computes, for every $k, m, n \in \mathbb{N}_0$,
\[ (k + m) + (n + 1) \overset{\text{D.1}}{=} (k + m) + S(n) \overset{\text{D.1}}{=} S((k + m) + n) \overset{\text{ind. hyp.}}{=} S(k + (m + n)) \]
\[ \overset{\text{D.1}}{=} k + S(m + n) \overset{\text{D.1}}{=} k + (m + S(n)) \]
\[ \overset{\text{D.1}}{=} k + (m + (n + 1)), \quad \text{(D.3b)} \]
completing the induction.

**Neutral Element of Addition:** That 0 is the neutral element of addition is immediate from (D.1).

**Commutativity of Addition:** We have to show
\[ \forall m, n \in \mathbb{N}_0 \quad m + n = n + m. \quad \text{(D.4a)} \]
The proof of (D.4a) is also carried out by induction on $n$. More precisely, we prove $n = 0$ separately, and then carry out the induction for $n \in \mathbb{N}$. The case $n = 0$ is proved by induction on $m$: The base case ($m = 0$) is the true statement $0 + 0 = 0 + 0$. For the induction step, one computes $(m + 1) + 0 = m + 1 = S(m) = S(m + 0) = S(0 + m) = 0 + S(m) = 0 + (m + 1)$. The base case for the induction on $n$, i.e. $n = 1$ is also proved by induction on $m$: The base case ($m = 0$) is the true statement $0 + 1 = S(0) = 1 = 1 + 0$. For the induction step, one computes, for every $m \in \mathbb{N}_0$,
\[ (m + 1) + 1 \overset{\text{D.1}}{=} S(m + 1) \overset{\text{ind. hyp.}}{=} S(1 + m) \overset{\text{D.1}}{=} (1 + m) + 1 \overset{\text{D.3a}}{=} 1 + (m + 1). \quad \text{(D.4b)} \]
Now, for the induction step of the induction on \(n\), one computes, for every \((m,n) \in \mathbb{N}_0 \times \mathbb{N}\),

\[
m + (n + 1) \overset{(D.1)}{=} m + S(n) \overset{(D.1)}{=} S(m + n) \overset{\text{ind. hyp.}}{=} S(n + m) \overset{(D.1)}{=} n + S(m) \overset{(D.1)}{=} n + (m + 1) \overset{\text{base case}}{=} n + (1 + m) \overset{(D.3a)}{=} (n + 1) + m, \tag{D.4c}
\]

completing the induction.

Neutral Element of Multiplication: That 1 is the neutral element of addition is immediate from (D.2).

Commutativity of Multiplication: We have to show

\[
\forall m,n \in \mathbb{N}_0 \quad m \cdot n = n \cdot m. \tag{D.5a}
\]

We start with some preparatory steps: We first show

\[
\forall m \in \mathbb{N}_0 \quad m = m \cdot 1 = 1 \cdot m. \tag{D.5b}
\]

We have \(m \cdot 1 = m\) for each \(m \in \mathbb{N}_0\) directly from (D.2). We prove \(1 \cdot m = m\) for each \(m \in \mathbb{N}_0\) via induction on \(m\): 1 \(\cdot 0 = 0\) and \(1 \cdot 1 = 1\) are immediate from (D.2). For the induction step, one computes, for every \(m \in \mathbb{N}\),

\[
1 \cdot (m + 1) \overset{(D.2)}{=} 1 \cdot m + 1 \overset{\text{ind. hyp.}}{=} m + 1.
\]

Next, we show

\[
\forall m,n \in \mathbb{N}_0 \quad n \cdot m + m = (n + 1) \cdot m \tag{D.5c}
\]

via induction on \(m\). For the base case \((m = 0)\), we note \((n + 1) \cdot 0 = 0\) by (D.2), and \(n \cdot 0 + 0 = 0\) by (D.1) and (D.2). For the induction step, we compute

\[
n \cdot (m + 1) + m + 1 \overset{(D.2)}{=} n \cdot m + n + m + 1 \overset{(D.4a)}{=} n \cdot m + m + n + 1 \overset{\text{ind. hyp.}}{=} (n + 1) \cdot m + n + 1 \overset{(D.2)}{=} (n + 1) \cdot (m + 1).
\]

We are now in a position to carry out the proof of (D.5a) by induction on \(n\). More precisely, we prove \(n = 0\) separately, and then carry out the induction for \(n \in \mathbb{N}\). Let \(n = 0\). We have \(m \cdot 0 = 0\) for each \(m \in \mathbb{N}_0\) directly from (D.2). We prove \(0 \cdot m = 0\) for each \(m \in \mathbb{N}_0\) via induction on \(m\): \(0 \cdot 0 = 0\) and \(0 \cdot 1 = 0\) are immediate from (D.2). For the induction step, one computes, for every \(m \in \mathbb{N}\),

\[
0 \cdot (m + 1) \overset{(D.2)}{=} 0 \cdot m + 0 \overset{\text{ind. hyp., (D.1)}}{=} 0.
\]

The base case for the induction on \(n \in \mathbb{N}\) is provided by (D.5b). For the induction step, one computes, for every \((m,n) \in \mathbb{N}_0 \times \mathbb{N}\),

\[
m \cdot (n + 1) \overset{(D.2)}{=} m \cdot n + m \overset{\text{ind. hyp.}}{=} n \cdot m + m \overset{(D.5c)}{=} (n + 1) \cdot m,
\]
D. CONSTRUCTION OF THE REAL NUMBERS

completing the proof of (D.5a).

Distributivity: As we have commutativity of multiplication, we only need to show
\[
∀ k, m, n ∈ \mathbb{N}_0 \quad (k + m) · n = k · n + m · n.
\] (D.6a)

The proof of (D.6a) is carried out by induction on \( n \). The base case \( (n = 0) \) follows from (D.2): \((k + m) · 0 = 0 = k · 0 + m · 0\). For the induction step, one computes, for every \( k, m, n ∈ \mathbb{N}_0 \),
\[
(k + m) · (n + 1) \quad \overset{(D.2)}=\quad (k + m) · n + k + m
\]
\[
\overset{\text{ind. hyp.}}= k · n + m · n + k + m
\]
\[
\overset{(D.4a)}= k · n + k · n + m + m
\]
\[
\overset{(D.2)}= k · (n + 1) + m · (n + 1),
\] (D.6b)

completing the induction.

Associativity of Multiplication: We have to show
\[
∀ k, m, n ∈ \mathbb{N}_0 \quad (k · m) · n = k · (m · n).
\] (D.7a)

The proof of (D.7a) is carried out by induction on \( n \). The base case \( (n = 0) \) follows from (D.2): \((k · m) · 0 = 0 = k · (m · 0)\) for every \( k, m ∈ \mathbb{N}_0 \). For the induction step, one computes, for every \( k, m, n ∈ \mathbb{N}_0 \),
\[
(k · m) · (n + 1) \quad \overset{(D.2)}=\quad (k · m) · n + k · m
\]
\[
\overset{\text{ind. hyp.}}= k · (m · n) + k · m
\]
\[
\overset{(D.6a)}= k · (m · n + m)
\]
\[
\overset{(D.2)}= k · (m · (n + 1)),
\] (D.7b)

completing the induction. ■

Lemma D.3. We have
\[
∀ m ∈ \mathbb{N}_0 \quad ∀ n ∈ \mathbb{N} \quad m + n \neq m.
\] (D.8)

Proof. We fix \( n ∈ \mathbb{N} \) and conduct an induction over \( m ∈ \mathbb{N}_0 \). The base case \( (m = 0) \) is clear, since \( 0 + n = n \neq 0 \). The induction step is also clear, since \( m + n \neq m \) implies \( m + 1 + n = m + n + 1 = S(m + n) \neq S(m) = m + 1 \), where we used that \( S \) is injective by Peano axiom P2. ■

Next, one defines an order \( \leq \) on \( \mathbb{N}_0 \):

Definition D.4. For each \( n, m ∈ \mathbb{N}_0 \), let
\[
n \leq m \quad :⇔ \quad ∃ k ∈ \mathbb{N}_0 \quad n + k = m.
\] (D.9)

Theorem D.5. The relation defined in (D.9) constitutes a well-order on \( \mathbb{N}_0 \) (in particular, a total order) that is compatible with addition and multiplication, i.e. it satisfies (4.3).
Proof. \( \leq \) is Reflexive: For each \( n \in \mathbb{N}_0 \), we have \( n + 0 = n \), showing \( n \leq n \).

\( \leq \) is Antisymmetric: If \( m \leq n \) and \( n \leq m \), then \( m + k = n \) and \( n + l = m \). Thus, \( n = m + k = n + l + k \). Thus, \( l + k = 0 \) by Lem. D.3, implying \( l = k = 0 \) (since \( 0 \not\in S(\mathbb{N}_0) \)) and \( m = n \).

\( \leq \) is Transitive: If \( n \leq m \) and \( m \leq l \), then there are \( k_n, k_m \in \mathbb{N}_0 \) such that \( n + k_n = m \) and \( m + k_m = l \). Then \( n + k_n + k_m = m + k_m = l \), showing \( m \leq l \).

\( \leq \) is Well-Order: We first show that, for each \( n \), there is a finite \( k_n, k_m \in \mathbb{N}_0 \) such that \( n + k_n = m \) and \( m + k_m = l \). Then \( n + k_n + k_m = m + k_m = l \), showing \( m \leq l \).

For the induction step, let \( m \in \mathbb{N}_0 \). If \( m \leq n \), then \( m + k = n \). If \( k = 0 \), then \( n \leq n + 1 = m + 1 \). If \( k \neq 0 \), then \( k = S(l) \), i.e. \( n = m + l + 1 = m + 1 + l \), showing \( m + 1 \leq n \). If \( n \leq m + 1 \), then \( n \leq m + 1 \) is immediate, completing the induction.

\( \leq \) is Total Order: We first show that, for each \( n \in \mathbb{N}_0 \), the set \( A_n := \{ m \in \mathbb{N}_0 : m \leq n \} \) is finite: Indeed, this follows by an induction over \( n \) if we can show \( A_{n+1} = A_n \cup \{n+1\} \). Indeed, if \( m \leq n \), then \( m \leq n \leq n + 1 \), i.e. \( m \leq n + 1 \) by transitivity, showing \( A_n \cup \{n+1\} \leq A_{n+1} \). If \( m \leq n + 1 \) and \( m \leq n \), then \( n < m \), i.e. \( n + k = m \) with \( k \neq 0 \), i.e. \( n + l = m \), showing \( n + 1 \leq m \), i.e. \( m = n + 1 \) and \( A_{n+1} \subseteq A_n \cup \{n+1\} \). One now finishes the proof that \( \leq \) is a well-order as in the proof of Th. 3.13(b): Let \( \emptyset \neq A \subseteq \mathbb{N}_0 \). We have to show \( A \) has a min. If \( A \) is finite, then \( A \) has a min by Th. 3.13(a). If \( A \) is infinite, let \( n \) be an element from \( A \). Then the finite set \( B := \{ k \in A : k \leq n \} = A_n \cap B \) must have a min \( m \) by Th. 3.13(a). Since \( m \leq x \) for each \( x \in B \) and \( m \leq n < x \) for each \( x \in A \setminus B \), we have \( m = \min A \).

Compatibility with Addition: We have to show
\[
\forall_{k,m,n \in \mathbb{N}_0} (k \leq m \implies k + n \leq m + n).
\]

To this end, note that \( k \leq m \) means that there is \( l \in \mathbb{N}_0 \) such that \( k + l = m \). But then \( k + n + l = k + l + n = m + n \), showing \( k + n \leq m + n \).

Compatibility with Multiplication: As \( 0 \leq n \) holds for every \( n \in \mathbb{N}_0 \), there is nothing to prove.

Lemma D.6. Let \( a, n \in \mathbb{N} \). Then the following holds:

(a) \( a + n \in \mathbb{N} \).

(b) \( a \cdot n \in \mathbb{N} \).

Proof. (a): Since \( a + n = 0 + (a + n) \), \( a + n \neq 0 \) is due to Lem. D.3.

(b): We conduct the proof via induction on \( n \): For \( n = 1 \), \( a \cdot 1 = a \in \mathbb{N} \) by hypothesis. For \( n \geq 1 \), \( a \cdot (n + 1) = a \cdot n + a \cdot 1 = a \cdot n + a \). Since \( a \cdot n \in \mathbb{N} \) by induction hypothesis, we have \( a \cdot n + a \in \mathbb{N} \) by (a).

Theorem D.7. (a) Monotonicity: Addition and multiplication on \( \mathbb{N}_0 \) are strictly increasing in the sense that, for each \( a, b \in \mathbb{N}_0 \) and each \( n \in \mathbb{N} \),

\[
\begin{align*}
a < b & \implies a + n < b + n \quad (D.11a) \\
(1) a < b & \implies a \cdot n < b \cdot n. \quad (D.11b)
\end{align*}
\]
Cancellation Laws: Given $a, b \in \mathbb{N}_0$ and $n \in \mathbb{N}$, addition and multiplication on $\mathbb{N}_0$ satisfy
\begin{align*}
a + n = b + n & \implies a = b \quad \text{(D.12a)} \\
a \cdot n = b \cdot n & \implies a = b. \quad \text{(D.12b)}
\end{align*}

Proof. Let $a, b \in \mathbb{N}_0$ with $a < b$. Then there exists $k \in \mathbb{N}$ with $a + k = b$. Let $n \in \mathbb{N}$.

(a): We have $b + n = a + k + n$. Since $k + n > 0$ by Lem. D.6(a), this shows $a + n < b + n$. We have $b \cdot n = (a + k) \cdot n = a \cdot n + k \cdot n$. Since $k \cdot n > 0$ by Lem. D.6(b), this shows $a \cdot n < b \cdot n$.

(b): Arguing via contraposition, both laws are immediate from (a).

\section*{D.2 Interlude: Orders on Groups}

In the succeeding sections, we will construct the set of integers $\mathbb{Z}$, the set of rational numbers $\mathbb{Q}$, and the set of real numbers $\mathbb{R}$. In each case, we will use the same method to define a total order on the constructed set, making use of the algebraic structure of its additive group. It is therefore economical as well as mathematically interesting, to study this construction once in its abstract form, which is the purpose of the present section.

Recall the definition of a group from Def. C.1.

\textbf{Theorem D.8.} Assume $(G, +)$ to be a group. Moreover, assume we have a disjoint decomposition
\begin{equation}
G = P \cup \{0\} \cup (-P), \quad -P := \{x \in G : -x \in P\}, \tag{D.13}
\end{equation}
where $-x$ denotes the inverse of $x$ with respect to $+$. If $P$ is closed under $+$ (i.e. $x, y \in P$ implies $x + y \in P$), then
\begin{equation}
y \leq x \iff x - y \in P \cup \{0\} \tag{D.14}
\end{equation}
defines a total order on $G$ that is compatible with addition, i.e. it satisfies (4.3a). Moreover, if a multiplication is also defined on $G$ and $P \cup \{0\}$ is closed under this multiplication, then $\leq$ is also compatible with multiplication, i.e. it satisfies (4.3b). Of course, one refers to the elements of $P$ as positive and to the elements of $-P$ as negative.

Proof. For each $x \in G$, one has $x-x = 0 \in P \cup \{0\}$, i.e. $x \leq x$ and the relation is reflexive. If $x, y \in G$, $x \leq y$ and $y \leq x$, then $x - y \in P \cup \{0\}$ and $-(x - y) = y - x \in P \cup \{0\}$, and the disjointness of the union in (D.13) implies $x - y = 0$, i.e. $x = y$, showing the relation is antisymmetric. If $x, y, z \in G$ with $x \leq y$ and $y \leq z$, then $y - x \in P \cup \{0\}$, $z - y \in P \cup \{0\}$, and $z - x = z - y + y - x \in P \cup \{0\}$ since $P$ is closed under $+$, showing the relation is transitive. So we have shown $\leq$ constitutes a partial order on $G$. It remains to show the order is total. However, given the decomposition in (D.13), for each $x, y \in G$, precisely one of the statements $x - y \in P$ (i.e. $y < x$), $x - y = 0$ (i.e. $x = y$), $x - y \in -P$ (i.e. $x < y$) must be true, proving that the order is total. To see $\leq$ satisfies (4.3a), let
D  CONSTRUCTION OF THE REAL NUMBERS 219

...x, y, z ∈ G. If x ≤ y, then y − x ∈ P ∪ {0}, i.e. y + z − (x + z) = y + z − x ∈ P ∪ {0}, showing x + z ≤ y + z. The proof is completed by noting (4.3b) is precisely the statement that P ∪ {0} is closed under multiplication.

D.3  Integers

As compared to our goal, the set of real numbers R, the set N₀ still has three deficiencies, namely the lack of inverse elements for addition, the lack of inverse elements for multiplication, and that the order ≤ lacks completeness. The construction of the integers will remedy (only) the first of the three deficiencies by providing the inverse elements of addition.

Definition and Remark D.9. The relation ∼ on N₀ × N₀ defined by

\[(a, b) \sim (c, d) :\iff a + d = b + c, \quad \text{(D.15)}\]

constitutes an equivalence relation on N₀ × N₀ (cf. Def. 2.23): Indeed, if a, b ∈ N₀, then \(a + b = b + a\) shows \((a, b) \sim (a, b)\), proving ∼ to be reflexive. If a, b, c, d ∈ N₀, then

\[(a, b) \sim (c, d) \Rightarrow a + d = b + c \Rightarrow c + b = d + a \Rightarrow (c, d) \sim (a, b),\]

proving ∼ to be symmetric. If a, b, c, d, e, f ∈ N₀, then

\[(a, b) \sim (c, d) \land (c, d) \sim (e, f) \Rightarrow a + d = b + c \land c + f = d + e \Rightarrow a + c + f = b + d + e \Rightarrow (a, b) \sim (e, f),\]

proving ∼ to be transitive and an equivalence relation.

Definition D.10. (a) Define the set of integers \(\mathbb{Z}\) as the set of equivalence classes of the equivalence relation ∼ defined in (D.15), i.e.

\[\mathbb{Z} := (\mathbb{N}_0 \times \mathbb{N}_0)/\sim = \{(a, b) \in \mathbb{N}_0 \times \mathbb{N}_0 : ([a, b])\} \quad \text{(D.16)}\]

is the quotient set of \(\mathbb{N}_0 \times \mathbb{N}_0\) with respect to ∼ (cf. Ex. 2.24(c)). To simplify notation, in the following, we will write

\[[a, b] := [(a, b)] \quad \text{(D.17)}\]

for the equivalence class of \((a, b)\) with respect to ∼.

(b) Addition on \(\mathbb{Z}\) is defined by

\[+: \mathbb{Z} \times \mathbb{Z} \rightarrow \mathbb{Z}, \quad ([a, b], [c, d]) \mapsto [a, b] + [c, d] := [a + c, b + d]. \quad \text{(D.18)}\]

Subtraction on \(\mathbb{Z}\) is defined by

\[-: \mathbb{Z} \times \mathbb{Z} \rightarrow \mathbb{Z}, \quad ([a, b], [c, d]) \mapsto [a, b] - [c, d] := [a, b] + [d, c]. \quad \text{(D.19)}\]
For the definitions in Def. D.10(b) to make sense, one needs to check that they do not depend on the chosen representatives of the equivalence classes. Moreover, one needs to convince oneself that these definitions yield the desired familiar operations of addition and subtraction. Let us start by verifying the independence of the representatives is the following Lem. D.11.

**Lemma D.11.** The definitions in Def. D.10(b) do not depend on the chosen representatives, i.e.

\[ \forall_{a,b,c,d,\tilde{a},\tilde{b},\tilde{c},\tilde{d}\in\mathbb{N}_0} \left( [a,b] = [\tilde{a},\tilde{b}] \wedge [c,d] = [\tilde{c},\tilde{d}] \Rightarrow [a+c,b+d] = [\tilde{a}+\tilde{c},\tilde{b}+\tilde{d}] \right) \] (D.20)

and

\[ \forall_{a,b,c,d,\tilde{a},\tilde{b},\tilde{c},\tilde{d}\in\mathbb{N}_0} \left( [a,b] = [\tilde{a},\tilde{b}] \wedge [c,d] = [\tilde{c},\tilde{d}] \Rightarrow [a,b]-[c,d] = [\tilde{a},\tilde{b}]-[\tilde{c},\tilde{d}] \right) \] (D.21)

**Proof.** (D.20): \([a,b] = [\tilde{a},\tilde{b}]\) means \(a + \tilde{b} = b + \tilde{a}\), \([c,d] = [\tilde{c},\tilde{d}]\) means \(c + \tilde{d} = d + \tilde{c}\), implying \(a + c + \tilde{b} + \tilde{d} = b + \tilde{a} + d + \tilde{c}\), i.e. \(a + c + b + d = [\tilde{a} + \tilde{c}, \tilde{b} + \tilde{d}]\).

(D.21) is just (D.19) combined with (D.20). □

**Theorem D.12.** The set of integers \(\mathbb{Z}\) forms a commutative group with respect to addition as defined in Def. D.10(b), where \([0,0]\) is the neutral element, \([b,a]\) is the inverse element of \([a,b]\) for each \(a,b \in \mathbb{N}_0\), and, denoting the inverse element of \([a,b]\) by \(-[a,b]\) in the usual way, \([a,b] - [c,d] = [a,b] + (-[c,d])\) for each \(a,b,c,d \in \mathbb{N}_0\).

**Proof.** To verify commutativity and associativity of addition on \(\mathbb{Z}\), let \(a,b,c,d,e,f \in \mathbb{N}_0\). Then

\[ [a,b] + [c,d] = [a + c, b + d] = [c + a, b + d] = [c,d] + [a,b], \]

proving commutativity, and

\[ [a,b] + ([c,d] + [e,f]) = [a,b] + [c + e, d + f] = [a + (c + e), b + (d + f)] \]
\[ = [(a + c) + e, (b + d) + f] = [a + c, b + d] + [e,f] \]
\[ = ([a,b] + [c,d]) + [e,f], \]

proving associativity. For every \(a,b \in \mathbb{N}_0\), one obtains \([a,b] + [0,0] = [a+0, b+0] = [a,b]\), proving neutrality of \([0,0]\), whereas \([a,b] + [b,a] = [a + b, b + a] = [a + b, a + b] = [0,0]\) (since \((a + b, a + b) \sim (0,0)) shows \([b,a] = -[a,b]\). Now \([a,b] - [c,d] = [a,b] + (-[c,d])\) is immediate from (D.19). □

**Remark D.13.** The map

\[ \iota : \mathbb{N}_0 \longrightarrow \mathbb{Z}, \quad \iota(n) := [n,0], \] (D.22)

is a monomorphism, i.e. it is injective (since \(\iota(m) = [m,0] = \iota(n) = [n,0]\) implies \(m + 0 = 0 + n\), i.e. \(m = n\)) and satisfies

\[ \forall_{m,n \in \mathbb{N}_0} \iota(m+n) = [m+n,0] = [m,0] + [n,0] = \iota(m) + \iota(n). \] (D.23)

It is customary to identify \(\mathbb{N}_0\) with \(\iota(\mathbb{N}_0)\), as it usually does not cause any confusion. One then just writes \(n\) instead of \([n,0]\) and \(-n\) instead of \([0,n]\) = \([-n,0]\).
Lemma D.14. We have the disjoint decomposition

\[ Z = \mathbb{N} \cup \{0\} \cup \mathbb{Z}^-, \quad \mathbb{Z}^- := -\mathbb{N} = \{n \in \mathbb{Z} : -n \in \mathbb{N}\}. \tag{D.24} \]

**Proof.** Note that, due to (D.15), an equivalence class remains the same if a natural number is added or subtracted in both components: \([a, b] = [a + m, b + m]\). Thus, for each \(x = [a, b] \in \mathbb{Z}\), if \(a > b\), then \(x = [a - b, 0] \in \mathbb{N}\); if \(a = b\), then \(x = [0, 0] = 0\); if \(a < b\), then \(x = [0, b - a] = -[b - a, 0] \in \mathbb{Z}^-\). It just remains to verify that the union in (D.24) is disjoint. However, if \([n, 0] = [0, m]\) with \(m, n \in \mathbb{N}_0\), then \(n + m = 0\), proving \(n = m = 0\), completing the proof. \(\blacksquare\)

**Remark D.15.** In the above construction, we obtained the commutative group \((\mathbb{Z}, +)\) from the commutative semigroup \((\mathbb{N}_0, +)\). It is worth pointing out that the same construction always works when, instead of with \(\mathbb{N}_0\), one starts with any commutative semigroup \((H, +)\) that satisfies the cancellation law \(a + c = b + c \Rightarrow a = b\), to obtain a commutative group \((G, +)\) and a monomorphism \(i : H \rightarrow G\).

---

To obtain the expected laws of arithmetic, multiplication on \(\mathbb{Z}\) needs to be defined such that \((a - b) \cdot (c - d) = (ac + bd) - (ad + bc)\), which leads to the following definition.

**Definition D.16.** Multiplication on \(\mathbb{Z}\) is defined by

\[ \cdot : \mathbb{Z} \times \mathbb{Z} \rightarrow \mathbb{Z}, \quad ([a, b], [c, d]) \mapsto [a, b] \cdot [c, d] := [ac + bd, ad + bc]. \tag{D.25} \]

**Lemma D.17.** The definition in Def. D.16 does not depend on the chosen representatives, i.e.

\[ \forall_{a, b, c, d, \tilde{a}, \tilde{b}, \tilde{c}, \tilde{d} \in \mathbb{N}_0} (\tilde{a}, \tilde{b}, \tilde{c}, \tilde{d} \in [a, b] \land [c, d] = [\tilde{c}, \tilde{d}] \Rightarrow [ac + bd, ad + bc] = [\tilde{a}c + \tilde{b}d, \tilde{a}d + \tilde{b}c]). \tag{D.26} \]

**Proof.** As mentioned before, due to (D.15), an equivalence class remains the same if a natural number is added or subtracted in both components. Thus, one computes

\[
\begin{align*}
[ac + bd, ad + bc] \overset{(D.15)}{=} \quad & [ac + bd + \tilde{b}c, ad + bc + \tilde{bc}] = [(a + \tilde{b})c + bd, ad + bc + \tilde{bc}] \\
& = [(a + b)c + bd, ad + bc + \tilde{bc}] \quad \overset{(D.15)}{=} \quad [\tilde{a}d + \tilde{a}c + bd, \tilde{a}d + ad + \tilde{bc}] \\
& = [\tilde{a}(d + c) + bd, \tilde{a}d + ad + \tilde{bc}] = [\tilde{a}(d + \tilde{c}) + bd, \tilde{a}d + ad + \tilde{bc}] \\
& = [\tilde{a}c + (a + b)d, \tilde{a}d + ad + \tilde{bc}] = [\tilde{a}c + (a + \tilde{b})d, \tilde{a}d + ad + \tilde{bc}] \\
& = [\tilde{a}c + \tilde{b}(d + c), \tilde{a}d + bc + \tilde{bc}] = [\tilde{a}c + \tilde{b}(d + \tilde{c}), \tilde{a}d + bc + \tilde{bc}] \\
& = [\tilde{a}c + \tilde{b}(\tilde{d} + c), \tilde{a}d + \tilde{bc} + \tilde{bc}] = [\tilde{a}c + \tilde{b}d, \tilde{a}d + \tilde{bc}], \quad \tag{D.27}
\end{align*}
\]

completing the proof. \(\blacksquare\)
Theorem D.18. The set of integers \( \mathbb{Z} \) is associative and commutative with respect to
the multiplication defined in Def. D.16. Moreover, distributivity, i.e. Def. C.7(iii) is
satisfied, \([1, 0]\) is the neutral element of multiplication, and there are no zero divisors,
i.e.

\[
\forall_{a,b,c,d \in \mathbb{N}_0} \left( [a, b] \cdot [c, d] = [ac + bd, ad + bc] = [0, 0] \Rightarrow [a, b] = [0, 0] \lor [c, d] = [0, 0] \right).
\]

(D.28)

Algebraically, the theorem can be summarized by saying that \((\mathbb{Z}, +, \cdot)\) constitutes a principal ideal domain.

Proof. Let \(a, b, c, d, e, f \in \mathbb{N}_0\). Then, using commutativity of addition and multiplication
on \(\mathbb{N}_0\),

\[
[a, b] \cdot [c, d] = [ac + bd, ad + bc] = [ca + db, cb + da] = [c, d] \cdot [a, b],
\]

proving commutativity of multiplication on \(\mathbb{Z}\); using distributivity and commutativity
of addition on \(\mathbb{N}_0\),

\[
[a, b] \cdot ([c, d] \cdot [e, f]) = [a, b] \cdot [ce + df, cf + de] = [ace + adf + bce + bdf, acf + ade + bce + bdf] = [ace + bde + adf + bc, acf + bdf + ade + bce] = [ac + bd, ad + bc] \cdot [e, f] = ([a, b] \cdot [c, d]) \cdot [e, f],
\]

proving associativity of multiplication on \(\mathbb{Z}\); again using distributivity and commutativity
of addition on \(\mathbb{N}_0\),

\[
[a, b] \cdot ([c, d] + [e, f]) = [a, b] \cdot [c + e, d + f] = [ac + ae + bd + bf, ad + af + bc + be] = [ac + bd + ae + bf, ad + bc + af + be] = [ac + bd, ad + bc] + [ae + bf, af + be] = [a, b] \cdot [c, d] + [a, b] \cdot [e, f],
\]

proving distributivity on \(\mathbb{Z}\). Next, \([a, b] \cdot [1, 0] = [a \cdot 1 + b \cdot 0, a \cdot 0 + b \cdot 1] = [a, b] \) proves neutrality of \([1, 0]\). It remains to prove (D.28). Note that, due to (D.15), the conclusion
is equivalent to \(a = b\) or \(c = d\). We assume \(0 \leq a < b\) and have to prove \(c = d\). According
to Def. D.5, \(a < b\) means \(b = a + k\) for some \(k \in \mathbb{N}\). Thus, \([ac + bd, ad + bc] = [0, 0]\)
implies

\[
ac + (a + k)d = ac + bd = ad + bc = ad + (a + k)c \Rightarrow kd = kc \quad \text{(D.12b)} \quad c = d, \quad \text{(D.29)}
\]

establishing the case.

Definition D.19. For each \(k, l \in \mathbb{Z}\), let

\[
l \leq k \quad \iff \quad k - l \in \mathbb{N}_0.
\]

(D.30)

Theorem D.20. (a) The relation defined in (D.30) constitutes a total order on \(\mathbb{Z}\) that
is compatible with addition and multiplication, i.e. it satisfies (4.3).
(b) The map $\iota$ from (D.22) is strictly increasing.

**Proof.** (a) follows from (D.30), (D.24), and Th. D.8 since $\mathbb{N}_0$ is closed under addition and multiplication.

(b): According to Def. D.5, if $m,n \in \mathbb{N}$ with $n < m$, then $m = n + k$ for some $k \in \mathbb{N}$. In consequence $\iota(m) = \iota(n) + \iota(k)$ by (D.23), i.e. $\iota(m) - \iota(n) = \iota(k) \in \mathbb{N}$, proving $\iota(n) < \iota(m)$. ■

### D.4 Rational Numbers

The remaining two deficiencies of the set of integers $\mathbb{Z}$ (as compared with $\mathbb{R}$) are the lack of inverse elements for multiplication and that the order $\leq$ lacks completeness.

We proceed to the construction of the rational numbers, which will provide the inverse elements for multiplication. The completion of the order will then be achieved in the last step in the next section.

**Definition and Remark D.21.** The relation $\sim$ on $\mathbb{Z} \times (\mathbb{Z} \setminus \{0\})$ defined by

$$ (a,b) \sim (c,d) \quad :\Leftrightarrow \quad ad = bc, \quad (D.31) $$

constitutes an equivalence relation on $\mathbb{Z} \times (\mathbb{Z} \setminus \{0\})$ (cf. Def. 2.23): Indeed, noting that (D.31) is precisely the same as (D.15) if + is replaced by $\cdot$, the proof from Def. and Rem. D.9 also shows that (D.31) does, indeed, define an equivalence relation on $\mathbb{Z} \times (\mathbb{Z} \setminus \{0\})$: One merely replaces each + with $\cdot$ and each $\mathbb{N}_0$ with $\mathbb{Z}$ or $\mathbb{Z} \setminus \{0\}$, respectively. The only modification needed occurs for $0 \in \{a,c,e\}$ in the proof of transitivity (in this case, the proof of Def. and Rem. D.9 yields $adf = 0 = bcde$, which does not imply $af = be$), where one now argues, for $a = 0$,

$$ (a,b) \sim (c,d) \land (c,d) \sim (e,f) \quad \Rightarrow \quad ad = 0 = bc \land cf = de $$

$\Rightarrow c = 0 \quad d \not\Rightarrow e = 0 \quad \Rightarrow \quad af = 0 = be \quad \Rightarrow \quad (a,b) \sim (e,f),$

for $c = 0$,

$$ (a,b) \sim (c,d) \land (c,d) \sim (e,f) \quad \Rightarrow \quad ad = 0 = bc \land cf = 0 = de $$

$\not\Rightarrow a = e = 0 \quad af = 0 = be \quad \Rightarrow \quad (a,b) \sim (e,f),$

and, for $e = 0$,

$$ (a,b) \sim (c,d) \land (c,d) \sim (e,f) \quad \Rightarrow \quad ad = bc \land cf = 0 = de $$

$\not\Rightarrow c = 0 \quad d \not\Rightarrow a = 0 \quad \Rightarrow \quad af = 0 = be \quad \Rightarrow \quad (a,b) \sim (e,f).$

**Definition D.22.** (a) Define the set of rational numbers $\mathbb{Q}$ as the set of equivalence classes of the equivalence relation $\sim$ defined in (D.31), i.e.

$$ \mathbb{Q} := (\mathbb{Z} \times (\mathbb{Z} \setminus \{0\}))/\sim = \{[(a,b)] : (a,b) \in \mathbb{Z} \times (\mathbb{Z} \setminus \{0\})\} \quad (D.32) $$
is the quotient set of $\mathbb{Z} \times (\mathbb{Z} \setminus \{0\})$ with respect to $\sim$ (cf. Ex. 2.24(c)). As is common, we will write

$$\frac{a}{b} := a/b := [(a, b)] \quad (D.33)$$

for the equivalence class of $(a, b)$ with respect to $\sim$.

(b) Addition on $\mathbb{Q}$ is defined by

$$+: \mathbb{Q} \times \mathbb{Q} \to \mathbb{Q}, \quad \left(\frac{a}{b}, \frac{c}{d}\right) \mapsto \frac{a}{b} + \frac{c}{d} := \frac{ad + bc}{bd}. \quad (D.34)$$

Multiplication on $\mathbb{Q}$ is defined by

$$\cdot: \mathbb{Q} \times \mathbb{Q} \to \mathbb{Q}, \quad \left(\frac{a}{b}, \frac{c}{d}\right) \mapsto \frac{a}{b} \cdot \frac{c}{d} := \frac{ac}{bd}. \quad (D.35)$$

For the definitions in Def. D.22(b) to make sense, one needs to check that they do not depend on the chosen representatives of the equivalence classes, and that the results of both addition and multiplication are always elements of $\mathbb{Q}$. All this is provided by the following lemma.

**Lemma D.23.** The definitions in Def. D.22(b) do not depend on the chosen representatives, i.e.

$$\forall_{a,c,\bar{a},\bar{c} \in \mathbb{Z}} \forall_{b,d,\bar{b},\bar{d} \in \mathbb{Z} \setminus \{0\}} \left(\frac{a}{b} = \frac{\bar{a}}{\bar{b}} \land \frac{c}{d} = \frac{\bar{c}}{\bar{d}} \Rightarrow \frac{ad + bc}{bd} = \frac{\bar{a}\bar{d} + \bar{b}\bar{c}}{\bar{b}\bar{d}}\right) \quad (D.36)$$

and

$$\forall_{a,c,\bar{a},\bar{c} \in \mathbb{Z}} \forall_{b,d,\bar{b},\bar{d} \in \mathbb{Z} \setminus \{0\}} \left(\frac{a}{b} = \frac{\bar{a}}{\bar{b}} \land \frac{c}{d} = \frac{\bar{c}}{\bar{d}} \Rightarrow \frac{ac}{bd} = \frac{\bar{a}\bar{c}}{\bar{b}\bar{d}}\right). \quad (D.37)$$

Furthermore, the results of both addition and multiplication are always elements of $\mathbb{Q}$.

**Proof.** (D.36) and (D.37): $a/b = \bar{a}/\bar{b}$ means $a\bar{b} = \bar{a}b$, $c/d = \bar{c}/\bar{d}$ means $c\bar{d} = \bar{c}d$, implying

$$(ad + bc)\bar{b}\bar{d} = bd(\bar{a}\bar{d} + \bar{b}\bar{c}), \quad \text{i.e.} \quad \frac{ad + bc}{bd} = \frac{\bar{a}\bar{d} + \bar{b}\bar{c}}{\bar{b}\bar{d}} \quad (D.38)$$

and

$$ac\bar{b}\bar{d} = bd\bar{a}\bar{c}, \quad \text{i.e.} \quad \frac{ac}{bd} = \frac{\bar{a}\bar{c}}{\bar{b}\bar{d}}. \quad (D.39)$$

That the results of both addition and multiplication are always elements of $\mathbb{Q}$ follows from (D.28), i.e. from the fact that $\mathbb{Z}$ has no zero divisors. In particular, if $b, d \neq 0$, then $bd \neq 0$, showing $(ad + bc)/(bd) \in \mathbb{Q}$ and $(ac)/(bd) \in \mathbb{Q}$. \[\blacksquare\]

**Theorem D.24.** (a) The set of rational numbers $\mathbb{Q}$ with addition and multiplication as defined in Def. D.22 forms a field, where $0/1$ and $1/1$ are the neutral elements with respect to addition and multiplication, respectively, $(-a/b)$ is the additive inverse to $a/b$, whereas $b/a$ is the multiplicative inverse to $a/b$ with $a \neq 0$.\[\blacksquare\]
(b) Defining subtraction and division in the usual way, for each \( r, s \in \mathbb{Q} \), by \( s - r := s + (-r) \) and \( s/r := sr^{-1} \), respectively, with \(-r\) denoting the additive inverse of \( r \) and \( r^{-1} \) denoting the multiplicative inverse of \( r \neq 0 \), all the rules stated in Th. C.10 are valid in \( \mathbb{Q} \).

(c) The map

\[
i : \mathbb{Z} \rightarrow \mathbb{Q}, \quad i(k) := \frac{k}{1}, \tag{D.40}
\]

is a monomorphism, i.e. it is injective and satisfies

\[
\forall_{k,l \in \mathbb{Z}} \quad i(k + l) = i(k) + i(l), \tag{D.41a}
\]

\[
\forall_{k,l \in \mathbb{Z}} \quad i(kl) = i(k) \cdot i(l). \tag{D.41b}
\]

It is customary to identify \( \mathbb{Z} \) with \( i(\mathbb{Z}) \), as it usually does not cause any confusion. One then just writes \( k \) instead of \( \frac{k}{1} \).

Proof. (a): We verify + and \( \cdot \) to be commutative and associative on \( \mathbb{Q} \): Let \( a, c, e \in \mathbb{Z} \) and \( b, d, f \in \mathbb{Z} \setminus \{0\} \). Then, using commutativity on \( \mathbb{Z} \), we compute

\[
\frac{c}{d} + \frac{a}{b} = \frac{cb + da}{db} = \frac{ad + bc}{bd} = \frac{a}{b} + \frac{c}{d}, \quad \frac{c}{d} \cdot \frac{a}{b} = \frac{ca}{db} = \frac{ac}{bd} = \frac{a}{b} \cdot \frac{c}{d},
\]

showing commutativity on \( \mathbb{Q} \). Using associativity and distributivity on \( \mathbb{Z} \), we compute

\[
\frac{a}{b} + \left( \frac{c}{d} + \frac{e}{f} \right) = \frac{a}{b} + \frac{cf + de}{df} = \frac{adf + b(cf + de)}{bdf} = \frac{(ad + bc)f + bde}{bdf}
\]

\[
= \frac{ad + bc}{bd} + \frac{e}{f} = \left( \frac{a}{b} + \frac{c}{d} \right) + \frac{e}{f},
\]

\[
\frac{a}{b} \cdot \left( \frac{c}{d} \cdot \frac{e}{f} \right) = \frac{a(ce)}{(bd)f} = \frac{(ac)e}{(bd)f} = \left( \frac{a}{b} \cdot \frac{c}{d} \right) \cdot \frac{e}{f},
\]

showing associativity on \( \mathbb{Q} \). We proceed to checking distributivity on \( \mathbb{Q} \): Using commutativity, associativity, and distributivity on \( \mathbb{Z} \), we compute

\[
\frac{a}{b} \cdot \left( \frac{c}{d} + \frac{e}{f} \right) = \frac{a(cf + de)}{bdf} = \frac{acf + dae}{bdf} = \frac{acbf + bdae}{bdf} = \frac{ac}{bd} + \frac{ae}{bf} = \frac{a}{b} \cdot \frac{c}{d} + \frac{a}{b} \cdot \frac{e}{f},
\]

proving distributivity on \( \mathbb{Q} \). We now check the claims regarding neutral and inverse elements:

\[
\frac{a}{b} + 0 = \frac{a \cdot 1 + b \cdot 0}{b \cdot 1} = \frac{a}{b},
\]

\[
\frac{a}{b} + \frac{-a}{b} = \frac{ab + b(-a)}{b^2} \quad \text{Def. C.7(iii) for } \mathbb{Z} \quad (a - a)b = 0 \quad \text{(D.31)} \quad 0 \quad \frac{1}{b^2} = \frac{1}{b^2} \quad \text{for } \mathbb{Z},
\]

\[
\frac{a}{b} \cdot \frac{1}{1} = \frac{a \cdot 1}{b \cdot 1} = \frac{a}{b},
\]

\[
\frac{a}{b} \cdot \frac{b}{a} = \frac{ab}{ba} = \frac{1}{1}.
\]
Thus, \((\mathbb{Q}, +, \cdot)\) is a ring and \((\mathbb{Q} \setminus \{0\}, \cdot)\) is a group, implying \(\mathbb{Q}\) to be a field.

(b) is a consequence of (a), since Th. C.10 and its proof are valid in every field.

(c): The map \(\iota\) is injective, as \(\iota(k) = k/1 = \iota(l) = l/1\) implies \(k \cdot 1 = l \cdot 1\), i.e. \(k = l\). Moreover,

\[
\iota(k) + \iota(l) = \frac{k}{1} + \frac{l}{1} = \frac{k \cdot 1 + l \cdot 1}{1} = \frac{k + l}{1} = \iota(k + l),
\]

(D.42a)

\[
\iota(k) \cdot \iota(l) = \frac{k}{1} \cdot \frac{l}{1} = \frac{kl}{1} = \iota(kl),
\]

(D.42b)

completing the proof.

\[\blacksquare\]

**Definition and Remark D.25.** Define

\[Q^+ := \left\{ r \in \mathbb{Q} : \exists_{a,b \in \mathbb{N}} r = \frac{a}{b} \right\}.\]  

(D.43)

We then have the decomposition

\[\mathbb{Q} = Q^+ \cup \{0\} \cup Q^- , \quad Q^- := -Q^+ = \{ r \in \mathbb{Q} : -r \in Q^+ \},\]  

(D.44)

since

\[
ap/b \in Q^+ \quad \iff \quad ((a > 0 \land b > 0) \lor (a < 0 \land b < 0)) ,
\]

(D.45a)

\[
ap/b = 0 \quad \iff \quad a = 0 ,
\]

(D.45b)

\[
ap/b \in Q^- \quad \iff \quad ((a > 0 \land b < 0) \lor (a < 0 \land b > 0)) .
\]

(D.45c)

**Definition D.26.** For each \(r, s \in \mathbb{Q}\), let

\[s \leq r \quad :\iff \quad r - s \in Q^+_0 := Q^+ \cup \{0\} .\]  

(D.46)

**Theorem D.27.** (a) The relation defined in (D.46) constitutes a total order on \(\mathbb{Q}\) that is compatible with addition and multiplication, i.e. it satisfies (4.3); in other words \((\mathbb{Q}, +, \cdot, \leq)\) constitutes a totally ordered field.

(b) All the rules stated in Th. 4.5 are valid in \(\mathbb{Q}\).

(c) The map \(\iota\) from (D.40) is strictly increasing.

**Proof.** (a) follows from (D.46), (D.44), and Th. D.8, since it is immediate from (D.34) and (D.35) that \(Q^+\) is closed under addition and multiplication.

(b) is a consequence of (a), since Th. 4.5 and its proof are valid in every totally ordered field.

(c): According to Def. D.27, if \(k, l \in \mathbb{Z}\) with \(l < k\), then \(n := k - l \in \mathbb{N}\). In consequence \(\iota(k) = \iota(l) + \iota(n)\) by (D.41a), i.e. \(\iota(k) - \iota(l) = \iota(n) = n/1 \in Q^+\), proving \(\iota(l) < \iota(k)\).  \[\blacksquare\]
D.5 Real Numbers

In the previous section, the construction of the rational numbers $\mathbb{Q}$ yielded a totally ordered field. However, the order on $\mathbb{Q}$ is not complete – for example, Rem. and Def. 7.62 shows that the set $M := \{ r \in \mathbb{Q} : r^2 < 2 \}$, which is bounded from above (for example by 2), has no supremum in $\mathbb{Q}$ (otherwise, we had a rational number $q = \sup M$ with $q^2 = 2$). Finally, in the present section, we will start out from $\mathbb{Q}$ to construct the set of real numbers $\mathbb{R}$ such that it becomes a complete totally ordered field. There are several different important constructions to obtain $\mathbb{R}$ from $\mathbb{Q}$. We will describe the construction that defines real numbers as equivalence classes of rational Cauchy sequences following [EHH+95, Ch. 2.3]. The construction using so-called Dedekind cuts can be found in [EHH+95, Ch. 2.2], the construction via nested intervals in [EHH+95, Ch. 2.4].

**Definition D.28.** (a) Let $\mathcal{S}$ denote the set of all Cauchy sequences in $\mathbb{Q}$, where we call a sequence $(r_n)_{n \in \mathbb{N}}$ in $\mathbb{Q}$ a Cauchy sequence if, and only if,

$$\forall \epsilon \in \mathbb{Q}^+ \exists N \in \mathbb{N} \forall n, m > N |r_n - r_m| < \epsilon,$$

which differs from (7.25) in that $\epsilon$ has to be from $\mathbb{Q}^+$ rather than from $\mathbb{R}^+$.

(b) **Addition** on $\mathcal{S}$ is defined by

$$+: \mathcal{S} \times \mathcal{S} \rightarrow \mathcal{S}, \quad ((r_n)_{n \in \mathbb{N}}, (s_n)_{n \in \mathbb{N}}) \mapsto (r_n)_{n \in \mathbb{N}} + (s_n)_{n \in \mathbb{N}} := (r_n + s_n)_{n \in \mathbb{N}}. \quad (D.48)$$

**Multiplication** on $\mathcal{S}$ is defined by

$$\cdot: \mathcal{S} \times \mathcal{S} \rightarrow \mathcal{S}, \quad ((r_n)_{n \in \mathbb{N}}, (s_n)_{n \in \mathbb{N}}) \mapsto (r_n)_{n \in \mathbb{N}} \cdot (s_n)_{n \in \mathbb{N}} := (r_n s_n)_{n \in \mathbb{N}}. \quad (D.49)$$

As a consequence of the following Lem. D.29, addition and multiplication are well-defined on $\mathcal{S}$.

**Lemma D.29.** If $(r_n)_{n \in \mathbb{N}}$ and $(s_n)_{n \in \mathbb{N}}$ are Cauchy sequences in $\mathbb{Q}$, so are $(r_n + s_n)_{n \in \mathbb{N}}$ and $(r_n s_n)_{n \in \mathbb{N}}$.

**Proof.** The proofs are analogous to the proofs of Th. 7.13(7.11b),(7.11c):

Given $\epsilon \in \mathbb{Q}^+$, there exists $N \in \mathbb{N}$ such that, for each $n, m > N$, $|r_n - r_m| < \epsilon/2$ and $|s_n - s_m| < \epsilon/2$, implying

$$\forall n, m > N |r_n + s_n - (r_m + s_m)| \leq |r_n - r_m| + |s_n - s_m| < \epsilon/2 + \epsilon/2 = \epsilon,$$

proving $(r_n + s_n)_{n \in \mathbb{N}}$ is Cauchy.

The proof of Th. 7.29 shows both $(r_n)_{n \in \mathbb{N}}$ and $(s_n)_{n \in \mathbb{N}}$ are bounded, i.e. there exists $M \in \mathbb{Q}^+$ that is an upper bound for the sets $\{|r_n| : n \in \mathbb{N}\}$ and $\{|s_n| : n \in \mathbb{N}\}$. Moreover,
given \( \epsilon \in \mathbb{Q}^+ \), there exists \( N \in \mathbb{N} \) such that, for each \( n, m > N \), \(|r_n - r_m| < \epsilon/(2M)\) and \(|s_n - s_m| < \epsilon/(2M)\), implying

\[
\forall \ n, m > N \quad \left| r_n s_n - r_m s_m \right| = \left| (r_n - r_m) s_n + r_m(s_n - s_m) \right| \\
\leq |s_n| \cdot |r_n - r_m| + |r_m| \cdot |s_n - s_m| < \frac{M \epsilon}{2M} + \frac{M \epsilon}{2M} = \epsilon,
\]

completing the proof of the lemma.

Theorem D.30. \((S, +)\) is a group and, in addition, \( S \) is associative and commutative with respect to multiplication. Moreover, distributivity also holds in \( S \). In algebraic terms, this can be summarized as the statement that \((S, +, \cdot)\) constitutes a commutative ring.

Proof. Note that, since the rational sequence \((r_n)_{n \in \mathbb{N}}\) is nothing but the function \( f : \mathbb{N} \to \mathbb{Q}, f(n) = r_n \), addition and multiplication as defined in Def. D.28(b) is analogous to the definition of addition and multiplication of real-valued functions in (6.1a), (6.1c), respectively. It is an easy exercise to verify that these function operations always inherit associativity, commutativity, and distributivity if these rules hold for the operations defined on the function’s codomain (i.e. for \(+\) and \(\cdot\) on \(\mathbb{Q}\) in our present situation of rational sequences). The constant sequence \((0, 0, \ldots)\) is the neutral element of addition on \(S\) and \(- (r_n)_{n \in \mathbb{N}} = (-r_n)_{n \in \mathbb{N}}\) is the additive inverse of \((r_n)_{n \in \mathbb{N}}\).

The reason that we need another step in our construction of \(\mathbb{R}\) is the fact that \(S\) is not a field: As soon as 0 occurs, even just once, in the sequence \((r_n)_{n \in \mathbb{N}} \in S\), the sequence does not have a multiplicative inverse (where the neutral element of multiplication is obviously the constant sequence \((1, 1, \ldots)\)). The solution to this problem consists of factoring out all sequences converging to 0.

Definition and Remark D.31. Let

\[
\mathcal{N} := \left\{ (r_n)_{n \in \mathbb{N}} \in S : \lim_{n \to \infty} r_n = 0 \right\},
\]

be the set of rational sequences converging to zero. The relation \(\sim\) on \(S\) defined by

\[
(r_n)_{n \in \mathbb{N}} \sim (s_n)_{n \in \mathbb{N}} \iff (r_n)_{n \in \mathbb{N}} - (s_n)_{n \in \mathbb{N}} \in \mathcal{N},
\]

constitutes an equivalence relation on \(S\) (cf. Def. 2.23): Indeed, \(\sim\) is reflexive, as \(f \in S\) implies \(f - f = 0 \in \mathcal{N}\); \(\sim\) is symmetric, since \(f, g \in S\) with \(f - g \in \mathcal{N}\) implies \(g - f \in \mathcal{N}\); \(\sim\) is transitive, as \(f, g, h \in S\) with \(f - g \in \mathcal{N}\) and \(g - h \in \mathcal{N}\) implies \(f - h = f - g + g - h \in \mathcal{N}\).

Definition D.32. (a) Define the set of real numbers \(\mathbb{R}\) as the set of equivalence classes of the equivalence relation \(\sim\) defined in (D.53), i.e.

\[
\mathbb{R} := S/\sim = \left\{ [(r_n)_{n \in \mathbb{N}}] : (r_n)_{n \in \mathbb{N}} \in S \right\}
\]

is the quotient set of \(S\) with respect to \(\sim\) (cf. Ex. 2.24(c)).
(b) Addition on $\mathbb{R}$ is defined by

\[ + : \mathbb{R} \times \mathbb{R} \rightarrow \mathbb{R}, \quad ([f], [g]) \mapsto [f] + [g] := [f + g]. \tag{D.55} \]

Multiplication on $\mathbb{R}$ is defined by

\[ \cdot : \mathbb{R} \times \mathbb{R} \rightarrow \mathbb{R}, \quad ([f], [g]) \mapsto [f] \cdot [g] := [fg]. \tag{D.56} \]

Once again, for the definitions in Def. D.32(b) to make sense, one needs to check that they do not depend on the chosen representatives of the equivalence classes, and once again, we provide a lemma providing this check:

**Lemma D.33.** The definitions in Def. D.32(b) do not depend on the chosen representatives, i.e.

\[ \forall f, g, \tilde{f}, \tilde{g} \in S \quad \left( f - \tilde{f} \in \mathcal{N} \land g - \tilde{g} \in \mathcal{N} \Rightarrow f + g - (\tilde{f} + \tilde{g}) \in \mathcal{N} \right) \tag{D.57} \]

and

\[ \forall f, g, \tilde{f}, \tilde{g} \in S \quad \left( f - \tilde{f} \in \mathcal{N} \land g - \tilde{g} \in \mathcal{N} \Rightarrow fg - (\tilde{f}\tilde{g}) \in \mathcal{N} \right). \tag{D.58} \]

**Proof.** Let $f = (r_n)_{n \in \mathbb{N}}$, $g = (s_n)_{n \in \mathbb{N}}$, $\tilde{f} = (\tilde{r}_n)_{n \in \mathbb{N}}$, $\tilde{g} = (\tilde{s}_n)_{n \in \mathbb{N}}$ be elements of $S$ such that $f - \tilde{f} \in \mathcal{N}$ and $g - \tilde{g} \in \mathcal{N}$, i.e. $\lim_{n \to \infty} (r_n - \tilde{r}_n) = \lim_{n \to \infty} (s_n - \tilde{s}_n) = 0$.

Then (7.11b) implies $0 = \lim_{n \to \infty} (r_n + s_n - (\tilde{r}_n + \tilde{s}_n))$, proving (D.57).

To prove (D.58), one computes

\[ \lim_{n \to \infty} (r_n s_n - \tilde{r}_n \tilde{s}_n) = \lim_{n \to \infty} (r_n(s_n - \tilde{s}_n) - \tilde{s}_n(r_n - \tilde{r}_n)) = 0, \tag{D.59} \]

where the last equality follows from the boundedness of $(r_n)_{n \in \mathbb{N}}$ and $(\tilde{s}_n)_{n \in \mathbb{N}}$ together with Prop. 7.11(b). \[\blacksquare\]

We will also use the following auxiliary result:

**Proposition D.34.** If $(r_n)_{n \in \mathbb{N}} \in S$, then precisely one of the following statements is correct:

\[ (r_n)_{n \in \mathbb{N}} \in \mathcal{N}, \quad \tag{D.60a} \]

\[ \exists \epsilon \in \mathbb{Q}^+ \# \{ n \in \mathbb{N} : r_n \leq \epsilon \} \in \mathbb{N}_0, \tag{D.60b} \]

\[ \exists \epsilon \in \mathbb{Q}^+ \# \{ n \in \mathbb{N} : r_n \geq -\epsilon \} \in \mathbb{N}_0. \tag{D.60c} \]

**Proof.** Let us first verify that the three statements in (D.60) are mutually exclusive. If (D.60a) holds, then, for every $\epsilon \in \mathbb{Q}^+$, $-\epsilon < r_n < \epsilon$ holds for almost all (in particular, for infinitely many) $n \in \mathbb{N}$, i.e. (D.60b) and (D.60c) are both false. If (D.60b) holds,
then (D.60a) must be false as we have just seen. Moreover, if \( r_n \leq \epsilon \) holds for at most finitely many \( n \in \mathbb{N} \), then \( r_n > \epsilon > 0 \) must hold for infinitely many \( n \in \mathbb{N} \), i.e. (D.60c) is false.

Now suppose (D.60a) and (D.60b) are false. We have to show that (D.60c) is true. Since (D.60a) is false, there exists \( \delta > 0 \) and an increasing sequence of indices \( (n_k)_{k \in \mathbb{N}} \) with \( |r_{n_k}| > \delta \) for each \( k \in \mathbb{N} \). Since (D.60b) is false, there is an increasing sequence of indices \( (m_k)_{k \in \mathbb{N}} \) with \( r_{m_k} < 1/k \). Thus, since \( (r_n)_{n \in \mathbb{N}} \) is a Cauchy sequence, only finitely many \( r_{n_k} > \delta \) and infinitely many \( r_{n_k} < -\delta \). Now, if \( N \in \mathbb{N} \) is such that \( |r_n - r_m| < \delta/2 \) for all \( n, m > N \) and \( k_0 \in \mathbb{N} \) such that \( n_{k_0} > N \), then \( r_n < -\delta/2 \) for each \( n > N \) (since \( |r_n - r_{n_{k_0}}| < \delta/2 \)). Thus, (D.60c) holds with \( \epsilon := \delta/2 \). □

**Theorem D.35.** (a) The set of real numbers \( \mathbb{R} \) with addition and multiplication as defined in Def. D.32 forms a field, where \([0,0,\ldots]\) and \([(1,1,\ldots)]\) are the neutral elements with respect to addition and multiplication, respectively.

(b) The map

\[
\iota : \mathbb{Q} \to \mathbb{R}, \quad \iota(r) := [(r, r, \ldots)],
\]

is a monomorphism, i.e. it is injective and satisfies

\[
\forall r, s \in \mathbb{Q}, \quad \iota(r + s) = \iota(r) + \iota(s),
\]

\[
\forall r, s \in \mathbb{Q}, \quad \iota(rs) = \iota(r) \cdot \iota(s).
\]

It is customary to identify \( \mathbb{Q} \) with \( \iota(\mathbb{Q}) \), as it usually does not cause any confusion. One then just writes \( r \) instead of \([(r, r, \ldots)]\).

**Proof.** (a): Clearly, Def. D.32(b) ensures the laws of associativity and commutativity of addition and multiplication valid in \( \mathcal{S} \) are preserved in \( \mathbb{R} \), and, likewise, the law of distributivity. It is also immediate from (D.55) and (D.56), respectively, that \([0,0,\ldots]\) and \([(1,1,\ldots)]\) are the respective neutral elements of addition and multiplication. Moreover, if \(-f\) is the additive inverse of \( f \in \mathcal{S} \), then \([-f]\) is the additive inverse of \([f] \in \mathbb{R}\). It remains to show that each \( x = [(r_n)_{n \in \mathbb{N}}] \neq [(0,0,\ldots)] \) has a multiplicative inverse \( x^{-1} \) in \( \mathbb{R} \). We claim \( x^{-1} = [(s_n)_{n \in \mathbb{N}}] \), where

\[
\forall n \in \mathbb{N}, \quad s_n := \begin{cases} r_n^{-1} & \text{for } r_n \neq 0, \\ 1 & \text{for } r_n = 0. \end{cases}
\]

We need to verify \([(s_n)_{n \in \mathbb{N}}] \in \mathbb{R} \), i.e. \((s_n)_{n \in \mathbb{N}}\) is a Cauchy sequence. We know \((r_n)_{n \in \mathbb{N}}\) is a Cauchy sequence that does not converge to 0. Thus, according to Prop. D.34, there exists \( \delta > 0 \) and \( M \in \mathbb{N} \) such that, for each \( n > M \), we have \( |r_n| > \delta \) (in particular, \( r_n \neq 0 \)). Let \( \epsilon > 0 \). As \((r_n)_{n \in \mathbb{N}}\) is a Cauchy sequence, there exists \( N \in \mathbb{N} \) such that \( N \geq M \) and, for each \( n, m > N \), \( |r_n - r_m| < \epsilon \delta^2 \). Thus,

\[
\forall n, m > N, \quad |s_n - s_m| = \left| \frac{1}{r_n} - \frac{1}{r_m} \right| = \left| \frac{r_n - r_m}{r_n r_m} \right| < \frac{\epsilon \delta^2}{\delta^2} = \epsilon,
\]

(D.64)
proving \((s_n)_{n\in\mathbb{N}}\) is a Cauchy sequence. Moreover,
\[
[(r_n)_{n\in\mathbb{N}}] \cdot [(s_n)_{n\in\mathbb{N}}] = [(r_n s_n)_{n\in\mathbb{N}}] = [(1, 1, \ldots)],
\]
(D.65)
since \(r_n s_n = 1\) for almost all \(n \in \mathbb{N}\), and the proof of (a) is complete.
(b): The map \(\iota\) is injective, since \(\iota(r) = [(r, r, \ldots)] = \iota(s) = [(s, s, \ldots)]\) implies
\[
\lim_{n \to \infty} (r - s) = 0,
\]
i.e. \(r = s\). Moreover,
\[
\begin{align*}
\iota(r) + \iota(s) &= [(r, r, \ldots)] + [(s, s, \ldots)] = [(r + s, r + s, \ldots)] = \iota(r + s), \\
\iota(r) \cdot \iota(s) &= [(r, r, \ldots)] \cdot [(s, s, \ldots)] = [(rs, rs, \ldots)] = \iota(rs),
\end{align*}
\]
completing the proof.

**Definition D.36.** We define \(\mathbb{R}^+\) to consist of all real numbers represented by sequences \((r_n)_{n\in\mathbb{N}}\) such that there exists \(\epsilon \in \mathbb{Q}^+\) satisfying \(r_n > \epsilon\) for almost all \(n \in \mathbb{N}\), i.e.
\[
\mathbb{R}^+ := \left\{ [(r_n)_{n\in\mathbb{N}}] \in \mathbb{R} : \exists \epsilon \in \mathbb{Q}^+ \#\{n \in \mathbb{N} : r_n \leq \epsilon\} \in \mathbb{N}_0 \right\}.
\]
(D.67)

**Proposition D.37.** (a) The definition in (D.67) does not depend on the chosen representatives \((r_n)_{n\in\mathbb{N}}\).

(b) We have the decomposition
\[
\mathbb{R} = \mathbb{R}^+ \cup \{0\} \cup \mathbb{R}^-,
\]
\[
\mathbb{R}^- := -\mathbb{R}^+ = \{x \in \mathbb{R} : -x \in \mathbb{R}^+\}.
\]
(D.68)

**Proof.** (a): If \((s_n)_{n\in\mathbb{N}}\) with \(\lim_{n \to \infty} (r_n - s_n) = 0\), then \(|r_n - s_n| < \epsilon/2\) for almost all \(n \in \mathbb{N}\). Thus, since \(|s_n| \geq |r_n| - |r_n - s_n|\), we obtain \(s_n > \epsilon/2\) for almost all \(n \in \mathbb{N}\), i.e.
\[
\#\{n \in \mathbb{N} : s_n \leq \frac{\epsilon}{2}\} \in \mathbb{N}_0.
\]
(b) is an immediate consequence of Prop. D.34. ■

**Definition D.38.** For each \(x, y \in \mathbb{R}\), let
\[
y \leq x \iff x - y \in \mathbb{R}^+_0 := \mathbb{R}^+ \cup \{0\}.
\]
(D.69)

**Theorem D.39.** (a) The relation defined in (D.69) constitutes a total order on \(\mathbb{R}\) that is compatible with addition and multiplication, i.e. it satisfies (4.3); in other words \((\mathbb{R}, +, \cdot, \leq)\) constitutes a totally ordered field.

(b) The map \(\iota\) from (D.61) is strictly increasing.

**Proof.** (a) follows from (D.69), (D.68), and Th. D.8, once we have shown that \(\mathbb{R}^+\) is closed under addition and multiplication. Let \((r_n)_{n\in\mathbb{N}} \in \mathcal{S}\), \((s_n)_{n\in\mathbb{N}} \in \mathcal{S}\). If \(r_n > \epsilon_1 \in \mathbb{Q}^+\) for almost all \(n \in \mathbb{N}\) and \(s_n > \epsilon_2 \in \mathbb{Q}^+\) for almost all \(n \in \mathbb{N}\), then \(r_n + s_n > \epsilon_1 + \epsilon_2\), showing \(\mathbb{R}^+\) is closed under addition. Moreover, \(r_n s_n > \epsilon_1 \epsilon_2\), showing \(\mathbb{R}^+\) is closed under multiplication.

(b): According to Def. D.39, if \(r, s \in \mathbb{Q}\) with \(s < r\), then \(q := r - s \in \mathbb{Q}^+\). In consequence \(\iota(r) = \iota(s) + \iota(q)\) by (D.62a), i.e. \(\iota(r) - \iota(s) = \iota(q) = [(q, q, \ldots)] \in \mathbb{R}^+\), proving \(\iota(s) < \iota(r)\). ■
Finally, we will show in Th. D.41 below that the order ≤ on \( \mathbb{R} \) is complete. However, we first need some additional auxiliary results.

**Proposition D.40.** (a) For each \( x \in \mathbb{R} \), there is \((r_n)_{n \in \mathbb{N}} \in \mathcal{S}\) satisfying \( \lim_{n \to \infty} r_n = x \).

(b) Every \((r_n)_{n \in \mathbb{N}} \in \mathcal{S}\) converges in \( \mathbb{R} \) – more precisely, \( \lim_{n \to \infty} r_n = [(r_n)_{n \in \mathbb{N}}] \).

(c) Every Cauchy sequence in \( \mathbb{R} \) converges in \( \mathbb{R} \).

**Proof.** (a) and (b): If \( x = [(r_n)_{n \in \mathbb{N}}] \) with \((r_n)_{n \in \mathbb{N}} \in \mathcal{S}\), then, given \( \epsilon > 0 \), choose \( N \in \mathbb{N} \) such that, for each \( m, n > N \), one has \( |r_n - r_m| < \epsilon/2 \). Then, for each \( k > N \), one has \( |x - r_k| = |[(r_n - r_k)_{n \in \mathbb{N}}]| < \epsilon \), since \( |r_n - r_k| < \epsilon/2 \) for all \( n \geq k \), showing \( \lim_{n \to \infty} r_n = x \).

(c): Let \((x_n)_{n \in \mathbb{N}}\) be a Cauchy sequence in \( \mathbb{R} \). According to (a), for each \( n \in \mathbb{N} \), there exists \( r_n \in \mathbb{Q} \) such that \( |x_n - r_n| < \frac{1}{n} \). Then \((r_n)_{n \in \mathbb{N}}\) is a Cauchy sequence: Given \( \epsilon > 0 \), choose \( k \in \mathbb{N} \) such that \( \frac{1}{k} < \frac{\epsilon}{3} \) and \( |x_n - x_m| < \frac{\epsilon}{3} \) for each \( n, m > k \). Then

\[
\forall \, n,m > k \, \mid r_n - r_m \mid < |r_n - x_n| + |x_n - x_m| + |x_m - r_m| < \frac{\epsilon}{3} + \frac{\epsilon}{3} + \frac{\epsilon}{3} = \epsilon,
\]

showing \((r_n)_{n \in \mathbb{N}}\) is Cauchy. Thus, from (b), we obtain \( x \in \mathbb{R} \) with \( \lim_{n \to \infty} r_n = x \). We can now show, \( \lim_{n \to \infty} x_n = x \) as well: Given \( \epsilon > 0 \), choose \( N \in \mathbb{N} \) such that \( \frac{1}{N} < \frac{\epsilon}{2} \) and \( |x - r_n| < \frac{\epsilon}{2} \) for each \( n > N \). Then

\[
\forall \, n > N \, |x - x_n| \leq |x - r_n| + |r_n - x_n| < \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon,
\]

showing \( \lim_{n \to \infty} x_n = x \) and completing the proof.

**Theorem D.41.** The order \( \leq \) on \( \mathbb{R} \) is complete, i.e. \( \mathbb{R}, +, \cdot, \leq \) constitutes a complete totally ordered field.

**Proof.** Let \( \emptyset \neq A \subseteq \mathbb{R} \) and let \( M \in \mathbb{R} \) be an upper bound for \( A \). We have to show that \( A \) has a supremum in \( \mathbb{R} \). To this end, we recursively construct two Cauchy sequences \((x_n)_{n\in\mathbb{N}}\) and \((y_n)_{n\in\mathbb{N}}\) in \( \mathbb{R} \) such that \((x_n)_{n\in\mathbb{N}}\) is increasing, \((y_n)_{n\in\mathbb{N}}\) is decreasing, \( x_n < y_n \), and \( \lim_{n \to \infty}(y_n - x_n) = 0 \). Let \( x_1 \in A \) be arbitrary and \( y_1 := M \). Define

\[
\forall \, n \in \mathbb{N} \begin{cases} x_{n+1} := \frac{(x_n + y_n)}{2} & \text{if } (x_n + y_n)/2 \text{ is not an upper bound for } A, \\ x_n & \text{otherwise}, \end{cases}
\]

\[
\forall \, n \in \mathbb{N} \begin{cases} y_{n+1} := \frac{(x_n + y_n)}{2} & \text{if } (x_n + y_n)/2 \text{ is an upper bound for } A, \\ y_n & \text{otherwise}. \end{cases}
\]

Then, clearly, the \( x_n \) are increasing, the \( y_n \) are decreasing, and \( x_n \leq y_n \) holds for each \( n \in \mathbb{N} \). Moreover, letting \( d := M - x_1 \geq 0 \), a simple induction shows \( y_n - x_n = d/2^{n-1} \) and \( \lim_{n \to \infty}(y_n - x_n) = 0 \). Also, for \( m > n \),

\[
x_m - x_n = \sum_{i=n}^{m-1} (x_{i+1} - x_i) \leq d \sum_{i=n}^{m-1} 2^{-i} = \frac{d}{2^n} \sum_{i=n}^{m-1} 2^{-i+n} = \frac{d}{2^n} \sum_{i=0}^{m-1-n} 2^{-i} \leq \frac{2d}{2^n},
\]

(D.73)
showing \((x_n)_{n \in \mathbb{N}}\) is a Cauchy sequence. Analogous, one sees that \((y_n)_{n \in \mathbb{N}}\) is a Cauchy sequence. By Prop. D.40(c), we obtain \(s \in \mathbb{R}\) such that \(s = \lim_{n \to \infty} x_n = \lim_{n \to \infty} (y_n - x_n + x_n) = \lim_{n \to \infty} y_n\). We claim \(s = \sup A\). If \(s < y\), then there is \(n \in \mathbb{N}\) with \(s \leq y_n < y\), showing \(y \notin A\), i.e. \(s\) is an upper bound for \(A\). If \(y < s\), then there is \(n \in \mathbb{N}\) with \(y < x_n \leq s\), showing \(y\) is not an upper bound for \(A\). Thus, \(s\) is the smallest upper bound for \(A\), i.e. \(s = \sup A\).

D.6 Uniqueness

We will show in Th. D.45 below that, up to a unique isomorphism, \(\mathbb{R}\) is the only complete totally ordered field.

**Notation D.42.** Let \((A, +, \cdot, \leq)\) be a complete totally ordered field. The neutral elements with respect to + and \(\cdot\), we denote with \(0_A\) and \(1_A\), respectively. We recursively define \((n + 1)_A := n_A + 1_A\) for each \(n \in \mathbb{N}\). Then \(\mathbb{N}_A := \{n_A : n \in \mathbb{N}\}\), \(\mathbb{Z}_A := \mathbb{N}_A \cup \{0_A\} \cup \{-n_A : n \in \mathbb{N}\}\), \(\mathbb{Q}_A := \{0_A\} \cup \{\frac{k}{l} : k, l \in \mathbb{Z}_A \setminus \{0_A\}\}\).

**Proposition D.43.** Let \((A, +, \cdot, \leq)\) and \((B, +, \cdot, \leq)\) be complete totally ordered fields. Moreover, let \(\phi : A \to B\) be a field isomorphism, i.e. a bijective map, satisfying

\[
\begin{align*}
\forall x, y \in A & \quad \phi(x + y) = \phi(x) + \phi(y), \\
\forall x, y \in A & \quad \phi(xy) = \phi(x)\phi(y).
\end{align*}
\]

(a) \(\phi(n_A) = n_B\) holds for each \(n \in \mathbb{N}\).

(b) \(\phi\) is strictly isotone, i.e.

\[
\forall x, y \in A \quad (x < y \Rightarrow \phi(x) < \phi(y)).
\]

**Proof.** (a): As (D.74a) and (D.74b) state \(\phi\) to be a group homomorphism with respect to addition and multiplication, respectively, Prop. C.4(a) yields \(\phi(0_A) = 0_B\) and \(\phi(1_A) = 1_B\). If \(n \in \mathbb{N}\), then (D.74a) implies \(\phi(n_A + 1_A) = \phi(n_A) + 1_B\) and, thus, an induction shows \(\phi(n_A) = n_B\) for each \(n \in \mathbb{N}\).

(b): If \(x, y \in A\) with \(x < y\), then, by Rem. and Def. 7.61, there exists a unique \(z \in A\) such that \(z^2 = y - x\). Thus, \((\phi(z))^2 = \phi(y) - \phi(x)\). By Th. 4.5(c), we have \((\phi(z))^2 > 0\) and, thus, \(\phi(x) < \phi(y)\), proving the strict isotonicity of \(\phi\).

**Proposition D.44.** Let \((A, +, \cdot, \leq)\) be a complete totally ordered field. If \(\phi : A \to A\) is (field) automorphism, i.e. a bijective map, satisfying (D.74), then \(\phi\) is the identity on \(A\).

**Proof.** From Prop. D.43(a), we already know \(\phi(n) = n\) for each \(n \in \mathbb{N}_A\). Next, if \(n \in \mathbb{N}_A\), then, using Prop. C.4(b), we obtain \(\phi(-n) = -\phi(n) = -n\), showing \(\phi(k) = k\) for each \(k \in \mathbb{Z}_A\). If \(k, l \in \mathbb{Z}_A \setminus \{0_A\}\), then \(\phi(k/l) = \phi(k \cdot l^{-1}) = \phi(k) \cdot (\phi(l))^{-1} = kl^{-1}\),
where Prop. C.4(b) was used again. Thus, we already have \( \phi(q) = q \) for each \( q \in \mathbb{Q}_A \).
From Prop. D.43(b), we know \( \phi \) to be strictly isotone. Finally, if \( x \in A \), then, by Th. 7.68(c), there exist a sequences \((r_n)_{n \in \mathbb{N}}\) in \( \mathbb{Q}_A \) and \((s_n)_{n \in \mathbb{N}}\) in \( \mathbb{Q}_A \) such that \((r_n)_{n \in \mathbb{N}}\) is strictly increasing, \((s_n)_{n \in \mathbb{N}}\) is strictly decreasing, and
\[
\lim_{n \to \infty} r_n = \lim_{n \to \infty} s_n = x.
\]
As \( \phi \) is isotone, we obtain
\[
\forall n \in \mathbb{N} \quad \phi(r_n) = r_n \leq f(x) \leq s_n = \phi(s_n).
\]
But then \( f(x) = x \) follows from Th. 7.16, proving \( \phi = \text{Id}_A \). \( \blacksquare \)

**Theorem D.45.** Let \((A, +, \cdot, \leq)\) and \((B, +, \cdot, \leq)\) be complete totally ordered fields. Then there exists a unique isomorphism \( \phi : A \to B \), i.e. a unique bijective \( \phi : A \to B \), satisfying (D.74a), (D.74b), and (D.75).

**Proof.** Uniqueness: Suppose, \( \phi : A \to B \) and \( \psi : A \to B \) are both isomorphisms. Then, according to Prop. D.44, \( \phi^{-1} \circ \psi = \text{Id}_A \), where Prop. C.4(c) was used as well.
However, this already shows \( \psi = \phi \).
Existence: Due to Prop. D.43(b), it suffices to show there exists a bijective \( \phi : A \to B \), satisfying (D.74). We define \( \phi \) and verify (D.74) in several steps. In the first step, set
\[
\forall n \in \mathbb{N}_0 \quad \phi(n_A) := n_B. \tag{D.76}
\]
Then \( \phi : \mathbb{N}_A \cup \{0_A\} \to \mathbb{N}_B \cup \{0_B\} \) is bijective with \( \phi^{-1} : \mathbb{N}_B \cup \{0_B\} \to \mathbb{N}_A \cup \{0_A\} \), \( \phi^{-1}(n_B) = n_A \). We first verify
\[
\forall n \in \mathbb{N}_0 \quad \phi(1_A + n_A) = \phi(1_A) + \phi(n_A) : \tag{D.77}
\]
Indeed,
\[
\phi(1_A + n_A) = \phi((n + 1)_A) = (n + 1)_B = 1_B + n_B = \phi(1_A) + \phi(n_A).
\]
Next, we verify
\[
\forall m, n \in \mathbb{N}_0 \quad \phi(m_A + n_A) = \phi(m_A) + \phi(n_A) \tag{D.78}
\]
via induction on \( m \): The case \( m = 0 \) holds, due to \( \phi(0_A + n_A) = \phi(n_A) = 0_B + n_B = \phi(0_A) + \phi(n_A) \). The case \( m = 1 \) holds, due to (D.77). For the induction step, we compute
\[
\phi((m + 1)_A + n_A) = \phi(m_A + 1_A + n_A) \overset{\text{ind. hyp.}}{=} \phi(m_A) + \phi(1_A + n_A) \overset{(D.77)}{=} \phi(m_A) + \phi(1_A) + \phi(n_A) \overset{(D.77)}{=} \phi((m + 1)_A) + \phi(n_A),
\]
proving (D.78). We now prove
\[
\forall m, n \in \mathbb{N}_0 \quad \phi(m_A n_A) = \phi(m_A) \phi(n_A) \tag{D.79}
\]
via induction on \( m \): The case \( m = 0 \) holds, due to

\[
\phi(0_A n_A) = \phi(0_A) = 0_B = \phi(0_A) \phi(n_A).
\]

For the induction step, we compute

\[
\phi((m + 1)_A n_A) = \phi((m + 1)_A n_A) = \phi(m_A n_A + n_A) \quad \text{(D.78)} \quad \phi(m_A n_A) + \phi(n_A)
\]

\[
\phi(m_A) \phi(n_A) + \phi(n_A) = (\phi(m_A) + 1_B) \phi(n_A) \quad \text{(D.78)}
\]

\[
\phi((m + 1)_A) \phi(n_A) = \phi((m + 1)_A) \phi(n_A).
\]

In particular, according to (D.78) and (D.79), (D.74) holds for each \( x, y \in \mathbb{N}_A \).

In the second step, set

\[
\forall n \in \mathbb{N} \quad \phi(-n_A) := -n_B.
\]

Then \( \phi : \mathbb{Z}_A \rightarrow \mathbb{Z}_B \) is still bijective, where, for each \( m, n \in \mathbb{N} \), we have \( \phi^{-1}(-n_B) = -n_A \).

Let \( m, n \in \mathbb{N}_0 \). If \( m \leq n \), then

\[
\phi(-m_A + n_A) = \phi((n - m)_A) = (n - m)_B = -m_B + n_B = \phi(-m_A) + \phi(n_A).
\]

If \( m > n \), then

\[
\phi(-m_A + n_A) = \phi(-(m - n)_A) = -(m - n)_B = -m_B + n_B = \phi(-m_A) + \phi(n_A).
\]

Now, for arbitrary \( m, n \in \mathbb{N}_0 \), \( \phi(m_A + (-n_A)) = \phi(-n_A + m_A) = \phi(-n_A) + \phi(m_A) = \phi(m_A) + \phi(-n_A) \) and \( \phi(-m_A + (-n_A)) = \phi(-(m_A + n_A)) = -(\phi(m_A + n_A)) = -(\phi(m_A) + \phi(n_A)) = -\phi(m_A) - \phi(n_A) = \phi(-m_A) + \phi(-n_A) \). We now consider multiplication, still for \( m, n \in \mathbb{N}_0 \):

\[
\phi((-m_A) n_A) = -\phi(m_A n_A) \quad \text{(D.79)} \quad -\phi(m_A) \phi(n_A) = \phi(-m_A) \phi(n_A).
\]

Then \( \phi(m_A (-n_A)) = \phi(m_A) \phi(-n_A) \) also follows and

\[
\phi((-m_A)(-n_A)) = \phi(m_A n_A) \quad \text{(D.79)} \quad \phi(m_A) \phi(n_A) = \phi(-m_A) \phi(-n_A).
\]

Thus, we have verified (D.74) for each \( x, y \in \mathbb{Z}_A \).

In the third step, set

\[
\forall k, l \in \mathbb{Z}_A \setminus \{0_A\} \quad \phi(k/l) := \phi(k)/\phi(l).
\]

We verify that (D.81) well-defines \( \phi \) for each \( q \in \mathbb{Q}_A \): If \( m, n, k, l \in \mathbb{Z}_A \) with \( n, l \neq 0_A \), then \( m/n = k/l \) implies \( ml = kn \) and \( \phi(m) \phi(l) = \phi(ml) = \phi(kn) = \phi(k) \phi(n) \). Thus,

\[
\phi(m/n) = \phi(m)/\phi(n) = \phi(k)/\phi(l) = \phi(k/l).
\]

We show \( \phi : \mathbb{Q}_A \rightarrow \mathbb{Q}_B \) to be bijective by providing the inverse map: Define \( \psi : \mathbb{Q}_B \rightarrow \mathbb{Q}_A \) by setting

\[
\forall k, l \in \mathbb{Z}_B \setminus \{0_B\} \quad \psi(k/l) := \phi^{-1}(k)/\phi^{-1}(l).
\]
We claim that \( \psi = \phi^{-1} \) on \( \mathbb{Q}_B \): Indeed, for each \( k, l \in \mathbb{Z}_A \setminus \{0_A\} \) and for each \( m, n \in \mathbb{Z}_B \setminus \{0_B\} \)

\[
\psi(\phi(k/l)) = \psi(\phi(k)/\phi(l)) = \phi^{-1}(\phi(k))/\phi^{-1}(\phi(l)) = k/l,
\]

\[
\phi(\psi(m/n)) = \phi(\phi^{-1}(m)/\phi^{-1}(n)) = \phi(\phi^{-1}(m))/\phi(\phi^{-1}(n)) = m/n.
\]

Moreover, we have

\[
\phi \left( \frac{m}{n} + \frac{k}{l} \right) = \phi \left( \frac{ml + kn}{nl} \right) = \frac{\phi(ml + kn)}{\phi(nl)} \quad \text{(D.74) for } Z_A \quad \phi(m)\phi(l) + \phi(k)\phi(n)
\]

\[
= \frac{\phi(m)}{\phi(n)} + \frac{\phi(k)}{\phi(l)} = \phi \left( \frac{m}{n} \right) + \phi \left( \frac{k}{l} \right)
\]

and

\[
\phi \left( \frac{m \cdot k}{n \cdot l} \right) = \phi \left( \frac{mk}{nl} \right) = \frac{\phi(mk)}{\phi(nl)} \quad \text{(D.74) for } Z_A \quad \phi(m)\phi(k)
\]

\[
= \frac{\phi(m)}{\phi(n)} \cdot \frac{\phi(k)}{\phi(l)} = \phi \left( \frac{m}{n} \right) \cdot \phi \left( \frac{k}{l} \right).
\]

Thus, we have verified (D.74) for each \( x, y \in \mathbb{Q}_A \).

We now show \( \phi : \mathbb{Q}_A \to \mathbb{Q}_B \) to be strictly isotone, i.e.

\[
\forall_{r, s \in \mathbb{Q}_A} \quad \left( r < s \implies \phi(r) < \phi(s) \right) \quad \text{(D.82)}
\]

Let \( r, s \in \mathbb{Q}_A \) such that \( r < s \). Then \( d := s - r > 0_A \), i.e. there are \( m, n \in \mathbb{N}_A \) satisfying \( d = \frac{m}{n} \). Then \( \phi(s) - \phi(r) = \phi(d) = \frac{\phi(m)}{\phi(n)} > 0_B \), proving \( \phi(r) < \phi(s) \).

In the fourth (and last) step, for each \( x \in \mathbb{A} \), we choose a sequence \( (r_n)_{n \in \mathbb{N}} \) in \( \mathbb{Q}_A \) such that \( x = \lim_{n \to \infty} r_n \) and set

\[
\phi(x) := \lim_{n \to \infty} \phi(r_n). \quad \text{(D.83)}
\]

To show that \( \phi \) is well-defined by (D.83), we have to verify that \( (\phi(r_n))_{n \in \mathbb{N}} \) does, indeed, converge in \( B \), and that \( \phi(x) \) does not depend on the chosen sequence \( (r_n)_{n \in \mathbb{N}} \). As \( (r_n)_{n \in \mathbb{N}} \) converges to \( x \), it has to be a Cauchy sequence by Th. 7.29. We show that \( (\phi(r_n))_{n \in \mathbb{N}} \) must be a Cauchy sequence as well: Let \( \epsilon \in B \), \( \epsilon > 0_B \) and choose \( \bar{\epsilon} \in \mathbb{Q}_B \) such that \( 0_B < \bar{\epsilon} < \epsilon \). Then

\[
\exists_{N \in \mathbb{N}} \quad \forall_{n, m > N} \quad |r_n - r_m| < \phi^{-1}(\bar{\epsilon}).
\]

As \( \phi \) is strictly isotone, we obtain

\[
\forall_{n, m > N} \quad |\phi(r_n) - \phi(r_m)| < \bar{\epsilon} < \epsilon,
\]

proving \( (\phi(r_n))_{n \in \mathbb{N}} \) to be a Cauchy sequence. Now Th. 7.29 implies the convergence of \( (\phi(r_n))_{n \in \mathbb{N}} \). Next, we show \( \lim_{n \to \infty} r_n = 0_A \) implies \( \lim_{n \to \infty} \phi(r_n) = 0_B \) for each sequence in \( \mathbb{Q}_A \): Indeed, as above, let \( \epsilon > 0_B \) and choose \( \bar{\epsilon} \in \mathbb{Q}_B \) such that \( 0_B < \bar{\epsilon} < \epsilon \). Then

\[
\exists_{N \in \mathbb{N}} \quad \forall_{n > N} \quad |r_n| < \phi^{-1}(\bar{\epsilon}).
\]
As \( \phi \) is strictly isotone, we obtain
\[
\forall n > N \quad |\phi(r_n)| < \epsilon,
\]
proving \( \lim_{n \to \infty} \phi(r_n) = 0_B \). Thus, if \( (r_n)_{n \in \mathbb{N}} \) and \( (s_n)_{n \in \mathbb{N}} \) are sequences in \( \mathbb{Q}_A \) such that \( \lim_{n \to \infty} r_n = x = \lim_{n \to \infty} s_n \), then
\[
\lim_{n \to \infty} \phi(r_n) = \lim_{n \to \infty} \phi(r_n - s_n + s_n) = 0_B + \lim_{n \to \infty} \phi(s_n) = \lim_{n \to \infty} \phi(s_n),
\]
showing \( \phi \) to be well-defined by (D.83). To see that \( \phi \) is injective, let \( x, y \in A \) with \( x < y \) and choose \( r, s \in \mathbb{Q}_A \) such that \( x < r < s < y \). If \( (r_n)_{n \in \mathbb{N}} \) and \( (s_n)_{n \in \mathbb{N}} \) are sequences in \( \mathbb{Q}_A \) such that \( x = \lim_{n \to \infty} r_n \) and \( y = \lim_{n \to \infty} s_n \), then
\[
\exists N \in \mathbb{N} \quad \forall n > N \quad (r_n < r < s < s_n).
\]
As \( \phi \) is strictly isotone, we obtain
\[
\forall n > N \quad \left( \phi(r_n) < \phi(r) < \phi(s) < \phi(s_n) \right),
\]
showing \( \phi(x) \neq \phi(y) \) and the injectivity of \( \phi \). To see that \( \phi \) is surjective, let \( b \in B \) and let \( (r_n)_{n \in \mathbb{N}} \) be an increasing sequence in \( \mathbb{Q}_B \) such that \( b = \lim_{n \to \infty} r_n \). Then \( (\phi^{-1}(r_n))_{n \in \mathbb{N}} \) is an increasing sequence in \( \mathbb{Q}_A \) that is bounded, i.e. it must converge to some \( a \in A \). Then
\[
\phi(a) = \lim_{n \to \infty} \phi(\phi^{-1}(r_n)) = \lim_{n \to \infty} r_n = b,
\]
showing \( \phi \) to be surjective. Finally, if \( x, y \in A \), then let \( (r_n)_{n \in \mathbb{N}} \) and \( (s_n)_{n \in \mathbb{N}} \) be sequences in \( \mathbb{Q}_A \) such that \( x = \lim_{n \to \infty} r_n \) and \( y = \lim_{n \to \infty} s_n \). Then
\[
\phi(x + y) = \lim_{n \to \infty} \phi(r_n + s_n) = \lim_{n \to \infty} \phi(r_n) + \lim_{n \to \infty} \phi(s_n) = \phi(x) + \phi(y)
\]
and
\[
\phi(xy) = \lim_{n \to \infty} \phi(r_n s_n) = \lim_{n \to \infty} \phi(r_n) \lim_{n \to \infty} \phi(s_n) = \phi(x)\phi(y).
\]
Thus, we have verified (D.74) for each \( x, y \in A \), and, thereby completed the proof. \( \blacksquare \)

### E Series: Additional Material

#### E.1 Riemann Rearrangement Theorem

Here, we provide the details for the proof of the Riemann rearrangement Th. 7.93, that was merely sketched in Sec. 7.3.3.

**Proof of Th. 7.93.** As already stated in the sketch, we define
\[
\forall k \in \mathbb{N} \quad x_k := \begin{cases} -k & \text{for } x = -\infty, \\ x & \text{for } x \in \mathbb{R}, \\ k & \text{for } x = \infty, \end{cases} \quad y_k := \begin{cases} -k & \text{for } y = -\infty, \\ y & \text{for } y \in \mathbb{R}, \\ k & \text{for } y = \infty, \end{cases}
\] (E.1)
noting \( x_k \leq y_k \) for almost all \( k \in \mathbb{N} \). Next, we observe

\[
\mathbb{N} = \mathbb{I}^+ \cup \mathbb{I}^-, \quad \text{where}
\]

\[
\mathbb{I}^+ := \{ j \in \mathbb{N} : a_j \geq 0 \},
\]

\[
\mathbb{I}^- := \{ j \in \mathbb{N} : a_j < 0 \}.
\]

We have to define a suitable bijective map \( \phi : \mathbb{N} \rightarrow \mathbb{N} \) such that

\[
\forall_{j \in \mathbb{N}} b_j := a_{\phi(j)},
\]

\[
\forall_{n \in \mathbb{N}} t_n := \sum_{j=1}^{n} b_j.
\]

The definition of \( \phi \) will be recursive, and we will also need to recursively define an auxiliary sequence \( (\sigma_j)_{j \in \mathbb{N}} \) taking values in \( \{-1, 1\} \), serving as an accounting tool to keep track if we are in the process of moving right (i.e. adding \( a_j^+ \)) or moving left (i.e. subtracting \( a_j^- \)). Moreover, we need a recursively defined auxiliary function \( \kappa : \mathbb{N} \rightarrow \mathbb{N} \) to update the left and right boundaries \( x_k \) and \( y_k \), respectively, to handle the first and third case of (E.1) if need be. The recursion is initialized by

\[
\phi(1) := 1,
\]

\[
\sigma_1 := \begin{cases} 
1 & \text{if } t_1 \leq y_1, \\
-1 & \text{if } t_1 > y_1,
\end{cases}
\]

\[
\kappa(1) := \begin{cases} 
1 & \text{if } t_1 \leq y_1, \\
2 & \text{if } t_1 > y_1,
\end{cases}
\]

and completed by

\[
\forall_{j > 1} \phi(j) := \begin{cases} 
\min (\mathbb{I}^+ \setminus \phi\{\ldots, j-1\}) & \text{if } \sigma_{j-1} = 1, \\
\min (\mathbb{I}^- \setminus \phi\{\ldots, j-1\}) & \text{if } \sigma_{j-1} = -1,
\end{cases}
\]

\[
\forall_{j > 1} \sigma_j := \begin{cases} 
1 & \text{if } \sigma_{j-1} = 1 \text{ and } t_j \leq y_{\kappa(j-1)}, \\
-1 & \text{if } \sigma_{j-1} = 1 \text{ and } t_j > y_{\kappa(j-1)}, \\
-1 & \text{if } \sigma_{j-1} = -1 \text{ and } t_j \geq x_{\kappa(j-1)}, \\
1 & \text{if } \sigma_{j-1} = -1 \text{ and } t_j < x_{\kappa(j-1)},
\end{cases}
\]

\[
\forall_{j > 1} \kappa(j) := \begin{cases} 
\kappa(j-1) & \text{if } \sigma_{j-1} = 1 \text{ and } t_j \leq y_{\kappa(j-1)}, \\
1 + \kappa(j-1) & \text{if } \sigma_{j-1} = 1 \text{ and } t_j > y_{\kappa(j-1)}, \\
\kappa(j-1) & \text{if } \sigma_{j-1} = -1 \text{ and } t_j \geq x_{\kappa(j-1)}, \\
1 + \kappa(j-1) & \text{if } \sigma_{j-1} = -1 \text{ and } t_j < x_{\kappa(j-1)}.
\end{cases}
\]

We note that \( \phi \) is well-defined, since, according to (7.86), both \( \mathbb{I}^+ \) and \( \mathbb{I}^- \) must have infinitely many elements. Moreover, \( \phi \) is injective, since, for \( j_1 < j_2 \), \( \phi(j_2) \neq \phi(j_1) \) is immediate from (E.5a). Finally, \( \phi \) is also surjective: Otherwise, there is a smallest
\( n \in \mathbb{N} \setminus \{1\} \) such that \( n \notin \phi(\mathbb{N}) \). Suppose \( n \in I^+ \). Then, according to (E.5a), there must be \( j_0 \in \mathbb{N} \) such that \( \sigma_j = -1 \) for every \( j > j_0 \), i.e., according to (E.5b) and (E.5c), \( t_j \geq x_{\kappa(j_0)} \in \mathbb{R} \) for each \( j > j_0 \), which is in contradiction to the \( \sum_{j=1}^{\infty} a_j^- = \infty \) part of (7.86). Analogously, \( n \in I^- \) leads to a contradiction to the \( \sum_{j=1}^{\infty} a_j^+ = \infty \) part of (7.86), completing the proof of surjectivity of \( \phi \). So we have shown that \( \sum_{j=1}^{\infty} b_j \) is a rearrangement of \( \sum_{j=1}^{\infty} a_j \) as desired. We still need to verify that \( \sum_{j=1}^{\infty} b_j \) (i.e. \( (t_n)_{n \in \mathbb{N}} \)) has precisely all elements of \([x,y]\) as cluster points. To this end, first note that, due to (7.86) and (E.1), \( \lim_{j \to \infty} x_{\kappa(j)} = -\infty \) holds if, and only if, \( x = -\infty \); and \( \lim_{j \to \infty} x_{\kappa(j)} = \infty \) holds if, and only if, \( x = \infty \); and likewise for the \( y_{\kappa(j)} \) and \( y \). If \( x = -\infty \), then \( \lim_{j \to \infty} x_{\kappa(j)} = -\infty \) and the bijectivity of \( \phi \) together with (E.5b) and (E.5c) implies

\[
\forall \ N \in \mathbb{N} \quad \exists \ j \in \mathbb{N} \quad t_j < x_{\kappa(j-1)} \leq -N,
\]

showing \( -\infty \) is a cluster point of \( (t_n)_{n \in \mathbb{N}} \). Analogously, if \( y = \infty \), then \( \lim_{j \to \infty} y_{\kappa(j)} = \infty \) and the bijectivity of \( \phi \) together with (E.5b) and (E.5c) implies

\[
\forall \ N \in \mathbb{N} \quad \exists \ j \in \mathbb{N} \quad t_j > y_{\kappa(j-1)} \geq N,
\]

showing \( \infty \) is a cluster point of \( (t_n)_{n \in \mathbb{N}} \). Now let \( \xi \in [x,y] \cap \mathbb{R} \) and \( \epsilon > 0 \). Due to \( \lim_{j \to \infty} a_j^+ = \lim_{j \to \infty} a_j^- = 0 \), we have

\[
\exists \ N \in \mathbb{N} \quad \forall \ j > N \quad t_j - t_{j-1} < \epsilon.
\]

Due to the bijectivity of \( \phi \) together with (E.5b) and (E.5c), for each \( j_0 \in \mathbb{N} \), there exists \( j > \max\{j_0, N\} \) such that \( t_{j-1} \leq \xi \leq t_j \), showing \( \xi \) is a cluster point of \( (t_n)_{n \in \mathbb{N}} \). On the other hand, if \( \xi \in ]-\infty, x[ \), then \( x \neq -\infty \). If \( x = \infty \), then \( \lim_{j \to \infty} t_j = \infty \) and \( \xi \) is not a cluster point of \( (t_n)_{n \in \mathbb{N}} \). If \( \xi < x < \infty \), then let \( \epsilon := (x - \xi)/2 \) and choose \( N \) as in (E.8). Then, by (E.5b) and (E.5c), for each \( j > N \), \( t_j > x - \epsilon = \xi + \epsilon \), showing \( \xi \) is not a cluster point of \( (t_n)_{n \in \mathbb{N}} \). Analogously, one sees that \( \xi \in ]y, \infty[ \) can not be a cluster point of \( (t_n)_{n \in \mathbb{N}} \). \( \blacksquare \)

### E.2 \( b \)-Adic Representations of Real Numbers

The main goal of this section is to provide a proof of Th. 7.99. We begin with some preparatory lemmas.

**Lemma E.1.** Given a natural number \( b \geq 2 \), consider the \( b \)-adic series given by (7.96). Then

\[
\sum_{\nu=0}^{\infty} d_{N-\nu} b^{N-\nu} \leq b^{N+1}, \tag{E.9}
\]

and, in particular, the \( b \)-adic series converges to some \( x \in \mathbb{R}_0^+ \). Moreover, equality in (E.9) holds if, and only if, \( d_n = b - 1 \) for every \( n \in \{N, N-1, N-2, \ldots\} \).
Proof. One estimates, using the formula for the value of a geometric series:
\[
\sum_{\nu=0}^{\infty} d_{N-\nu} b^{N-\nu} \leq \sum_{\nu=0}^{\infty} (b-1) b^{N-\nu} = (b-1)b^N \sum_{\nu=0}^{\infty} b^{-\nu} = (b-1)b^N \frac{1}{1-\frac{1}{b}} = b^{N+1}. \tag{E.10}
\]
Note that (E.10) also shows that equality is achieved if all \(d_n\) are equal to \(b-1\). Conversely, if there is \(n \in \{N, N-1, N-2, \ldots\}\) such that \(d_n < b-1\), then there is \(\bar{n} \in \mathbb{N}\) such that \(d_{N-\bar{n}} < b-1\) and one estimates
\[
\sum_{\nu=0}^{\infty} d_{N-\nu} b^{N-\nu} < \sum_{\nu=0}^{\bar{n}-1} d_{N-\nu} b^{N-\nu} + (b-1)b^{N-\bar{n}} + \sum_{\nu=\bar{n}+1}^{\infty} d_{N-\nu} b^{N-\nu} \leq b^{N+1}, \tag{E.11}
\]
showing that the inequality in (E.9) is strict. \(\blacksquare\)

**Lemma E.2.** Given a natural number \(b \geq 2\), consider two \(b\)-adic series
\[
x := \sum_{\nu=0}^{\infty} d_{N-\nu} b^{N-\nu} = \sum_{\nu=0}^{\infty} e_{N-\nu} b^{N-\nu}, \tag{E.12}
\]
\(N \in \mathbb{Z}\) and \(d_n, e_n \in \{0, \ldots, b-1\}\) for each \(n \in \{N, N-1, N-2, \ldots\}\). If \(d_N < e_N\), then \(e_N = d_N + 1, d_n = b-1\) for each \(n < N\) and \(e_n = 0\) for each \(n < N\).

**Proof.** By subtracting \(d_N b^N\) from both series, one can assume \(d_N = 0\) without loss of generality. From Lem. E.1, we know
\[
x = \sum_{\nu=0}^{\infty} d_{N-\nu} b^{N-\nu} = \sum_{\nu=0}^{\infty} d_{N-\nu} b^{N-\nu-1} \leq b^N. \tag{E.13a}
\]
On the other hand:
\[
x = \sum_{\nu=0}^{\infty} e_{N-\nu} b^{N-\nu} \geq b^N. \tag{E.13b}
\]
Combining (E.13a) and (E.13b) yields \(x = b^N\). Once again employing Lem. E.1, (E.13a) also shows that \(d_n = b-1\) for each \(n \leq N-1\) as claimed. Since \(e_N > 0\) and \(e_n \geq 0\) for each \(n\), equality in (E.13b) can only occur for \(e_N = 1\) and \(e_n = 0\) for each \(n < N\), thereby completing the proof of the lemma. \(\blacksquare\)

**Notation E.3.** For each \(x \in \mathbb{R}\), we let
\[
[x] := \max\{k \in \mathbb{Z} : k \leq x\} \tag{E.14}
\]
denote the **integral part** of \(x\) (also called **floor** of \(x\) or \(x\) **rounded down**).

**Proof of Th. 7.99.** We start by constructing numbers \(N\) and \(d_n, n \in \{N, N-1, N-2, \ldots\}\), such that (7.97) holds. For \(x = 0\), one chooses an arbitrary \(N \in \mathbb{Z}\) and \(d_n = 0\) for each \(n \in \{N, N-1, N-2, \ldots\}\). Thus, for the remainder of the proof, fix \(x > 0\). Let
\[
N := \max\{n \in \mathbb{Z} : b^n \leq x\}. \tag{E.15}
\]
The numbers $d_{N-n} \in \{0, \ldots, b-1\}$ and $x_n \in \mathbb{R}^+$, $n \in \mathbb{N}_0$, are defined inductively by letting
\[
\begin{align*}
d_N &:= \left\lfloor \frac{x}{b^N} \right\rfloor, & x_0 &:= d_N b^N, \\
d_{N-n} &:= \left\lfloor \frac{x - x_{n-1}}{b^{N-n}} \right\rfloor, & x_n &:= x_{n-1} + d_{N-n} b^{N-n} \quad \text{for } n \geq 1.
\end{align*}
\]

(E.16a) (E.16b)

Claim 3. One can verify by induction on $n$ that the numbers $d_{N-n}$ and $x_n$ enjoy the following properties for each $n \in \mathbb{N}_0$:
\[
\begin{align*}
d_{N-n} &\in \{0, \ldots, b-1\}, \\
0 &< x_n = \sum_{\nu=0}^{n} d_{N-\nu} b^{N-\nu} \leq x, \\
x - x_n &< b^{N-n}.
\end{align*}
\]

(E.17a) (E.17b) (E.17c)

Proof. The induction is carried out for all three statements of (E.17) simultaneously. From (E.15), we know $b^N \leq x < b^{N+1}$, i.e. $1 \leq \frac{x}{b^N} < b$. Using (E.16a), this yields $d_N \in \{1, \ldots, b-1\}$ and $0 < x_0 = d_N b^N = b^N d_N \leq b^N \frac{x}{b^N} = x$ as well as $x - x_0 = x - d_N b^N = b^N (\frac{x}{b^N} - d_N) < b^N$. For $n \geq 1$, by induction, one obtains $0 \leq x - x_{n-1} < b^{1+N-n}$, i.e. $0 \leq \frac{x - x_{n-1}}{b^{N-n}} < b$. Using (E.16b), this yields $d_{N-n} \in \{0, \ldots, b-1\}$ and $x_n = x_{n-1} + d_{N-n} b^{N-n} \leq x_{n-1} + b^{N-n} \frac{x - x_{n-1}}{b^{N-n}} = x$. Moreover, by induction, $0 < x_{n-1} = \sum_{\nu=0}^{n-1} d_{N-\nu} b^{N-\nu}$, such that (E.16b) implies $x_n = x_{n-1} + d_{N-n} b^{N-n} \geq x_{n-1} > 0$ and $x_n = x_{n-1} + d_{N-n} b^{N-n} = d_{N-n} b^{N-n} + \sum_{\nu=0}^{n-1} d_{N-\nu} b^{N-\nu} = d_{N-n} b^{N-n} + \sum_{\nu=0}^{n-1} d_{N-\nu} b^{N-\nu}$. Finally, $x - x_n = x - x_{n-1} - d_{N-n} b^{N-n} = b^{N-n} (\frac{x - x_{n-1}}{b^{N-n}} - d_{N-n}) \leq b^{N-n}$, completing the proof of the claim. \hfill \square

Since, for each $n \in \mathbb{N}_0$,
\[
0 \leq x - x_n < b^{N-n},
\]

(E.18)

and $\lim_{n \to \infty} b^{N-n} = 0$, we have $\lim_{n \to \infty} x_n = x$, thereby establishing (7.97).

It remains to verify the equivalence of (i) – (iv).

(ii) $\Rightarrow$ (i) is trivial.

“(iii) $\Rightarrow$ (i)”: Assume (iii) holds. Without loss of generality, we can assume that $n_0$ is the largest index such that $d_n = 0$ for each $n \leq n_0$. We distinguish two cases. If $n_0 < N - 1$ or $d_N \neq 1$, then
\[
\sum_{\nu=0}^{N-n_0-2} d_{N-\nu} b^{N-\nu} + (d_{n_0+1} - 1) b^{n_0+1} + \sum_{\nu=N-n_0}^{\infty} (b-1) b^{N-\nu}
\]
is a different $b$-adic representation of $x$ and its first coefficient is nonzero. If $n_0 = N - 1$ and $d_N = 1$, then
\[
\sum_{\nu=1}^{\infty} (b-1) b^{N-\nu} = \sum_{\nu=0}^{\infty} (b-1) b^{N-1-\nu}
\]
is a different $b$-adic representation of $x$ and its first coefficient is nonzero.

“(iv) $\Rightarrow$ (i)”: Assume (iv) holds. Without loss of generality, we can assume that $n_0$ is the largest index such that $d_n = b - 1$ for each $n \leq n_0$. Then

$$\sum_{n=0}^{N-n_0-1} d_{N-n} b^{N-n} + (d_{n_0} + 1) b^{n_0+1} + \sum_{n=n_0}^{\infty} 0 b^{N-n}$$

is a different $b$-adic representation of $x$ and its first coefficient is nonzero.

We will now show that, conversely, (i) implies (ii), (iii), and (iv). To that end, let $x > 0$ and suppose that $x$ has two different $b$-adic representations

$$x = \sum_{\nu=0}^{\infty} d_{N_1-\nu} b^{N_1-\nu} = \sum_{\nu=0}^{\infty} e_{N_2-\nu} b^{N_2-\nu} \quad (E.19)$$

with $N_1, N_2 \in \mathbb{Z}$; $d_n, e_n \in \{0, \ldots, b - 1\}$; and $d_{N_1}, e_{N_2} > 0$. This implies

$$x \geq b^{N_1}, \quad x \geq b^{N_2}. \quad (E.20a)$$

Moreover, Lem. E.1 yields

$$x \leq b^{N_1+1}, \quad x \leq b^{N_2+1}. \quad (E.20b)$$

If $N_2 > N_1$, then (E.20) imply $N_2 = N_1 + 1$ and $b^{N_2} \leq x \leq b^{N_1+1} = b^{N_2}$, i.e. $x = b^{N_2} = b^{N_1+1}$. Since $e_{N_2} > 0$, one must have $e_{N_2} = 1$, and, in turn, $e_n = 0$ for each $n < N_2$. Moreover, $x = b^{N_1+1}$ and Lem. E.1 imply that $d_n = b - 1$ for each $n \in \{N_1, N_1 - 1, \ldots \}$. Thus, for $N_2 > N_1$, the value of $N_1$ is determined by $N_2$ and the values of all $d_n$ and $e_n$ are also completely determined, showing that there are precisely two $b$-adic representations of $x$. Moreover, the $d_n$ have the property required in (iv) and the $e_n$ have the property required in (iii). The argument also shows that, for $N_1 > N_2$, one must have $N_1 = N_2 + 1$ with the $e_n$ taking the values of the $d_n$ and vice versa. Once again, there are precisely two $b$-adic representations of $x$; now the $d_n$ have the property required in (iii) and the $e_n$ have the property required in (iv).

It remains to consider the case $N := N_1 = N_2$. Since, by hypothesis, the two $b$-adic representations of $x$ in (E.19) are not identical, there must be a largest index $n \leq N$ such that $d_n \neq e_n$. Thus, (E.19) implies

$$y := \sum_{\nu=0}^{\infty} d_{n-\nu} b^{n-\nu} = \sum_{\nu=0}^{\infty} e_{n-\nu} b^{n-\nu}. \quad (E.21)$$

Now Lem. E.2 shows that there are precisely two $b$-adic representations of $x$, one having the property required in (iii) and the other having property required in (iv).

Thus, in each case ($N_2 > N_1$, $N_1 > N_2$, and $N_1 = N_2$), we find that (i) implies (ii), (iii), and (iv), thereby concluding the proof of the theorem. ■

In most cases, it is understood that we work only with decimal representations such that there is no confusion about the meaning of symbol strings like 101.01. However,
in general, 101.01 could also be meant with respect to any other base, and, the number represented by the same string of symbols does obviously depend on the base used. Thus, when working with different representations, one needs some notation to keep track of the base.

**Notation E.4.** Given a natural number $b \geq 2$ and finite sequences

\[
(d_{N_1}, d_{N_1-1}, \ldots, d_0) \in \{0, \ldots, b - 1\}^{N_1+1}, \quad (E.22a)
\]

\[
(e_1, e_2, \ldots, e_{N_2}) \in \{0, \ldots, b - 1\}^{N_2}, \quad (E.22b)
\]

\[
(p_1, p_2, \ldots, p_{N_3}) \in \{0, \ldots, b - 1\}^{N_3}, \quad (E.22c)
\]

$N_1, N_2, N_3 \in \mathbb{N}_0$ (where $N_2 = 0$ or $N_3 = 0$ is supposed to mean that the corresponding sequence is empty), the respective string

\[
(d_{N_1}d_{N_1-1}\ldots d_0)_b \quad \text{for} \ N_2 = N_3 = 0,
\]

\[
(d_{N_1}d_{N_1-1}\ldots d_0.e_1 \ldots e_{N_2}p_1 \ldots p_{N_3})_b \quad \text{for} \ N_2 + N_3 > 0
\]

represents the number

\[
\sum_{\nu=0}^{N_1} d_\nu b^\nu + \sum_{\nu=1}^{N_2} e_\nu b^{-\nu} + \sum_{\alpha=0}^{\infty} \sum_{\nu=1}^{N_3} p_\nu b^{-N_2-\alpha N_3-\nu}. \quad (E.24)
\]

**Example E.5.** For the number from (7.95), we get

\[
x = (131.\overline{6})_{10} = (10000011.100)_{2} = (83.\overline{A})_{16} \quad (E.25)
\]

(for the hexadecimal system, it is customary to use the symbols 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, A, B, C, D, E, F).

One frequently needs to convert representations with respect to one base into representations with respect to another base. When working with digital computers, conversions between bases 10 and 2 and vice versa are the most obvious ones that come up. Converting representations is related to the following elementary remainder theorem and the well-known long division algorithm.

**Theorem E.6.** For each pair of numbers $(a, b) \in \mathbb{N}^2$, there exists a unique pair of numbers $(q, r) \in \mathbb{N}_0^2$ satisfying the two conditions $a = qb + r$ and $0 \leq r < b$.

**Proof.** Existence: Define

\[
q := \max\{n \in \mathbb{N}_0 : nb \leq a\}, \quad (E.26a)
\]

\[
r := a - qb. \quad (E.26b)
\]

Then $q \in \mathbb{N}_0$ by definition and (E.26b) immediately yields $a = qb + r$ as well as $r \in \mathbb{Z}$. Moreover, from (E.26a), $qb \leq a = qb + r$, i.e. $0 \leq r$, in particular, $r \in \mathbb{N}_0$. Since (E.26a) also implies $(q + 1)b > a = qb + r$, we also have $b > r$ as required.

Uniqueness: Suppose $(q_1, r_1) \in \mathbb{N}_0$, satisfying the two conditions $a = q_1b + r_1$ and $0 \leq r_1 < b$. Then $q_1b = a - r_1 \leq a$ and $(q_1 + 1)b = a - r_1 + b > a$, showing $q_1 = \max\{n \in \mathbb{N}_0 : nb \leq a\} = q$. This, in turn, implies $r_1 = a - q_1b = a - qb = r$, thereby establishing the case.
F  Cardinality of $\mathbb{R}$ and Some Related Sets

Theorem F.1. (a) The set of natural numbers $\mathbb{N}$ is countable.

(b) The set of integers $\mathbb{Z}$ is countable: $\#\mathbb{Z} = \#\mathbb{N}$.

(c) The set of rational numbers $\mathbb{Q}$ is countable: $\#\mathbb{Q} = \#\mathbb{N}$.

Proof. (a): The identity $\text{Id} : \mathbb{N} \to \mathbb{N}$ shows $\mathbb{N}$ is countable.

(b): Using (D.24), the map

$$
\phi : \mathbb{N} \to \mathbb{Z}, \quad \phi(n) := \begin{cases} 
\frac{n}{2} & \text{if } n \text{ is even}, \\
0 & \text{if } n = 1, \\
-\frac{n-1}{2} & \text{if } n \text{ is odd},
\end{cases}
$$

(F.1)
is clearly bijective, proving $\#\mathbb{Z} = \#\mathbb{N}$.

(c): According to (b), $\mathbb{Z}$ and $\mathbb{Z} \setminus \{0\}$ are countable. Then Th. 3.16 implies that $A := \mathbb{Z} \times (\mathbb{Z} \setminus \{0\})$ is countable and there is a bijective map $f : \mathbb{N} \to A$. It is then immediate from Def. D.22(a) that the map

$$
\phi : \mathbb{N} \to \mathbb{Q}, \quad \phi(n) := \left[ f(n) \right],
$$

(F.2)
where $[f(n)]$ denotes the equivalence class of $f(n)$ with respect to $\sim$ from (D.31), is surjective. Thus, $\mathbb{Q}$ is countable by Prop. 3.15.

In the following theorem and its two corollaries, we will see that the set $\mathbb{R}$ of real numbers is not countable, but has the same cardinality as the power set of $\mathbb{N}$. Moreover, the same is true for every nontrivial interval of real numbers.

Theorem F.2. Let $a, b \in \mathbb{R}$ with $a < b$. Recalling the notations $\mathcal{F}(\mathbb{N}, \{0, 1\}) = \{0, 1\}^\mathbb{N}$ for the set of sequences in $\{0, 1\}$, we obtain the following equalities of cardinalities:

$$
\#\mathbb{R} = \#]a, b[ = \#\{0, 1\}^\mathbb{N} = \#\mathcal{P}(\mathbb{N}).
$$

(F.3)

Proof. We divide the proof into the following steps:

(i) $\#\{0, 1\}^\mathbb{N} = \#\mathcal{P}(\mathbb{N})$.

(ii) $\#]0, 1[ = \#\{0, 1\}^\mathbb{N}$.

(iii) $\#] -1, 1[ = \#\mathbb{R}$.

(iv) $\#]a, b[ = \#]0, 1[$.
(i): To prove \( \#\{0,1\}^N = \#\mathcal{P}(N) \), we have to show the existence of a bijective map \( f : \{0,1\}^N \rightarrow \mathcal{P}(N) \). Given \( \sigma \in \{0,1\}^N \), i.e. \( \sigma \) is a function \( \sigma : \mathbb{N} \rightarrow \{0,1\} \), define

\[
\sigma(\sigma) := \sigma^{-1}\{1\} = \{n \in \mathbb{N} : \sigma(n) = 1\}. \tag{F.4}
\]

Then, indeed, \( f : \{0,1\}^N \rightarrow \mathcal{P}(N) \). It remains to show \( f \) is bijective. To verify \( f \) is injective, consider \( \sigma, \tau \in \{0,1\}^N \). If \( \sigma \neq \tau \), then there exists \( n \in \mathbb{N} \) with \( \sigma(n) \neq \tau(n) \). If \( \sigma(n) = 1 \), then \( \tau(n) = 0 \), i.e. \( n \in f(\sigma) \), but \( n \notin f(\tau) \), showing \( f(\sigma) \neq f(\tau) \). Analogously, if \( \sigma(n) = 0 \), then \( \tau(n) = 1 \), i.e. \( n \in f(\tau) \), but \( n \notin f(\sigma) \), again showing \( f(\sigma) \neq f(\tau) \), concluding the proof that \( f \) is injective. To verify \( f \) is surjective, for each \( A \in \mathcal{P}(\mathbb{N}) \), define

\[
\sigma_A : \mathbb{N} \rightarrow \{0,1\}, \quad \sigma_A(n) := \begin{cases} 1 & \text{if } n \in A, \\ 0 & \text{if } n \notin A. \end{cases} \tag{F.5}
\]

Then \( \sigma_A \in \{0,1\}^N \) and \( f(\sigma_A) = \sigma_A^{-1}\{1\} = A \), proving \( f \) is surjective.

(ii): To prove \( \#\{0,1\}^N \neq \#\mathbb{N} \), we have to show the existence of a bijective map \( f : \{0,1\}^N \rightarrow [0,1] \). The map

\[
g : \{0,1\}^N \rightarrow [0,1], \quad g((x_i)_{i \in \mathbb{N}}) := \sum_{i=1}^{\infty} x_i 2^{-i}, \tag{F.6}
\]

is well-defined by Lem. E.1 (i.e. \( 0 \leq g \leq 1 \)). Moreover, according to Th. 7.99, \( g \) is surjective, but not injective, as there are numbers \( x \in [0,1] \), that have two different dual (i.e. 2-adic) representations. However, as there are only countably many such numbers, we can use a modification to obtain our desired \( f \). In preparation, we define, for each \( n \in \mathbb{N} \), the sequences \( e_n := (e_{ni})_{i \in \mathbb{N}} \) and \( f_n := (f_{ni})_{i \in \mathbb{N}} \), where

\[
\forall_{n,i \in \mathbb{N}} e_{ni} := \begin{cases} 1 & \text{for } i = n, \\ 0 & \text{for } i \neq n, \end{cases} \tag{F.7a}
\]

\[
\forall_{n,i \in \mathbb{N}} f_{ni} := \begin{cases} 1 & \text{for } i > n, \\ 0 & \text{for } i \leq n, \end{cases} \tag{F.7b}
\]

and we note

\[
g((0,0,\ldots)) = 0, \tag{F.8a}
\]

\[
g((1,1,\ldots)) = 1, \tag{F.8b}
\]

\[
g(e_n) = g(f_n) = 2^{-n} \text{ for each } n \in \mathbb{N}. \tag{F.8c}
\]

We are now in a position to define

\[
f : \{0,1\}^N \rightarrow [0,1], \quad f((x_i)_{i \in \mathbb{N}}) := \begin{cases} 2^{-1} & \text{if } (x_i)_{i \in \mathbb{N}} = (0,0,\ldots), \\ 2^{-2} & \text{if } (x_i)_{i \in \mathbb{N}} = (1,1,\ldots), \\ 2^{-2(n+1)} & \text{if } x_i = e_{ni} \text{ for each } i \in \mathbb{N}, \\ 2^{-2(n+2)} & \text{if } x_i = f_{ni} \text{ for each } i \in \mathbb{N}, \\ \sum_{i=1}^{\infty} x_i 2^{-i} & \text{otherwise}. \tag{F.9}
\]

and
Introducing the auxiliary sets

\[ A := \{(0,0,\ldots), (1,1,\ldots)\} \cup \{e_n : n \in \mathbb{N}\} \cup \{f_n : n \in \mathbb{N}\}, \quad (F.10a) \]

\[ B := \{2^{-n} : n \in \mathbb{N}\}, \quad (F.10b) \]

it follows from Th. 7.99 that (the following restrictions of \( f \) which, to simplify notation, we also denote by \( f \))

\[ f : \{0,1\}^\mathbb{N} \setminus A \rightarrow [0,1[ \setminus B, \quad (F.11a) \]

and

\[ f : A \rightarrow B \quad (F.11b) \]

are bijective, i.e. the full \( f \) of (F.9) is itself bijective, completing the proof of (ii).

(iii): To prove \( \#[1,1[ \rightarrow \#\mathbb{R} \), we have to show the existence of a bijective map \( f : \mathbb{R} \rightarrow ]-1,1[. \) Since we know from Def. and Rem. 8.27 that \( \arctan : \mathbb{R} \rightarrow ]-\pi/2,\pi/2[ \) is bijective, we can define

\[ f : \mathbb{R} \rightarrow ]0,1[, \quad f(x) := \frac{2 \arctan x}{\pi}. \quad (F.12) \]

However, even though this provides a valid proof, \( \arctan \) is a somewhat complicated function (as it is defined via \( \sin \) and \( \cos \), which are defined via power series). Thus, it might be desirable to see an alternative proof, using a more elementary \( f \). We claim that

\[ f : \mathbb{R} \rightarrow ]-1,1[, \quad f(x) := \frac{x}{|x| + 1}, \quad (F.13) \]

is also bijective. Since \( f \) is clearly continuous, according to the intermediate value Th. 7.57, it suffices to show

\[ \forall \epsilon \in ]0,1[ \exists x_1,x_2 \in \mathbb{R} f(x_1) < -1 + \epsilon < 1 - \epsilon < f(x_2). \quad (F.14) \]

However, for each \( \epsilon \in ]0,1[ \),

\[ x_1 < \frac{-1 + \epsilon}{\epsilon} = -\epsilon^{-1} + 1 \quad \Rightarrow \quad x_1 < x_1 - 1 - \epsilon x_1 + \epsilon \quad \Rightarrow \quad f(x_1) = \frac{x_1}{-x_1 + 1} < -1 + \epsilon, \]

\[ x_2 > \frac{1 - \epsilon}{\epsilon} = \epsilon^{-1} - 1 \quad \Rightarrow \quad x_2 > 1 + x_2 - \epsilon - \epsilon x_2 \quad \Rightarrow \quad f(x_2) = \frac{x}{x + 1} > 1 - \epsilon, \]

proving (F.14) and the surjectivity of \( f \). To verify \( f \) is injective, it suffices to show that \( f \) is strictly increasing. Since

\[ x_1 \leq 0 \leq x_2 \wedge x_1 < x_2 \quad \Rightarrow \quad f(x_1) = \frac{x_1}{-x_1 + 1} \leq 0 \leq \frac{x_2}{x_2 + 1} = f(x_2) \]

\[ f(x_1) < f(x_2), \]

\[ x_1 < x_2 \leq 0 \quad \Rightarrow \quad -x_1 x_2 + x_1 < -x_1 x_2 + x_2 \]

\[ \Rightarrow \quad f(x_1) = \frac{x_1}{-x_1 + 1} < \frac{x_2}{-x_2 + 1} = f(x_2), \]

\[ 0 \leq x_1 < x_2 \quad \Rightarrow \quad x_1 x_2 + x_1 < x_1 x_2 + x_2 \]

\[ \Rightarrow \quad f(x_1) = \frac{x_1}{x_1 + 1} < \frac{x_2}{x_2 + 1} = f(x_2), \]
showing \( f \) is strictly increasing and, hence, injective.

(iv): To prove \( \# [a, b[ = \# ]0, 1[ \), we have to show the existence of a bijective map \( f : [a, b[ \rightarrow ]0, 1[ \). Such a bijective map is given by the (restriction of an) affine map

\[
f : [a, b[ \rightarrow ]0, 1[ , \quad f(x) := \frac{x - a}{b - a} .
\]

The proof that \( f \) is bijective can be conducted analogous to (but much simpler than) the proof in (iii), or one can use (for example, from Linear Algebra) that every nonconstant affine map from \( \mathbb{R} \) into \( \mathbb{R} \) is bijective.

\[\text{Corollary F.3.} \quad \# \mathbb{R} = \# \mathcal{P}(\mathbb{N}) \quad \text{in particular,} \quad \mathbb{R} \text{ is not countable.}\]

\[\text{Proof.} \quad \# \mathbb{R} = \# \mathcal{P}(\mathbb{N}) \text{ was proved in Th. F.2 and } \mathcal{P}(\mathbb{N}) \text{ is uncountable by Th. A.69.} \]

\[\text{Corollary F.4.} \quad \text{If } a, b \in \mathbb{R} \text{ with } a < b, \text{ then } \# (\mathbb{Q} \cap [a, b[) = \# \mathbb{N} \text{ and } \# ([a, b[ \setminus \mathbb{Q}) = \# \mathbb{R}, \text{ i.e. } [a, b[ \text{ contains countably many rational and uncountably many irrational numbers.}\]

\[\text{Proof.} \quad \text{Since } \mathbb{Q} \cap [a, b[ \subseteq \mathbb{Q}, \text{ the claim } \# (\mathbb{Q} \cap [a, b[) = \# \mathbb{N} \text{ follows from Th. F.1(c), Prop. 3.14, and Th. 7.68(a).} \]

\[\text{To prove } \# ([a, b[ \setminus \mathbb{Q}) = \# \mathbb{R}, \text{ a bijection between } [a, b[ \setminus \mathbb{Q} \text{ and } \mathbb{R} \text{ can be constructed analogous to the construction of } f \text{ in step (ii) of the proof of Th. F.2, making use of the fact that } \# [a, b[ = \# \mathbb{R} \text{ and } \# \mathbb{Q} = \# \mathbb{N}. \]

\[\text{Theorem F.5.} \quad \text{The set of complex numbers } \mathbb{C} = \mathbb{R} \times \mathbb{R} \text{ has the same cardinality as } \mathbb{R}: \]

\[\# (\mathbb{R} \times \mathbb{R}) = \# \mathbb{R} = \# \mathcal{P}(\mathbb{N}).\]

\[\text{Proof.} \quad \text{Let } A := \{0, 1\}^\mathbb{N}. \]

\[\text{By an application of Th. F.2, it suffices to prove } \# A = \# (A \times A), \text{ which is accomplished by showing the existence of a bijective map } f : A \rightarrow A \times A. \text{ We define}
\]

\[
f : A \rightarrow A \times A , \quad f((x_j)_{j \in \mathbb{N}}) := ((y_j)_{j \in \mathbb{N}}, (z_j)_{j \in \mathbb{N}}) .
\]

\[\text{where}
\]

\[\forall j \in \mathbb{N} \quad y_j := x_{2j-1} , \quad \text{(F.17b)}
\]

\[\forall j \in \mathbb{N} \quad z_j := x_{2j} , \quad \text{(F.17c)}
\]

\[\text{and}
\]

\[g : A \times A \rightarrow A , \quad g((y_j)_{j \in \mathbb{N}}, (z_j)_{j \in \mathbb{N}}) := (x_j)_{j \in \mathbb{N}} , \quad \text{(F.18a)}
\]

\[\text{where}
\]

\[\forall j \in \mathbb{N} \quad x_j := \begin{cases} y_{(j+1)/2} & \text{for } j \text{ odd,} \\ z_{j/2} & \text{for } j \text{ even.} \end{cases} \quad \text{(F.18b)}
\]

\[\text{Clearly, } g = f^{-1}, \text{ proving that } f \text{ is bijective as desired.} \]
G Partial Fraction Decomposition

We consider $\mathbb{C}$-valued rational functions of the form

$$z \mapsto R(z) := \frac{P(z)}{Q(z)},$$

(G.1)

where $P, Q : \mathbb{C} \rightarrow \mathbb{C}$ are polynomials such that $\deg(P) < \deg(Q) =: n$. Using Cor. 8.33 as well as Rem. 6.7, we write $Q$ in the form

$$Q(z) = c \prod_{j=1}^{k} (z - \lambda_j)^{m_j},$$

(G.2)

where $c \in \mathbb{C}$, and $\lambda_1, \ldots, \lambda_k \in \mathbb{C}$, $k \in \{1, \ldots, n\}$, are the distinct zeros of $Q$, $m_j \in \mathbb{N}$ with $\sum_{j=1}^{k} m_j = n$ being their respective multiplicities.

In can be useful to write $R$ as a linear combination of the so-called partial fractions

$$\frac{1}{z - \lambda_j}, \frac{1}{(z - \lambda_j)^2}, \ldots, \frac{1}{(z - \lambda_j)^{m_j}} \quad (j = 1, \ldots, k)$$

(G.3)

(for example, for the computation of the antiderivative of $R$, cf. Ex. 10.22(b)). The following Th. G.1 guarantees this is always possible:

**Theorem G.1.** Let $P, Q : \mathbb{C} \rightarrow \mathbb{C}$ be polynomials such that $\deg(P) < \deg(Q) =: n$. Moreover, let $N(Q)$ denote the set of zeros of $Q$, and assume $Q$ to have the form of (G.2). Then there exists a unique family of coefficients

$$a_{jl} \in \mathbb{C} \quad (j = 1, \ldots, k, \ l = 1, \ldots, m_j),$$

(G.4)

such that

$$\forall z \in \mathbb{C} \setminus N(Q) \quad R(z) = \frac{P(z)}{Q(z)} = \sum_{j=1}^{k} \sum_{l=1}^{m_j} \frac{a_{jl}}{(z - \lambda_j)^l}.$$  

(G.5)

**Proof.** We first prove the existence of the decomposition (G.5) via induction on $n = \deg(Q)$. If $n = 1$, then $P$ must be constant and there is nothing to prove. For the induction step, consider $n \geq 2$. Let $\zeta$ be a zero of $Q$ with multiplicity $m \in \{1, \ldots, n\}$. Then, according to Rem. 6.7, there exists a polynomial $S : \mathbb{C} \rightarrow \mathbb{C}$ such that $Q(z) = (z - \zeta)^m S(z)$ and $S(\zeta) \neq 0$. Noting

$$\tilde{R}(z) := \frac{P(z)}{S(z)} - \frac{P(\zeta)}{S(\zeta)} = \frac{P(z)S(\zeta) - S(z)P(\zeta)}{S(z)S(\zeta)}.$$  

(G.6)

and that $P(z)S(\zeta) - S(z)P(\zeta)$ vanishes for $z = \zeta$, there exists a polynomial $T : \mathbb{C} \rightarrow \mathbb{C}$, $\deg T \leq n - 2$, such that

$$\tilde{R}(z) = \frac{(z - \zeta)T(z)}{S(z)}.$$  

(G.7)
Thus, for each \( z \in \mathbb{C} \setminus \mathcal{N}(Q) \), we have

\[
R(z) - \frac{P(\zeta)}{(z - \zeta)^m S(\zeta)} = \frac{\tilde{R}(z)}{(z - \zeta)^m} = \frac{T(z)}{(z - \zeta)^{m-1} S(z)}.
\]

(G.8)

We will now apply (G.8) with \( \zeta = \lambda_k \) and \( m = m_k \). Since \( \text{deg}(T) < n - 1 = \text{deg}((z - \zeta)^{m-1} S(z)) \), the induction hypothesis applies to the function in (G.8), yielding coefficients \( a_{jl} \in \mathbb{C}, j = 1, \ldots, k, l = 1, \ldots, m_j \) for \( j < k, l = 1, \ldots, m_j - 1 \) for \( j = k \), satisfying

\[
R(z) - \frac{P(\lambda_k)}{(z - \lambda_k)^{m_k} S(\lambda_k)} = \sum_{j=1}^{k-1} \sum_{l=1}^{m_j} \frac{a_{jl}}{(z - \lambda_j)^l} + \sum_{l=1}^{m_k-1} \frac{a_{kl}}{(z - \lambda_k)^l},
\]

(G.9)

thereby completing the induction for the existence proof.

It remains to prove the uniqueness of the coefficients \( a_{jl} \) in (G.5). Thus, suppose one has \( b_{jl} \in \mathbb{C}, j = 1, \ldots, k, l = 1, \ldots, m_j \), such that

\[
\sum_{j=1}^{k} \sum_{l=1}^{m_j} \frac{a_{jl}}{(z - \lambda_j)^l} = \sum_{j=1}^{k} \sum_{l=1}^{m_j} \frac{b_{jl}}{(z - \lambda_j)^l}.
\]

(G.10)

We fix \( j_0 \) and prove \( a_{j_0l} = b_{j_0l} \) via induction on \( l = 1, \ldots, m_{j_0} \): Let \( l \in \{1, \ldots, m_{j_0}\} \) and assume \( a_{j_0\alpha} = b_{j_0\alpha} \) has already been shown for each \( \alpha > l \) (the induction does, indeed, start at \( l = m_{j_0} \), working itself down to \( l = 1 \)). Then (G.10) implies

\[
\sum_{j=1}^{l} \frac{a_{j_0\beta}}{(z - \lambda_{j_0})^\beta} + \sum_{j=1}^{k} \sum_{\beta=1}^{m_j} \frac{a_{j\beta}}{(z - \lambda_j)^\beta} = \sum_{j=1}^{l} \frac{b_{j_0\beta}}{(z - \lambda_{j_0})^\beta} + \sum_{j=1}^{k} \sum_{\beta=1}^{m_j} \frac{b_{j\beta}}{(z - \lambda_j)^\beta}.
\]

(G.11)

One now multiplies (G.11) by \( (z - \lambda_{j_0})^l \). Then taking the limit for \( z \to \lambda_{j_0} \) on both sides yields \( a_{j_0l} = b_{j_0l} \) as desired.

If, in (G.1), \( P, Q : \mathbb{R} \to \mathbb{R} \), then the partial fraction decomposition (G.5) of Th. G.1 is not quite satisfactory, since, even though \( P \) and \( Q \) are both real, the \( a_{jl} \) will typically be nonreal elements of \( \mathbb{C} \). As the following Th. G.2 shows, if \( P, Q \) are real, then it is always possible to obtain a partial fraction decomposition with only real coefficients, however its form is somewhat more complicated.

We start by using the real factorization of \( Q : \mathbb{R} \to \mathbb{R}, \) \( \text{deg}(Q) = n \in \mathbb{N} \), according to (8.58), where, as in (G.2), we combine identical factors, obtaining

\[
Q(x) = c \prod_{j=1}^{k_1} (x - \lambda_j)^{m_j} \prod_{j=1}^{k_2} (x^2 + \alpha_j x + \beta_j)^{n_j},
\]

(G.12)

where \( c \in \mathbb{R}; \lambda_1, \ldots, \lambda_{k_1} \in \mathbb{R}, k_1 \in \{0, \ldots, n\} \), are the distinct real zeros of \( Q \) (if any), \( m_j \in \mathbb{N} \) being their respective multiplicities; and \( (\alpha_1, \beta_1), \ldots, (\alpha_{k_2}, \beta_{k_2}) \in \mathbb{R}^2 \)
If $k_2 \in \{0, \ldots, n\}$, are distinct pairs of real numbers, each pair arising from combining two conjugate nonreal zeros of $Q$ according to (8.60), $n_j \in \mathbb{N}$ being their respective multiplicities;

$$\sum_{j=1}^{k_1} m_j + 2 \sum_{j=1}^{k_2} n_j = n. \quad (G.13)$$

**Theorem G.2.** Let $P, Q : \mathbb{R} \rightarrow \mathbb{R}$ be polynomials such that $\deg(P) < \deg(Q) = n$. Moreover, let $\mathcal{N}(Q)$ denote the set of zeros of $Q$, and assume $Q$ to have the form of (G.12). Then there exist families of coefficients

$$a_{jl} \in \mathbb{R} \quad (j = 1, \ldots, k_1, \ l = 1, \ldots, m_j), \quad (G.14a)$$
$$b_{jl} \in \mathbb{R} \quad (j = 1, \ldots, k_2, \ l = 1, \ldots, n_j), \quad (G.14b)$$
$$c_{jl} \in \mathbb{R} \quad (j = 1, \ldots, k_2, \ l = 1, \ldots, n_j) \quad (G.14c)$$

such that

$$\forall x \in \mathbb{R} \setminus \mathcal{N}(Q) \quad R(x) = \frac{P(x)}{Q(x)} = \sum_{j=1}^{k_1} \sum_{t=1}^{m_j} \frac{a_{jt}}{(x - \lambda_j)^t} + \sum_{j=1}^{k_2} \sum_{t=1}^{n_j} \frac{b_{jt} x + c_{jt}}{(x^2 + \alpha_j x + \beta_j)^t}. \quad (G.15)$$

**Proof.** We show that, if $P, Q$ are real, where $Q$ is as in (G.12), then (G.5) can be rewritten in the form (G.15): First, consider $\lambda_{j_0} \in \mathbb{R}$ to be a real zero of $Q$. Then all corresponding coefficients in (G.5) are real: We prove $\alpha > l$ such that

$\lambda_\alpha \neq \lambda_{j_0} \in \mathbb{C}$ to be a real zero of $Q$. Then all corresponding coefficients in (G.5) are conjugate: We prove $\nu_{j_0 \alpha} \in \mathbb{R}$ via induction on $l = 1, \ldots, m_{j_0}$: Let $l \in \{1, \ldots, m_{j_0}\}$ and assume $\nu_{j_0 \alpha} \in \mathbb{R}$ has already been shown for each $\alpha > l$. Then (G.5) (with $z$ replaced by $x$) implies

$$\forall x \in \mathbb{R} \setminus \mathcal{N}(Q) \quad S(x) := R(x) - \sum_{\beta=l+1}^{m_j} a_{j_0 \beta} \frac{(x - \lambda_{j_0})^\beta}{(x - \lambda_{j_0})^\beta} = \sum_{\beta=1}^{l} \frac{a_{j_0 \beta}}{(x - \lambda_{j_0})^\beta} + \sum_{\beta=1}^{k} \sum_{j \neq j_0} \frac{a_{j \beta}}{(x - \lambda_j)^\beta} \in \mathbb{R}. \quad (G.16)$$

One now multiplies (G.16) by $(x - \lambda_{j_0})^l$. Then taking the limit for $x \to \lambda_{j_0}$ on both sides yields $a_{j_0 l} \in \mathbb{R}$ as desired (the limit on the right-hand side is clearly $a_{j_0 l}$ and all values and, thus, the limit on the left-hand side are clearly in $\mathbb{R}$).

Thus, the summands corresponding to real zeros of $Q$ are identical in (G.5) and (G.15). It remains to show that terms in (G.5), corresponding to conjugate nonreal zeros of $Q$, can be combined to result in the summands involving the $b_{jl}$ and $c_{jl}$ in (G.15). To this end, consider $\lambda_{j_0}, \lambda_{j_1} \in \mathbb{C}$ to be conjugate nonreal zeros of $Q$, $\lambda_{j_1} = \overline{\lambda}_{j_0}$. Then all corresponding coefficients in (G.5) are conjugate: We prove $\nu_{j_1 \alpha} = \overline{\nu_{j_1 \alpha}}$ via induction on $l = 1, \ldots, m_{j_0} = m_{j_1}$: Let $l \in \{1, \ldots, m_{j_0}\}$ and assume $\nu_{j_0 \alpha} \in \mathbb{R}$ has already been shown for each $\alpha > l$. We once again have the formula (G.16) for $S(x)$ (even for each $x \in \mathbb{C} \setminus \mathcal{N}(Q)$, but we can no longer expect $S(x) \in \mathbb{R}$). As before, after multiplying (G.16) by $(x - \lambda_{j_0})^l$, we obtain

$$\lim_{x \to \lambda_{j_0}} (S(x)(x - \lambda_{j_0})^l) = a_{j_0 l}. \quad (G.17)$$
G \hspace{1em} \text{PARTIAL FRACTION DECOMPOSITION} \hspace{1em} 251

Analogously, we also have

\[ R(x) - \sum_{\beta=1}^{l} \frac{a_{j_{1} \beta}}{(x - \lambda_{j_{1}})^{\beta}} = \sum_{\beta=1}^{l} \frac{a_{j_{1} \beta}}{(x - \lambda_{j_{1}})^{\beta}} + \sum_{j=1, \beta=1}^{m_{j}} \frac{a_{j \beta}}{(x - \lambda_{j})^{\beta}}. \]  

(G.18)

Taking complex conjugates in (G.16) and using the induction hypothesis as well as \( R(x) = R(\bar{x}) \) (since the coefficients of \( P, Q \) are real) yields

\[ \forall x \in \mathbb{C}\setminus\mathcal{N}(Q) \quad \overline{S(x)} = R(\bar{x}) - \sum_{\beta=1}^{l} \frac{a_{j_{1} \beta}}{(\bar{x} - \lambda_{j_{1}})^{\beta}} = \sum_{\beta=1}^{l} \frac{a_{j_{1} \beta}}{(\bar{x} - \lambda_{j_{1}})^{\beta}} + \sum_{j=1, \beta=1}^{m_{j}} \frac{a_{j \beta}}{(\bar{x} - \lambda_{j})^{\beta}}. \]  

(G.19)

If we multiply (G.19) by \((\bar{x} - \lambda_{j_{1}})^{l}\), we obtain

\[ \lim_{x \to \lambda_{j_{1}}} (\overline{S(x)}(\bar{x} - \lambda_{j_{1}})^{l}) = a_{j_{1} l}. \]  

(G.20)

Thus,

\[ a_{j_{1} l} \overset{(G.17)}{=} \lim_{x \to \lambda_{j_{0}}} (S(x)(x - \lambda_{j_{0}})^{l}) = \lim_{x \to \lambda_{j_{0}}} (\overline{S(x)}(\bar{x} - \lambda_{j_{0}})^{l}) \overset{(G.20)}{=} a_{j_{1} l}, \]  

(G.21)

as needed.

We now combine two corresponding summands of (G.5) (for \( x \in \mathbb{R} \setminus \mathcal{N}(Q) \)):

\[ \sigma_{l} := \frac{a_{j_{0} l}}{(x - \lambda_{j_{0}})^{l}} + \frac{\bar{a}_{j_{0} l}}{(x - \lambda_{j_{0}})^{l}} = \frac{a_{j_{0} l}(x - \lambda_{j_{0}})^{l} + \bar{a}_{j_{0} l}(x - \lambda_{j_{0}})^{l}}{(x^{2} - 2x \Re \lambda_{j_{0}} + |\lambda_{j_{0}}|^{2})^{l}} = \frac{a(x - \bar{x})^{l} + \bar{a}(x - \lambda)^{l}}{(x^{2} + bx + c)^{l}}, \]  

(G.22)

where we have set

\[ a := a_{j_{0} l}, \quad \lambda := \lambda_{j_{0}}, \quad b := -2 \Re \lambda_{j_{0}}, \quad c := |\lambda_{j_{0}}|^{2}, \]  

(G.23)

to simplify notation. To finish the proof of (G.15), it remains to show there are real coefficients \( s_{1 l}, \ldots, s_{ll} \) and \( t_{1 l}, \ldots, t_{ll} \) such that

\[ \forall \quad \sigma_{l} = \sum_{\beta=1}^{l} \frac{s_{\beta l} x + t_{\beta l}}{(x^{2} + bx + c)^{\beta}}, \]  

(G.24)

which we prove via induction on \( l \): For \( l = 1 \), we have

\[ \sigma_{1} = \frac{a(x - \bar{x}) + \bar{a}(x - \lambda)}{x^{2} + bx + c} = \frac{(a + \bar{a})x - (\bar{a} \lambda + a \bar{x})}{x^{2} + bx + c}, \]  

(G.25)

which fits the requirements of (G.24). For the induction step, we consider, for \( l = 1, \ldots, m_{j_{0}} - 1 \),

\[ \sigma_{l+1} = \frac{a(x - \bar{x})^{l+1} + \bar{a}(x - \lambda)^{l+1}}{(x^{2} + bx + c)^{l+1}}. \]  

(G.26)
The numerator can be rewritten as

\[
\begin{align*}
a(x - \overline{\lambda})^{l+1} + \overline{a}(x - \lambda)^{l+1} &= a(x - \overline{\lambda})^{l+1} + \overline{a}(x - \lambda)^{l+1}(x - \overline{\lambda}) \\
&\quad - \overline{a}(x - \lambda)^{l}(x - \overline{\lambda}) - a(x - \overline{\lambda})^{l}(x - \lambda) \\
&\quad + a(x - \overline{\lambda})^{l}(x - \lambda) + \overline{a}(x - \lambda)^{l+1}.
\end{align*}
\] (G.27)

Thus, \(\sigma_{l+1} = S_1 + S_2 + S_3\), where

\[
\begin{align*}
S_1 &= \frac{(a(x - \overline{\lambda})^l + \overline{a}(x - \lambda)^l)(x - \overline{\lambda})}{(x^2 + bx + c)^{l+1}} = \frac{\sigma_l(x - \overline{\lambda})}{x^2 + bx + c}, \quad (G.28) \\
S_2 &= \frac{-(\overline{a}(x - \lambda)^{l-1} + a(x - \overline{\lambda})^{l-1})(x - \lambda)(x - \overline{\lambda})}{(x^2 + bx + c)^{l+1}} = \frac{T_{l-1}}{x^2 + bx + c}, \quad (G.29) \\
S_3 &= \frac{(a(x - \overline{\lambda})^l + \overline{a}(x - \lambda)^l)(x - \lambda)}{(x^2 + bx + c)^{l+1}} = \frac{\sigma_l(x - \lambda)}{x^2 + bx + c}, \quad (G.30)
\end{align*}
\]

where, for (G.29) to hold for \(l = 1\), we set \(\sigma_0 := a + \overline{a} \in \mathbb{R}\). Using the induction hypothesis, \(S_2\) clearly has the form required by (G.24). Using the induction hypothesis together with the elementary equality

\[
\frac{sx^2}{x^2 + bx + c} = s - \frac{sbx + sc}{x^2 + bx + c}, \quad (G.31)
\]

we also obtain

\[
\begin{align*}
S_1 + S_3 &= \frac{\sigma_l(2x + b)}{x^2 + bx + c} = \frac{2x + b}{x^2 + bx + c} \sum_{\beta = 1}^{l} \frac{s_{\beta}x + t_{\beta}}{(x^2 + bx + c)^{\beta}} \\
&= \sum_{\beta = 1}^{l} \frac{2s_{\beta}x^2 + (bs_{\beta} + 2t_{\beta})x + bt_{\beta}}{(x^2 + bx + c)^{\beta}(x^2 + bx + c)} \\
&\quad \quad \quad + \sum_{\beta = 1}^{l} \frac{1}{(x^2 + bx + c)^{\beta}} \left( \frac{2s_{\beta}x^2 - 2s_{\beta}(bx + c)}{x^2 + bx + c} \right) \\
&\quad \quad \quad + \sum_{\beta = 1}^{l} \frac{(bs_{\beta} + 2t_{\beta})x + bt_{\beta}}{(x^2 + bx + c)^{\beta+1}} \quad (G.32)
\end{align*}
\]

to have the form required by (G.24), thereby finishing the induction and the proof of the theorem. \(\blacksquare\)

**Remark G.3.** Given a rational function \(R = P/Q\) as in Th. G.1 (or Th. G.2), there remains the question of how to actually compute the coefficients \(a_{\beta j}\) of the partial fraction decomposition (G.5) (or \(a_{\beta j}, b_{\beta j}, c_{\beta j}\) of the partial fraction decomposition (G.15) in the real case)? First, one always needs to obtain the zeros \(\lambda_j\) and their respective multiplicities \(m_j\), which, for \(\deg(Q)\) large, can be very difficult. Then there are basically three different possibilities to proceed, where, in practise, the most efficient way in a concrete situation might be to mix the three strategies:
(a) Linear System: To determine \( k \) unknown coefficients, one can plug \( k \) different values for \( z \) into (G.5) (or for \( x \) into (G.15)) to obtain a linear system for the unknown coefficients.

(b) One can multiply (G.5) (or (G.15)) by \( Q \), obtaining a polynomial on both sides of the equation. As the polynomials need to be equal, the coefficients of equal powers need to be equal on both sides, yielding a system of equations for the unknown coefficients.

(c) Multiplying (G.5) (or (G.15)) by \((z-\lambda_j)^m_j\) and setting \( z = \lambda_j \) yields the coefficient of \( \frac{1}{(z-\lambda_j)^m_j} \), etc.

\[ H \quad \text{Irrationality of } e \text{ and } \pi \]

\[ H.1 \quad \text{Irrationality of } e \]

The following Prop. H.1, which will then be used to prove the irrationality of \( e \) in Th. H.2, shows, in particular, that the series (8.26) can be used to efficiently compute accurate approximations of \( e \).

Proposition H.1. Defining

\[
R_n(z) := e^z - \sum_{j=0}^{n-1} \frac{z^j}{j!}, \quad (H.1)
\]

we have

\[
\forall n \in \mathbb{N} \quad \forall z \in \mathbb{C} \quad \left( |z| \leq 1 \Rightarrow |R_n(z)| \leq \frac{2|z|^n}{n!} \right), \quad (H.2)
\]

i.e. the error made when approximating \( e^z \) by the partial sum (for \( |z| \leq 1 \)) is at most as large as twice the modulus of the first missing summand.

Proof. One estimates, for each \( n \in \mathbb{N} \) and each \( z \in \mathbb{C} \) with \( |z| \leq 1 \),

\[
|R_n(z)| \leq \sum_{j=n}^{\infty} \frac{|z|^j}{j!} = \frac{|z|^n}{n!} \left( 1 + \frac{|z|}{n+1} + \frac{|z|^2}{(n+1)(n+2)} + \ldots \right)
\]

\[
\leq \frac{|z|^n}{n!} \left( 1 + \frac{1}{2} + \frac{1}{2^2} + \ldots \right) \leq \frac{2|z|^n}{n!}, \quad (H.3)
\]

which establishes the case.

Theorem H.2. Euler’s number \( e \) is irrational.
Proof. Seeking a contradiction, we assume $e$ to be rational. Then there exist $m,n \in \mathbb{N}$ with $n \geq 2$ such that $e = \frac{m}{n}$. Then $n! e \in \mathbb{N}$ and, thus,

$$n! R_{n+1}(1) \overset{(H.1)}{=} n! e - n! \sum_{j=0}^{n} \frac{1}{j!} \in \mathbb{Z}, \quad (H.4)$$

in contradiction to $0 < |n! R_{n+1}(1)| < \frac{2}{n+1} < 1$, which holds according to (H.2) (recalling $n \geq 2$).

\[\blacksquare\]

H.2 Irrationality of $\pi$

Theorem H.3. $\pi^2$ is irrational (then, in particular, $\pi$ must be irrational as well).

Proof. Seeking a contradiction, we assume $\pi^2$ to be rational. Then

$$\exists \ a, b \in \mathbb{N} \ \pi^2 = \frac{a}{b}. \quad (H.5)$$

We can then choose some even $n \in \mathbb{N}$ satisfying

$$0 < \frac{\pi a^n}{n!} < 1. \quad (H.6)$$

We now consider the function

$$f : \mathbb{R} \rightarrow \mathbb{R}, \quad f(x) := \frac{x^n (1-x)^n}{n!} \overset{(*)}{=} \frac{1}{n!} \sum_{k=n}^{2n} (-1)^k \binom{n}{k-n} x^k, \quad (H.7)$$

where the equality at $(*)$ is proved by

$$\frac{x^n (1-x)^n}{n!} = \frac{x^n}{n!} \sum_{k=0}^{n} (-1)^k \binom{n}{k} x^k \overset{n \text{ even}}{=} \frac{1}{n!} \sum_{k=n}^{2n} (-1)^k \binom{n}{k-n} x^k. \quad (H.8)$$

Thus, for the polynomial $f$, we obtain the derivatives

$$f^{(j)}(0) = \begin{cases} 0 & \text{for } 0 \leq j < n, \\ \frac{j!}{n!} (-1)^j \binom{n}{j-n} & \text{for } n \leq j \leq 2n, \\ 0 & \text{for } 2n < j. \end{cases} \quad (H.9)$$

In consequence, since, for $n \leq j \leq 2n$, $\frac{j!}{n!} \in \mathbb{N}$ and $\binom{n}{j-n} \in \mathbb{N}$,

$$\forall \ j \in \mathbb{N}_0 \ f^{(j)}(0) \in \mathbb{Z}. \quad (H.10)$$

Moreover, since $f(1-x) = f(x)$ for each $x \in \mathbb{R}$, and, thus, $f^{(j)}(1-x) = (-1)^j f^{(j)}(x)$ for each $x \in \mathbb{R}$, we also have

$$\forall \ j \in \mathbb{N}_0 \ f^{(j)}(1) \in \mathbb{Z}. \quad (H.11)$$
Next, we consider another polynomial, namely

\[ g : \mathbb{R} \rightarrow \mathbb{R}, \quad g(x) := b^n \sum_{k=0}^{n} (-1)^k \pi^{2(n-k)} f^{(2k)}(x). \] (H.12)

Due to (H.5), (H.10), (H.11), and (H.12), we have

\[ \forall j \in \mathbb{N}_0 \left( g(0) \in \mathbb{Z} \land g(1) \in \mathbb{Z} \right). \] (H.13)

For each \( x \in \mathbb{R} \), one calculates

\[ g''(x) + \pi^2 g(x) = b^n \sum_{k=0}^{n} (-1)^k \pi^{2(n-k)} f^{(2k+1)}(x) + b^n \sum_{k=0}^{n} (-1)^k \pi^{2(n-(k-1))} f^{(2k)}(x) \]
\[ = b^n \sum_{k=1}^{n+1} (-1)^k \pi^{2(n-(k-1))} f^{(2k)}(x) + b^n \sum_{k=0}^{n} (-1)^k \pi^{2(n-(k-1))} f^{(2k)}(x) \]
\[ = b^n (-1)^n f^{(2n+2)}(x) + b^n \pi^{2n+2} f(x) = b^n \pi^{2n+2} f(x), \] (H.14)

and, thus, for

\[ h : \mathbb{R} \rightarrow \mathbb{R}, \quad h(x) := g'(x) \sin(\pi x) - \pi g(x) \cos(\pi x), \] (H.15)

one obtains, for each \( x \in \mathbb{R} \),

\[ h'(x) = g''(x) \sin(\pi x) + \pi g'(x) \cos(\pi x) - \pi g'(x) \cos(\pi x) + \pi^2 g(x) \sin(\pi x) \]
\[ = \left( g''(x) + \pi^2 g(x) \right) \sin(\pi x) \overset{(H.14)}{=} b^n \pi^{2n+2} f(x) \sin(\pi x) \]
\[ = \pi^2 a^n f(x) \sin(\pi x), \] (H.16)

implying the function \( h \) is the antiderivative of the function \( x \mapsto \pi^2 a^n f(x) \sin(\pi x) \). This, together with the fundamental theorem of calculus in the form Th. 10.20(b) implies

\[ I := \frac{\pi^2 a^n}{\pi} \int_{0}^{1} f(x) \sin(\pi x) \, dx = \frac{h(1) - h(0)}{\pi} = \frac{\pi g(1) + \pi g(0)}{\pi} = g(1) + g(0) \in \mathbb{Z}. \] (H.17)

On the other hand, the definition of \( f \) in (H.7) yields

\[ \forall 0 < x < 1 \quad 0 < f(x) < \frac{1}{n!}, \] (H.18)

and, thus, by (10.30) (i.e. by the monotonicity of the integral),

\[ 0 < I < \frac{\pi a^n}{n!} \overset{(H.6)}{<} 1. \] (H.19)

The contradiction between (H.19) and (H.17) establishes the case. ■
I Trigonometric Functions

I.1 Additional Trigonometric Formulas

Proposition I.1. We have the following identities:

\[ \forall z \in \mathbb{C} \quad \sin(2z) = 2\sin z \cos z, \]  
(I.1a)

\[ \forall z \in \mathbb{C} \quad \cos(2z) = (\cos z)^2 - (\sin z)^2, \]  
(I.1b)

\[ \forall z \in \mathbb{C} \quad \frac{1 - \cos z}{2} = \left(\sin \frac{z}{2}\right)^2, \]  
(I.1c)

\[ \forall z \in \mathbb{C} \setminus \{(2k+1)\pi : k \in \mathbb{Z}\} \quad \tan \frac{z}{2} = \frac{\sin z}{\cos z + 1}. \]  
(I.1d)

\[ \forall z \in \mathbb{C} \setminus \{(2k+1)\pi : k \in \mathbb{Z}\} \quad \cos z = \frac{1 - (\tan \frac{z}{2})^2}{1 + (\tan \frac{z}{2})^2}. \]  
(I.1e)

Proof. (I.1a) is immediate from (8.44c), (I.1b) is immediate from (8.44d).

(I.1c): For each \( z \in \mathbb{C} \), one computes
\[
\frac{1 - \cos z}{2} = \frac{1 - (\cos \frac{z}{2})^2 + (\sin \frac{z}{2})^2}{2} = \frac{2(\sin \frac{z}{2})^2}{2} = \left(\sin \frac{z}{2}\right)^2,
\]  
thereby establishing the case.

(I.1d): Note that, according to (8.47d), it is
\[
\cos \frac{z}{2} = 0 \quad \Leftrightarrow \quad \exists k \in \mathbb{Z} \quad z = (2k + 1)\pi.
\]  
(I.3)

Thus, for each \( z \in \mathbb{C} \setminus \{(2k+1)\pi : k \in \mathbb{Z}\} \), one computes
\[
\tan \frac{z}{2} = \frac{2 \sin \frac{z}{2} \cos \frac{z}{2}}{2(\cos \frac{z}{2})^2} = \frac{\sin z}{(\cos \frac{z}{2})^2 - (\sin \frac{z}{2})^2 + 1} = \frac{\sin z}{\cos z + 1},
\]  
thereby establishing the case.

(I.1e): Once again, using (I.3), one computes for each \( z \in \mathbb{C} \setminus \{(2k+1)\pi : k \in \mathbb{Z}\} \):
\[
\cos z = \frac{(\cos \frac{z}{2})^2 - (\sin \frac{z}{2})^2}{(\cos \frac{z}{2})^2 + (\sin \frac{z}{2})^2} = \frac{1 - (\tan \frac{z}{2})^2}{1 + (\tan \frac{z}{2})^2},
\]  
(I.5)

as claimed.

\[ \blacksquare \]

J Differential Calculus

J.1 Continuous, But Nowhere Differentiable Functions

The following Ex. J.1 provides functions from \( f : \mathbb{R} \to \mathbb{R} \) that are continuous, but nowhere differentiable.
Example J.1. We start by defining the triangle wave function

\[
g : \mathbb{R} \to \mathbb{R}, \quad g(x) := \begin{cases} 
x - k & \text{for } k \leq x \leq k + \frac{1}{2}, k \in \mathbb{Z}, \\
-x + k + 1 & \text{for } k + \frac{1}{2} \leq x \leq k + 1, k \in \mathbb{Z}.
\end{cases}
\]

Then \( g \) is well-defined and continuous, since, clearly, \( g \) is piecewise affine, \( k + \frac{1}{2} - k = \frac{1}{2} = -k - \frac{1}{2} + k + 1 \), and \( -(k + 1) + k + 1 = 0 = k + 1 - (k + 1) \). Moreover, for each \( k \in \mathbb{Z} \), \( g \) is clearly strictly increasing on \( [k,k + \frac{1}{2}] \) and clearly strictly decreasing on \( [k + \frac{1}{2},k + 1] \), implying

\[
\forall x \in \mathbb{R} \quad 0 \leq g(x) \leq \frac{1}{2}.
\]

Clearly, (J.1) implies \( g \) to be periodic with period 1, i.e.

\[
\forall x \in \mathbb{R} \quad g(x + 1) = g(x).
\]

Now fix \( q \in \mathbb{R}, a \in \mathbb{N} \) such that

\[
0 < q < 1 \quad \land \quad a \geq 4 \quad \land \quad aq > 2 \quad (J.4)
\]

(clearly, \( q = \frac{1}{2} \) and \( a = 5 \) satisfy (J.4), and there are (uncountably) many other admissible choices for \( a \) and \( q \)). We now claim that

\[
f : \mathbb{R} \to \mathbb{R}, \quad f(x) := \sum_{n=0}^{\infty} q^n g(a^n x),
\]

is continuous and nowhere differentiable. We first note that, as \( \sum_{n=0}^{\infty} q^n \) converges and

\[
\forall n \in \mathbb{N} \quad \forall x \in \mathbb{R} \quad g(a^n x) \leq q^n \leq q^n \frac{aq}{2} < q^n, \quad (J.6)
\]

Cor. 8.7(b) implies the series in (J.5) to converge uniformly. Then, since each function \( f_n : \mathbb{R} \to \mathbb{R}, f_n(x) := q^n g(a^n x) \), is continuous, \( f \) must be continuous by Cor. 8.7(c).

In preparation for showing \( f \) to be nowhere differentiable, we have to further investigate the properties of \( g \). We proceed by showing \( g \) to be Lipschitz continuous with Lipschitz constant 1, i.e.

\[
\forall x, y \in \mathbb{R} \quad |g(x) - g(y)| \leq |x - y| : \quad (J.7)
\]

If \( |x - y| \geq \frac{1}{2} \), then, using (J.2),

\[
|g(x) - g(y)| \leq \frac{1}{2} \leq |x - y|.
\]

If \( |x - y| < \frac{1}{2} \), then we distinguish four cases, where, without loss of generality, we let \( x \) denote the smaller of the two points and \( y \) the larger, i.e. \( x \leq y \). Case (i): There is \( k \in \mathbb{Z} \) such that \( k \leq x, y \leq k + \frac{1}{2} \). Then

\[
|g(x) - g(y)| = |x - k - y + k| = |x - y|.
\]
Case (ii): There is $k \in \mathbb{Z}$ such that $k + \frac{1}{2} \leq x, y \leq k + 1$. Then
\[
|g(x) - g(y)| = | - x + k + 1 + y - k - 1| = |x - y|.
\]

Case (iii): There is $k \in \mathbb{Z}$ such that $k \leq x \leq k + \frac{1}{2}$ and $k + \frac{1}{2} \leq y \leq k + 1$. Then $x - k - \frac{1}{2} \leq 0$ and $y - k - \frac{1}{2} \geq 0$, implying
\[
|g(x) - g(y)| = | - x + k + 1 + y - k - 1| = |x - y|.
\]

Case (iv): There is $k \in \mathbb{Z}$ such that $k - \frac{1}{2} \leq x \leq k$ and $k \leq y \leq k + \frac{1}{2}$. Then $- x + k \geq 0$ and $- y + k \leq 0$, implying
\[
|g(x) - g(y)| = | - x + k + y - k| \leq |x - k + y - k| = |x - y|,
\]
finishing the proof of (J.7).

For each $c \in \mathbb{R}$, we also consider the following modified versions of $g$:
\[
g_c : \mathbb{R} \rightarrow \mathbb{R}, \quad g_c(x) := g(cx) = \begin{cases} cx - k & \text{for } k \leq cx \leq k + \frac{1}{2}, k \in \mathbb{Z}, \\ -cx + k + 1 & \text{for } k + \frac{1}{2} \leq cx \leq k + 1, k \in \mathbb{Z}. \end{cases} \quad (J.8)
\]

Then, for $c \neq 0$, $g_c$ is periodic with period $c^{-1}$:
\[
\forall x \in \mathbb{R} \quad g_c(x + c^{-1}) = g(cx + 1) = g(cx) = g_c(x). \quad (J.9)
\]

Moreover, $g_c$ is Lipschitz continuous with Lipschitz constant $|c|$:
\[
\forall x,y \in \mathbb{R} \quad |g_c(x) - g_c(y)| = |g(cx) - g(cy)| \leq |cx - cy| = |c| |x - y|. \quad (J.10)
\]

To show that $f$ is nowhere differentiable, we will now study suitable difference quotients. Let $(h_k)_{k \in \mathbb{N}}$ be a sequence such that
\[
\forall k \in \mathbb{N} \quad h_k = \pm \frac{1}{a^{k+1}}. \quad (J.11)
\]

Then (J.4) implies $\lim_{k \to \infty} h_k = 0$. Let $x \in \mathbb{R}$ be arbitrary. Define
\[
\forall k,n \in \mathbb{N} \quad \delta_k g(a^n x) := \frac{g(a^n(x + h_k)) - g(a^n x)}{h_k}. \quad (J.12)
\]

Then
\[
\forall k,n \in \mathbb{N} \quad |\delta_k g(a^n x)| \leq \frac{a^n |h_k|}{|h_k|} = a^n. \quad (J.13)
\]
and, recalling \( a \in \mathbb{N} \),

\[
\forall \ n > k \in \mathbb{N} \quad \delta_k g(a^n x) = \frac{g_{a^n}(x \pm \frac{1}{a^{k+1}}) - g_{a^n}(x)}{h_k} = \frac{g_{a^n}(x \pm \frac{a^{n-(k+1)}}{a^n}) - g_{a^n}(x)}{h_k} \quad \text{(J.9)}
\]

Thus, for each \( k \in \mathbb{N} \), we obtain

\[
\delta_k f := \frac{f(x + h_k) - f(x)}{h_k} = \sum_{n=0}^{k-1} q^n \delta_k g(a^n x) + q^k \delta_k g(a^k x) \quad \text{(J.15)}
\]

and estimate, recalling \( aq > 2 \),

\[
\left| \sum_{n=0}^{k-1} q^n \delta_k g(a^n x) \right| \leq \sum_{n=0}^{k-1} q^n \left| \delta_k g(a^n x) \right| \leq \sum_{n=0}^{k-1} q^n a^n = \frac{1 - q^k a^k}{1 - qa} < \frac{q^k a^k}{qa - 1}. \quad \text{(J.16)}
\]

We rewrite (J.16) as

\[
\left| \sum_{n=0}^{k-1} q^n \delta_k g(a^n x) \right| < \eta q^k a^k, \quad \text{where} \quad \eta := \frac{1}{aq - 1}, \quad 0 < \eta < 1. \quad \text{(J.17)}
\]

According to (J.8), \( g_{a^k} \) is affine on intervals of length \( \frac{1}{2a^k} \). Thus, since \( a \geq 4 \) implies

\[
\forall \ k \in \mathbb{N} \quad \frac{1}{a^{k+1}} \leq \frac{1}{4a^k}, \quad \text{(J.18)}
\]

we can always choose the sign of \( h_k \) such that

\[
|\delta_k g(a^k x)| = \left| \frac{a^k(x + h_k - x)}{|h_k|} \right| = a^k. \quad \text{(J.19)}
\]

Using this choice for the \( h_k \), we combine our estimates to obtain

\[
\forall \ k \in \mathbb{N} \quad \left| \delta_k f \right| = \left| \sum_{n=0}^{k-1} q^n \delta_k g(a^n x) + q^k \delta_k g(a^k x) \right| \geq (1 - \eta) q^k a^k. \quad \text{(J.20)}
\]

Thus, as \( qa > 2 \), \( \lim_{k \to \infty} |\delta_k f| = \infty \), proving that \( f \) is not differentiable in \( x \). As \( x \in \mathbb{R} \) was arbitrary, \( f \) is nowhere differentiable.

**References**

REFERENCES


