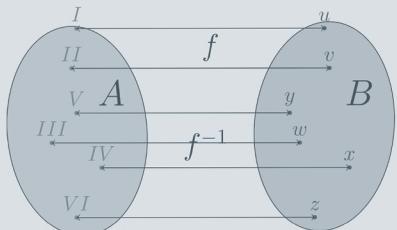
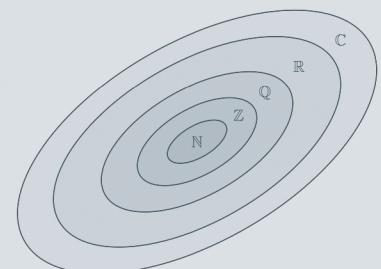
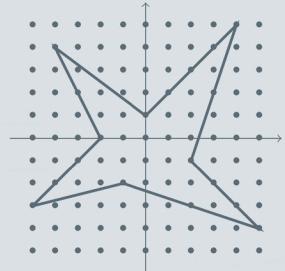




A Bridge to Advanced Mathematics

From Natural to Complex Numbers

Sebastian M. Cioabă
Werner Linde



A Bridge to Advanced Mathematics

From Natural to
Complex Numbers



Pure and Applied
UNDERGRADUATE TEXTS • 58

A Bridge to Advanced Mathematics

From Natural to
Complex Numbers

Sebastian M. Cioabă
Werner Linde



Providence, Rhode Island USA

EDITORIAL COMMITTEE

Giuliana Davidoff Tara S. Holm
Steven J. Miller Maria Cristina Pereyra
Gerald B. Folland (Chair)

2020 *Mathematics Subject Classification.* Primary 00-01, 00A05, 00A06, 05-01.

For additional information and updates on this book, visit
www.ams.org/bookpages/amstext-58

Library of Congress Cataloging-in-Publication Data

Names: Cioabă, Sebastian M., author. | Linde, Werner, 1947- author.

Title: A bridge to advanced mathematics : from natural to complex numbers / Sebastian M. Cioabă, Werner Linde.

Description: Providence, Rhode Island : American Mathematical Society, [2023] | Series: Pure and applied undergraduate texts, 1943-9334 ; volume 58 | Includes bibliographical references and index.

Identifiers: LCCN 2022034218 | ISBN 9781470471484 (paperback) | 9781470472139 (ebook)

Subjects: LCSH: Numeration--Textbooks. | Numbers, Natural--Textbooks. | Numbers, Complex--Textbooks. | AMS: General -- Instructional exposition (textbooks, tutorial papers, etc.). | General -- General and miscellaneous specific topics -- General mathematics. | General -- General and miscellaneous specific topics -- Mathematics for nonmathematicians (engineering, social sciences, etc.). | Combinatorics -- Instructional exposition (textbooks, tutorial papers, etc.).

Classification: LCC QA141 .C48 2023 | DDC 513.2--dc23/eng20221014

LC record available at <https://lccn.loc.gov/2022034218>

Copying and reprinting. Individual readers of this publication, and nonprofit libraries acting for them, are permitted to make fair use of the material, such as to copy select pages for use in teaching or research. Permission is granted to quote brief passages from this publication in reviews, provided the customary acknowledgment of the source is given.

Republication, systematic copying, or multiple reproduction of any material in this publication is permitted only under license from the American Mathematical Society. Requests for permission to reuse portions of AMS publication content are handled by the Copyright Clearance Center. For more information, please visit www.ams.org/publications/pubpermissions.

Send requests for translation rights and licensed reprints to reprint-permission@ams.org.

© 2023 by the authors. All rights reserved.

Printed in the United States of America.

∞ The paper used in this book is acid-free and falls within the guidelines established to ensure permanence and durability.

Visit the AMS home page at [https://www.ams.org/](http://www.ams.org/)

10 9 8 7 6 5 4 3 2 1 28 27 26 25 24 23

Contents

| | |
|--|-----|
| Preface | ix |
| 1. The Content of the Book | ix |
| 2. How to Use This Book? | x |
| Chapter 1. Natural Numbers \mathbb{N} | 1 |
| 1.1. Basic Properties | 1 |
| 1.2. The Principle of Induction | 10 |
| 1.3. Arithmetic and Geometric Progressions | 29 |
| 1.4. The Least Element Principle | 34 |
| 1.5. There are 10 Kinds of People in the World | 45 |
| 1.6. Divisibility | 51 |
| 1.7. Counting and Binomial Formula | 65 |
| 1.8. More Exercises | 78 |
| Chapter 2. Integer Numbers \mathbb{Z} | 81 |
| 2.1. Basic Properties | 81 |
| 2.2. Integer Division | 84 |
| 2.3. Euclidean Algorithm Revisited | 88 |
| 2.4. Congruences and Modular Arithmetic | 104 |
| 2.5. Modular Equations | 124 |
| 2.6. The Chinese Remainder Theorem | 130 |
| 2.7. Fermat and Euler Theorems | 135 |
| 2.8. More Exercises | 156 |

| | |
|--|-----|
| Chapter 3. Rational Numbers \mathbb{Q} | 159 |
| 3.1. Basic Properties | 159 |
| 3.2. Not Everything Is Rational | 166 |
| 3.3. Fractions and Decimal Representations | 173 |
| 3.4. Finite Continued Fractions | 193 |
| 3.5. Farey Sequences and Pick's Formula | 205 |
| 3.6. Ford Circles and Stern–Brocot Trees | 220 |
| 3.7. Egyptian Fractions | 228 |
| 3.8. More Exercises | 233 |
| Chapter 4. Real Numbers \mathbb{R} | 237 |
| 4.1. Basic Properties | 237 |
| 4.2. The Real Numbers Form a Field | 244 |
| 4.3. Order and Absolute Value | 247 |
| 4.4. Completeness | 254 |
| 4.5. Supremum and Infimum of a Set | 257 |
| 4.6. Roots and Powers | 264 |
| 4.7. Expansion of Real Numbers | 272 |
| 4.8. More Exercises | 288 |
| Chapter 5. Sequences of Real Numbers | 291 |
| 5.1. Basic Properties | 291 |
| 5.2. Convergent and Divergent Sequences | 297 |
| 5.3. The Monotone Convergence Theorem and Its Applications | 310 |
| 5.4. Subsequences | 319 |
| 5.5. Cauchy Sequences | 325 |
| 5.6. Infinite Series | 329 |
| 5.7. Infinite Continued Fractions | 352 |
| 5.8. More Exercises | 362 |
| Chapter 6. Complex Numbers \mathbb{C} | 367 |
| 6.1. Basic Properties | 367 |
| 6.2. The Conjugate and the Absolute Value | 375 |
| 6.3. Polar Representation of Complex Numbers | 379 |
| 6.4. Roots of Complex Numbers | 385 |
| 6.5. Geometric Applications | 397 |
| 6.6. Sequences of Complex Numbers | 400 |
| 6.7. Infinite Series of Complex Numbers | 405 |
| 6.8. More Exercises | 416 |

| | |
|--|-----|
| Epilogue | 419 |
| Appendix. Sets, Functions, and Relations | 421 |
| A.1. Logic | 421 |
| A.2. Sets | 430 |
| A.3. Functions | 438 |
| A.4. Cardinality of Sets | 455 |
| A.5. Relations | 472 |
| A.6. Proofs | 485 |
| A.7. Peano's Axioms and the Construction of Integers | 495 |
| A.8. More Exercises | 505 |
| Bibliography | 515 |
| Index | 517 |

Preface

Of all the aids which the human mind has yet created to simplify its life — that is, to simplify the work in which thinking consists — none is so momentous and so inseparably bound up with the mind's most inward nature as the concept of number. Arithmetic, whose sole object is this concept, is already a science of immeasurable breadth, and there can be no doubt that there are absolutely no limits to its further development; and the domain of its application is equally immeasurable, for every thinking person, even if he does not clearly realize it, is a person of numbers, an arithmetician.

Richard Dedekind¹

1. The Content of the Book

This book is a journey through various number systems, starting with the most familiar ones, natural numbers, and finishing with complex numbers. We tried to achieve a balance between having the readers learn and practice writing proofs related to new abstract concepts such as numbers, sets, and functions and enabling the readers to understand and assimilate new mathematical content. As Jordan Ellenberg eloquently puts it², the first task is analogous to practicing for the soccer game while the second is akin to playing in a real match. We believe that all students should practice and play mathematics.

This is a fine line to walk, and we have chosen to follow the story of numbers starting with the natural numbers that everyone knows. Here, one can learn about mathematical induction, arithmetic and geometric progressions, basic number theory, and counting and then progress to studying the integers where some concepts and results extend from the natural numbers and some do not.

¹undated fragment in Dedekind's inheritance

²See his wonderful book [10].

Expanding on the number theory notions done with natural numbers, we present the basics of congruences and more advanced number theory such as Fermat's and Euler's theorems and the Chinese remainder theorem as well as applications such as Costas arrays, Latin squares, coding theory, and basic cryptography.

We study rational numbers and their various representations including familiar ones such as decimal representations of fractions or less familiar ones such as finite continued fractions, Farey sequences, Ford circles, Stern–Brocot trees, or Egyptian fractions. We discuss some basic notions of groups, rings, and fields (which are introduced and discussed through the concrete examples of integers, rational numbers, and modular arithmetic) in order to give the students a glimpse of what they may see in abstract algebra courses.

In our journey to complex numbers, we continue with geometric reasoning using the Pythagorean theorem that leads to the investigation of irrational numbers. We study the notions of lower bound and upper bound, infimum and supremum, and limits of sequences and prime the students for a subsequent analysis course.

We progress to studying complex numbers and their basic properties including geometric interpretations, De Moivre's formula, and the fundamental theorem of algebra. Finally, we include an appendix containing some basic notions of logic, sets, relations, functions, and proofs for the interested readers.

We also show that mathematics is a dynamic and active research area where proofs and counterexamples continue to appear and where, despite intense investigation from many people, there are still many mathematical statements, some very simple to state and understand, that have eluded the search for a proof. To the best of our knowledge, we include historical details involving the mathematicians responsible for the results described.

We believe that examples and exercises are fundamental for understanding mathematics. Whenever possible, we introduce a new notion by presenting examples first and then proceeding to the more abstract content. At the end of each section, we include exercises related to the content of that section, and at the end of each chapter, we include exercises that use the content presented in the entire chapter. This approach enhances the readers' ability to practice and retain the knowledge presented in this book.

2. How to Use This Book?

For over 2,500 years, mathematical statements and their proofs have fascinated amateur and professional mathematicians. The first proofs were discovered by Thales of Miletus who lived around 600 BCE. Thales was an astronomer and mathematician who apparently successfully predicted a solar eclipse in 585 BCE. He is the first scholar who suggested that geometric statements should not be accepted unless they have a logical and rigorous proof. This is quite a legacy! Among the results proved by Thales are that a diameter splits a circle into two equal parts, an angle inscribed in a semicircle is a right angle, and the angle sum of a triangle equals two right angles. According to various sources, he was so absorbed studying the sky and pondering mathematics that he

fell into a well. Thales escaped from this event unscathed and continued to think about mathematics. We can only hope that this is the fate of our students and readers, as in recent years proofs have also terrified and caused anxiety among students, especially at undergraduate levels.

From teaching courses in mathematics at various universities, both authors independently noticed that for young mathematics students, the most difficult and most complicated transition in their careers is that from high school to the university or from mainly calculation-driven courses, such as calculus, to proof-based and conceptually advanced abstract courses such as abstract algebra and analysis.

Too many students fail, lose their interest, and give up. There are several reasons why many students have such serious problems in mathematics during their first semesters at the university. On one hand, the volume of the topics taught at universities is incommensurately bigger. In general, topics taught in one year at high school are presented in a math class at the university in a few weeks. Another difficult obstacle for new students in mathematics is the difference in rigor between the kind of mathematics taught at the high school level and the way mathematics is presented at the university. Why do mathematical notions have to be defined so precisely? What are axioms good for and where do they come from? And the most often asked question of students is: Why do mathematical statements need a proof, even if they look obvious? For example, why does one have to verify that the set of natural numbers is infinite? This seems to be self-evident and well known!

And after the students understand that only those statements which can be proved correctly are true, the next problem for them is to understand what a rigorous mathematical proof is. How often does it happen that students start a proof with the assertion they want to verify, make some transformations, end up with something which is obviously true, and believe that this also verifies the validity of the assertion? Or how difficult is it for several students to understand the principle of mathematical induction?

A lot of these failures may be avoided by making the transformation from high school to university smoother. The previous observations triggered us to write a book which helps students with their entrance into modern mathematics. But what is the best way of doing this? In our experience **numbers** are the optimal entrance door for most of the students: natural numbers, integers, rational numbers, and later on, real numbers and complex numbers. The first three kinds of numbers are topics that the students are familiar with, and they may act as guide and anchor for the students' explorations through the universe of higher mathematics.

The present book arose from the lectures given by the authors in an *Introduction to Proof* course at the University of Delaware (UD) between 2010 and 2020. This course is a prelude to more abstract mathematics that one expects every math major to know and our UD students are required to take this course as a preparation for more advanced undergraduate courses such as *Abstract Algebra* and *Real Analysis*. The UD transition course progression is similar to many other universities³.

³According to a recent survey paper by Erika J. Davis and Dov Zazkis, *Characterizing introduction to proof courses: A survey of R1 and R2 course syllabi in International Journal of Mathematical Education in Science and Technology*, there are over 170 such transition courses at the R1 and R2 research universities in the United States.

Our book may be used in a course similar to the course at UD or in other introduction to proofs or transition to abstract mathematics courses depending on the knowledge of the students and which chapters the instructor focuses on. For a course similar to the one at UD, a typical structure for a one-semester course (14 weeks, 40 lectures of 50 minutes each) is as follows:

- Introduction (4 lectures; numbers, sets, functions, proofs).
- Natural Numbers (6 lectures; induction, counting, divisibility, primes).
- Integer Numbers (6 lectures; modular arithmetic, Fermat's and Euler's theorems, Chinese remainder theorem).
- Rational Numbers (6 lectures; rational numbers as equivalence classes, decimal representation, continued fractions, irrational numbers).
- Real Numbers (9 lectures; upper/lower bound, supremum/infimum, limits of sequences, bounded sequences, Bolzano–Weierstrass theorem).
- Complex Numbers (6 lectures; roots of unity, geometric meaning and applications, De Moivre's formula, fundamental theorem of algebra).
- Topics chosen by the instructor (3 lectures).

This structure may be customized depending on the audience and the instructor's preferences. The material from this book can be readily adapted to other transition courses. We included an appendix describing the basic notions of logic, sets, functions, and proofs that may be used in the beginning of such a course. Our choice of topics gives an instructor many possibilities for what to include and to spend more time on, such as counting, graph theory, number theory, continued fractions, Farey sequences, Pick's formula, sequences and limits, and complex numbers. The material can be used for a more advanced transition course by spending more time on rational, real, and complex numbers topics. We believe that it is important for students to see in one place the big picture and the connections between topics and number systems that otherwise may appear piecemeal in various courses.

Our book is also useful to high school students or undergraduate students in mathematics, physics, or engineering looking to sharpen their problem-solving skills and to expand their mathematical knowledge. The level and the knowledge of beginners in math is very heterogeneous. There are students who may be familiar with logic, sets, or functions. For those, we suggest starting the present book with Chapter 1, that is, going directly into the heart of the matter. But there are also always students whose foreknowledge of mathematics is poor, in general not by their fault but by certain negative circumstances during their time in high school. For those students, we suggest starting with the appendix, and later on, after knowing the basic facts presented there, switching to Chapter 1.

One pedagogical approach that works well in our courses is to start each class with a hook problem or example to stimulate the students' curiosity. This approach is not new and can be found in many places⁴. We are in agreement with Albert Einstein's statement:

Example isn't another way to teach, it is the only way to teach.

Concrete examples are fundamental to learning, and we discuss them first before introducing more abstract notions and results.

Exercises are very important for understanding the material, and we have included about 700 exercises of various degrees of difficulty⁵ in the book. We believe that once the students are curious and focused, they can actively work through the topics discussed and can absorb more abstract content with better longer-term outcomes. Our goal is to increase the students' knowledge of mathematics and familiarity with proof techniques, as well as prepare them for more abstract content⁶.

Like most mathematics books, studying this book with pencil, paper, and an open mind is the recommended way and readers should keep in mind Maria Montessori's quote: *Knowledge is best given where there is eagerness to learn.*

We thank our students Anna Maria Bella, John Byrne, Brendan Coffey, Arianna Cordrey, Himanshu Gupta, Nathaniel Kim, Natalie Konschnik, Hava Marneweck, Alexis Mathew, Lauren Rosica, Jake Sitison, Nathanael Urmoneit, who have carefully read various versions of our book. We are grateful to Felix Lazebnik for suggesting the topic of this book and for his teaching notes on various topics that have been a source of inspiration for one of us, to the AMS editors for their comments and suggestions, and to Eriko Hironaka for her tremendous help and support in completing this project.

The authors are grateful to their families for their patience and support during the writing of this book. Without them, this book would never have been finished.

Sebastian M. Cioabă and Werner Linde

⁴See Paul Halmos et al.'s article [14].

⁵More advanced exercises are marked by (★).

⁶In order to advance one's mathematical creativity, it is worth recalling George Pólya's remark: *It is hard to have a good idea if we have little knowledge on the subject, and impossible to have it if we have no knowledge.*

Natural Numbers \mathbb{N}

This is exactly how mathematics proceeds. You throw down some rules and begin to play.
Ben Orlin¹

1.1. Basic Properties

The appearance and use of numbers was a slow process. Our ancestors first learned to distinguish between one, two, and many. A long time passed before people learned to count and understood how to use larger numbers. In some languages, the names used for 3,4,5 were constructed by combining the names of previous used numbers: 3 is 2 and 1, 4 is 2 and 2, 5 is 2 and 3. In some cultures, numbers were called different things depending on what was being counted. For example, 10 boats had a different name than 10 coconuts. As commerce and economy grew, there was a need to keep track of larger quantities and that led to the appearance of various number systems. One ancient way of representing a natural number is to put down a mark for each object on a bone or stick.

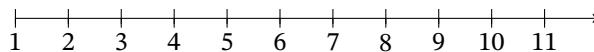


Figure 1.1.1. A representation of the first natural numbers.

The Greeks used geometry to see the natural numbers as points on a line. One can think about the usual operations of addition, subtraction, and multiplication in a geometric/visual way. To add the quantities in two adjacent boxes, we put the two boxes together and count the content of their union. This is something that we consider obvious when dealing with natural numbers. Another fact we take for granted is that order does not matter when adding two numbers: $4 + 3 = 3 + 4 = 7$. Note that this property, which has the fancy name commutativity, may not hold in other realms. Try

¹*Math with Bad Drawings: Illuminating the Ideas That Shape Our Reality*

getting out of the car and opening your umbrella versus opening your umbrella and then getting out of the car.

$$\boxed{\bullet \bullet \bullet \bullet} + \boxed{\bullet \bullet \bullet} = \boxed{\bullet \bullet \bullet \bullet \bullet \bullet \bullet \bullet}$$

Figure 1.1.2. Addition.

Subtraction can be represented by peeling off the smaller part from the larger part.

$$\boxed{\bullet \bullet \bullet \bullet} - \boxed{\bullet \bullet \bullet} = \boxed{\bullet}$$

Figure 1.1.3. Subtraction.

Multiplication means arranging the boxes in a rectangular shape as in Figure 1.1.4. This operation is commutative and $4 \times 3 = 3 \times 4$. A simple and intuitive explanation for this fact is that counting dots row by row is the same as counting them column by column. These interpretations of addition and multiplication are the basics of our counting techniques and will be described later in this chapter in the language of sets.

$$\boxed{\bullet \bullet \bullet \bullet} \times \boxed{\bullet \bullet \bullet} = \begin{array}{|c|c|c|} \hline \bullet & \bullet & \bullet \\ \hline \end{array}$$

Figure 1.1.4. Multiplication.

Many results such as the Pythagorean theorem or various algebraic identities that we use today were seen by the Greeks through the lens of geometry.

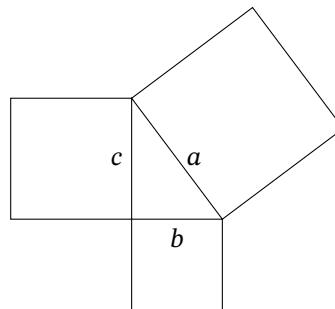


Figure 1.1.5. Pythagorean theorem: the area of the square of length a equals the sum of the areas of the squares of length b and c .

The first major advance in abstraction was the use of numerals to represent numbers. This allowed systems to be developed for recording large numbers. The ancient Egyptians invented a powerful system of numerals with distinct hieroglyphs for 1, 10,

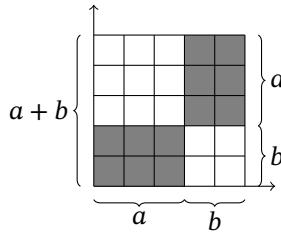


Figure 1.1.6. Geometric proof $(a + b)^2 = a^2 + b^2 + 2ab$.

and all the powers of 10 up to over one million. The number systems differed between different cultures: the Mayans used base 20 and the Mesopotamian people used base 60. We can assume that 20 was used as we have 20 fingers and toes while 60 was popular due to its divisibility properties. Nowadays, we all use natural numbers in our day-to-day life. These numbers are given by $1, 2, 3, \dots, 2021, 2022, \dots$ and they go on forever. We use them to count things and to play games. A young kids' game² is to name the largest possible number. Of course, if the first person chooses 100, then the second person can pick $100 + 1 = 101$ or $2 \times 100 = 200$ or any other number larger than 100 to win. In general, 1 is a natural number and if n is a natural number, then $n+1$ is also a natural number. For now, we take it for granted that the reader is familiar with the natural numbers and their basic properties. We will do a quick overview of these properties in the next few pages. Later on, based on Peano's axioms, presented in Section A.7, we will give a more abstract approach to natural numbers.

If you are puzzled by the absence of zero from the collection of natural numbers, we assure you that zero has not been forgotten. The invention of 0 is an important event in mathematics and life and is a highly nontrivial matter. It required a high degree of abstraction. For example, the ancient Greeks had troubles with it and asked themselves

How can nothing be something?

In 2nd century BCE, the Indian mathematician Pingala used the Sanskrit word *śūnya* (translated as emptiness) to refer to zero.

Definition 1.1.1. A number that is greater than 0 is called **positive**. A number that is smaller than 0 is called **negative**.

Some authors consider 0 to be a natural number and others do not. Given the higher level of abstraction needed to come up with such a number, our convention will be to consider 0 as being outside the set of natural numbers.

Definition 1.1.2. The set of all natural numbers is denoted by \mathbb{N} .

Informally, we can think of \mathbb{N} as a large infinite yard where we keep all the natural numbers:

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

²This game was popular with the first author's kids.

We will also use the set obtained from \mathbb{N} by appending 0:

$$\mathbb{N}_0 = \{0, 1, 2, 3, \dots\}.$$

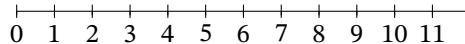


Figure 1.1.7. A representation of the first numbers in \mathbb{N}_0 .

We summarize below the basic properties of the natural numbers. There are two binary operations on \mathbb{N} :

- **addition:** $(a, b) \mapsto a + b$,
- **multiplication:** $(a, b) \mapsto a \cdot b$,

that have the following properties.

- (1) Both operations are **associative**, meaning that for all natural numbers a, b , and c (abbreviated as $\forall a, b, c \in \mathbb{N}$, cf. Section A.1),

$$(a + b) + c = a + (b + c) \quad \text{as well as} \quad (ab)c = a(bc).$$

- (2) Both operations are **commutative**,

$$(\forall a, b \in \mathbb{N})[a + b = b + a \text{ and } a \cdot b = b \cdot a].$$

- (3) Addition and multiplication are connected by the **distributive law**.

$$(\forall a, b, c \in \mathbb{N})[a \cdot (b + c) = a \cdot b + a \cdot c].$$

- (4) The number 1 is a **unit** with respect to multiplication:

$$(\forall a \in \mathbb{N})[1 \cdot a = a \cdot 1 = a].$$

- (5) The following **cancellation laws** are satisfied.

$$(\forall a, b, c \in \mathbb{N})[(a + b = a + c) \Rightarrow (b = c) \text{ and } (a \cdot b = a \cdot c) \Rightarrow (b = c)].$$

- (6) There is an order on \mathbb{N} which may be defined by

$$(1.1.1) \quad (\forall a, b \in \mathbb{N})[(a < b) \Leftrightarrow (\exists c \in \mathbb{N})(b = a + c)]$$

and

$$(a \leq b) \Leftrightarrow (a < b \text{ or } a = b).$$

What does formula (1.1.1) express? It means the following: Given arbitrary natural numbers a and b , we say that a is less than b , written $a < b$, if and only if there exists (abbreviated as \exists , cf. Section A.1) a natural number c for which $b = a + c$.

It is easy to see that

$$(1.1.2) \quad \begin{aligned} & (\forall a \in \mathbb{N})(a \leq a) \\ & (\forall a, b \in \mathbb{N})[(a \leq b \text{ and } b \leq a) \Rightarrow (a = b)] \quad \text{and} \\ & (\forall a, b, c \in \mathbb{N})[(a \leq b \text{ and } b \leq c) \Rightarrow (a \leq c)]. \end{aligned}$$

Moreover, the order is compatible with the algebraic operations.

$$(\forall a, b, c \in \mathbb{N})[(a \leq b) \Rightarrow (a + c \leq b + c \text{ and } a \cdot c \leq b \cdot c)].$$

We are now ready for our first formal proof. We take small steps and prove the following result.

Proposition 1.1.1. *There is no largest number in \mathbb{N} .*

Proof: Let us assume that there is a number M that is the largest in \mathbb{N} . This means that M is a natural number and $M \geq k$ for any natural number k . However, the number $M + 1$ is also a natural number, but $M < M + 1$. This contradicts the previous assumption. Hence, the number M cannot exist. ■

The proof we described is an example of a proof by contradiction. This is a powerful method in mathematics. Informally, it can be used whenever we do not really know how to start our argument and can be seen as a *devil's advocate* type of method. According to some historians of mathematics, the first proof done by Thales of Miletus over 2500 years ago was proof by contradiction. Thales proved that any diameter of a circle splits the circle into two equal parts.

We finish this section with some examples with geometric meaning. Many students learn in high school or before the formula for solving quadratic equations. It is often memorized like a song or mantra and its origins are seldom understood.

Example 1.1.1. Say we have a rectangle of a given perimeter, say 16 units, and given area, say 15 units squared. What are the dimensions of the rectangle?

If we denote the height of the rectangle by x and the width by y , then the previous information means that

$$x + y = 8 \text{ and } xy = 15.$$

One way to solve this system of equations is to substitute the value $y = 8 - x$ into the second equation, get the equation $x(8 - x) = 15$ which is the same as the quadratic $x^2 - 8x + 15 = 0$ and then use the quadratic formula. However, there is a slicker way. The average or the arithmetic mean of x and y is half their sum and is 4 in this case. If x is the larger one of the two numbers x and y and x exceeds the average by z , then y must be below the average by z :

$$x = 4 + z \text{ and } y = 4 - z.$$

Since $xy = 15$, we get that $(4 + z)(4 - z) = 15$ and therefore, $16 - z^2 = 15$. Since $z \geq 0$, we deduce that $z = 1$ and $x = 5, y = 3$.

This example gives us the excuse to mention the following result which informally states that in the grades of any class, there is someone above the average. Equivalently, not everybody can be strictly below the average.

Proposition 1.1.2. *Let n be a natural number. If x_1, \dots, x_n are real numbers whose average or arithmetic mean $\frac{x_1 + \dots + x_n}{n}$ equals M , then there exists j between 1 and n such that $x_j \geq M$ and there exists k between 1 and n such that $x_k \leq M$.*

Proof: Take the maximum of the numbers x_1, \dots, x_n and denote its index by j . If the maximum is attained by several indices, let us take the smallest such index to be j . Now $x_j \geq x_\ell$ for any $1 \leq \ell \leq n$. Adding up these n inequalities, we get that

$nx_j \geq x_1 + \dots + x_n$. Dividing by n gives us that $x_j \geq \frac{x_1 + \dots + x_n}{n}$. This proves the first part. The second part is similar, and we leave it as an exercise. ■

This simple result is the basis of a powerful method in mathematics called the **pigeonhole principle** or the **box principle**. These names are derived from the following result.

Proposition 1.1.3 (Pigeonhole Principle). *Let n be a natural number. If we have n holes or boxes in which we put $n+1$ pigeons, then there is at least one hole containing two or more pigeons.*

Proof: We follow the method of proof for the previous result. For ℓ between 1 and n , denote by x_ℓ the number of pigeons in hole ℓ . Then $x_1 + \dots + x_n = n+1$ and therefore, by the previous proposition, there exists an index j such that we have $x_j \geq \frac{n+1}{n} = 1 + \frac{1}{n}$. Because x_j is a natural number strictly greater than 1, it must be that $x_j \geq 2$. Hence, the hole j contains at least two pigeons. ■

This proposition can be generalized (see Exercise 1.1.10). In applications, it is useful to figure out first what are the holes and what are the pigeons. We illustrate the meaning of this statement by some examples.

Example 1.1.2. Among any three natural numbers, there are two that are both even or both odd (have the same parity). For this problem, consider the box E which consists of all even numbers and the box O which contains the odd numbers. We have 3 numbers or *pigeons* in 2 boxes and therefore, by Proposition 1.1.3, one of the boxes contains at least two numbers which must have the same parity.

Example 1.1.3. Let n be a natural number. We claim that no matter how we pick $n+1$ numbers from the set $\{1, \dots, 2n\}$, there are two consecutive numbers among them. To prove this, define for ℓ between 1 and n , the ℓ -th hole or box is defined as consisting of the numbers $2\ell - 1$ and 2ℓ . Thus, when $n = 3$, the boxes are $\{1, 2\}$, $\{3, 4\}$, and $\{5, 6\}$. We have $n+1$ numbers or *pigeons* in n boxes and by Proposition 1.1.3, there must be a box containing two numbers. Those numbers must be consecutive by our definition of boxes.

Example 1.1.4. Among any 13 people, there are two born in the same month. In this situation, the boxes are the twelve months of a year and the *pigeons* are the thirteen people. A box contains a pigeon if the respective person is born in that month. By Proposition 1.1.3 for $n = 12$, there is a box containing two pigeons or a month containing the birthdays of two people.

We leave the boxes and the pigeons and switch gears to a different type of problem.

Example 1.1.5. Assume that we have a rectangular plot of land whose perimeter is given, let us say it equals 20 units. What dimensions will give the maximum area? Of course, there are several concrete dimensions one can try, and they can be plotted as in Figure 1.1.8.

If we move from the rectangle with dimensions 1 and 9 whose opposite corners are $(0, 0)$ and $(1, 9)$ and compare its area with the rectangle whose corners are $(0, 0)$

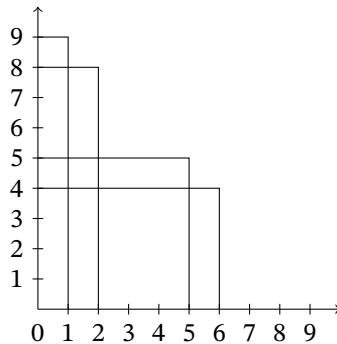


Figure 1.1.8. Various rectangles of perimeter 20.

and $(2, 8)$, we see in Figure 1.1.8 that they overlap over a rectangle of dimension 1 and 8 with the first rectangle having a 1×1 square of its own while the second rectangle has a 1×8 rectangle of its own. It appears that the area of the rectangle increases as we decrease the height from 9 and increase the width from 1. This way, we start with rectangles of areas 9, 16 and we end up in the *middle* of the rectangle/square with opposite corners $(0, 0)$ and $(5, 5)$ which has area 25. If we continue sliding down from $(5, 5)$ towards $(9, 1)$, the area of the rectangle will now decrease by symmetry as we will see the same rectangles we saw before with height and width interchanged. Now let us prove that the square will have the maximum area among all rectangles with this perimeter.

Proposition 1.1.4. *Let a and b be two nonnegative numbers. Then*

$$\left(\frac{a+b}{2}\right)^2 \geq ab.$$

Equality happens if and only if $a = b$.

Proof: We will give geometric and algebraic proofs of this result. For the geometric proof, Figure 1.1.9 shows how a square of length $a + b$ can be decomposed into 4 rectangles of dimensions $a \times b$ and a square of length $a - b$ in the center.

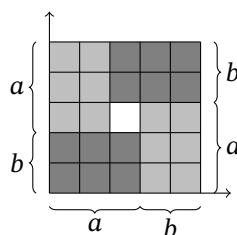


Figure 1.1.9. Geometric proof that $(a + b)^2 \geq 4ab$.

Let us do the algebraic proof. We have seen earlier that

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

Subtracting $4ab$ from both sides, we get that

$$(a + b)^2 - 4ab = a^2 + b^2 - 2ab = (a - b)^2.$$

Now $(a - b)^2 \geq 0$ since if $a - b \geq 0$ we can write $(a - b)^2 = (a - b) \cdot (a - b)$ as the product of two nonnegative numbers and if $a - b < 0$, we can do the same: $(a - b)^2 = (b - a) \cdot (b - a)$. Therefore, $(a + b)^2 \geq 4ab$ which is the same as what we wanted to prove. Just divide by 4. We deal now with the equality case. If $a = b$, then it is easy to see that $\left(\frac{a+b}{2}\right)^2 = ab$. If $\left(\frac{a+b}{2}\right)^2 = ab$, then the previous argument implies that $(a - b)^2 = 0$ and therefore $a = b$. \blacksquare

Taking square roots of both sides of the inequality in the previous proposition, we obtain an equivalent formulation:

$$\frac{a + b}{2} \geq \sqrt{ab},$$

for any nonnegative real numbers a and b . The quantity $\frac{a+b}{2}$ is the **arithmetic mean** of the numbers a and b and its definition extends to all numbers. The quantity \sqrt{ab} is the **geometric mean** of a and b and here we need $ab \geq 0$ in order to take the square root. These definitions extend to more than two numbers.

Definition 1.1.3. Let n be a natural number. Given n numbers a_1, \dots, a_n , their **arithmetic mean** is defined as

$$\frac{a_1 + \dots + a_n}{n}.$$

When a_1, \dots, a_n are nonnegative, define their **geometric mean** as

$$\sqrt[n]{a_1 \cdot \dots \cdot a_n}.$$

Proposition 1.1.4 implies that the arithmetic mean of two nonnegative real numbers is at least their geometric mean and Exercise 1.1.13 guides the reader through a proof of the similar result for three nonnegative numbers. In the next chapter, we will present a proof that the arithmetic mean of any n nonnegative real numbers is always at least their geometric mean for every n .

Exercise 1.1.1. Use Figure 1.1.10 to prove that $a^2 - b^2 = (a - b)(a + b)$ for any two nonnegative numbers a and b .

Exercise 1.1.2. The Pythagorean theorem states that the square of the hypotenuse of a right-angle triangle equals the sum of the squares of the other two sides. Use Figure 1.1.11 to prove this theorem.

Exercise 1.1.3. Prove that

$$(a + b + c)^2 = a^2 + b^2 + c^2 + 2ab + 2bc + 2ac,$$

for any numbers a , b , and c . Give both algebraic and geometric proofs.

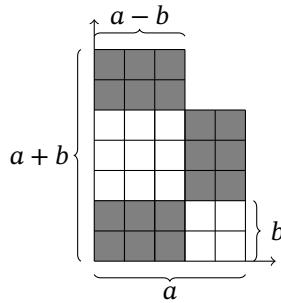


Figure 1.1.10. Geometric proof $a^2 - b^2 = (a - b)(a + b)$.

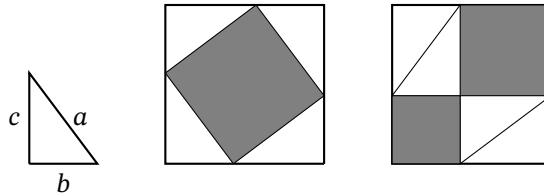


Figure 1.1.11. Pythagorean theorem: $a^2 = b^2 + c^2$.

Exercise 1.1.4. Show that for any two numbers a and b ,

$$a^3 - b^3 = (a - b)(a^2 + ab + b^2).$$

Give both algebraic and geometric proofs.

Exercise 1.1.5. Prove the properties of the order on \mathbb{N} stated in (1.1.2).

Exercise 1.1.6. A natural number n is called even if $n = 2k$ for some natural number k and is called odd, otherwise. For m and n natural numbers, prove the following:

- (1) The number n is odd if and only if $n = 2\ell - 1$ for some natural number ℓ .
- (2) If m and n are even, then $m + n$ is even.
- (3) If m and n are odd, then $m + n$ is even.
- (4) If $m + n$ is odd, then exactly one of m and n is odd.
- (5) If mn odd, then m is odd and n is odd.

Exercise 1.1.7. Let n be a natural number. Show that n^2 can be written as $4t$ or $4t + 1$ for some natural number $t \in \mathbb{N}_0$.

Exercise 1.1.8. Determine the number of ways to make change for one dollar (100 cents) using quarters (25 cents), dimes (10 cents), and nickels (5 cents).

Exercise 1.1.9. What is the number of ways to make change for one dollar using quarters, dimes, and nickels, when using at least one coin of each type?

Exercise 1.1.10. If we have p pigeons in h holes, then prove that there is a hole that contains at least $[p/h]$ pigeons, where $[p/h]$ denotes the smallest integer that is greater than or equal to p/h .

Exercise 1.1.11. Show that among any 25 people, there must be three born in the same month.

Exercise 1.1.12. Let n be a natural number. Prove that if $n + 1$ distinct numbers are chosen from the set $\{1, \dots, 2n\}$, then there must be at least two among them whose sum is $2n$.

Exercise 1.1.13. Prove that for any three numbers a, b , and c ,

$$a^3 + b^3 + c^3 - 3abc = (a + b + c)(a^2 + b^2 + c^2 - ab - bc - ca).$$

Show that

$$a^2 + b^2 + c^2 \geq ab + bc + ca,$$

and that equality happens if and only if $a = b = c$.

Exercise 1.1.14. If $x, y, z \geq 0$, prove that

$$\frac{x + y + z}{3} \geq \sqrt[3]{xyz}.$$

1.2. The Principle of Induction

We mentioned in Section 1.1 a crucial property of natural numbers: For each natural number n there exists $n + 1$ with $n < n + 1$. This procedure never ends. One simple question remains unanswered. Suppose we start with $n = 1$, we get next that $2 = 1 + 1$, then $3 = 2 + 1 = (1 + 1) + 1$ and so on. Do we get by this construction **all** natural numbers? This is not as easy to answer as it looks like. Consider, for example, the collection of numbers $A = \{1, 1.5, 2, 2.5, 3, 3.5, \dots\}$. If a number x belongs to A , then so does $x + 1$. But, starting either at 1 or at 1.5, we never get all members of A by arbitrarily often executing the procedure $x \rightarrow x + 1$. This tells us that we have to make sure that the only natural numbers are $1, 1 + 1, 1 + 1 + 1, \dots$ and nothing more. That is, for any natural number n , no natural number can ever be between n and $n + 1$.

This is ensured by the **principle of induction** which cannot be derived from the other properties of \mathbb{N} .

Axiom 1.2.1. *The natural numbers \mathbb{N} possess the following property:*

$$(1.2.1) \quad (\forall A \subseteq \mathbb{N})[(1 \in A \text{ and } (n \in A \Rightarrow n + 1 \in A)) \Rightarrow A = \mathbb{N}]$$

Another way to express the previous axiom is as follows: There is no collection of natural numbers different from \mathbb{N} which contains 1 and with each n also $n + 1$. In particular, there do **not** exist $n, m \in \mathbb{N}$ with $n < m < n + 1$.

This induction property of \mathbb{N} is a very powerful tool for proving mathematical statements or propositions.

Proposition 1.2.1 (Principle of Mathematical Induction). *For $n \in \mathbb{N}$, let $P(n)$ be a mathematical statement that depends on n . Suppose that*

- (1) *The proposition $P(1)$ is true.*
- (2) *If $P(n)$ is true for some $n \geq 1$, then $P(n + 1)$ is true.*

*Then $P(n)$ is true for **all** $n \in \mathbb{N}$.*

Proof: Set $A := \{n \in \mathbb{N} : P(n) \text{ is true}\}$. By assumption, $1 \in A$, and, moreover, if $n \in A$, then $n + 1 \in A$. Consequently, the principle of induction lets us conclude that $A = \mathbb{N}$, meaning that $P(n)$ is true for all $n \geq 1$. \blacksquare

The first item above “ $P(1)$ true” is usually called **the base case** of the induction proof. The second item “ $P(n)$ implies $P(n + 1)$ ” or $P(n) \Rightarrow P(n + 1)$ is often referred to as **the induction step** and in its proof, $P(n)$ is called **the induction hypothesis**.

We can use dominoes to make the induction idea more concrete. Imagine that you have infinitely many domino pieces: one for each natural number. Initially all the pieces are standing up vertically. The domino piece at number n will fall down if our proposition $P(n)$ is true. The base case $P(1)$ ensures that our first domino falls. The induction step $P(1) \Rightarrow P(2)$ tells us that the first domino piece will make the second domino piece fall, and then will tell us that $P(2) \Rightarrow P(3)$ meaning that the second piece will fall and make the third one fall and so on. In the end, all the pieces will fall meaning that $P(n)$ is true for every n .

A more heroic interpretation of the proof by induction is to imagine climbing a ladder whose rungs correspond to the propositions $P(n)$ for $n \geq 1$. In order to clear all the rungs, we must get on the ladder and prove $P(1)$, and after that, once we get on the rung $P(n)$, we should figure out how to get to $P(n + 1)$, for any $n \geq 1$.

Proposition 1.2.2. *For any natural number n , the sum of the first n natural numbers equals*

$$(1.2.2) \quad 1 + \dots + n = \frac{n(n + 1)}{2}.$$

Proof: Let n be a natural number. We denote by $P(n)$ the statement given in equation (1.2.2). The proposition $P(n)$ is true if equation (1.2.2) is valid. Otherwise, $P(n)$ is false.

For example,

$$P(3) : 1 + 2 + 3 = \frac{3(3 + 1)}{2}$$

and

$$P(2021) : 1 + 2 + \dots + 2021 = \frac{2021(2021 + 1)}{2}.$$

Our job is proving that $P(n)$ is true for every natural number n . Of course, we can check that $P(3)$ is true and $P(2021)$ is true by direct calculation (if you are brave), but that is not sufficient as there are still many other numbers n requiring verification.

To start our induction proof, first note that $P(1)$ is true because $1 = \frac{1(1+1)}{2}$. This shows that the base case is true.

For the induction step, suppose that for some $n \in \mathbb{N}$, $P(n)$ is true. That means

$$(1.2.3) \quad 1 + \dots + n = \frac{n(n + 1)}{2}$$

is true. We have to prove that $P(n + 1)$ is true, that is we have to show that

$$(1.2.4) \quad 1 + \dots + n + (n + 1) = \frac{(n + 1)(n + 2)}{2}.$$

To avoid confusion, one should start with what is known to be true which is equation (1.2.3). That is the starting point of our proof journey and equation (1.2.4) is the final

destination. When comparing (1.2.3) and (1.2.4), we observe that the left-hand sides of these equations look very similar. A natural thing to do is to start from (1.2.3), tweak the left-hand side to make it look like in (1.2.4) and see what happens. More precisely, from equation (1.2.3) (which we know to be true), we add $n + 1$ to both sides, and we get

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

At this point, we obtained the left-hand side of (1.2.4) and so there is no point doing anything on that side. The only thing we can do is to simplify the expression on the right-hand side

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

which is the same as

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

Putting these things together, we end up

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

We have reached our destination, equation (1.2.4), and therefore, $P(n + 1)$ is true. By the principle of induction, this proves our identity for all natural numbers n . ■

We note here that the result above has a shorter solution that does not require induction. Carl Friedrich Gauss (1777–1855) was a German mathematician who made important contributions to mathematics and science. He is considered to be one of the greatest mathematicians in history. We describe below essentially the proof that a seven years old Gauss gave his teacher when he was asked to add all the natural numbers from 1 to 100. The general idea works as follows. Denote

$$S = 1 + \cdots + n.$$

Now write S in two different ways as:

$$S = 1 + \cdots + n$$

$$S = n + \cdots + 1.$$

Adding up these two equations, we will get $2S$ on the left-hand side. When we add up the right-hand side, we will do it column by column and will get

$$(n + 1) + (n + 1) + \cdots + (n + 1)$$

with the sum having n terms and therefore, being equal to $n(n + 1)$. Thus, $2S = n(n + 1)$ which coincides with our result above $S = \frac{n(n+1)}{2}$. This is a neat proof, and we will apply it again in the next section to arithmetic progressions. The argument above has a geometric interpretation given by Figure 1.2.1. We make a quick remark before moving on. The numbers of the form $\frac{n(n+1)}{2}$ are called triangular numbers and the explanation for this name is given in Figure 1.2.2. The sequence of triangular numbers appears as sequence OA000217 in the *Online Encyclopedia of Integer Sequences*, <https://oeis.org/A000217>. Like most people, mathematicians like comfort and saving effort when possible. Therefore, notations have been invented to simplify and

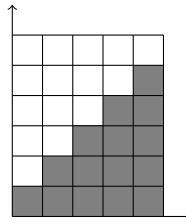


Figure 1.2.1. A geometric proof of $1 + \dots + n = \frac{n(n+1)}{2}$.

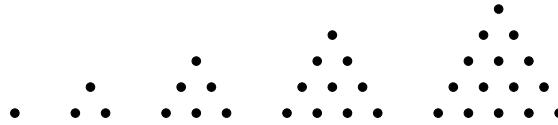


Figure 1.2.2. The first triangular numbers $\frac{n(n+1)}{2}$ for $n \leq 5$.

reduce the size of various mathematical expressions. For sums like $1 + \dots + n$, it is common to use $\sum_{k=1}^n k$ as notation. Similarly, for the sum of the squares of the first n natural numbers, one writes $\sum_{k=1}^n k^2$. More general, we define the following very useful abbreviations.

Definition 1.2.1. Given n numbers a_1, \dots, a_n , then their sum as well as their product will be written as

$$\sum_{k=1}^n a_k = a_1 + \dots + a_n \quad \text{and} \quad \prod_{k=1}^n a_k = a_1 \cdot \dots \cdot a_n.$$

So, for example

$$\sum_{k=1}^n \frac{1}{k} = 1 + \frac{1}{2} + \dots + \frac{1}{n} \quad \text{and} \quad \prod_{k=1}^n k = 1 \cdot 2 \cdot \dots \cdot n := n!.$$

In Proposition A.3.10 of the Appendix one can find some useful properties of this notation for finite sums and products, respectively.

The trick presented above of reversing the sum will not work for all sequences. For example, try to come up with a formula for

$$\sum_{k=1}^n k^2 = 1^2 + 2^2 + \dots + n^2$$

and you will see that you run into trouble when applying the previous method. However, you can use the principle of induction to show that the sum above equals $\frac{n(n+1)(2n+1)}{6}$ (see Exercise 1.2.2).

Example 1.2.1. Here is another example of an exercise which has both an inductive and a noninductive proof. Experimenting with sums of consecutive numbers, we can try summing up the first n consecutive even numbers:

$$2 + 4 + \dots + 2n.$$

This sum turns out to be easy to evaluate since we can take 2 as a common factor and use our previous result to get that

$$\begin{aligned} 2 + 4 + \cdots + 2n &= 2(1 + 2 + \cdots + n) \\ &= 2 \cdot \frac{n(n+1)}{2} = n(n+1). \end{aligned}$$

We may be discouraged to do the same trick for the sum of the first n consecutive odd numbers given that 2 does not factor into an odd number. Experimenting with some small values of n , we come up with the following:

$$\begin{aligned} 1 &= 1 \\ 1 + 3 &= 4 \\ 1 + 3 + 5 &= 9 \\ 1 + 3 + 5 + 7 &= 16 \\ 1 + 3 + 5 + 7 + 9 &= 25. \end{aligned}$$

Perhaps the skeptical reader may need more examples to guess a pattern here, but it is likely that most of us are convinced that the general pattern of

$$1 + 3 + \cdots + (2n - 1) = n^2$$

must be true. This is the case indeed, and it can be proved in various ways including induction on n . We also include below a geometric proof that does not use induction.

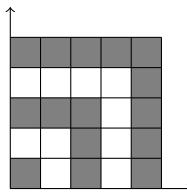


Figure 1.2.3. A geometric proof of $1 + 3 + \cdots + (2n - 1) = n^2$.

Despite our disparaging comment above about taking 2 as a common factor, the gritty reader may be able to pull this through, and we also leave this as an exercise. Our examples above may seem a bit pedantic, but we believe that it is important to know different proofs for a single mathematical statement.

Example 1.2.2. Say we have n distinct points on a circle, and we draw all the chords between them. Assume that no three chords pass through the same point inside the circle. In how many regions have we partitioned the interior of the circle? Let R_n denote this number. It is easy to see that $R_2 = 2$ (duh!) and five seconds of drawing and counting will convince us that $R_3 = 4$ and $R_4 = 8$. When $n = 5$, it takes a bit more work and care to make sure no three chords intersect inside the circle, but one can draw perhaps a picture similar to the one below and convince ourselves that $R_5 = 16$. It is a pattern that seems familiar: 2, 4, 8, 16 so a natural guess would be that $R_6 = 32$ and perhaps that $R_n = 2^{n-1}$ for any $n \geq 6$ in the case of a bold reader. To the surprise of many, it turns out that $R_6 = 31$ and that R_n is much smaller than 2^{n-1} for larger n . We will return to this problem in Section 1.7 where we will determine R_n for any $n \geq 2$.

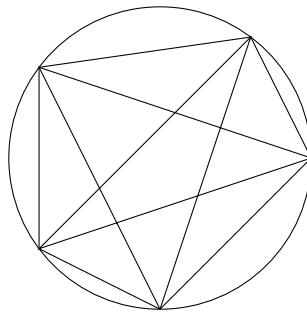


Figure 1.2.4. Five points on a circle create 16 regions.

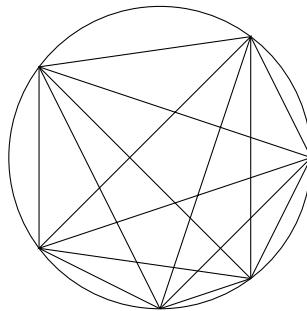


Figure 1.2.5. Six points on a circle create 31 regions.

We have seen that appearances can be deceiving and there exist problems where a solution that fits can be found for several small cases, but will not work in general. We now turn to a different type of problem that can be solved by induction, a problem where a certain pattern does not start at $n = 1$, but at some larger value of n .

Example 1.2.3. Let us consider two sequences of numbers: an exponential $(2^n)_{n \geq 1}$ and a quadratic $(4n^2)_{n \geq 2}$. If we compare these sequences for the first few values, we may get a table such Table 1.2.1. At this point, having checked more cases than in our points on the circle problem, one may guess that $4n^2$ beats 2^n . However, that is not true, and these small cases are deceiving us. Already for $n = 8$, things begin to turn around as $2^8 = 256 = 4 \cdot 8^2$ and when $n = 9$, things change as $2^9 = 512$ while $4 \cdot 9^2 = 324$. As you may be aware from your calculus classes, exponential growth beats any polynomial after a certain point. This seems to be the case here, namely that 2^n is strictly larger than $4n^2$ for $n \geq 9$. So it looks like an induction type of problem, where the only difference is that the statement we are trying to prove does not hold from 1 onwards, but from a larger value (in this case 9) onwards.

Table 1.2.1. Comparison of 2^n and $4n^2$ for small n .

| n | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|--------|---|----|----|----|-----|-----|-----|
| 2^n | 2 | 4 | 8 | 16 | 32 | 64 | 128 |
| $4n^2$ | 4 | 16 | 36 | 64 | 100 | 144 | 196 |

Proposition 1.2.1 can be refined to deal with such problems as follows.

Proposition 1.2.3 (Principle of Mathematical Induction). *Let $n_0 \in \mathbb{N}_0$. For each natural number $n \geq n_0$ let $P(n)$ be a proposition that depends on n . Suppose that*

- (1) *The proposition $P(n_0)$ is true.*
- (2) *If $P(n)$ is true for some $n \geq n_0$, then $P(n + 1)$ is true.*

Then $P(n)$ is true for all $n \geq n_0$.

The proof of this result is similar to the one of Proposition 1.2.1, and we leave it as an exercise.

With this method in mind, let us prove that

$$(1.2.5) \quad 2^n > 4n^2$$

for any $n \geq 9$.

Denote by $P(n)$ the inequality $2^n > 4n^2$. As we have seen before, $P(n)$ is false for $1 \leq n \leq 8$. To prove (1.2.5), we must first figure out our n_0 which is fairly easy to do from the statement of the problem. Usually, these problems will give you n_0 in their statements. In our case, $n_0 = 9$.

Now the base case is $P(9) : 2^9 > 4 \cdot 9^2$ which is true as we have checked earlier that $2^9 = 512 > 4 \cdot 9^2 = 324$. For the induction step, let $n \geq 9$ be a natural number and assume that $P(n)$ is true, namely that $2^n > 4n^2$. We want to prove that $P(n + 1)$ is true which is the same as $2^{n+1} > 4(n + 1)^2$. As usual, we start with the things that we know, namely that $2^n > 4n^2$. If we multiply both sides by 2, we get that $2^{n+1} = 2 \cdot 2^n > 8n^2$. It is always useful to keep an eye on the prize which in our case is $P(n + 1) : 2^{n+1} > 4(n + 1)^2$. Inequalities are tricky because they sometimes require you to take a leap of faith or trust your intuition. We know that $2^{n+1} > 8n^2$ and we would like to show that $2^{n+1} > 4(n + 1)^2$. If it were true that $8n^2 > 4(n + 1)^2$, then we would be done. It feels like stumbling in the darkness which is quite what we are doing in some sense.

To see if $8n^2 > 4(n + 1)^2$, we can try to see if this is equivalent to some other inequality that may be easier to figure out. It is not too hard to see that

$$8n^2 > 4(n + 1)^2 \Leftrightarrow 2n^2 > (n + 1)^2 \Leftrightarrow (n - 1)^2 > 2.$$

A general rule is to try to simplify the inequality as much as possible by replacing it with equivalent, but simpler inequalities at each step. It is good to do such things in small steps in order not to make errors. The solutions may diverge between students as for example the inequality $n^2 > 2n + 1$ is also equivalent to the inequality $n^2 - 1 > 2n$ which is the same as $(n - 1)(n + 1) > 2n$. If things are done correctly, all these inequalities are equivalent to each other and deciding the truth value of one will give you the truth

value of each of them. Let us consider the inequality $(n - 1)^2 > 2$, for example. Since our n was at least 9, we know that $n - 1 \geq 8$ and therefore $(n - 1)^2 \geq 8^2 = 64$ which is certainly larger than 2. Thus, the inequality $(n - 1)^2 > 2$ is true for $n \geq 9$ which means that $8n^2 > 4(n + 1)^2$ holds. Therefore, $2^{n+1} > 8n^2 > 4(n + 1)^2$ which means that $P(n + 1)$ is true and finishes our proof.

Example 1.2.4. Inequalities are tricky and when proving them by induction, require a bridge between things that we know and things that we wish were true to give us the required answer. Let us give one example to make this statement more concrete. Consider the sum

$$\sum_{k=1}^n \frac{1}{k^2},$$

for small values of n . You may convince yourself that perhaps

$$\sum_{k=1}^n \frac{1}{k^2} < 2,$$

for any $n \geq 1$. Let us try to prove it by induction on n . It is clear that the base case is true since for $n = 1$, $\sum_{k=1}^n \frac{1}{k^2} = 1 < 2$. For the induction step, assume that $n \geq 1$ is a natural number and $\sum_{k=1}^n \frac{1}{k^2} < 2$. We would like to prove that $\sum_{k=1}^{n+1} \frac{1}{k^2} < 2$. The only thing we have to work with is the inequality

$$\sum_{k=1}^n \frac{1}{k^2} < 2.$$

The obvious and pretty much the only step we can take is to add $\frac{1}{(n+1)^2}$ to both parts, giving us the inequality

$$(1.2.6) \quad \sum_{k=1}^{n+1} \frac{1}{k^2} < 2 + \frac{1}{(n+1)^2}.$$

This is where we are now. The inequality we want to get to is

$$(1.2.7) \quad \sum_{k=1}^{n+1} \frac{1}{k^2} < 2.$$

Reflecting for a moment on our situation, we see that inequality (1.2.6) is weaker than (1.2.7) and will not be able to lead us to it. This can be a disappointing moment for a student that has worked hard for a solution, but you can be assured that most mathematicians have experienced such moments and that is useful to know what methods do not work as well as the proofs that work.

The paradox of this problem is that in order to solve it, we will actually prove a stronger inequality. Let us prove that

$$(1.2.8) \quad \sum_{k=1}^n \frac{1}{k^2} \leq 2 - \frac{1}{n},$$

for any $n \geq 1$. It should be clear that (1.2.8) implies (1.2.7).

To prove (1.2.8), we use induction on n . The base case $n = 1$ is the same as $\sum_{k=1}^1 \frac{1}{k^2} \leq 2 - \frac{1}{1}$ which is certainly true as both sides are 1. For the induction step, let $n \geq 1$ be a natural number and assume that

$$(1.2.9) \quad \sum_{k=1}^n \frac{1}{k^2} \leq 2 - \frac{1}{n}.$$

We would like to prove that

$$(1.2.10) \quad \sum_{k=1}^{n+1} \frac{1}{k^2} \leq 2 - \frac{1}{n+1}.$$

Starting with what we know, namely (1.2.9), we can add $\frac{1}{(n+1)^2}$ to both sides and get

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

This inequality has the left-hand side as our goal (1.2.10), but the right-hand sides are different.

If we could prove that $2 - \frac{1}{n} + \frac{1}{(n+1)^2} \leq 2 - \frac{1}{n+1}$, then that would show that (1.2.9) (which we know is true) implies (1.2.10) (which we want to show is true) and would finish our induction proof.

We use our method from before as writing the inequality in simpler but equivalent forms:

$$\begin{aligned} 2 - \frac{1}{n} + \frac{1}{(n+1)^2} \leq 2 - \frac{1}{n+1} &\Leftrightarrow \frac{1}{(n+1)^2} \leq \frac{1}{n} - \frac{1}{n+1} \\ &\Leftrightarrow n(n+1) \leq (n+1)^2 \\ &\Leftrightarrow n \leq n+1. \end{aligned}$$

This last inequality is certainly true (it is the same as $0 \leq 1$) and it gives us the desired proof by induction of inequality (1.2.8).

One may of course ask how do you come up with such stronger inequalities. For the problem above, one may wish to read the section on arithmetic and geometric progressions where similar sums are considered and also later on the section on real numbers and limits. It turns out that as n gets larger, the sum $\sum_{k=1}^n \frac{1}{k^2}$ approaches $\frac{\pi^2}{6}$. This is called the Basel problem and the evaluation of $\sum_{k=1}^n \frac{1}{k^2}$ as n gets large, was proposed by the Italian mathematician Pietro Mengoli (1626–1686) in 1650 and solved by Euler in 1734. Leonhard Euler (1707–1783) was a Swiss mathematician, considered one of the greatest mathematicians in the world with numerous contributions in many areas of mathematics. You can learn more about his work from the wonderful book [9].

Our next example of using induction shows that the arithmetic mean of any n real nonnegative numbers is at least their geometric mean. This is known as the arithmetic mean–geometric mean inequality or the AM–GM inequality.

Proposition 1.2.4 (AM–GM inequality). *Let $n \in \mathbb{N}$. For any n real nonnegative numbers a_1, \dots, a_n ,*

$$(1.2.11) \quad \frac{a_1 + \cdots + a_n}{n} \geq \sqrt[n]{a_1 \cdots a_n}.$$

Proof: For $n = 1$, the inequality is equality so let us assume $n \geq 2$. We first prove the following result which will imply our inequality above. For any n real nonnegative numbers x_1, \dots, x_n whose product equals one,

$$(1.2.12) \quad x_1 + \cdots + x_n \geq n.$$

To see how inequality (1.2.12) implies (1.2.11), consider a_1, \dots, a_n real nonnegative numbers. If one of them is zero, then their geometric mean is zero and their arithmetic mean is at least zero, so we are done in this case. If all the numbers a_1, \dots, a_n are positive, define

$$x_1 = \frac{a_1}{\sqrt[n]{a_1 \cdots a_n}}, \quad x_2 = \frac{a_2}{\sqrt[n]{a_1 \cdots a_n}}, \dots, \quad x_n = \frac{a_n}{\sqrt[n]{a_1 \cdots a_n}}.$$

Therefore, the product of x_1, \dots, x_n is one and by the (yet unproven) inequality (1.2.12), we get that $x_1 + \cdots + x_n \geq n$. Substituting x_1, \dots, x_n by their formulas yields the arithmetic mean–geometric mean inequality (1.2.11).

To prove inequality (1.2.12), we use induction on n . The base case $n = 2$ follows from Proposition 1.1.4. We have seen the case $n = 3$ before as well (see Exercise 1.1.14).

For the induction step, let $n \in \mathbb{N}$ such that $n \geq 2$. Assume that inequality (1.2.12) is true for any n real nonnegative numbers x_1, \dots, x_n with $x_1 \cdots x_n = 1$. We will show that for any $n + 1$ real nonnegative numbers y_1, \dots, y_n, y_{n+1} with $y_1 \cdots y_n \cdot y_{n+1} = 1$, the following inequality is true: $y_1 + \cdots + y_n + y_{n+1} \geq n + 1$.

If all the numbers y_1, \dots, y_n, y_{n+1} are equal to one, then the inequality (1.2.12) is equality and is true. Otherwise, since the product of the numbers y_1, \dots, y_n, y_{n+1} is one, there must be at least one of these numbers that is greater than one and at least one of these numbers that is smaller than one. Without loss of generalization, assume that $y_n > 1 > y_{n+1}$. Therefore, $(y_n - 1)(1 - y_{n+1}) > 0$ which means that $y_n + y_{n+1} > 1 + y_n y_{n+1}$. Because the product of the $n + 1$ numbers $y_1, y_2, \dots, y_n, y_{n+1}$ is one, we have that the product of the n numbers $y_1, y_2, \dots, y_{n-1}, y_n y_{n+1}$ is one. Therefore, by our induction hypothesis, we must have that

$$y_1 + y_2 + \cdots + y_{n-1} + y_n y_{n+1} \geq n.$$

Because $y_n y_{n+1} < y_n + y_{n+1} - 1$, we deduce that $y_1 + y_2 + \cdots + y_n + y_{n+1} > n + 1$. This finishes our proof. ■

Remark 1.2.1. We mention briefly here that Proposition 1.2.4 may be proved using calculus. The function $\log : (0, +\infty) \rightarrow \mathbb{R}, x \mapsto \log x$ is concave down as its second derivative is negative. A result called Jensen's inequality implies that

$$\log\left(\frac{a_1 + \cdots + a_n}{n}\right) \geq \frac{1}{n} \sum_{j=1}^n \log a_j = \log\left(\sqrt[n]{a_1 \cdots a_n}\right),$$

which implies (1.2.11) because the logarithm function is increasing.

The arithmetic and the geometric mean are two of the three Pythagorean means. The third one is the so-called harmonic mean which is defined as follows.

Definition 1.2.2. Let n be a natural number and a_1, \dots, a_n be n positive numbers. The **harmonic mean** (HM) of the numbers a_1, \dots, a_n is defined as

$$\frac{n}{\frac{1}{a_1} + \dots + \frac{1}{a_n}}.$$

Example 1.2.5. The harmonic, geometric, and arithmetic mean of $a_1 = 2$, $a_2 = 3$, and $a_3 = 5$ equal

$$\text{HM} = \frac{3}{\frac{1}{2} + \frac{1}{3} + \frac{1}{5}} = \frac{90}{31} \approx 2.903, \quad \text{GM} = \sqrt[3]{30} \approx 3.1072 \quad \text{and} \quad \text{AM} = \frac{10}{3} \approx 3.333.$$

Example 1.2.6. Suppose n cars drive from city A to city B with average speeds v_1, \dots, v_n . Then the average speed v of all n cars together is given by the harmonic mean of the single speeds. That is,

$$v = \frac{n}{\frac{1}{v_1} + \dots + \frac{1}{v_n}}.$$

Why is this so? If s is the distance between A and B and if the times the cars need from A to B are t_1, \dots, t_n , then one has

$$v_1 = \frac{s}{t_1}, \dots, v_n = \frac{s}{t_n}.$$

Hence we get

$$v = \frac{ns}{t_1 + \dots + t_n} = \frac{n}{\frac{t_1}{s} + \dots + \frac{t_n}{s}} = \frac{n}{\frac{1}{v_1} + \dots + \frac{1}{v_n}}.$$

In Figure 1.2.6, we give a geometric proof of the AM-GM-HM inequality for two numbers. The points A, B , and C are on the circle with center at O and radius $\frac{a+b}{2}$ with the points B and C being diametrically opposite. The distance between C and D is a and the distance between D and B is b . The perpendicular at D on the line BC touches the circle at A and the length of the segment AD is \sqrt{ab} (prove this). The perpendicular from D to OA intersects OA at the point E . The reader should also prove that the length of the segment DE is $\frac{2}{\frac{1}{a} + \frac{1}{b}} = \frac{2ab}{a+b}$.

A more general result is true for any n . We state it now and prove it below.

Proposition 1.2.5 (AM-GM-HM inequality). *The three Pythagorean means are related by*

$$\text{HM} \leq \text{GM} \leq \text{AM}.$$

Equivalently, for all positive a_1, \dots, a_n it follows

$$(1.2.13) \quad \frac{n}{\frac{1}{a_1} + \dots + \frac{1}{a_n}} \leq \sqrt[n]{a_1 \cdots a_n} \leq \frac{a_1 + \dots + a_n}{n}.$$

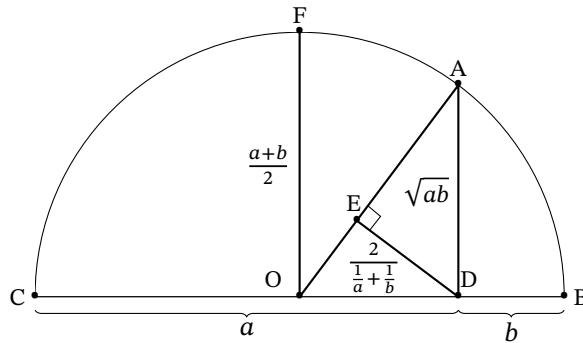


Figure 1.2.6. A geometric proof of the AM-GM-HM inequality for $n = 2$.

Proof: Let a_1, \dots, a_n be given positive real numbers. An application of (1.2.11) to $1/a_1, \dots, 1/a_n$ leads to

$$(1.2.14) \quad \frac{1}{\sqrt[n]{a_1 \cdots a_n}} \leq \frac{\frac{1}{a_1} + \cdots + \frac{1}{a_n}}{n}.$$

If we invert both sides of (1.2.14), we get that

$$\frac{n}{\frac{1}{a_1} + \cdots + \frac{1}{a_n}} \leq \sqrt[n]{a_1 \cdots a_n}$$

as claimed in (1.2.13). Since we already proved the right-hand estimate in (1.2.13), this completes the proof. ■

We finish the section with some examples from graph theory where the method of induction is often used. Graphs are abstract models of networks and are useful in many applications.

Definition 1.2.3. A graph $G = (V, E)$ consists of a set of vertices V , a collection of edges E and a correspondence that associates to each edge two (not necessarily distinct) vertices called its **endpoints**.

The vertices can be represented as points in a plane (or other space if you like) and each edge is a curve or segment joining its endpoints. A **loop** is an edge whose endpoints are the same. **Multiple edges** are edges with the same endpoints. A graph with neither loops nor multiple edges is usually called a **simple graph**.

The graph in Figure 1.2.7 has vertex set $V = \{v_1, v_2, v_3, v_4, v_5\}$ and edge set $E = \{e_j : 1 \leq j \leq 9\}$. The correspondence between the edges and pairs of vertices is clear from the picture. The edge e_7 is a loop with v_5 and v_5 as endpoints and the edges e_1 and e_2 are multiple edges with the same endpoints v_1 and v_4 . If x and y are endpoints of an edge e , we say that x and y are **adjacent** and that the vertex x (or y) is **incident with** or **contained** in e .

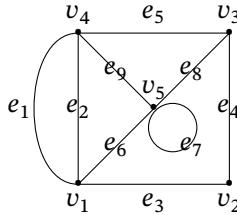


Figure 1.2.7. A graph with 5 vertices and 9 edges.

The **degree** $d(x)$ of a vertex $x \in V$ equals the number of edges that contain it. In the case that the vertex is contained in loops, each loop contributes two to the degree of that vertex. For the graph in Figure 1.2.7, Table 1.2.2 gives the degrees of its vertices.

Table 1.2.2. The degrees of the graph in Figure 1.2.7.

| vertex | v_1 | v_2 | v_3 | v_4 | v_5 |
|--------|-------|-------|-------|-------|-------|
| degree | 4 | 2 | 3 | 4 | 5 |

Each edge contributes one to each degree of its endpoints if it is not a loop or contribute two to the degree of a vertex, otherwise. The following result uses this idea and gives a relation between the degrees of the vertices and the number of edges of a graph. It has been given the name of Handshaking lemma as one can interpret the vertices of G as people and any edge of G as a handshake between its endpoints (or perhaps a pat on a back if the edge is a loop).

Proposition 1.2.6 (Handshaking Lemma). *For any graph, the sum of the degrees of its vertices is twice the number of its edges.*

Proof: When we sum the degrees of the vertices of G , each nonloop edge is counted twice (once for each of its endpoints). Each loop edge is counted twice according to our definition of the degree. Therefore, the sum of the degrees will be twice the number of edges. ■

To make the idea of the proof a bit more concrete, consider a rectangular array or a matrix M whose rows are labeled by the vertices of G and whose columns are indexed by the edges of G with entries defined as follows:

$$M(u, e) = \begin{cases} 2 & \text{if } e \text{ is a loop incident with } u \\ 1 & \text{if } e \text{ is a nonloop edge incident with } u \\ 0 & \text{if } e \text{ is not incident with } u. \end{cases}$$

The matrix M is usually called an **incidence matrix** of G . For our example in Figure 1.2.7, it has 5 rows and 9 columns (labeled for the convenience of the reader): The row corresponding to a vertex x tells us which edges are incident with x and the sum of its entries equals the degree of x . Thus, the sum of the entries in M is the sum of the degrees of the vertices of G . On the other hand, the column corresponding to any edge

Table 1.2.3. An incidence matrix for the graph in Figure 1.2.7.

| vertex/edge | e_1 | e_2 | e_3 | e_4 | e_5 | e_6 | e_7 | e_8 | e_9 |
|-------------|-------|-------|-------|-------|-------|-------|-------|-------|-------|
| v_1 | 1 | 1 | 1 | 0 | 0 | 1 | 0 | 0 | 0 |
| v_2 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 |
| v_3 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 1 |
| v_4 | 1 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 1 |
| v_5 | 0 | 0 | 0 | 0 | 0 | 1 | 2 | 1 | 1 |

will contain two entries of 1 (and the rest of 0) if the edge is not a loop or will contain one entry of 2 (and the rest of 0) if the edge is a loop. In either case, the sum of the entries of any column is 2. Hence, the sum of the entries of M is twice the number of edges. This is the statement of the Handshaking lemma.

Definition 1.2.4. A **walk** in a graph G is an alternating sequence of (not necessarily distinct) vertices and edges: $x_0 h_0 x_1 \dots x_{k-1} h_k x_k$ for some $k \geq 1$ such that h_j is an edge with endpoints x_j and x_{j+1} for any $0 \leq j \leq k - 1$. Such a walk is called **closed** if $x_0 = x_k$.

The sequence $v_1 e_1 v_4 e_1 v_1 e_6 v_5 e_7 v_5$ is a walk that is different from the walk $v_1 e_1 v_4 e_2 v_1 e_6 v_5 e_7 v_5$.

Definition 1.2.5. A **path** in a graph G is a walk whose vertices x_0, \dots, x_k are all distinct. We can say that this is an x_0, x_k -path.

In Figure 1.2.7, $v_1 e_2 v_4 e_5 v_3 e_8 v_5$ is a path, but $v_1 e_6 v_5 e_7 v_5 e_8 v_2$ is not.

Definition 1.2.6. A **cycle** is a closed walk where all the edges are distinct and all the vertices are distinct except for $x_0 = x_k$. The **length** of a walk, path, or cycle equals the number of its edges.

The sequences $v_1 e_2 v_4 e_1 v_1$, $v_5 e_7 v_5$ and $v_1 e_1 v_4 e_5 v_3 e_4 v_2 e_3 v_1$ are cycles (of length 2, 1, and 4, respectively) while $v_1 e_6 v_5 e_7 v_5 e_9 v_4 e_2 v_1$ and $v_1 e_1 v_4 e_2 v_1 e_6 v_5 e_8 v_3 e_4 v_2 e_3 v_1$ are not.

Definition 1.2.7. A graph G is **connected** if for any two distinct vertices x and y , there exists an x, y -path.

The graph in Figure 1.2.7 is connected. It is likely that some readers have seen graphs before. For example, some chemical molecules can be represented as certain graphs that are connected and acyclic (have no cycles). Such graphs are called trees.

Definition 1.2.8. A **tree** is a connected and acyclic graph.

We determine some basic properties of trees below. A leaf in a tree is a vertex of degree one.

Proposition 1.2.7. Any tree with $n \geq 2$ vertices contains at least two leaves.

Proof: Let T be a tree with $n \geq 2$ vertices. Consider a path P of maximum length $x_0 x_1 \dots x_k$ in T for some $k \geq 1$. The vertex x_0 cannot be adjacent to any vertex outside

$\{x_1, \dots, x_k\}$ because if it were, then we would extend the path P with an edge containing x_0 and this would contradict the fact that P has maximum length. If x_0 had a neighbor x_j among x_2, \dots, x_k , then $x_0, x_1, \dots, x_j, x_0$ would be a cycle. However, T is a tree and has no cycles. We conclude that the only neighbor of x_0 is x_1 and x_0 must have degree one. By a similar argument, one can show that x_k has degree one, and we leave the details as an exercise. ■

The next result shows that the number of edges of a tree is determined by the number of its vertices.

Proposition 1.2.8. *If a tree has n vertices, then it has $n - 1$ edges.*

Proof: We use induction on the number of vertices. The base case is $n = 1$. In this situation, our tree consists of a single vertex and no edges. Since $0 = 1 - 1$, the result is true for $n = 1$. For the induction step, let $n \geq 2$. Let T be a tree with n vertices. Assume that any tree on $n - 1$ vertices contains exactly $n - 2$ edges. By Proposition 1.2.7, T contains at least one leaf a . Denote by T' the graph obtained by deleting a and the edge containing it from T . The graph T' has $n - 1$ vertices since we deleted exactly one vertex from T . Also, T' contains no cycles as deleting vertices and edges cannot create cycles. Finally, T' is connected since if $b \neq c$ are vertices of T' , any path connecting them in T will survive in T' . Such a path is not using a nor the edge incident with it since a has degree one and every internal vertex of the path has degree two. Thus, T' is a tree with $n - 1$ vertices. By the induction hypothesis, T' has $n - 2$ edges. Now T has one more edge than T' and therefore, contains $n - 2 + 1 = n - 1$ edges. This finishes our proof. ■

The proof of the following corollary will be left as an exercise.

Corollary 1.2.9. *A connected graph with n vertices has at least $n - 1$ edges.*

The astute reader may have noticed something about the last two figures, namely that no two edges in them crossed.

Definition 1.2.9. A connected graph G is called **planar** if its vertices can be drawn as points in the plane and its edges as curves joining their endpoints such that any two edges cannot intersect each other except at common endpoints. Any such drawing is called a plane drawing of the planar graph G .

For example, the graph in Figure 1.2.7 and the three trees in Figure 1.2.8 are planar.

Definition 1.2.10. Given a drawing of a planar graph G , the plane is partitioned into regions called **faces**, one of which is **unbounded**. The **length** of a face in a planar drawing is the number of edges in the closed walk of G bounding the face.



Figure 1.2.8. Three trees on 5 vertices.

The graph in Figure 1.2.7 has 6 faces:

$$\begin{aligned}F_1 &= \{v_1, e_1, v_4, e_5, v_3, e_4, v_2, e_3, v_1\} \\F_2 &= \{v_1, e_1, v_4, e_2, v_1\}, \\F_3 &= \{v_1, e_2, v_4, e_9, v_5, e_6, v_1\}, \\F_4 &= \{v_3, e_5, v_4, e_9, v_5, e_8, v_3\}, \\F_5 &= \{v_5, e_7, v_5\} \\F_6 &= \{v_1, e_6, v_5, e_7, v_5, e_8, v_3, e_4, v_2, e_3, v_1\}.\end{aligned}$$

The face F_1 is the infinite face and has length 4. Each tree in Figure 1.2.8 has one face (the infinite face) of length 8.

It turns out that the number of faces in a plane drawing of a planar graph is determined by its number of vertices and edges. This is a famous result of Euler obtained in the context of polyhedra. A **polyhedron** is a 3-dimensional region of the space which is the intersection of finitely many half-spaces. For example, the cube is the intersection of 6 half-spaces, one for each of its faces.

Theorem 1.2.10 (Euler). *If G is a connected and planar graph with v vertices and e edges. If f is the number of faces of a plane drawing of G , then*

$$(1.2.15) \quad v - e + f = 2.$$

Proof: We give a proof by induction on the number of edges e . The base case is $e = 0$ and in this situation, the graph has no edges, so it must consist of a single vertex. Then $v = 1$ and $f = 1$ and $v - e + f = 1 - 0 + 1 = 2$. For our peace of mind, we can also settle $e = 1$. In this case the graph consists of two vertices connected by one edge and will have $v = 2$ and $f = 1$. Therefore, $v - e + f = 2 - 1 + 1 = 2$. For the induction step, let $v \geq 2$ and assume that G is a connected planar graph with v vertices, e edges and f faces. Assume that the formula (1.2.15) is true for any planar connected graph with $e-1$ edges. If G is a tree, then $e = v-1$ and $f = 1$. Therefore, $v - e + f = v - (v-1) + 1 = 2$ and we are done. If G is not a tree, then G contains a cycle. Delete one edge of that cycle and call the resulting graph H . Because we deleted an edge from a cycle, H is still connected. Also, deleting edges leaves the graph planar so H is connected and planar. Moreover, it has v vertices, $e-1$ edges and $f-1$ faces. By the induction hypothesis, the formula (1.2.15) is true for H and therefore, $v - (e-1) + (f-1) = 2$ which gives $v - e + f = 2$. This finishes our proof. ■

In Figure 1.2.9 we give two drawings of the cube: the one on the left is planar and the other is not. In the planar drawing, there are 6 faces. Euler's formula has many consequences, and we describe some of them here.

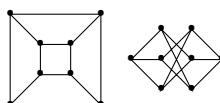


Figure 1.2.9. Two drawings of the cube.

Definition 1.2.11. The complete graph K_n with n vertices is a graph with n vertices in which any two distinct vertices are adjacent.

The complete graph K_n has the largest number of edges ($\binom{n}{2}$) among all simple graphs with n vertices (why?).

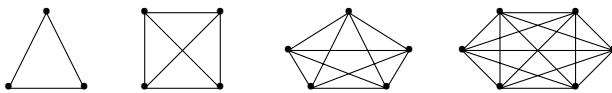


Figure 1.2.10. The complete graphs on 3, 4, 5, or 6 vertices.

Note that the complete graph with four vertices is given a nonplanar drawing in Figure 1.2.10, but can have a planar drawing as shown in Figure 1.2.11.



Figure 1.2.11. A planar drawing of the complete graph K_4 .

Proposition 1.2.11. The complete graph K_5 is not planar.

Proof: We use proof by contradiction. Assume that there is a plane drawing of K_5 . Using the notation from (1.2.15), for K_5 we have that $v = 5$ and $e = 10$. Therefore, the number of faces f must be $e - v + 2 = 7$. Each face has at least 3 edges and if we add up the number of edges in these 7 faces, we will get at least 21 edges. At the same time, each edge is involved in two faces and therefore, when we sum up all the edges in the faces of this drawing, we should get twice the number of edges which is 20. This is a contradiction which proves that K_5 is not planar. ■

The proof of the next corollary is left to the reader.

Corollary 1.2.12. If $n \geq 5$ is a natural number, then the complete graph K_n is not planar.

Another important class of graphs are the bipartite graphs. They arise in job assignments and other optimization problems.

Definition 1.2.12. A graph $G = (V, E)$ is called **bipartite** if there is a way to split its vertices into two nonempty sets A and B such that any edge of the graph has one endpoint in A and the other in B . The subsets A and B are called the **partite sets** of the bipartite graph G .

A consequence of the definition is that in a bipartite graph G with partite sets A and B , there are no edges with both endpoints in A or with both endpoints in B . The reader can prove that the complete graph K_n is not bipartite for $n \geq 3$.

Definition 1.2.13. A **cycle** graph is a simple graph whose vertices can be placed on a circle such that vertices are adjacent whenever they are consecutive on the circle. For n natural number, denote by C_n the cycle with n vertices.

The cycle C_1 consists of one vertex with one loop containing it. The cycle C_2 consists of two vertices joined by two multiple edges. The cycle C_3 is the same as the complete graph K_3 , but that is not true for $n \geq 4$, the cycle C_n is not the same as K_n . An algebraic way of describing the cycle C_n is being the graph with vertex set $\{1, \dots, n\}$ where there is one edge between k and $k + 1$ for any $1 \leq k \leq n - 1$ and one edge between 1 and n .

Proposition 1.2.13. *The cycle C_n is bipartite if and only if n is even.*

Proof: Assume that n is even. Let A consist of all the vertices from $\{1, \dots, n\}$ that are even and let B be the rest. Clearly, $1 \in B$ and $n \in A$ so the edge between 1 and n has one endpoint in A and the other in B . For $1 \leq k \leq n - 1$, one of the numbers k and $k + 1$ is even and the other is odd. Thus, the edge between k and $k + 1$ has one endpoint in A and the other in B . Hence, C_n is bipartite if n is even.

Assume that C_n is bipartite. There exists a partition of its vertex set $A \cup B$ such that each edge has one endpoint in A and the other in B . Suppose that $1 \in A$ (the case when $1 \in B$ is similar and left as an exercise). Because 1 is adjacent to 2 and n , it follows that $2, n \in B$. Because $2 \in B$, we get that $3 \in A$ which implies that $4 \in B$ and so on. One can make this formal and use induction on k to show that $k \in A$ and $k + 1 \in B$ for any odd k such that $1 \leq k \leq n - 1$. We leave the details of this part to the reader as well. Since $n \in B$, we must have that $n - 1 \in A$ and $n - 1$ is odd, therefore n is even. ■

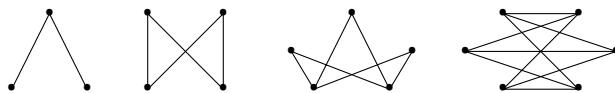


Figure 1.2.12. The complete bipartite graphs $K_{1,2}, K_{2,2}, K_{2,3}$, and $K_{3,3}$.

Definition 1.2.14. Let a, b be natural numbers. The **complete bipartite graph** $K_{a,b}$ is the bipartite graph with partite sets A and B such that $|A| = a, |B| = b$ with the property that every vertex in A is adjacent to every vertex in B .

The complete bipartite graph $K_{1,b}$ is planar for any $b \geq 1$. We give planar drawings below for $K_{2,2}$ and $K_{2,3}$ and invite the reader to discover planar drawings of $K_{2,b}$ for any $b \geq 4$ and prove that such graphs are planar.

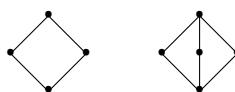


Figure 1.2.13. Planar drawings of $K_{2,2}$ and $K_{2,3}$.

We now give another application of Euler's formula (Theorem 1.2.10).

Proposition 1.2.14. *The complete bipartite graph $K_{3,3}$ is not planar.*

Proof: Assume that there is a planar drawing of $K_{3,3}$. The graph $K_{3,3}$ has $v = 6$ and $e = 9$. By Euler's formula, the number of faces in the planar drawing must be $f = e - v + 2 = 5$. The graph $K_{3,3}$ does not contain any cycles of length 3 or less so any face must have length 4 or more. Therefore, adding up the lengths of the faces in this planar drawing will give us at least $5 \cdot 4 = 20$ edges. At the same time, each edge is contained in two faces and this count should be the same as twice the number of edges which is 18. Hence, we arrive at the contradiction that $18 \geq 20$. This shows that our assumption was false and $K_{3,3}$ is not planar. ■

Exercise 1.2.1. Assume that n points are placed on a circle. Show that there are $\frac{n(n-1)}{2}$ chords passing through these points.

Exercise 1.2.2. Using induction, show that for any natural number n ,

$$1^2 + \dots + n^2 = \frac{n(n+1)(2n+1)}{6}.$$

Exercise 1.2.3. Using induction, show that for any natural number n ,

$$1^3 + \dots + n^3 = \frac{n^2(n+1)^2}{4}.$$

Exercise 1.2.4. Let n be a natural number. Show that

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

Exercise 1.2.5. (Bernoulli's Inequality³) Prove that for any number $x \geq -1$ and any natural number n ,

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

Exercise 1.2.6. Let $n \geq 3$ be a natural number. If P is a convex n -gon, show that the sum of the angles of P equals $(n-2)\pi$ (in radians) or $(n-2)180^\circ$ (in degrees).

Exercise 1.2.7. What is the maximum number of slices of pizza that can be obtained by making n cuts into a circular pizza? A cut does not have to go through the center of the pizza.

Exercise 1.2.8. If $n \in \mathbb{N}_0$, define $n!$ as follows: $0! = 1$ and if $n \geq 1$, then $n!$ equals the product of the first n natural numbers, also denoted as $\prod_{k=1}^n k = 1 \cdot \dots \cdot n$. A permutation of $1, \dots, n$ is any rearrangement of these numbers where the order matters. When $n = 2$, the permutations are $1, 2$ and $2, 1$ for example. Show that the number of permutations of $1, \dots, n$ equals $n!$. The expression $n!$ is called *n factorial*, the exclamation sign at the end is just a notation to signify joy.

Exercise 1.2.9. Show that the number of subsets of $\{1, \dots, n\}$ equals 2^n for any natural number n .

Exercise 1.2.10. Write down the values of $n!$ and 2^n side by side for $0 \leq n \leq 7$. Using induction, show that $n! > 2^n$ for $n \geq 4$.

³This is named after the Swiss mathematician Jacob Bernoulli (1654–1705). Note that there are several (related) mathematicians with last name Bernoulli: Daniel, Jacob, Johann, and Nikolaus.

Exercise 1.2.11. Let $a_1, \dots, a_n, n \geq 2$, be positive real numbers. Prove the equivalence of the following properties.

- (1) The arithmetic and the geometric mean of the a_j s coincide. That is,

$$\sqrt[n]{a_1 \cdot \dots \cdot a_n} = \frac{a_1 + \dots + a_n}{n}.$$

- (2) The geometric and the harmonic mean of the a_j s are equal. In other words,

$$\frac{n}{\frac{1}{a_1} + \dots + \frac{1}{a_n}} = \sqrt[n]{a_1 \cdot \dots \cdot a_n}.$$

- (3) The numbers a_1, \dots, a_n are equal.

Consequently, the AM-GM-HM inequality in Proposition 1.2.5 becomes a strict estimate if and only if $n \geq 2$ and moreover, there are $1 \leq i < j \leq n$ such that $a_i \neq a_j$.

1.3. Arithmetic and Geometric Progressions

Human minds have an amazing ability to see and guess patterns. Tell children to continue the sequence 2, 5, 8, 11 and almost surely they would guess the next terms 14, 17, 20, and so on. Similarly, ask someone what comes after 2, 4, 8 and you will likely hear 16, 32, 64. Problems involving such sequences appear in the Rhind Papyrus which dates to around 1550 BCE. These sequences are examples of arithmetic and geometric progressions and their properties can be discovered by experiment and proof. This is a place where inductive proofs are quite useful.

Definition 1.3.1. A sequence $(a_n)_{n \geq 1}$ of numbers is called an arithmetic progression if there exists a number d such that $a_{n+1} - a_n = d$ for any $n \geq 1$. The number d is called the common difference of the arithmetic progression $(a_n)_{n \geq 1}$.

In other words, our sequence starts with a_1 and then the next terms are given by $a_1 + d, a_1 + d + d, \dots$. We prove some basic properties of such sequences below. Note that the second item of the next proposition indicates the reason why such sequences are called arithmetic progressions.

Proposition 1.3.1. Let $(a_n)_{n \geq 1}$ be an arithmetic progression with common difference d . For any $n \geq 1$, the following assertions are true:

- (1) $a_n = a_1 + (n - 1)d$.
- (2) If $n \geq 2$ and t is any natural number between 1 and $n - 1$, then a_n is the arithmetic mean of a_{n-t} and a_{n+t} , meaning that $a_n = \frac{a_{n-t} + a_{n+t}}{2}$.
- (3) $a_1 + \dots + a_n = \frac{n(a_1 + a_n)}{2} = \frac{n(2a_1 + (n-1)d)}{2}$.

Proof: For the first part, we will use induction on n to prove that $a_n = a_1 + (n - 1)d$ for any $n \geq 1$. The base case when $n = 1$ is clear as $a_1 = a_1 + (1 - 1)d$. Let $n \geq 1$ now and assume that our statement is true for n , namely that $a_n = a_1 + (n - 1)d$. We wish to prove that $a_{n+1} = a_1 + nd$. Starting with what we know is true, $a_n = a_1 + (n - 1)d$, we add d to both sides, and we obtain that $a_n + d = a_1 + (n - 1)d + d = a_1 + nd$. Now

since our sequence is an arithmetic progression, we know that $a_{n+1} = a_n + d$. Putting these two equations together, we get what we wanted, namely $a_{n+1} = a_1 + nd$.

For the second part, we can make use of our good work in the first part. Let $n > t \geq 1$ be two natural numbers. From the first part, we already know that $a_{n-t} = a_1 + (n-t-1)d$, $a_n = a_1 + (n-1)d$ and $a_{n+t} = a_1 + (n+t-1)d$. Therefore,

$$\begin{aligned}\frac{a_{n-t} + a_{n+t}}{2} &= \frac{a_1 + (n-t-1)d + a_1 + (n+t-1)d}{2} = \frac{2a_1 + (2n-2)d}{2} \\ &= a_1 + (n-1)d = a_n.\end{aligned}$$

For the third part, we note that the statement can be proved using induction on n and we leave this as an exercise for the reader. We will give a noninduction proof which is essentially the same as the one used in calculating the sum of the first n numbers. To see this, if S denotes the sum of the first n terms of our sequence, then

$$S = a_1 + a_2 + \cdots + a_n$$

$$S = a_n + a_{n-1} + \cdots + a_1.$$

Adding up these equations, we get $2S$ on the left-hand side. To calculate the sum on the right-hand side we add up the terms column by column and using the previous part of our proposition, we observe that each of these n terms equals $a_1 + a_n$. Therefore, $2S = n(a_1 + a_n)$ which gives the first formula $S = \frac{n(a_1 + a_n)}{2}$. The second formula is obtained by replacing a_n by $a_1 + (n-1)d$. ■

Definition 1.3.2. A sequence $(b_n)_{n \geq 1}$ of numbers is called a geometric progression if there exists a number r such that $b_{n+1} = rb_n$ for any $n \geq 1$. The number r is called the common ratio of the geometric progression $(b_n)_{n \geq 1}$.

Our geometric sequence starts with b_1 and then the next terms are $rb_1, rrb_1, rrrb_1$, and so on. When $r \neq 0$ and $b_1 \neq 0$, the condition $b_{n+1} = rb_n$ is equivalent $b_{n+1}/b_n = r$ and this explains the name of common ratio used for r . Before proving some properties of these sequences, note that a geometric progression of common ratio 1 is a constant sequence and a geometric sequence of common ratio 0 consists of an initial value b_1 and all its values after that are 0. Thus, geometric progressions with common ratio 1 or 0 lead to easy sequences.

Proposition 1.3.2. Let $(b_n)_{n \geq 1}$ be a geometric progression with common ratio r . For any $n \geq 1$, the following assertions are true:

- (1) $b_n = b_1 r^{n-1}$.
- (2) If $b_1 \geq 0$, $r \geq 0$ and $n \geq 2$, then for any natural number t between 1 and $n-1$, b_n is the geometric mean of b_{n-t} and b_{n+t} , meaning that $b_n = \sqrt{b_{n-t} b_{n+t}}$.
- (3) If $r \neq 1$, then $b_1 + \cdots + b_n = \frac{b_1 - b_{n+1}}{1-r} = \frac{b_1(1-r^n)}{1-r}$.

Proof: For the first part, we use induction on n . The base case when $n = 1$ is true since $b_1 = b_1 r^{1-1}$. For the induction step, let $n \geq 1$ and assume that our statement is true for n , namely $b_n = b_1 r^{n-1}$. We wish to prove that $b_{n+1} = b_1 r^n$. We start with our hypothesis $b_n = b_1 r^{n-1}$ and multiply both sides by r , therefore obtaining that $b_n r = b_1 r^n$. As our sequence is a geometric progression, we know that $b_{n+1} = b_n r$.

Combining these two equations, we obtain that $b_{n+1} = b_n r = b_1 r^n$ which finishes this proof.

For the second part, we use our work done in the first part and observe that $b_{n-t} = b_1 r^{n-t-1}$, $b_n = b_1 r^{n-1}$ and $b_{n+t} = b_1 r^{n+t-1}$. Therefore,

$$b_{n-t} b_{n+t} = b_1 r^{n-t-1} b_1 r^{n+t-1} = b_1^2 r^{2n-2} = b_n^2.$$

Since $b_1, r \geq 0$, we have that $b_n = b_1 r^{n-1} \geq 0$ and therefore, taking square roots in the previous equation, we get that $b_n = \sqrt{b_{n-t} b_{n+t}}$.

For the last part, we leave the proof of the result using induction on n as an exercise for the reader. We give a direct proof without induction as follows. The attentive reader will have noticed already that our previous trick from arithmetic progressions will not work. However, there is something else we can do. Denote

$$T = b_1 + b_2 + \cdots + b_{n-1} + b_n.$$

In this sum, each term is obtained from the previous term by multiplication by r . Let us try to multiply the whole sum by r ,

$$\begin{aligned} rT &= rb_1 + rb_2 + \cdots + rb_{n-1} + rb_n \\ &= b_2 + b_3 + \cdots + b_n + b_{n+1}. \end{aligned}$$

It appears that the indices in our sum have shifted up by one. If we collect T and rT next to each other, we get the following

$$\begin{aligned} T &= b_1 + b_2 + \cdots + b_{n-1} + b_n \\ rT &= 0 + b_2 + \cdots + b_{n-1} + b_n + b_{n+1}. \end{aligned}$$

The two left-hand sides have $n - 1$ terms in common. We can perform a magic trick to make them disappear by subtracting the second equation from the first:

$$T - rT = b_1 - b_{n+1},$$

which implies our formula $T = \frac{b_1 - b_{n+1}}{1-r}$. Since $b_{n+1} = b_1 r^n$ from our first part, we get the second formula immediately. ■

Let us see some applications of this result.

Example 1.3.1. For the first one, we can go back in history to the following problem:

Suppose we place a grain of rice on a square of a chessboard, then two grains on the next square, four on the next one and so on. How much rice would be on the chessboard at the end of this procedure?

The legend is much more interesting than we are describing it here, and it allegedly involves a bet between a king and a math savvy person. We can leave that aside for now and just focus on calculating the number of grains in total. It is fairly clear that this number equals

$$1 + 2^1 + 2^2 + \cdots + 2^{63}.$$

This is the sum of a geometric progression with initial term 1 and common ratio 2. By our previous work, it is fairly easy to see that

$$1 + 2^1 + 2^2 + \cdots + 2^{63} = 2^{64} - 1.$$

Each of us is likely to imagine that the number $2^{64} - 1$ is a large one, but can we actually quickly approximate it? With the powers of 2, it helps to know that $2^{10} = 1024$ which is slightly larger than $1000 = 10^3$. Therefore, $2^{64} - 1$ will be at least $2^4 \cdot 10^{18} = 16 \cdot 10^{18}$. This is not a bad approximation for such a small amount of effort. The actual value of $2^{64} - 1$ is

$$18,446,744,073,709,551,615 > 18 \cdot 10^{18}.$$

Example 1.3.2. Our second problem also has a long history. It is one of Zeno's paradoxes from Ancient Greece devised around the 5th century BCE. Say you are standing 1 meter away from a wall, and you throw a piece of chalk at the wall. Assume that you are standing at the point of coordinates $(0, 0)$ and the wall is at the point of coordinates $(1, 0)$. The chalk traverses half of the distance, then half of the remaining half and so on. Does the chalk ever arrive at the wall?

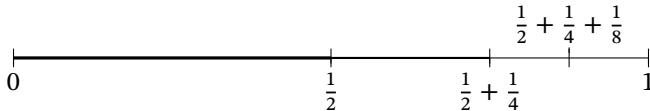


Figure 1.3.1. $\frac{1}{2} + \frac{1}{4} + \frac{1}{8} + \dots$.

There are many things that we can debate regarding this story, but one thing is clear, namely that it leads to the calculation of sums of the form $\frac{1}{2}, \frac{1}{2} + \frac{1}{4}, \frac{1}{2} + \frac{1}{4} + \frac{1}{8},$ and more generally of the forms

$$\frac{1}{2} + \frac{1}{2^2} + \dots + \frac{1}{2^n}.$$

This is the sum of the first n terms of a geometric progression with initial term $\frac{1}{2}$ and common ratio $\frac{1}{2}$. By our previous proposition, we deduce that it equals $1 - \frac{1}{2^n}$. This tells us that as n gets larger, this sum approaches 1. We will come back to this topic later in the book when discussing limits.

The previous two examples are applications of the formula giving the sum of the first n terms of a geometric progression. They are particular examples of the following result. For any number r and any natural number n ,

$$(1.3.1) \quad 1 - r^n = (1 - r)(r^0 + \dots + r^{n-1}) = (1 - r)(1 + \dots + r^{n-1}).$$

When $r \neq 1$, this is equivalent to

$$\frac{1 - r^n}{1 - r} = r^0 + \dots + r^{n-1} = 1 + \dots + r^{n-1}.$$

Furthermore, if in (1.3.1), we replace r by the ratio of two numbers a and $b \neq 0$, we get the following identity:

$$(1.3.2) \quad a^n - b^n = (a - b)(a^{n-1} + a^{n-2}b + \dots + ab^{n-2} + b^{n-1}).$$

Note that this identity also holds for $b = 0$ so it is true for any numbers a and b . When $n = 2k + 1$ is odd, replacing b by $-b$ gives us another useful identity:

$$(1.3.3) \quad a^{2k+1} + b^{2k+1} = (a+b)(a^{2k} - a^{2k-1}b + a^{2k-2}b^2 - \dots + a^2b^{2k-2} - ab^{2k-1} + b^{2k}).$$

Exercise 1.3.1. Let a , b , and c be three nonzero numbers such that a^2 , b^2 , and c^2 are distinct and form an arithmetic progression. Show that the numbers

$$\frac{1}{b+c}, \frac{1}{c+a}, \frac{1}{a+b}$$

form an arithmetic progression.

Exercise 1.3.2. Let $(a_n)_{n \geq 1}$ be an arithmetic progression such that for any natural number n , the sum of its first n terms equals $n^2 + 4n$. Determine a_n for any $n \geq 1$.

Exercise 1.3.3. Let $(b_n)_{n \geq 1}$ be a geometric progression such that for any natural number n , the sum of its first n terms equals $2 \cdot 3^n - 2$. Determine b_n for any $n \geq 1$.

Exercise 1.3.4. Let $(a_n)_{n \geq 1}$ be a sequence of numbers such that for any natural number $n \geq 2$, a_n is the arithmetic mean of a_{n-1} and a_{n+1} . Show that $(a_n)_{n \geq 1}$ is an arithmetic progression.

Exercise 1.3.5. Let $(a_n)_{n \geq 1}$ be an arithmetic progression of nonzero numbers. Prove that for any natural number n ,

$$\sum_{k=1}^n \frac{1}{a_k a_{k+1}} = \frac{n}{a_1 a_{n+1}}.$$

Exercise 1.3.6. Let $(x_n)_{n \geq 1}$ be a sequence of nonzero numbers such that for any natural number n ,

$$\sum_{k=1}^n \frac{1}{x_k x_{k+1}} = \frac{n}{x_1 x_{n+1}}.$$

Show that $(x_n)_{n \geq 1}$ is an arithmetic progression.

Exercise 1.3.7. Let $(b_n)_{n \geq 1}$ be a sequence of nonnegative numbers such that for any natural number $n \geq 2$, b_n is the geometric mean of b_{n-1} and b_{n+1} . Show that $(b_n)_{n \geq 1}$ is a geometric progression.

Exercise 1.3.8. Let k be a natural number. Show that the sum

$$1^2 - 2^2 + 3^2 - \dots - (2k)^2 + (2k+1)^2$$

can be written a sum of an arithmetic progression. Using this result, find a simple formula for the sum

$$\sum_{j=1}^n (-1)^{j-1} j^2,$$

for any natural number n .

Exercise 1.3.9. We know that for any real number $x \neq 1$ and for any natural number n ,

$$1 + x + \dots + x^n = \frac{1 - x^{n+1}}{1 - x}.$$

By considering each side of this equation as a function in x and taking derivatives, prove that

$$1 + 2x + \dots + (n - 1)x^{n-1} = \frac{nx^{n+1} - (n + 1)x^n + 1}{(1 - x)^2}.$$

Give a proof of this identity without using derivatives.

Exercise 1.3.10. Take a piece of paper which is 1 millimeter thick. Fold it in the middle. After that one obtains a piece of paper which is 2 millimeter thick and half the original length. Execute this procedure all in all 12 times. Evaluate the thickness of the folded paper after this procedure⁴. If the original length of the paper is 4,000 ft, which length does the paper possess after 12 folds?

Exercise 1.3.11. Let $(a_n)_{n \geq 1}$ be an arithmetic progression with nonzero common difference. Let $(b_n)_{n \geq 1}$ be a geometric progression with positive common ratio. Prove that there exist numbers α and β such that for any natural number n ,

$$b_n = \alpha\beta^{a_n}.$$

1.4. The Least Element Principle

For historic events such as printing a book, having a phone conversation, or landing on the Moon, there are many years when that particular event occurred. However, there is always a (unique) year when such an event happened for the first time. We often take such a fact for granted and this corresponds to an important property of natural numbers that is called the **well-ordering property** of \mathbb{N} or the **least element principle** for subsets of \mathbb{N} .

Theorem 1.4.1 (The Well-Ordering or Least Element Principle). *The set of natural numbers \mathbb{N} is well-ordered, meaning that any nonempty subset $A \subseteq \mathbb{N}$ has a least element.*

$$(\forall A \subseteq \mathbb{N}, A \neq \emptyset)(\exists a_{\min} \in A)[(\forall a \in A)(a_{\min} \leq a)]$$

Proof: If $1 \in A$, the assertion is true as $1 \leq a$ for any $a \in \mathbb{N}$ (and consequently, for any $a \in A$). In this case, $a_{\min} = 1$. Now suppose $1 \notin A$ and assume that A does not possess a least element. We are going to show that $1, \dots, n \notin A$ always implies $n + 1 \notin A$. Why is this so? Because $1, \dots, n \notin A$ and $n + 1 \in A$ would imply that $n + 1$ is a minimal element of A , and this contradicts the assumption about A .

Let us make this now more precise. Define the set $S \subseteq \mathbb{N}$ as follows:

$$S := \{n \in \mathbb{N} : \{1, \dots, n\} \cap A = \emptyset\}.$$

From our assumption that $1 \notin A$, we get that $1 \in S$.

⁴In guinnessworldrecords.com/world-records the following can be found: "It was an accepted belief that folding a piece of paper in half more than 8 times was impossible. On 27 January 2002, high school student, Britney Gallivan, of Pomona, California, USA, folded a single piece of paper in half 12 times and was the first person to fold a single piece of paper in half 9, 10, 11, and 12 times. The tissue paper used was 4,000 ft (1,219 m; 0.75 miles) long."

We show that if $n \in S$, then $n + 1 \in S$. Suppose this is not true and there is an $n \in S$ such that $n + 1 \notin S$. Using the definition of S , we deduce that $n + 1 \in A$, but $1, \dots, n \notin A$. Consequently, $n + 1$ is a least element in A . But our assumption was that A does not have a least element. Therefore, $n + 1 \in S$.

Now the induction principle applies and yields $S = \mathbb{N}$. But this is only possible if $A = \emptyset$, contradicting our assumption $A \neq \emptyset$. This completes the proof. ■

Remark 1.4.1. Note that a least element of a nonempty subset $A \subseteq \mathbb{N}$ is unique. Indeed, if there would be two minimal elements $a_1, a_2 \in A$, then $a_1 \leq a_2$ as well as $a_2 \leq a_1$ which implies $a_1 = a_2$.

Given that 0 is smaller than any natural number, an immediate corollary is the following result.

Corollary 1.4.2. *Every subset of \mathbb{N}_0 has a least element.*

Remark 1.4.2. The least element principle does not hold for the set \mathbb{Z} of integer numbers. For example, the subset of even integers does not have a least element.

The least element principle does not hold for the set \mathbb{Q} of rational numbers. Take the subset $\{1/n : n \in \mathbb{N}\}$ for example. For any number of the form $1/k$ with $k \in \mathbb{N}$, one can find a smaller number of the same form.

The least element principle does not hold for the set of all positive real numbers. For example, the interval $(0, 1)$ has no smallest element. For any element x in $(0, 1)$ we can always find a smaller element in $(0, 1)$, for example, $x/2$.

Given $m \in \mathbb{Z}$, the least element principle also holds for $\{m, m+1, m+2, \dots\}$. When $m = 0$, this gives the previous corollary.

When one analyzes the proof of Theorem 1.4.1 one observes that the well-ordering property is deduced from the principle of induction. We will show that these two principles are in fact equivalent and that if we assume that \mathbb{N} is well-ordered, then we may derive the principle of induction. Thus, instead of adding Axiom 1.2.1 we could also have added the well-ordering property.

Proposition 1.4.3. *The well-ordering property of \mathbb{N} implies the induction principle.*

Proof: Assume that every nonempty subset of \mathbb{N} has a minimal element. Take $S \subseteq \mathbb{N}$ with $1 \in S$ and such that $n \in S$ always implies $n + 1 \in S$. We want to show that $S = \mathbb{N}$.

Assume that $S \neq \mathbb{N}$. Therefore, the complement set

$$A := \{n \in \mathbb{N} : n \notin S\}$$

is nonempty. By the well-ordering property, the set A has a least element a_{\min} . Because $1 \in S$, $a_{\min} \neq 1$. Thus, there is an $k \in \mathbb{N}$ such that $a_{\min} = k + 1$. Because a_{\min} is the least element of A , $k \notin A$. Now $k \notin A$ implies that $k \in S$. Thus, we found a natural number k with $k \in S$ but $k + 1 \in A$, hence $k + 1 \notin S$. This contradicts the assumed induction property of S . ■

The well-ordering property allows us to state and to prove another version of mathematical induction.

Proposition 1.4.4 (Strong Induction). *Let $P(n)$ be a proposition depending on $n \in \mathbb{N}$. If the following conditions are satisfied:*

- (1) *The proposition $P(1)$ is true.*
- (2) *For any $n \geq 1$ the following implication is valid:*

If $P(k)$ is true for all $1 \leq k \leq n$, then $P(n + 1)$ is true.

Under these assumptions proposition $P(n)$ is true for any $n \geq 1$.

Proof: Let us assume that there is at least one $n \in \mathbb{N}$ for which $P(n)$ is not true. Set

$$S := \{n \in \mathbb{N} : P(n) \text{ is not true}\}.$$

By assumption, S is nonempty and $1 \notin S$. Thus, the well-ordering principle applies to S , and there is some least element in it. Call it n_0 . First note that $n_0 \geq 2$, and, moreover, by the definition of S it follows that $P(k)$ is true for all integers $k = 1, \dots, n_0 - 1$. Now the basic assumption about property P applies and yields that $P((n_0 - 1) + 1) = P(n_0)$ is true as well. But this contradicts $n_0 \in S$ and completes the proof. ■

Remark 1.4.3. To clarify the name *strong* that is applied to this method, we have to compare the conditions about P in Proposition 1.2.1 and Proposition 1.4.4. In both cases one has to ensure that the induction starts at 1. But while in Proposition 1.2.1 we suppose

$$(1.4.1) \quad (\forall n \in \mathbb{N})[(P(n) \text{ true}) \Rightarrow (P(n + 1) \text{ true})],$$

the assumption about P in Proposition 1.4.4 is as follows:

$$(1.4.2) \quad (\forall n \in \mathbb{N})[(P(1) \text{ true}, \dots, P(n) \text{ true}) \Rightarrow (P(n + 1) \text{ true})],$$

Since any property P satisfying (1.4.1) also satisfies (1.4.2), Proposition 1.4.4 has a stronger induction hypothesis than Proposition 1.2.1.

As for the usual induction (see Proposition 1.2.1 and Proposition 1.2.3), one can state a similar result when trying to prove that $P(n)$ is true for any $n \geq n_0$ for some $n_0 \in \mathbb{N}_0$.

Proposition 1.4.5 (Strong Induction). *Let $n_0 \in \mathbb{N}_0$. Suppose that $P(n)$ is a proposition depending on $n \in \mathbb{N}$, $n \geq n_0$. If the following conditions are satisfied:*

- (1) *The proposition $P(n_0)$ is true.*
- (2) *For any $n \geq n_0$ the following is satisfied:*

If $P(k)$ is true for all $n_0 \leq k \leq n$, then also $P(n + 1)$ is true,

then $P(n)$ is true for all $n \geq n_0$.

The proof of this result is similar to one for Proposition 1.4.4, and we leave it as an exercise for the reader. When facing a problem, a student may be confused choosing between the induction principle from the previous section and the strong induction principle above. Generally speaking, if the proposition $P(n + 1)$ that we are trying to prove today depends only on the proposition $P(n)$ that happened yesterday, then the usual induction should be okay. However, there are cases in mathematics and life when

what happens today, or the proposition $P(n+1)$ we are trying to prove, does not depend only on yesterday (proposition $P(n)$), but also on other days that are deeper in our past (possibly all of $P(1), P(2), \dots, P(n)$).

In our previous examples of usual induction, the base case was a simple formality and one can expect it to be that way in general. One has to be careful with the base case of strong induction. We present an example of the use of strong induction and how things can be slightly different from the usual induction.

Example 1.4.1. Say one can buy stamps in packages of 3 or packages of 5 stamps. Is it possible to buy any quantity of stamps this way? Examining small values, it is clear that we cannot buy 1, 2, 4, 7 stamps, but we can buy 3, 5, 6 = $2 \cdot 3$, 8 = $5 + 3$, 9 = $3 \cdot 3$ or 10 = $2 \cdot 5$ stamps. Actually, with a bit more patience, we can see that 11, 12, and 13 work: 11 = $8 + 3 = 2 \cdot 3 + 5$, 12 = $9 + 3 = 4 \cdot 3$, 13 = $10 + 3 = 3 + 2 \cdot 5$. It is natural to conjecture that any number $n \geq 8$ can be attained this way, meaning there exists a packages of 3 stamps and b of 5 stamps such that $n = 3a + 5b$. Of course, the numbers a and b are in \mathbb{N}_0 and depend on n . To relate this problem to the general template above, first note that $n_0 = 8$. Clearly, $P(8)$ is true as $8 = 3 + 5$. Moving to the induction step, does $P(8)$ imply $P(9)$? Since we know $P(8)$ is true, this is equivalent to $P(9)$ being true which it is as $9 = 3 \cdot 3$. Moving another step forward, do $P(8)$ true and $P(9)$ true imply that $P(10)$ true? Again, this is equivalent to $P(10)$ being true which it is since $10 = 2 \cdot 5$. These two extra cases $P(9)$ and $P(10)$ are sometimes lumped together with $P(8)$ into a larger base case. Take $n \geq 10$ now and consider $P(n+1)$. The key observation to which we alluded to earlier ($11 = 8 + 3, 12 = 9 + 3, 13 = 10 + 3$) is the following: $P(n-2)$ implies $P(n+1)$: if $n-2 = 3a + 5b$ with $a, b \in \mathbb{N}_0$, then $n+1 = 3(a+1) + 5b$ with $a+1, b \in \mathbb{N}_0$. We will write this proof formally below.

Proposition 1.4.6. Every natural number $n \geq 8$ can be written as

$$n = 3a_n + 5b_n$$

where $a_n, b_n \in \mathbb{N}_0$.

Proof: We will use strong induction to prove our result. The base case consists of three situations:

$$\begin{aligned} 8 &= 3 + 5 \\ 9 &= 3 \cdot 3 \\ 10 &= 2 \cdot 5. \end{aligned}$$

Moving to the induction step, let $n \geq 10$. Assume that any number k between 8 and n can be written as $3a_k + 5b_k$, for some $a_k, b_k \in \mathbb{N}_0$. We want to prove that $n+1$ can be written as $3a + 5b$ for some $a, b \in \mathbb{N}_0$. Because $n+1 - 3 = n - 2 \geq 8$, our induction hypothesis tells us that there are $a_{n-2}, b_{n-2} \in \mathbb{N}_0$ such that $n-2 = 3a_{n-2} + 5b_{n-2}$. Adding 3 back to both sides, we get that $n+1 = 3(a_{n-2}+1)+5b_{n-2}$. Since $a_{n-2}+1 \in \mathbb{N}$ and $b_{n-2} \in \mathbb{N}_0$, this proves our induction step. ■

There is no general result that can predict the number of situations one has to deal with in the base case when solving a problem by strong induction. Doing more

problems will develop the skills and experience to deal with any new situation. We give more examples below that will hopefully help in this regard.

Example 1.4.2. Say we want to write each natural number as a sum of 1s and 2s. For example, when $n = 1$ there is one way $1 = 1$, when $n = 2$ there are two ways as $n = 1 + 1 = 2$, when $n = 3$ we get that $3 = 1 + 2 = 1 + 1 + 1 = 2 + 1$ and for $n = 4$ we can do $4 = 1 + 3 = 1 + 2 + 1 = 1 + 1 + 2 = 2 + 2 = 2 + 1 + 1$. In how many ways can this be done for a general n ? A geometric way of looking at this problem is tiling a $1 \times n$ rectangle with smaller pieces of the form 1×1 and 1×2 . When $n = 1$, we can only use one 1×1 tile, but when $n = 2$, we can tile our 1×2 piece by two 1×1 tiles or by one 1×2 tiles. How about $n = 3$? How about for larger n ?



Figure 1.4.1. Tiling a 1×3 board with 1×1 and 1×2 tiles.

Figure 1.4.1 gives us the answer for $n = 3$. So far, we have 1 tiling for $n = 1$, 2 tilings for $n = 2$ and 3 tilings for $n = 3$. Previous examples have taught us to embrace patterns, but also be skeptical and treat them with respect, so perhaps we can try $n = 4$ before making any conjectures. Figure 1.4.2 shows us 5 tilings of a 1×4 board with 1×1 and 1×2 tiles. At this point, we can introduce some notation to make our life a bit easier. For $n \geq 2$, let F_n denote the number of ways of tiling a $1 \times (n - 1)$ board with 1×1 and 1×2 tiles. There is a historic reason for this slight shift between the sequence index and the length of the board so do not worry about it for now. Just remember that $F_2 = 1, F_3 = 2, F_4 = 3$ (Figure 1.4.1) and $F_5 = 5$ (Figure 1.4.2).



Figure 1.4.2. Tiling a 1×4 board with 1×1 and 1×2 tiles.

Let $n \geq 3$ be a natural number. When we calculate F_{n+1} we are counting the number of ways of tiling a $1 \times n$ board. As you can see in Figure 1.4.1 and Figure 1.4.2, the first piece from the left could be a 1×1 tile or could be a 1×2 tile. Therefore, we can split the F_{n+1} possible tilings of a $1 \times n$ board into two groups: the ones whose first piece on the left is a 1×1 tile and the ones starting with a 1×2 tile. For example, when $n = 3$, the first group has two tilings and the second group has one (see Figure 1.4.1) and when $n = 4$, the first group consists of three tilings and the second group contains two tilings (see Figure 1.4.2). In general, the number of tilings that start with a 1×1 tile will be the same as the number of tilings of the remaining $1 \times (n - 1)$ board which is F_n and the number of tilings that start with a 1×2 tile will equal the number of tilings of the remaining $1 \times (n - 2)$ board which is F_{n-1} . Hence, $F_{n+1} = F_n + F_{n-1}$. This is an example of a recurrence relation which describes the dependence between various members of a sequence of numbers. It is associated with a famous sequence in mathematics.

The Fibonacci sequence $(F_n)_{n \geq 0}$ is defined as follows: $F_0 = 0, F_1 = 1$ and

$$(1.4.3) \quad F_{n+1} = F_n + F_{n-1}$$

for $n \geq 1$. This is the sequence we were describing above in the tiling problem with the additional terms F_0 and F_1 included such that (1.4.3) works for $n \geq 1$. This sequence is named after Leonardo of Pisa or Fibonacci (meaning the son of Bonacci) (1170–1240), whose 1202 book *Liber Abaci (Book of Calculation)* popularized the Indian/Arabic numerals used in the Western World. It turns out that the Fibonacci sequence was known to Indian mathematicians such as Pingala in the 2nd century BCE in connection to the enumeration of patterns of two letters and one letter syllables in Sanskrit poems. The Fibonacci sequence appears as sequence OA000045 in the *Online Encyclopedia of Integer Sequences*, <https://oeis.org/A000045>. We list below the first Fibonacci numbers:

Table 1.4.1. The Fibonacci numbers F_n for $0 \leq n \leq 14$.

| n | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 |
|-------|---|---|---|---|---|---|---|----|----|----|----|----|-----|-----|-----|
| F_n | 0 | 1 | 1 | 2 | 3 | 5 | 8 | 13 | 21 | 34 | 55 | 89 | 144 | 233 | 377 |

The following result will likely seem mysterious at first, but let us just try to prove it using strong induction for now as this was our initial motivation to discussing this problem. Later in this section, we will describe some general principles for coming up with such formulas.

Proposition 1.4.7. *For $n \geq 0$, the n -th term of the Fibonacci sequence equals*

$$(1.4.4) \quad F_n = \frac{1}{\sqrt{5}} \left[\left(\frac{1 + \sqrt{5}}{2} \right)^n - \left(\frac{1 - \sqrt{5}}{2} \right)^n \right].$$

Proof: First, we observe that our recurrence relation (1.4.3) tells us that F_{n+1} depends on the previous two terms F_n and F_{n-1} . Thus, F_2 depends on F_1 and F_0 , F_3 depends on F_2 and F_1 and so on. This is a sign of having a statement amenable to be proved by strong induction.

We will use strong induction on n . For the base case, we need to check both $n = 0$ and $n = 1$. When $n = 0$, both sides of (1.4.4) equal 0. When $n = 1$, both sides of (1.4.4) are 1.

For the induction step, let $n \geq 1$ and assume that (1.4.4) is true for any k with $1 \leq k \leq n$:

$$(1.4.5) \quad F_k = \frac{1}{\sqrt{5}} \left[\left(\frac{1 + \sqrt{5}}{2} \right)^k - \left(\frac{1 - \sqrt{5}}{2} \right)^k \right].$$

We want to prove that (1.4.4) is true for $n + 1$, namely that

$$(1.4.6) \quad F_{n+1} = \frac{1}{\sqrt{5}} \left[\left(\frac{1 + \sqrt{5}}{2} \right)^{n+1} - \left(\frac{1 - \sqrt{5}}{2} \right)^{n+1} \right].$$

What we know now are (1.4.3) and (1.4.5) and we apply them as follows. Replacing F_n and F_{n-1} by the formulas above in $F_{n+1} = F_n + F_{n-1}$, we get that

$$\begin{aligned} F_{n+1} &= F_n + F_{n-1} \\ &= \frac{1}{\sqrt{5}} \cdot \left(\frac{1+\sqrt{5}}{2} \right)^{n-1} \cdot \frac{3+\sqrt{5}}{2} - \frac{1}{\sqrt{5}} \cdot \left(\frac{1-\sqrt{5}}{2} \right)^{n-1} \cdot \frac{3-\sqrt{5}}{2}. \end{aligned}$$

Now come two mysterious substitutions which arise from the following calculations:

$$\begin{aligned} \left(\frac{1+\sqrt{5}}{2} \right)^2 &= \frac{6+2\sqrt{5}}{4} = \frac{3+\sqrt{5}}{2} \\ \left(\frac{1-\sqrt{5}}{2} \right)^2 &= \frac{6-2\sqrt{5}}{4} = \frac{3-\sqrt{5}}{2}. \end{aligned}$$

Replacing $\frac{3+\sqrt{5}}{2}$ by $\left(\frac{1+\sqrt{5}}{2} \right)^2$ and $\frac{3-\sqrt{5}}{2}$ by $\left(\frac{1-\sqrt{5}}{2} \right)^2$ in the previous calculation involving F_{n+1} , we get that

$$\begin{aligned} F_{n+1} &= \frac{1}{\sqrt{5}} \cdot \left(\frac{1+\sqrt{5}}{2} \right)^{n+1} - \frac{1}{\sqrt{5}} \cdot \left(\frac{1-\sqrt{5}}{2} \right)^{n+1} \\ &= \frac{1}{\sqrt{5}} \left[\left(\frac{1+\sqrt{5}}{2} \right)^{n+1} - \left(\frac{1-\sqrt{5}}{2} \right)^{n+1} \right], \end{aligned}$$

which is exactly our goal (1.4.6). ■

Whew, that was a bit of work! To recap, strong induction allowed us to use the formulas for F_n and F_{n-1} and get the desired formula for F_{n+1} in the induction step. The usual induction method would not have been able to deliver this result since in its induction step, the hypothesis would only consist of the formula for F_n and not the one for F_{n-1} .

Note that $\varphi := \frac{1+\sqrt{5}}{2}$ is a famous number in mathematics called the *golden ratio*⁵. Two numbers $a > b$ are said to be in golden ratio if their ratio a/b is the same as the ratio of their sum to the larger number $(a+b)/a$. Several artists used this number in their works as lengths approximately in this proportion were considered aesthetically pleasing.

The Fibonacci sequence is an example of a linear recurrence sequence. A sequence $(x_n)_{n \geq 0}$ is said to satisfy a linear recurrence of second order if there exist two coefficients α and β such that

$$(1.4.7) \quad x_{n+1} = \alpha x_n + \beta x_{n-1}$$

⁵Also known as **golden section** or **golden mean** or **divine proportion**.

for $n \geq 1$. So the sequence starts with two initial values x_0, x_1 and its values are calculated recursively from the previous two:

$$x_2 = \alpha x_1 + \beta x_0, \quad x_3 = \alpha x_2 + \beta x_1 = \alpha(\alpha x_1 + \beta x_0) + \beta x_1, \dots$$

and so on. In the case of the Fibonacci sequence we had $x_0 = 0, x_1 = 1$ and $\alpha = \beta = 1$.

We will describe how one can find a closed formula for x_n in two situations, with the third situation being treated in the chapter on complex numbers (see Section 6.4, p. 393). The characteristic equation of this sequence is the quadratic equation

$$(1.4.8) \quad r^2 = \alpha r + \beta.$$

The following are our possible situations:

- (1) The equation (1.4.8) has two distinct real solutions or $\alpha^2 - 4\beta > 0$.
- (2) The equation (1.4.8) has one real solution or $\alpha^2 - 4\beta = 0$.
- (3) The equation (1.4.8) has no real solutions or $\alpha^2 - 4\beta < 0$.

In the first case, let r_1 and r_2 be two distinct solutions of (1.4.8). Thus, $r_j^2 = \alpha r_j + \beta$ and consequently, $r_j^{n+1} = \alpha r_j^n + \beta r_j^{n-1}$ for $1 \leq j \leq 2$ and any $n \geq 1$. Hence, each of the sequences $(r_1^n)_{n \geq 0}$ and $(r_2^n)_{n \geq 0}$ satisfies the recurrence relation (1.4.7). It turns out that x_n will have the following form:

$$x_n = ar_1^n + br_2^n$$

for some a and b to be determined. The sequence $(ar_1^n + br_2^n)_{n \geq 1}$ satisfies the same recurrence relation as the sequence $(x_n)_{n \geq 1}$. In order for the two sequences to match, we need to find a and b such that the sequences are the same on the first two values:

$$\begin{aligned} a + b &= x_0 \\ ar_1 + br_2 &= x_1. \end{aligned}$$

Because $r_1 \neq r_2$, this system of equations has a unique solution (a, b) . And if we define with these numbers

$$x_n = ar_1^n + br_2^n, \quad n \geq 0,$$

then the sequence $(x_n)_{n \geq 0}$ satisfies the initial conditions as well as

$$x_{n+1} = \alpha x_n + \beta x_{n-1}, \quad n = 1, 2, \dots$$

A first test for the general approach is the Fibonacci sequence already investigated in Proposition 1.4.7.

Example 1.4.3. Consider the sequence $(F_n)_{n \geq 0}$ which satisfies

$$F_{n+1} = F_n + F_{n-1}, \quad n \geq 1,$$

as well as $F_0 = 0$ and $F_1 = 1$. Its characteristic equation is

$$r^2 = r + 1,$$

with solutions

$$r_1 = \frac{1 + \sqrt{5}}{2} \quad \text{and} \quad r_2 = \frac{1 - \sqrt{5}}{2}.$$

If $a = 1/\sqrt{5}$ and $b = -1/\sqrt{5}$, then, of course, $a + b = 0$. Moreover,

$$ar_1 + br_2 = \frac{1}{\sqrt{5}} \left(\frac{1 + \sqrt{5}}{2} - \frac{1 - \sqrt{5}}{2} \right) = 1.$$

Consequently, the quested sequence $(F_n)_{n \geq 0}$ is given by

$$F_n = ar_1^n + br_2^n = \frac{1}{\sqrt{5}} \left[\left(\frac{1 + \sqrt{5}}{2} \right)^n - \left(\frac{1 - \sqrt{5}}{2} \right)^n \right], \quad n \geq 0.$$

This is the same formula for the Fibonacci numbers as stated in Proposition 1.4.7.

Example 1.4.4. There is a related sequence worth mentioning here, the Lucas sequence $(L_n)_{n \geq 0}$ which has the same recurrence relation as the Fibonacci sequence: $L_{n+1} = L_n + L_{n-1}$ for $n \geq 1$, but starts with different initial values: $L_0 = 2, L_1 = 1$ as opposed to $F_0 = 0, F_1 = 1$. We list below the first numbers of the Lucas sequence in the same table as the one of Fibonacci numbers used earlier.

Table 1.4.2. The Fibonacci and Lucas numbers F_n for $0 \leq n \leq 14$.

| n | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 |
|-------|---|---|---|---|---|----|----|----|----|----|-----|-----|-----|-----|-----|
| F_n | 0 | 1 | 1 | 2 | 3 | 5 | 8 | 13 | 21 | 34 | 55 | 89 | 144 | 233 | 377 |
| L_n | 2 | 1 | 3 | 4 | 7 | 11 | 18 | 29 | 47 | 76 | 123 | 199 | 322 | 521 | 843 |

The characteristic equation and its solutions r_1 and r_2 , are exactly as in Example 1.4.3. But here $L_0 = 2$ and $L_1 = 1$, hence the unknown coefficients a and b have now to satisfy

$$a + b = 2 \quad \text{and} \quad ar_1 + br_2 = 1.$$

It is not difficult to see that $a = b = 1$ is a solution for both equations. So we get for the sequence of Lucas numbers the representation

$$L_n = r_1^n + r_2^n = \left(\frac{1 + \sqrt{5}}{2} \right)^n + \left(\frac{1 - \sqrt{5}}{2} \right)^n, \quad n \geq 0.$$

The Lucas sequence⁶ is sequence OA000032 in the *Online Encyclopedia of Integer Sequences*, <https://oeis.org/A000032>.

Before dealing with the case $\alpha^2 - 4\beta = 0$, let us give another interesting example with $\alpha^2 - 4\beta > 0$.

Example 1.4.5. Consider the sequence $(x_n)_{n \geq 0}$ satisfying $x_0 = 1, x_1 = 4$ and with $x_{n+1} = 5x_n - 6x_{n-1}$ for $n \geq 1$. The characteristic equation is given by $r^2 = 5r - 6$ or $r^2 - 5r + 6 = 0$ which has two distinct solutions $r_1 = 3$ and $r_2 = 2$. We are looking for x_n of the form $a \cdot 3^n + b \cdot 2^n$, where a and b satisfy the equations:

$$a + b = 1$$

$$3a + 2b = 4.$$

We can solve this system by substituting one variable in terms of the other, and we get that $a = 2$ and $b = -1$. This means that $x_n = 2 \cdot 3^n - 2^n$ for any $n \geq 0$.

⁶Investigated by the French mathematician François Édouard Anatole Lucas (1842–1891).

We now turn to the case $\alpha^2 - 4\beta = 0$. In this situation, the equation 1.4.8 has one multiple solution, let us call it r_0 . Note that $(r - r_0)^2 = r^2 - \alpha r - \beta$ which gives us that $r_0^2 = -\beta$ and $2r_0 = \alpha$. Therefore, $\alpha r_0 + 2\beta = 2r_0 r_0 - 2r_0^2 = 0$ which will be used in the next paragraph.

Similar to the previous case, the sequence $(r_0^n)_{n \geq 0}$ satisfies the recurrence relation (1.4.7) because $r_0^2 = \alpha r_0 + \beta$. The new ingredient here is that the sequence $(nr_0^n)_{n \geq 0}$ also satisfies the recurrence relation (1.4.7). This statement is equivalent to $nr_0^n = \alpha(n-1)r_0^{n-1} + \beta(n-2)r_0^{n-2}$ which is the same as

$$nr_0^n = n(\alpha r_0^{n-1} + \beta r_0^{n-2}) - r_0^{n-2}(\alpha r_0 + 2\beta).$$

Because $\alpha r_0 + 2\beta = 0$, this assertion is true. We will look for x_n of the form $ar_0^n + bnr_0^n$. The values a and b can be determined as before from the initial values of the sequence:

$$\begin{aligned} a &= x_0 \\ ar_0 + br_0 &= x_1. \end{aligned}$$

Let us finish this part with an example.

Example 1.4.6. Consider the sequence $(x_n)_{n \geq 0}$ given by $x_0 = 1$, $x_1 = 6$, and with $x_{n+1} = 4x_n - 4x_{n-1}$ for $n \geq 1$. The characteristic equation is $r^2 = 4r - 4$ which has only one solution $r_0 = 2$. We are looking for x_n of the form $a \cdot 2^n + b \cdot n2^n$ for $n \geq 0$. The values a and b can be found from:

$$\begin{aligned} a &= 1 \\ 2a + 2b &= 6 \end{aligned}$$

and they are $a = 1$ and $b = 2$. Therefore, $x_n = 2^n + 2n2^n = (2n+1)2^n$.

For the third case when $\alpha^2 - 4\beta < 0$ we have to deal with complex numbers, and we will cover it in Section 6.4, page 393.

To conclude this section let us state another result where strong induction turns out to be very useful. A more general result will be proved in Section 1.6.

Proposition 1.4.8. *For each natural number n there are an odd number $q \geq 1$ and an integer $m \geq 0$ such that*

$$(1.4.9) \quad n = q \cdot 2^m.$$

Proof: The result is true for $n = 1$ with $q = 1$ and $m = 0$.

Now suppose it is true for all $1 \leq k \leq n$. That is, each natural number $k \leq n$ admits a suitable representation (1.4.9).

If $n+1$ is odd, we are done. Indeed, then $n+1 = q \cdot 2^0$ with $q = n+1$ odd. On the other hand, if $n+1$ is even, let $k = (n+1)/2$. Because of $1 \leq k \leq n$, by assumption $k = q \cdot 2^m$ for a suitable odd number q and some $m \geq 0$. Of course, this implies

$$n+1 = q \cdot 2^{m+1},$$

hence $n+1$ admits a representation (1.4.9). The strong induction principle implies that every natural number n admits a representation (1.4.9). ■

Exercise 1.4.1. Let $(a_n)_{n \geq 1}$ be a sequence of numbers such that $a_1 = 1$, $a_2 = 2$, $a_3 = 3$ and

$$a_n = a_{n-1} + a_{n-2} + 2a_{n-3},$$

for any $n \geq 4$. What are a_4 and a_5 ? Prove that $a_n < 2^n$ for any $n \geq 1$.

Exercise 1.4.2. Show that any natural number $n \geq 18$ can be written as

$$n = 4a_n + 7b_n$$

where $a_n, b_n \in \mathbb{N}_0$.

Exercise 1.4.3. Show that for any $n \geq 1$, the n -th Lucas number L_n is equal to $F_{n+1} + F_{n-1} = F_n + 2F_{n-1}$.

Exercise 1.4.4. Prove that for any $n \geq 1$,

$$\begin{aligned} F_{2n-1} &= F_n^2 + F_{n-1}^2, \\ F_{2n} &= F_n(F_n + 2F_{n-1}), \\ (-1)^n &= F_{n+1}F_{n-1} - F_n^2. \end{aligned}$$

Exercise 1.4.5. Show that for any $n \geq 1$,

$$F_1^2 + \cdots + F_n^2 = F_n F_{n+1}.$$

Exercise 1.4.6. Show that any natural number n can be written as a sum of distinct Fibonacci numbers:

$$n = F_{k_1} + \cdots + F_{k_t},$$

where $t \geq 1$ and $k_1 > \dots > k_t \geq 1$.

Exercise 1.4.7. Let $(x_n)_{n \geq 0}$ be a sequence of numbers defined recursively as $x_0 = 1$, $x_1 = 6$ and

$$x_{n+1} = 7x_n - 12x_{n-1}$$

for $n \geq 1$. Determine x_n in terms of n for any $n \geq 0$.

Exercise 1.4.8. Let $(y_n)_{n \geq 0}$ be a sequence of numbers defined recursively as $y_0 = 2$, $y_1 = 9$ and

$$y_{n+1} = 6y_n - 9y_{n-1}$$

for $n \geq 1$. Determine y_n in terms of n for any $n \geq 0$.

Exercise 1.4.9. There are $n \geq 2$ towns in a country such that between any two of them, there is a unique one-way road. Show that there is a route that goes through every town exactly once.

Exercise 1.4.10. Prove that every natural number can be written as a sum of distinct powers of two.

1.5. There are 10 Kinds of People in the World

The digits and numbers we use these days appeared for the first time in India and were introduced to Europe by Arabic merchants. That is why they are often referred to as Arabic numerals. As witnessed by the well-watched Super Bowl games, there are such things as Roman numerals where I stands for 1, V for 5, X for 10, L for 50, C for 100, D for 500, and M for 1000. The Kansas City Chiefs won Super Bowl LIV in 2020 and the Tampa Bay Buccaneers won Super Bowl LV in 2021. We describe a brief correspondence between Roman numerals and Arabic numbers in the table below. The reader may have other examples, but the following can serve perhaps to remind us how nice it is to calculate $17 + 73 = 90$ instead of $XVII+LXXIII=XC$ or write 888 instead of $DCCCLXXXVIII$.

The number system we use these days has all kinds of advantages including the fact that the position of a digit in a number is important. Since our youth, we are conditioned to regard our natural numbers as

$$27 = 2 \cdot 10^1 + 7 \cdot 10^0$$

or

$$532 = 5 \cdot 10^2 + 3 \cdot 10 + 2 \cdot 10^0$$

or

$$2022 = 2 \cdot 10^3 + 0 \cdot 10^2 + 2 \cdot 10 + 2 \cdot 10^0,$$

meaning that we write and think about the natural numbers in relation to the powers of 10. Looking more carefully, we see that all the representations above have terms involving consecutive powers of 10 with each such power being multiplied by a number between 0 and 9.

Table 1.5.1. Some Roman and Arabic numerals.

| I | II | IV | V | VI | VIII | IX | X | XI | XXIX | XLVII | XCIV |
|---|----|----|---|----|------|----|----|----|------|-------|------|
| 1 | 2 | 4 | 5 | 6 | 8 | 9 | 10 | 11 | 29 | 47 | 94 |

Let us play a bit more and try to do a reverse engineering of this process. If we are given the natural number 319, how exactly can we break it down into powers of 10 as above? Here is one way of doing this. First, we can try to fit as many 10s into 319 as possible, and we end up with 31 of them and something left over:

$$319 = 31 \cdot 10 + 9.$$

Everything seems to be okay except that the power of 10 is multiplied by 31 which is not between 0 and 9. However, we can see how many 10s we can fit into 31 and get that

$$31 = 3 \cdot 10 + 1$$

Since 3 and 1 are digits, it looks like we might not be able to push this procedure further. Starting from the top, we get that

$$\begin{aligned} 319 &= 31 \cdot 10 + 9 = (3 \cdot 10 + 1) \cdot 10 + 9 \\ &= 3 \cdot 10^2 + 1 \cdot 10^1 + 9 \cdot 10^0. \end{aligned}$$

Some natural questions arise. Would such a representation work if we change 10 to a different number: 2, 5, 16, ...? We will show that the answer is yes, and moreover, given a natural number $b \geq 2$, which we call a base, any natural number has a unique representation in base b that looks similar to the ones above. Also, why did 10 become so important in our natural number system? As mentioned before, the Mayans actually used 20 as the base of their number system while the Mesopotamian culture used 60 as the base. It is likely that having ten fingers helped promote 10 to the center in our number system and having 20 fingers and toes lead to the base 20.

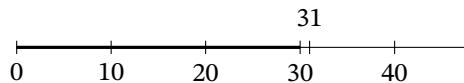


Figure 1.5.1. Dividing 31 by 10 gives $31 = 3 \cdot 10 + 1$.

The procedure we described above involving filling up numbers with as many multiples of 10 as possible is a general result involving natural numbers.

Proposition 1.5.1 (Euclidean Division or Division with Remainder). *Let a and b be two natural numbers. There exist unique numbers q and r in \mathbb{N}_0 such that*

$$a = bq + r \text{ and } 0 \leq r \leq b - 1.$$

Definition 1.5.1. The number q is called the **quotient** and the number r is called the **remainder** of the division of a by b . When $r = 0$, we say that b **divides** a and we write $b \mid a$. When $r > 0$, then b **does not divide** a which we write $b \nmid a$.

Proof of Proposition 1.5.1: The proof we give now follows the same principle as the procedure we followed above when dividing 31 or 19 by 10 with a twist. Instead of piling up b inside the a , we first show that we can always pile up enough b 's to exceed a .

More precisely, let $K = \{k \in \mathbb{N}_0 : kb > a\}$. Informally, K keeps track of the multiples of b that are above a . The first question is whether there are any such multiples. This is where the fact that b is nonzero comes in play. Because $b \neq 0$, we have that $b \geq 1$ and if we multiply both sides by a nonnegative number t , we get that $tb \geq t$. If we take $k = a + 1$, then $(a + 1)b \geq (a + 1) > a$. Hence, K is a nonempty subset of \mathbb{N}_0 . Therefore, K has a least element. Such element must be unique and let us denote it by ℓ . Note that ℓ cannot be 0 as otherwise, a would be 0. Thus, ℓ is a natural number, and we can write it as $\ell = q + 1$, where $q = \ell - 1 \in \mathbb{N}_0$. Also, $q + 1 \in K$ and $q \notin K$ imply that

$$qb \leq a < (q + 1)b.$$

Rearranging terms will give us

$$0 \leq a - qb < b.$$

We can now denote $a - qb$ by r and deduce that $a = bq + r$ and $0 \leq r < b$. The fact that b and r are unique follows from the uniqueness of $\ell = q + 1$. ■

The result we just proved is actually true for integer numbers as we will see in the next chapter. We will apply it right away to the problem of writing natural numbers in different bases.

Definition 1.5.2. Let $b \geq 2$ be a natural number. We say that $(x_k x_{k-1} \dots x_1 x_0)_b$ is the representation in base b of the natural number n if

$$n = x_k b^k + x_{k-1} b^{k-1} + \dots + x_1 b^1 + x_0 b^0$$

with $0 \leq x_k, \dots, x_0 \leq b - 1$ and $x_k \neq 0$.

The numbers x_k, \dots, x_0 are called the digits of n in base b . For example, when $b = 2$, the table below gives the representations of the numbers between 0 and 9 in base 2: The readers who might have been puzzled by the title of our section *There are 10 kinds of people* will get a reward for their patience by getting an explanation here. The common phrase

*There are 10 kinds of people in the world:
those who understand binary and those who don't!*

simply refers to the representation of 2 in base 2 which is 10. The base 2 is usually called binary base and is used in the computer representation of data.

Table 1.5.2. The binary representation of some numbers.

| n | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|--------|---|---|----|----|-----|-----|-----|-----|------|------|
| base 2 | 0 | 1 | 10 | 11 | 100 | 101 | 110 | 111 | 1000 | 1001 |

The following results show that this definition makes sense and any natural number n has a unique representation in a given base $b \geq 2$.

Theorem 1.5.2. Let $b \geq 2$ be a natural number. Any natural number has a unique representation in base b .

Proof: First, we prove that every natural number has a representation in given base b . The argument presented earlier for 319 and $b = 10$ contains the key ideas that will work in a more general setting. Let $b \geq 2$ be a natural number. We will use strong induction on $n \in \mathbb{N}$ to show that any natural number has a representation in base b .

For the base case, let n be any natural number between 1 and $b - 1$. In this case, n can be written as $n = x_0 \cdot b^0$ with $1 \leq x_0 = n \leq b - 1$. This means that n is represented by a single digit (the number n itself) in base b .

For the induction step, assume that $n \geq b$ is a natural number and that any natural number strictly less than n has a representation in base b . Dividing n by b , we deduce that

$$n = bq + r$$

where $q, r \in \mathbb{N}_0$ such that $1 \leq q < n$ and $0 \leq r \leq b - 1$. The fact that $b \geq 2$ comes into play here and implies that q must be strictly less than n . Also, $n \geq b$ implies that $q \geq 1$. Thus, q is a natural number that is strictly less than n . By the induction hypothesis, q has a representation $(x_k \dots x_1)_b$ in base b :

$$q = x_k b^{k-1} + \dots + x_1$$

where $k \geq 1$, $0 \leq x_k, \dots, x_1 \leq b - 1$ and $x_k \neq 0$. Plugging in q in $n = bq + r$ gives us that

$$n = bq + r = b(x_k b^{k-1} + \dots + x_1) + r = x_k b^k + \dots + x_1 b^1 + r.$$

If we let $x_0 = r \in \{0, \dots, b - 1\}$, we conclude that $(x_k \dots x_1 x_0)_b$ is a representation of n in base b .

For the uniqueness part, we show how given natural numbers n and $b \geq 2$, the digits of a representation of n in base b are determined uniquely. If $0 \leq n \leq b - 1$, then n has a one digit representation $(x_0)_b$ in base b with $x_0 = n$. Otherwise, if $n \geq b$, then

$$n = (x_k \dots x_1 x_0)_b = x_k b^k + \dots + x_1 b^1 + x_0 = b(x_k b^{k-1} + \dots + x_1) + x_0,$$

with $k \geq 1$. Since $0 \leq x_0 \leq b - 1$, x_0 must be the remainder of the division of n by b . The equation above also implies that $x_k b^{k-1} + \dots + x_1$ is the quotient of the division of n by b . If it is less than or equal to $b - 1$, then $k = 1$ and n must have the representation $(x_1 x_0)_b$. Otherwise, the procedure continues as before:

$$x_k b^{k-1} + \dots + x_2 b^1 + x_1 = b(x_k b^{k-2} + \dots + x_2) + x_1$$

with x_1 being the remainder of the division of $x_k b^{k-1} + \dots + x_2 b^1 + x_1$. This procedure continues until the quotient of the division is strictly less than b . ■

Another way to determine the digits of the representation of n in base b is by figuring out k first in terms of n and b . Using $0 \leq x_k, \dots, x_0 \leq b - 1$ and $x_k \geq 1$, some simple manipulations give us the following inequalities:

$$n = x_k b^k + \dots + x_1 b^1 + b_0 \geq b^k,$$

and

$$\begin{aligned} n &= x_k b^k + \dots + x_1 b^1 + b_0 \leq (b-1)b^k + \dots + (b-1)b^1 + (b-1) \\ &= (b-1)(b^k + \dots + b^1 + 1) = b^{k+1} - 1 < b^{k+1}. \end{aligned}$$

Thus, $b^k \leq n < b^{k+1}$. Taking logarithms in base b , we get that $k \leq \log_b(n) < k+1$. Thus, k is the largest natural number that is at most $\log_b(n)$. This is called the **floor** of $\log_b(n)$ and is denoted by $\lfloor \log_b(n) \rfloor$. Hence, the number of digits of n in base b is $k+1 = \lfloor \log_b(n) \rfloor + 1$.

Proposition 1.5.3. *The number of digits of $n \in \mathbb{N}$ in base $b \geq 2$ is $\lfloor \log_b(n) \rfloor + 1$.*

Note that in many situations, the value of k can be obtained without the use of a calculator computing logarithms, simply by using the inequalities $b^k \leq n < b^{k+1}$. For example, for $n = 247$ and $b = 2$, one can observe that $2^7 = 128 < 247 < 256 = 2^8$. This implies that 247 has 8 digits in base 2. We leave it as an exercise for the reader to show that $247 = 11110111_2$.

Refining this argument, we can obtain x_k as follows:

$$n = x_k b^k + \dots + x_1 b^1 + b_0 \geq x_k b^k,$$

and

$$\begin{aligned} n &= x_k b^k + \cdots + x_1 b^1 + b_0 \leq x_k b^k + (b-1)(b^{k-1} + \cdots + 1) \\ &= x_k b^k + b^k - 1 = (x_k + 1)b^k - 1 \\ &< (x_k + 1)b^k. \end{aligned}$$

Hence, $x_k b^k \leq n < (x_k + 1)b^k$. Thus, x_k is the largest natural number such that $x_k b^k \leq n$. The digit x_{k-1} and other subsequent can be determined similarly by noticing that $n - x_k b^k = x_{k-1} b^{k-1} + \cdots + x_1 b^1 + x_0$ and so on.

Let us illustrate these results with some examples.

Example 1.5.1. Take $n = 19$ and $b = 2$. If we try to determine the digits of n in base 2 representations starting from the right, we do a sequence of integer divisions:

$$\begin{aligned} 19 &= 9 \cdot 2 + 1 \\ 9 &= 4 \cdot 2 + 1 \\ 4 &= 2 \cdot 2 + 0 \\ 2 &= 2 \cdot 1 + 0 \\ 1 &= 2 \cdot 0 + 1. \end{aligned}$$

The representations of the numbers of the left can be obtained by concatenating the remainders from going from bottom up with

$$2 = 10_2, 4 = 100_2, 9 = 1001_2, 19 = 10011_2.$$

If we first wish to determine the number of digits of 19 in base 2, then we start by noticing that $2^4 = 16 \leq 19 < 32 = 2^5$. This means that $k = 4$ and 19 has 5 digits in base 2. The value of $x_4 = 1$ can be found immediately and all that is left is to determine the representation of $19 - x_4 \cdot 2^4 = 19 - 16 = 3$ in base 2. It is not too hard to see that $3 = 11_2$. Putting these things together, we get that $19 = 1 \cdot 2^4 + 0 \cdot 2^3 + 0 \cdot 2^2 + 1 \cdot 2^1 + 1 \cdot 2^0 = 11001_2$ as before.

Example 1.5.2. Let $n = 249$ and $b = 16$. The base 16 is commonly used in computer science and is usually called hexadecimal base. The digits in base 16 are $0, 1, \dots, 9, A, B, C, D, E, F$ where $A = 10, B = 11, C = 12, D = 13, E = 14$ and $F = 15$. As a simple example, 19 in base 16 would be 13_{16} . Using integer division, we get that

$$249 = 15 \cdot 16 + 9 \text{ and } 15 = 0 \cdot 16 + 15.$$

With $F = 15$, we get that $249 = F9_{16}$.

To complete this section, let us discuss how to add two numbers n and m both represented with respect to a given base $b \geq 2$. Thus, assume

$$n = \sum_{j=0}^k x_j b^j \quad \text{while} \quad m = \sum_{j=0}^{\ell} y_j b^j$$

for some $0 \leq x_j, y_j < b$. By choosing of $x_j = 0$ or $y_j = 0$ if $\min\{k, \ell\} < j \leq \max\{k, \ell\}$ we may assume $k = \ell$, which makes the calculations a bit easier. At a first glance one

may guess that

$$n + m = \sum_{j=0}^k (x_j + y_j)b^j$$

is the desired representation of $n + m$ in base $b \geq 2$. Indeed, this is so as long as all $x_j + y_j < b$. But what happens if this is not valid? We know that

$$n + m = \sum_{j=0}^{k+1} z_j b^j \quad \text{with} \quad 0 \leq z_j < b.$$

We calculate the z_j 's inductively. Choose

$$z_0 = \begin{cases} x_0 + y_0 & : x_0 + y_0 < b \\ x_0 + y_0 - b & : x_0 + y_0 \geq b. \end{cases}$$

Then we obtain

$$n + m = \begin{cases} \sum_{j=1}^{k+1} (x_j + y_j)b^j + z_0 & : x_0 + y_0 < b \\ \sum_{j=2}^{k+1} (x_j + y_j)b^j + (x_1 + y_1 + 1)b + z_0 & : x_0 + y_0 \geq b \end{cases}$$

with $0 \leq z_0 < b$. We proceed now further in this way. But note that we now have to treat four different cases.

$$z_1 = \begin{cases} x_1 + y_1 & : x_0 + y_0 < b, \quad x_1 + y_1 < b \\ x_1 + y_1 - b & : x_0 + y_0 < b, \quad x_1 + y_1 \geq b \\ x_1 + y_1 + 1 & : x_0 + y_0 \geq b, \quad x_1 + y_1 + 1 < b \\ x_1 + y_1 + 1 - b & : x_0 + y_0 \geq b, \quad x_1 + y_1 + 1 \geq b. \end{cases}$$

At the end we have either $z_{k+1} = 0$ if $n + m \leq b^k$ or $z_{k+1} = 1$ otherwise.

Example 1.5.3. Suppose we have two integers represented in base 2. Let

$$n = 10011010011_2 \quad \text{and} \quad m = 10000101011_2.$$

How to get their sum? Since $x_0 = y_0 = 1$, it follows that $z_0 = x_0 + y_0 - 2$, hence $z_0 = 0$. Because of $x_1 + y_1 + 1 = 3 > 2$ we get $z_1 = 1 + 1 + 1 - 2 = 1$. Next we look for $x_2 + y_2 + 1 = 0 + 0 + 1 < 2$, hence $z_2 = 0$. In the next step investigate $x_3 + y_3 = 1$, thus, $z_3 = 1$. Proceeding further we finally end up with

$$n + m = 10001111110_2,$$

$$\begin{array}{r} 1 & 0 & 0 & 1 & 1 & 0 & 1 & 0 & 0 & 1 & 1 \\ + & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 1 \\ \hline 1 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 0 \end{array}$$

By the way, $n = 1235$, $m = 1067$, hence $n + m = 2302$.

Example 1.5.4. Let $b = 16$ and we want to add $5C0F_{16}$ and $3A2B_{16}$. Recall that the letters A, \dots, F are used as symbols for $10, \dots, 15$,

$$\begin{array}{r} 5 & C & 0 & F \\ + & 3 & A & 2 & B \\ \hline 9 & 6 & 3 & A. \end{array}$$

Here $x_0 + y_0 = F + B = 26 > 16$, hence $z_0 = x_0 + y_0 - 16 = 10 = A$, hence $z_1 = 0 + 2 + 1 = 3$. Similarly, $x_2 + y_2 = C + A = 22 > 16$, thus $z_2 = C + A - 16 = 6$ and $z_3 = 5 + 3 + 1 = 9$. Note that in base 10 we have $n = 23567$, $m = 14891$, therefore $n + m = 38458$.

Exercise 1.5.1. Transform $15_2, 18_3, 47_4, 247_5, 1001_7, 2019_{16}$ into base 10.

Exercise 1.5.2. Transform $3_7, 21_5, 39_{10}, 2A_{16}$ into base 2.

Exercise 1.5.3. Calculate $1110110_2 - 1001001_2, 101_2 \cdot 110_2, 4A3B_{16} - 2F5C_{16}$.

Exercise 1.5.4. Let m and n be two natural numbers. If $b \geq 2$ is a natural number such that the representations of m and n in base b have the same number of digits:

$$\begin{aligned}m &= (x_k \dots x_0)_b \\n &= (y_k \dots y_0)_b,\end{aligned}$$

show that $m > n$ if and only if $x_k > y_k$ or there exist $0 \leq \ell \leq k-1$ such that $x_\ell > y_\ell$ and $x_t = y_t$ for any $\ell < t \leq k$.

Exercise 1.5.5. Extend the criterion in Exercise 1.5.4 to the case that m and n in base $b \geq 2$ may have a different number of digits.

Exercise 1.5.6. Let n be a natural number. Show that n is even if and only if its last digit in its binary representation is 0.

Exercise 1.5.7. If n has exactly k digits in base 2, how many digits will $2n$ have in base 2?

Exercise 1.5.8. Let k and $b \geq 2$ be two natural numbers. What is the largest number that has exactly k digits in base b ?

Exercise 1.5.9. Let k and $b \geq 2$ be two natural numbers. What is the smallest number that has exactly k digits in base b ?

Exercise 1.5.10. Let k and $b \geq 2$ be two natural numbers. How many numbers have exactly k digits in base b ?

Exercise 1.5.11. We have a balance and three weights: 1, 3, and 9 pounds. Prove that we can correctly weigh any object whose weight is a natural number between 1 and 13. For example to weigh an object of weight 5, we can put 9 pounds on one side and 1 + 3 pounds and the object on the other side. Prove that one can do the same thing with four weights: 1, 3, 9, and 27 and any object whose weight is between 1 and 40. Generalize to $k \geq 5$ weights.

1.6. Divisibility

In ancient times, some numbers acquired special significance due to their connection to celestial bodies or the human body. For example, 7 appears as the number of days in the creation or the number of heavenly bodies: the Sun, the Moon, Mercury, Venus, Mars, Jupiter, and Saturn. We have 5 orifices in our head: two ears, two nostrils and one mouth. Also, 5 was special because there are exactly five Platonic solids: tetrahedron, cube, octahedron, dodecahedron, and icosahedron which *corresponded* according to

ancient Greeks to the 5 classical elements: fire, earth, air, ether, and water. As we come closer to modern times, the numbers lost their association with mysticism, but their properties still fascinate mathematicians and there are many things that we do not know how to prove.

In the previous section, we have seen how one may try to divide a natural number a by another natural number b . Intuitively, one would try to fit as many b 's as possible inside a . Removing this multiple of b from a gives the remainder of the division of a by b which is always a number between 0 and $b - 1$. When the remainder is 0, it means that b divides a or that a is a multiple of b .

Definition 1.6.1. Let $a \in \mathbb{N}_0$ and $b \in \mathbb{N}$. We say that b **divides** a which we denote by $b | a$ if there exists $k \in \mathbb{N}$ such that $a = kb$.

$$(b \text{ divides } a) \Leftrightarrow (\exists k \in \mathbb{N})(a = b \cdot k) \Leftrightarrow (a/b \in \mathbb{N})$$

If b does not divide a , then we write $b \nmid a$. For example, $3 | 6$, but $3 \nmid 8$. Note that every natural number divides 0, but 0 does not divide anything.

Divisibility is an example of a binary relation (cf. Definition A.5.1). That is, the integer b is in relation with integer a whenever $b | a$.

Proposition 1.6.1. *The following holds for any $a, b, c \in \mathbb{N}$:*

- (1) **Reflexive** $a | a$.
- (2) **Anti-symmetric** If $a | b$ and $b | a$, then $a = b$.
- (3) **Transitive** If $a | b$ and $b | c$, then $a | c$.

This makes divisibility over \mathbb{N} an example of a **partial order** as introduced in Definition A.5.4.

Proof: Let $a, b, c \in \mathbb{N}$. For the first property (reflexivity), note that $a = a \cdot 1$ and $1 \in \mathbb{N}$ imply that $a | a$.

For the second property (anti-symmetry), if $a | b$, then $b = ak$ with $k \in \mathbb{N}$. Also, if $b | a$, then $a = b\ell$ with $\ell \in \mathbb{N}$. Combining these two equations, we obtain that $b = ak = (b\ell)k = b\ell k$. Dividing both sides by $b \neq 0$, we get that $1 = \ell k$. Since $\ell, k \in \mathbb{N}$, $\ell \geq 1$ and $k \geq 1$. Therefore, $\ell k \geq 1 \cdot 1 = 1$ with equality if and only if $\ell = k = 1$. Thus, $\ell k = 1$ implies $\ell = k = 1$ and $b = ak = a$ as desired.

For the third property (transitivity), if $a | b$, then $b = ak$ with $k \in \mathbb{N}$. Also, if $b | c$, then $c = bu$ with $u \in \mathbb{N}$. Combining these two equations, we get that $c = bu = (ak)u = a(ku)$. Because $ku \in \mathbb{N}$, we deduce that $a | c$. This finishes our proof. ■

In addition, there is another simple property of divisibility that will be used throughout the book. Informally, it says that if a natural number divides several other numbers, then it will divide any linear combination of them.

Proposition 1.6.2. Let $d, k \in \mathbb{N}$ and $a_1, \dots, a_k \in \mathbb{N}_0$ be such that d divides each number a_1, \dots, a_k . Then

$$d \mid \sum_{j=1}^k \lambda_j a_j,$$

for any $\lambda_1, \dots, \lambda_k \in \mathbb{N}_0$.

Proof: For each $1 \leq j \leq k$, $d \mid a_j$ implies that $a_j = db_j$ for some $b_j \in \mathbb{N}_0$. Therefore,

$$\sum_{j=1}^k \lambda_j a_j = \sum_{j=1}^k \lambda_j db_j = d \sum_{j=1}^k \lambda_j b_j.$$

Because $\sum_{j=1}^k \lambda_j b_j \in \mathbb{N}_0$, we get the desired result. ■

There is a special family of numbers whose only divisors are 1 and themselves, the primes.

Definition 1.6.2. A natural number $p \geq 2$ is called **prime** if the only natural divisors of p are 1 and p .

$$(p \geq 2 \text{ prime}) \Leftrightarrow [(\forall a \in \mathbb{N}) (a \mid p \Rightarrow (a = 1 \text{ or } a = p))]$$

A natural number that is not prime is called **composite**. Prime and composite numbers have fascinated people since ancient times and one way of producing primes was devised by Eratosthenes in ancient Alexandria around 200 BCE. He invented a famous method, called the **Sieve of Eratosthenes**, which strains out the prime numbers by removing the composite numbers from the list. More precisely, if we list the numbers from 2 to 101 as we do below, recognize that 2 is a prime, but any larger multiple of 2 on our list will be composite, and we cross it off the list: ~~4, 6, ...~~ and so on. We then do the same thing with 3 which is a prime, but any larger multiple of it must be composite and must be crossed off the list: ~~6, 9, ...~~ and so on. The procedure continues with 5 and its multiples: ~~10, 15, ...~~. The numbers that survive these slashes will be the prime numbers. Another amazing thing that Eratosthenes did was to approximate the circumference of the Earth to about 24,466 miles where the current accepted length is 24,800 miles⁷.

Table 1.6.1. The sieve of Eratosthenes up to 97.

| 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 |
|----|----|----|----|----|----|----|----|----|----|----|----|
| 14 | 15 | 16 | 17 | 18 | 19 | 20 | 21 | 22 | 23 | 24 | 25 |
| 26 | 27 | 28 | 29 | 30 | 31 | 32 | 33 | 34 | 35 | 36 | 37 |
| 38 | 39 | 40 | 41 | 42 | 43 | 44 | 45 | 46 | 47 | 48 | 49 |
| 50 | 51 | 52 | 53 | 54 | 55 | 56 | 57 | 58 | 59 | 60 | 61 |
| 62 | 63 | 64 | 65 | 66 | 67 | 68 | 69 | 70 | 71 | 72 | 73 |
| 74 | 75 | 76 | 77 | 78 | 79 | 80 | 81 | 82 | 83 | 84 | 85 |
| 86 | 87 | 88 | 89 | 90 | 91 | 92 | 93 | 94 | 95 | 96 | 97 |

⁷See [8] to learn more about this and other beautiful mathematical results.

It sometimes feels that prime numbers get all the glory, but composite numbers are also interesting. An immediate observation is that 2 is the only even prime number and is left as an exercise. The study of primes sits at the core of number theory and is full of fascinating questions. We start our investigation with a simple and natural result which states that any natural number $n \geq 2$ has a prime factor.

Proposition 1.6.3. *For any natural number $n \geq 2$, there exists a prime number p such that p divides n .*

Proof: We use strong induction on n . The base case $n = 2$ is true since 2 is a prime number and $2 \mid 2$. Let $n \geq 2$ be a natural number. Assume that any natural number k such that $2 \leq k \leq n$, has at least one prime divisor. We want to prove that the number $n + 1$ also has a prime divisor. If $n + 1$ is prime, then we are lucky: $n + 1$ divides itself and is a prime, and we are done. Otherwise, if $n + 1$ is not a prime, then $n + 1$ must be divisible by some number ℓ such that $1 < \ell < n + 1$. By our induction hypothesis, ℓ must be divisible by a prime number. By transitivity, such a prime number will divide $n + 1$. This finishes our proof. ■

Another way of interpreting this proof is that if a natural number $n \geq 2$ is not prime, then we can always *peel off* a prime divisor p and be left with a smaller number. Our first theorem of this section is one of the most important results in number theory and appeared in Euclid's Elements around 300 BCE.

Theorem 1.6.4. *There are infinitely many prime numbers.*

Proof: We give a proof by contradiction. Assume that there are finitely many primes: p_1, \dots, p_m for some natural number m . Consider the number $p_1 \dots p_m + 1$. This is obviously a natural number. Also, it is a number that is larger than any of the primes p_1, \dots, p_m . We are now at a crossroads: is $p_1 \dots p_m + 1$ prime or not? If it is prime, then we have obtained a prime number that was not in our list and this contradicts our assumption that the only primes are p_1, \dots, p_m . If $p_1 \dots p_m + 1$ is not a prime, then by our previous proposition, it should be divisible by a prime that is smaller. Since our list of primes is p_1, \dots, p_m , this means that there is some j between 1 and m such that p_j divides $p_1 \dots p_m + 1$. Since p_j also divides $p_1 \dots p_m$, we deduce that p_j must divide the difference $p_1 \dots p_m + 1 - p_1 \dots p_m = 1$. Therefore, p_j must be one which is impossible since p_j is a prime. Regardless of the status of $p_1 \dots p_m + 1$, we obtain a contradiction. Hence, our assumption was false and there are actually infinitely many primes. ■

We can reformulate the gist of the previous theorem as follows. Say we can write down the first few prime numbers: 2 and 3. To come up with another prime number, we can calculate $2 \cdot 3 + 1 = 7$ which happens to be a prime number. Now we have a list, albeit not complete since it does not contain 5, of three prime numbers 2, 3, 7. If we repeat the procedure, $2 \cdot 3 \cdot 7 + 1 = 43$ happens to be a prime again. Our list of primes is 2, 3, 7, 43 and the next step would give us $2 \cdot 3 \cdot 7 \cdot 43 + 1 = 1807$ which is a bit trickier to figure out. A bit of work gives $1807 = 13 \cdot 139$ and 13 is a prime factor that was not on our list. This method is definitely not the most practical or thorough way of generating primes, but its main value is conceptual and lies in giving a rigorous proof of the infinity of primes. The first prime numbers are listed below. The reader

Table 1.6.2. The table of primes up to 991.

| | | | | | | | | | | |
|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| 2 | 3 | 5 | 7 | 11 | 13 | 17 | 19 | 23 | 29 | 31 |
| 37 | 41 | 43 | 47 | 53 | 59 | 61 | 67 | 71 | 73 | 79 |
| 83 | 89 | 97 | 101 | 103 | 107 | 109 | 113 | 127 | 131 | 139 |
| 149 | 151 | 157 | 163 | 167 | 173 | 179 | 181 | 191 | 193 | 197 |
| 199 | 211 | 223 | 227 | 229 | 233 | 239 | 241 | 251 | 257 | 263 |
| 269 | 271 | 277 | 281 | 283 | 293 | 307 | 311 | 313 | 317 | 331 |
| 337 | 347 | 349 | 353 | 359 | 367 | 373 | 379 | 383 | 389 | 397 |
| 401 | 409 | 419 | 421 | 431 | 433 | 439 | 443 | 449 | 457 | 461 |
| 463 | 467 | 479 | 487 | 491 | 499 | 503 | 509 | 521 | 523 | 541 |
| 547 | 557 | 563 | 569 | 571 | 577 | 587 | 593 | 599 | 601 | 607 |
| 613 | 617 | 619 | 631 | 641 | 643 | 647 | 653 | 659 | 661 | 673 |
| 677 | 683 | 691 | 701 | 709 | 719 | 727 | 733 | 739 | 743 | 751 |
| 757 | 761 | 769 | 773 | 787 | 797 | 809 | 811 | 821 | 823 | 827 |
| 829 | 839 | 853 | 857 | 859 | 863 | 877 | 881 | 883 | 887 | 907 |
| 911 | 929 | 937 | 941 | 947 | 953 | 967 | 971 | 977 | 983 | 991 |

may be perhaps surprised to know that in 1762, Euler compiled a table of large primes and the largest prime number he mentioned was 2232037.

Our next result expresses our intuition that the primes are the atoms of the natural numbers realm and that any natural number can be written as a unique product of primes. Here uniqueness refers to a possible reordering of the factors in the product. For example, $60 = 2 \cdot 2 \cdot 3 \cdot 5$ but also equals $3 \cdot 2 \cdot 5 \cdot 2$ or $5 \cdot 2 \cdot 3 \cdot 2$. The meaning of the above sentence is that all such factorizations will be considered the same. This result is so important that it is usually referred to as the **fundamental theorem of arithmetic**.

Theorem 1.6.5. *Every natural number $n \geq 2$ can be written as a product of primes. This representation is unique up to the ordering of the prime factors.*

Proof: We deal with the existence and the uniqueness of prime factorization separately.

For the existence of prime factorization, we again use strong induction to show that every natural number $n \geq 2$ can be written as a product of primes. Our argument is similar to the one used for Proposition 1.6.3, and essentially we take that argument one step further. The base case is $n = 2$ and our statement is true in this situation as 2 is a prime number. Let $n \geq 2$ be a natural number and assume that any natural number k between 2 and n can be written as a product of primes. Consider now the number $n + 1$. If $n + 1$ is prime, then we are done. Otherwise, by Proposition 1.6.3, there exists a prime p that divides $n + 1$. Therefore, $n + 1 = p \cdot k$, where k is a natural number with $1 < k < n + 1$. Because $k < n + 1$, by our induction hypothesis, k is a product of primes. Since p is a prime, it follows that $n + 1 = p \cdot k$ is also a product of primes. This finishes the existence part of our proof.

For the uniqueness of prime factorization, we give a proof by contradiction. Assume that there exists at least one natural number that can be written as a product of primes in two different ways. By the least element principle of \mathbb{N} (see Section 1.4), there

exist a smallest such number, let us call it n for which we have that

$$n = p_1 \dots p_s = q_1 \dots q_t,$$

where $s, t \in \mathbb{N}$ with $p_1 \leq \dots \leq p_s$ and $q_1 \leq \dots \leq q_t$ being primes. Our first observation is that none of the p_1, \dots, p_s can equal any of the q_1, \dots, q_t . If that were the case and $p_i = q_j$ for some $1 \leq i \leq s$ and $1 \leq j \leq t$, then we can divide both sides above by $p_i = q_j$ and obtain that the smaller number

$$\prod_{k=1: k \neq i}^s p_k = \prod_{\ell=1: \ell \neq j}^t q_\ell < n$$

can be written in two different ways as a product of primes. This would contradict the minimality of n . A consequence of this observation is that $p_1 \neq q_1$. We may assume that $p_1 < q_1$ (if $p_1 > q_1$ the proof is similar and left as an exercise, just switch the p 's and the q 's). Consider now the number:

$$n' := n - p_1 q_2 \dots q_t.$$

This is a natural number because $q_1 > p_1$. Substituting $n = p_1 \dots p_s = q_1 \dots q_t$ into this equation, we get two expressions for n' :

$$\begin{aligned} n' &= p_1 \dots p_s - p_1 q_2 \dots q_t = p_1(p_2 \dots p_s - q_2 \dots q_t) \\ &= q_1 \dots q_t - p_1 q_2 \dots q_t = (q_1 - p_1)q_2 \dots q_t. \end{aligned}$$

The number n' is a natural number smaller than n . Since n was the smallest natural number that does not have a unique factorization into primes, it follows that the number n' must have a unique factorization into primes. From the first expression for n' above: $n' = p_1(p_2 \dots p_s - q_2 \dots q_t)$, we deduce that n' must contain the prime factor p_1 in its factorization into primes. Therefore, p_1 must appear in the factorization of the 2nd expression of n' above: $n' = (q_1 - p_1)q_2 \dots q_t$. Hence, p_1 must divide $(q_1 - p_1)q_2 \dots q_t$. Because $p_1 < q_1 \leq q_2 \leq \dots \leq q_t$, we deduce that we will not find p_1 in the factorization $q_2 \dots q_t$. The only possibility left is that p_1 divides $q_1 - p_1$. This implies that p_1 must divide $(q_1 - p_1) + p_1 = q_1$. However, this is impossible since p_1 and q_1 are distinct primes. Our job is done, no natural number can be written in two different ways as a product of primes. ■

As the name states, this is a fundamental result with numerous important consequences. One may wonder how we constructed the previous table of primes. The idea is that if a number n is composite, then the smallest prime p divisor of n must satisfy the inequality $p^2 \leq n$ (see Exercise 1.6.7). Hence, when we encounter a natural number n , we can check if any prime p with $p^2 \leq n$ divides n . If we find such p , then n is composite. Otherwise, if we find no prime p such that $p^2 \leq n$ and $p \mid n$, then we can conclude that n is prime and add it to our table. The interested reader can use this idea to determine some of the primes that appear after 991.

Another consequence of prime factorization is the following result which is usually called Euclid's lemma.

Corollary 1.6.6 (Euclid's Lemma). *Let p be a prime. If a and b are natural numbers such that $p \mid ab$, then $p \mid a$ or $p \mid b$.*

Proof: We can do a proof by contradiction as follows. Assume that p divides ab , but p does not divide a and p does not divide b . Because $p \mid ab$, it follows that p appears in the unique factorization of ab into primes. Now because $p \nmid a$ and $p \nmid b$, p will not appear in the prime factorization of a nor will appear in the prime factorization of b . Hence, p cannot appear in the prime factorization of ab which is obtained by combining the factorization of a and b . This gives us a contradiction and proves our statement. ■

In the next chapter, we will prove a more general result. For now, we remark that the result above is not true if we drop the assumption of p being a prime. For example $6 \mid 3 \cdot 4$, but $6 \nmid 3$ and $6 \nmid 4$.

Another consequence of the fundamental theorem of arithmetic is the following result which informally states that if we know the prime factorization of a natural number, then we can describe all of its divisors.

Corollary 1.6.7. *Let $n \geq 2$ be a natural number whose prime factorization is $n = p_1^{e_1} \dots p_s^{e_s}$, where $p_1 < \dots < p_s$ are prime numbers and e_1, \dots, e_s are natural numbers. If d is a divisor of n , then $d = p_1^{f_1} \dots p_s^{f_s}$, where $0 \leq f_j \leq e_j$ for each $1 \leq j \leq s$.*

With these results in our toolbox, let us turn now to another important notion that has a long history dating back to Euclid 300 BCE and is important in both theoretical and practical aspects of number theory: the greatest common divisor.

Definition 1.6.3. Let a and b be two natural numbers. The **greatest common divisor** of a and b , denoted by $\gcd(a, b)$, is the largest natural number that divides both a and b .

$$(d = \gcd(a, b)) \Leftrightarrow [(d \mid a \text{ and } d \mid b) \text{ and } (\text{if } d' \mid a \text{ and } d' \mid b, \text{ then } d' \leq d)].$$

Note that this definition makes sense. First, a and b have at least one common divisor because 1 divides both of them. A divisor of a natural number a must lie between 1 and a and a divisor of b must lie between 1 and b . Therefore, the set of common divisors of a and b consists of natural numbers that lie between 1 and $\min(a, b)$. Such a set has to have a largest element and that is $\gcd(a, b)$. Note that the notion of the greatest common divisor $\gcd(a, b)$ can be extended for the case when one of a and b is 0. If $b = 0$ and $a \in \mathbb{N}$, then $\gcd(a, b) = \gcd(a, 0) = a$ is the greatest common divisor of a and 0. However, the notion of the greatest common divisor of a and b cannot be defined when both a and b are 0 because in that case, any natural number divides 0 and therefore, there is no smallest common divisor of 0 and 0. In the next chapter, we will expand the definition of the greatest common divisor to integer numbers.

Definition 1.6.4. Let a and b be two natural numbers. If $\gcd(a, b) = 1$, then a and b are called **coprime** or **relatively prime**.

The following proposition states that peeling off the greatest common divisor of two numbers yields coprime numbers.

Proposition 1.6.8. *Let $a, b \in \mathbb{N}$. If $d = \gcd(a, b)$ with $a = da'$ and $b = db'$, $a', b' \in \mathbb{N}$, then a' and b' are coprime.*

Proof: We give a proof by contradiction. Assume that $d = \gcd(a, b)$ and that $\gcd(a', b') \neq 1$. There exists $k \geq 2$ such that $k \mid a'$ and $k \mid b'$. Therefore, $a' = ka''$ and $b' = kb''$ for some $a'', b'' \in \mathbb{N}$. Then $a = da' = (dk)a''$ and $b = db' = (dk)b''$. This implies that dk is a common divisor of a and b which is a contradiction with $d = \gcd(a, b)$ and $dk > d$.

■

Let us summarize the crucial properties of the greatest common divisor.

Proposition 1.6.9. *Let a, b and c be any natural numbers and p be a prime. Then*

- (1) $1 \leq \gcd(a, b) \leq \min\{a, b\}$,
- (2) $\gcd(a, b) = \gcd(b, a)$
- (3) $\gcd(a, 0) = a$,
- (4) $\gcd(a, 1) = 1$,
- (5) $\gcd(a, a) = a$
- (6) $\gcd(a, p) = \begin{cases} p & : \text{if } p \mid a \\ 1 & : \text{if } p \nmid a \end{cases}$
- (7) $\gcd(a, b + ca) = \gcd(a, b)$,
- (8) $\gcd(c a, c b) = c \gcd(a, b)$.

Proof: Property (1) has already been proved. We leave (2)-(6) as exercises for the reader, but will give a proof of (7) because it illustrates the key idea used in the Euclidean algorithm described later. Let $d = \gcd(a, b)$ and $d' = \gcd(a, b + ca)$. We first show that $d \leq d'$. Because $d = \gcd(a, b)$, d must divide both a and b . By Proposition 1.6.2, d must divide the linear combination $b + ca$ and therefore, is a common divisor of a and $b + ca$. However, d' is the greatest common divisor of a and $b + ca$ and this implies that $d \leq d'$.

We now show that $d' \leq d$. Because $d' = \gcd(a, b + ca)$, we get that $d' \mid a$ and $d' \mid b + ca$. Therefore, $a = d'k$ and $b + ca = d'\ell$ for some $k, \ell \in \mathbb{N}$. Substituting $a = d'k$ into $b + ca = d'\ell$, we get that $b + cd'k = d'\ell$ and therefore $b = d'\ell - d'kc = d'(\ell - kc)$. Because $b \in \mathbb{N}$, we deduce that $\ell - kc \in \mathbb{N}$ (since $d' \in \mathbb{N}$) and therefore, $d' \mid b$. Hence, d' is a common divisor of a and b . However, d is the greatest common divisor of a and b . Thus, $d' \leq d$ which finishes this proof.

For the proof of property (8) we refer to Exercise 1.6.3. ■

Let us turn to some practical things and try to calculate $\gcd(a, b)$ for some numbers a and b . When a and b are relatively small, one can find $\gcd(a, b)$ fairly quickly using their prime factorization (see Exercise 1.6.3).

Example 1.6.1. When $a = 48$ and $b = 30$, a few seconds of thought can give us that $a = 2^4 \cdot 3$, $b = 30 = 2 \cdot 3 \cdot 5$ and therefore $\gcd(48, 30) = 2 \cdot 3 = 6$.

However, things get a bit more complicated when the numbers a and b are large and their prime factorization is not easily computable. Take for example $a = 7596$ and $b = 3242$ or $a = 1238471412$ and $b = 9241483$. Fortunately for us, there is a fast and slick way for calculating $\gcd(a, b)$ that has been known for over 2000 years. This is known as the Euclidean algorithm for determining the greatest common divisor.

The basis of the Euclidean algorithm is the following result which essentially reduces the problem of calculating the greatest common divisor of a pair of natural numbers to the same problem for smaller numbers.

Proposition 1.6.10. *Let a and b be two natural numbers. If $a = bq + r$ with $q, r \in \mathbb{N}_0$ and $0 \leq r < b$, then $\gcd(a, b) = \gcd(b, r)$.*

Proof: The proof is similar to the property (7) above. Let $d_1 = \gcd(a, b)$ and $d_2 = \gcd(b, r)$. First, we show that $d_1 \leq d_2$. Because $d_1 = \gcd(a, b)$, it means that $d_1 \mid a$ and $d_1 \mid b$. Therefore, d_1 must divide $a - bq = r$. Thus, $d_1 \mid b$ and $d_1 \mid r$. Since d_2 is the greatest common divisor of b and r and d_1 is a common divisor of b and r , we get that $d_1 \leq d_2$.

To finish the proof, we prove now that $d_1 \geq d_2$. Because $d_2 = \gcd(b, r)$, we get that $d_2 \mid b$ and $d_2 \mid r$. Therefore, d_2 divides $bq + r = a$. Hence, d_2 is a common divisor of a and b . Since d_1 is the greatest common divisor of a and b , we deduce that $d_1 \geq d_2$ and we are done. ■

Informally, the Euclidean algorithm uses Proposition 1.6.10 to kick the can of computing $\gcd(a, b)$ down the road, meaning replacing by the easier problem of calculating $\gcd(b, r)$. This is a useful principle in all mathematics and life that can be summarized by the words of George Pólya:

If you cannot solve a problem, then there is an easier problem that you can solve: find it.

George Pólya (1887–1985) was a famous Hungarian mathematician with many important contributions to combinatorics, number theory and probability. In his book [28] he describes many approaches one can take when solving mathematical problems and is a classic.

We describe below the Euclidean algorithm for a specific pair of natural numbers.

Example 1.6.2. Consider the problem of calculating $\gcd(a, b)$ when $a = 7596$ and $b = 3242$. The previous proposition tells us that because of

$$7596 = 3242 \cdot 2 + 1112 \quad \text{it follows that} \quad \gcd(7596, 3242) = \gcd(3242, 1112).$$

Our initial problem of calculating $\gcd(7596, 3242)$ is replaced by another problem, computing $\gcd(3242, 1112)$, with smaller numbers than before. It is natural to use Proposition 1.6.10 for $\gcd(3242, 1112)$ and doing so, we obtain the following.

$$3242 = 1112 \cdot 2 + 1018 \quad \text{and} \quad \gcd(3242, 1112) = \gcd(1112, 1018).$$

Now we are literally on a roll, and we can repeat this procedure further.

$$1112 = 1018 \cdot 1 + 94 \quad \text{and} \quad \gcd(1112, 1018) = \gcd(1018, 94)$$

$$1018 = 94 \cdot 10 + 78 \quad \text{and} \quad \gcd(1018, 94) = \gcd(94, 78)$$

$$94 = 78 \cdot 1 + 16 \quad \text{and} \quad \gcd(94, 78) = \gcd(78, 16)$$

$$78 = 16 \cdot 4 + 14 \quad \text{and} \quad \gcd(78, 16) = \gcd(16, 14)$$

$$16 = 14 \cdot 1 + 2 \quad \text{and} \quad \gcd(16, 14) = \gcd(14, 2)$$

$$14 = 2 \cdot 7 + 0 \quad \text{and} \quad \gcd(14, 2) = \gcd(2, 0) = 2.$$

Working backward, we deduce that $\gcd(7596, 3242) = 2$. Of course, the algorithm does not have to proceed until one of the numbers is 0 as one can stop it earlier, for example when $\gcd(16, 14)$ or $\gcd(14, 2)$ appear.

The general form of the Euclidean algorithm is as follows. Assume that we are given two numbers $a, b \in \mathbb{N}_0$ with $b > 0$. If $a = 0$, then $\gcd(a, b) = \gcd(0, b) = b$ and we are done. Otherwise, we do integer division of a by b and use Proposition 1.6.10 to get that

$$a = bq_1 + r_1, 0 \leq r_1 < b, \gcd(a, b) = \gcd(b, r_1).$$

The problem of calculating $\gcd(a, b)$ has been reduced to the one of calculating $\gcd(b, r_1)$. If $r_1 = 0$, then $\gcd(b, r_1) = b$ and we are done. Otherwise, if $r_1 > 0$, then we repeat our previous argument: do integer division of b by r_1 and use Proposition 1.6.10 to obtain that

$$b = r_1 q_2 + r_2, 0 \leq r_2 < r_1, \gcd(b, r_1) = \gcd(r_1, r_2).$$

Hence, $\gcd(a, b) = \gcd(b, r_1) = \gcd(r_1, r_2)$. If $r_2 = 0$, then $\gcd(r_1, r_2) = \gcd(r_1, 0) = r_1$ and we are done. Otherwise, we have that $b > r_1 > r_2 > 0$ and we have to calculate $\gcd(r_1, r_2)$. We do integer division of r_1 by r_2 and get that

$$r_1 = r_2 q_3 + r_3, 0 \leq r_3 < r_2, \gcd(r_1, r_2) = \gcd(r_2, r_3).$$

Notice that at each step (integer division performed), the remainder strictly decreases. Therefore, in at most b steps, the remainder of our integer division will be 0. Hence, there is some natural number $n \geq 3$ and a sequence of integer divisions such that

$$a = bq_1 + r_1, 0 \leq r_1 < b, \gcd(a, b) = \gcd(b, r_1)$$

$$b = r_1 q_2 + r_2, 0 \leq r_2 < r_1, \gcd(b, r_1) = \gcd(r_1, r_2)$$

...

$$r_{n-2} = r_{n-1} q_n + r_n, 0 \leq r_n < r_{n-1}, \gcd(r_{n-2}, r_{n-1}) = \gcd(r_{n-1}, r_n)$$

$$r_{n-1} = r_n q_{n+1} + 0, \gcd(r_{n-1}, r_n) = \gcd(r_n, 0) = r_n,$$

with $b > r_1 > r_2 > \dots > r_n > 0 = r_{n+1}$. Hence, $\gcd(a, b) = \gcd(b, r_1) = \dots = \gcd(r_n, 0) = r_n$ which is the last nonzero remainder in our sequence of divisions. The interested readers can implement the Euclidean algorithm in their favorite programming language using the following loop:

```

while (a ≠ 0 and b ≠ 0)
    if (a > b), then a := remainder(a, b);
    else b := remainder(b, a);
    end while;
    return max(a, b);

```

In the code above, $\text{remainder}(a, b)$ gives the remainder of the integer division of a by b and $\max(a, b)$ is the maximum between a and b .

In Example 1.6.2, $a = 7596$, $b = 3242$ and $n+1 = 9$ divisions. It is interesting to say something more about the number of divisions, let us call it $D(a, b)$, of the Euclidean algorithm involved in calculating $\gcd(a, b)$ for $a > b$. Recall that F_n denotes the n -th Fibonacci number (see equation (1.4.3)). The following result is due to the French mathematician Gabriel Lamé (1795–1870) and implies that the number of divisions of the Euclidean algorithm for $a > b$ is at most⁸ $\log_{10} a / \log_{10} \varphi$, where $\varphi = \frac{1+\sqrt{5}}{2}$ is the golden ratio.

⁸Note that $x = \log_{10} a$ is the unique number satisfying $10^x = a$ for a given $a > 0$.

Proposition 1.6.11. *If k is a natural number, then $D(F_{k+2}, F_{k+1}) = k + 1$. If a and b are natural numbers such that $F_{k+2} \geq a > b$, then $D(a, b) \leq k + 1$ with equality if and only if $a = F_{k+2}$ and $b = F_{k+1}$.*

Proof: For the first part, note that the sequence of integer divisions required by the Euclidean algorithm to calculate $\gcd(F_{k+2}, F_{k+1})$ is below:

$$\begin{aligned} F_{k+2} &= F_{k+1} + F_k, \\ F_{k+1} &= F_k + F_{k-1}, \\ \vdots &= \vdots + \vdots, \\ F_2 &= F_1 + F_0, \quad F_0 = 0. \end{aligned}$$

This shows that $\gcd(F_{k+2}, F_{k+1}) = F_1 = 1$ and $D(F_{k+2}, F_{k+1}) = k + 1$.

For the second part, let $a > b$ be two natural numbers such that $F_{k+2} \geq a$ and consider the last two steps of the Euclidean algorithm when applied to a and b :

$$\begin{aligned} r_{n-2} &= r_{n-1}q_n + r_n, \quad 0 < r_n < r_{n-1}, \\ r_{n-1} &= r_nq_{n+1}. \end{aligned}$$

Because $r_{n-1} > r_n$, we deduce that $q_{n+1} \geq 2$. Also, $q_j \geq 1$ for any $0 \leq j \leq n$. Working our way backward from the last two integer divisions of the Euclidean algorithm, we deduce that $r_n \geq 1 = F_1$ and $r_{n-1} \geq 2r_n = 2 = F_2$. From the next division $r_{n-2} = r_{n-1}q_n + r_n$, we get that $r_{n-2} \geq 2 \cdot 1 + 1 = 3 = F_3$. Using strong induction on j , we can prove that $r_{n+1-j} \geq F_j$ for $1 \leq j \leq n+1$. The base case is true as the inequality is true for $j \in \{1, 2, 3\}$ as shown above. For the induction step, let j be a natural number between 1 and n and assume that $r_{n+1-\ell} \geq F_\ell$ for any $1 \leq \ell \leq j-1$. This means that $r_{n+1-(j-1)} \geq F_{j-1}$ and $r_{n+1-(j-2)} \geq F_{j-2}$. Combining these inequalities with the equation $r_{n+1-j} = r_{n+2-j}q_{n+3-j} + r_{n+3-j}$, we deduce that $r_{n+1-j} \geq r_{n+2-j} + r_{n+3-j} \geq F_{j-1} + F_{j-2} = F_j$. This proves our assertion. Hence, $b = r_0 \geq F_{n+1}$ and $a \geq F_{n+2}$. Because $F_{k+2} \geq a$, we get that $n \leq k$ which proves that $D(a, b) = n + 1 \leq k + 1$. The equality case can be analyzed using the arguments above and is left as an exercise. ■

The fundamental theorem of arithmetic is an important result that opens up new questions. If the primes are the atoms of numbers and there are infinitely many primes, then it is of great interest to construct or find some with bare hands and not so much effort. We are not the only ones to have thought this way and here are some attempts to constructing primes from scratch.

Let us consider first numbers of the form $2^n - 1$ for $n \geq 2$ (see also Exercise 1.6.5). The next result gives us a necessary condition for such number to be a prime.

Proposition 1.6.12. *If $2^n - 1$ is a prime, then n must be a prime number.*

Proof: We can prove this result using the contrapositive method. An implication $P \Rightarrow Q$ is equivalent to its contrapositive $\neg Q \Rightarrow \neg P$. Let $n \geq 2$. We will show that if n is composite, then $2^n - 1$ is composite. If n is composite, then $n = k\ell$ for some natural numbers k and ℓ with $1 < k, \ell < n$. Therefore, by (1.3.2),

$$2^n - 1 = 2^{k\ell} - 1 = (2^k - 1)[(2^k)^{\ell-1} + \cdots + (2^k)^1 + 1].$$

Thus, $2^k - 1$ divides $2^n - 1$. Since $k > 1$, $2^k - 1 > 1$ and since $k < n$, $2^k - 1 < 2^n - 1$. Therefore, $2^n - 1$ has a proper factor (not 1 nor itself) and must be a composite number. ■

With this result in mind, we can try things out and write down the first four numbers of the form $2^p - 1$ with p prime:

$$2^2 - 1 = 3, \quad 2^3 - 1 = 7, \quad 2^5 - 1 = 31, \quad 2^7 - 1 = 127.$$

All these numbers are prime. However, the next one $2^{11} - 1 = 2047$ is not prime as $2047 = 23 \cdot 89$. It turns out that many people have studied this problem. If p is a prime and $2^p - 1$ is a prime, then $2^p - 1$ is called a Mersenne prime. These primes get their name from Marin Mersenne (1588–1648), a French scholar and monk who studied them in the 17th century. It turns out that the following numbers of this form:

$$2^{13} - 1 = 8191, \quad 2^{17} - 1 = 131071, \quad \text{and } 2^{19} - 1 = 524287,$$

are also primes. This may be a bit tedious to prove by hand, but with the help of Exercise 1.6.10 and a bit of patience, it can be done. The Mersenne primes sequence is sequence OA000396 in the *Online Encyclopedia of Integers Sequences*, <https://oeis.org/A000396>. At present time, it is not known whether there are infinitely many primes of the form $2^p - 1$. The primes of these form appeared in Euclid's Elements in connection to the perfect numbers.

Definition 1.6.5. A natural number $n \geq 2$ is called **perfect** if it equals the sum of its proper divisors (all its divisors excluding itself).

For example, 4 is not perfect since $4 \neq 1 + 2$, but 6 is perfect because $6 = 1 + 2 + 3$. Perfection seems to be rare and one can check that the next perfect numbers are 28, 496 and 8128. Amazingly enough, these numbers were known to the Greeks who also observed the following pattern in their factorizations:

$$\begin{aligned} 6 &= 2^1(2^2 - 1) = 2 \cdot 3, \\ 28 &= 2^2(2^3 - 1) = 4 \cdot 7, \\ 496 &= 2^4(2^5 - 1) = 16 \cdot 31, \\ 8128 &= 2^6(2^7 - 1) = 64 \cdot 127. \end{aligned}$$

Comparing these numbers with the ones from the Mersenne primes above, a pattern seems to emerge, and we will prove it in the next proposition.

Proposition 1.6.13. *If p is a prime such that $2^p - 1$ is a prime, then $2^{p-1}(2^p - 1)$ is perfect.*

Proof: Because $2^p - 1$ is a prime, Proposition 1.6.7 implies that the proper divisors of $2^{p-1}(2^p - 1)$ are $1 = 2^0, 2^1, \dots, 2^{p-1}$ and $2^p - 1, 2(2^p - 1), \dots, 2^{p-2}(2^p - 1)$. Using the

geometric progressions summation formula, we can calculate their sum:

$$\begin{aligned} (1 + 2^1 + \cdots + 2^{p-1}) + [(2^p - 1) + 2(2^p - 1) + \cdots + 2^{p-2}(2^p - 1)] &= \\ 2^p - 1 + (2^{p-1} - 1)(2^p - 1) &= \\ = 2^{p-1}(2^p - 1). \end{aligned}$$

This shows that $2^{p-1}(2^p - 1)$ is perfect. ■

Euler proved the converse result that if an even number is perfect, then it must have this form $2^{p-1}(2^p - 1)$, where p is a prime and $2^p - 1$ is a Mersenne prime. Many people believe that there are no odd perfect numbers, but nobody has been able to prove such a statement.

Pierre de Fermat (1607–1665) was a French lawyer and mathematician who is the most famous due to his associations to what we know these days as Fermat's little theorem and Fermat's last theorem in number theory. In 1640, Fermat wrote to Marin Mersenne stating that he believed that the numbers in the sequence:

$$3, 5, 17, 257, 65537, \dots$$

are primes (see also Exercise 1.6.6). This is interpreted in modern times as Fermat conjecturing that every number of the form

$$2^{2^n} + 1$$

is a prime for $n \in \mathbb{N}_0$. Indeed, for $0 \leq n \leq 4$, we have that

$$\begin{aligned} 3 &= 2^{2^0} + 1 \text{ prime} \\ 5 &= 2^{2^1} + 1 \text{ prime} \\ 17 &= 2^{2^2} + 1 \text{ prime} \\ 257 &= 2^{2^3} + 1 \text{ prime} \\ 65537 &= 2^{2^4} + 1 \text{ prime}. \end{aligned}$$

What happens at $n = 5$? What can one say about $2^{2^5} + 1$? In 1732, Euler proved that

$$2^{2^5} + 1 = 2^{32} + 1 = 4294967297 = 641 \cdot 6700417.$$

Euler's result was not pure luck, and it was based on some modular arithmetic we will discuss in the next chapter. Fermat's reputation was perhaps affected, but nevertheless, a number of this form $2^{2^n} + 1$ is called a **Fermat number** these days. Amazingly, other than the five primes above, there are no other Fermat primes known.

The Fermat numbers form the sequence A000215 <https://oeis.org/A000215> in the *Online Encyclopedia of Integer Sequences*, <https://oeis.org/A000215>. They can be used to give another proof of the fact that there are infinitely many primes (see Exercise 1.8.16).

There are still questions posed hundreds of years ago for which no proofs have been found as of today. In 1742, the Prussian mathematician Christian Goldbach (1670–1764) wrote to Leonhard Euler (1707–1783) and proposed several conjectures regarding

primes. From their correspondence, the following emerged as the

Conjecture 1.6.14 (Goldbach's Conjecture). *Every even natural number greater than 2 is a sum of two primes.*

For examples, $4 = 2 + 2$, $6 = 3 + 3$, $8 = 3 + 5$, $10 = 5 + 5$, $12 = 5 + 7$, $14 = 7 + 7$, $16 = 5 + 11 = 3 + 13$ and so on. Actually, this conjecture has been checked for all even numbers less than $4 \cdot 10^{18}$ and is believed to be true by most people. Euler also thought it to be true, but no one has been able to prove it yet.

On the other hand, it has been observed that one can find primes in arithmetic progressions such as 3, 5, 7 or 5, 11, 17, 23, 29 for example. The reader should try to find other examples using the table of primes given earlier. It has been conjectured, perhaps as early as 1770, that there should be arbitrarily long arithmetic progressions whose terms are primes. This was proved in this century by Ben Green and Terry Tao⁹.

Theorem 1.6.15 (Green and Tao 2008). *The primes contain arbitrarily long arithmetic progressions.*

Another problem with a long history involving the primes is the Twin Prime Conjecture.

Definition 1.6.6. Pairs of primes of the form p and $p + 2$ are called **twin primes**.

This is as close as two primes could get since except for 2 and 3, there cannot be any consecutive primes (why?). Checking small numbers, one can observe some twin primes¹⁰:

$$(3, 5), (5, 7), (11, 13), (17, 19), (29, 31), (41, 43), (59, 61).$$

In 1849, the French mathematician Alphonse de Polignac (1826–1863) conjectured that there are infinitely many twin primes.

Conjecture 1.6.16 (Twin Prime Conjecture). *There are infinitely many primes p such that $p + 2$ is also a prime.*

This conjecture is still open, but in 2013, the Chinese-born American mathematician Yitang Zhang proved that there are infinitely many pairs of primes whose difference is at most 70 million¹¹. Later on, this gap was reduced significantly to 600 (see [20]), but as of today, the Twin Prime Conjecture is still open (compare Remark 5.6.2 for a surprising property of twin primes). Mathematics is an active field of research where new results are proved every day and progress on difficult questions, some posed hundreds of years ago, is made every year.

Exercise 1.6.1. Use the Euclidean algorithm to find $\gcd(324, 36)$, $\gcd(987, 1597)$ and $\gcd(2020, 1246)$.

Exercise 1.6.2. Let $a \geq 2$ be a natural number. If $a = p_1^{e_1} \dots p_s^{e_s}$ is the factorization of a into primes, where p_1, \dots, p_s are distinct primes and $e_1, \dots, e_s \in \mathbb{N}$, then prove that the number of divisors of a equals $(e_1 + 1) \dots (e_s + 1)$. What number between 1 and 100 has the largest number of divisors? How about between 1 and 200?

⁹Their paper appeared in *Annals of Mathematics*, one of the most prestigious mathematics journals, in 2008; see [12].

¹⁰Look at Table 1.6.2 to find more pairs of primes.

¹¹This result was published in *Annals of Mathematics* in 2014; see [33].

Exercise 1.6.3. Let $a, b \geq 2$ be two natural numbers. The **least common multiple** $\text{lcm}(a, b)$ is the smallest natural number that is divisible by a and b . If $a = p_1^{e_1} \dots p_s^{e_s}$ and $b = p_1^{e'_1} \dots p_s^{e'_s}$ are factorizations of a and b where p_1, \dots, p_s are distinct primes and $e_1, \dots, e_s, e'_1, \dots, e'_s \geq 0$, then prove that

$$\gcd(a, b) = p_1^{\min(e_1, e'_1)} \dots p_s^{\min(e_s, e'_s)}$$

and

$$\text{lcm}(a, b) = p_1^{\max(e_1, e'_1)} \dots p_s^{\max(e_s, e'_s)}.$$

Exercise 1.6.4. Let a and b be two natural numbers. Prove that

$$\gcd(a, b) \cdot \text{lcm}(a, b) = ab.$$

Exercise 1.6.5. Let a and $n \geq 2$ be two natural numbers such that $a^n - 1$ is a prime. Prove that $a = 2$ and n is a prime.

Exercise 1.6.6. Let $n \geq 2$ be a natural number such that $2^n + 1$ is a prime. Prove that n is a power of 2.

Exercise 1.6.7. Let a and b be two natural numbers such that $\gcd(a, b) = 1$. If n is a natural number such that $a \mid n$ and $b \mid n$, show that $ab \mid n$.

Exercise 1.6.8. Let $n \geq 3$ be a natural number. Show that the sequence of $n - 1$ consecutive numbers starting with $n! + 2$ and ending $n! + n$ does not contain any prime. This shows that there are sequences of consecutive numbers of any length that do not contain primes.

Exercise 1.6.9. Let $(F_n)_{n \geq 0}$ denote the sequence of Fibonacci numbers.

- (1) If $n \in \mathbb{N}_0$, prove that $\gcd(F_n, F_{n+1}) = 1$.
- (2) If m and n are natural numbers, show that $F_{m+n} = F_m F_{n+1} + F_{m-1} F_n$.
- (3) If m and n are natural numbers, prove that $\gcd(F_m, F_n) = F_{\gcd(m, n)}$.

Exercise 1.6.10. Let n be a composite number. Show that there exists a prime number p such that $p \mid n$ and $p^2 \leq n$.

1.7. Counting and Binomial Formula

We all start counting at an early age, and we count toys, fingers, fruits, and so on. What exactly do we mean when we say that we have 5 apples on the table? This means that we can *label* the apples as a_1, a_2, a_3, a_4, a_5 so that we create a one to one correspondence between the numbers 1, 2, 3, 4, 5 and our apples with the property that each apple has a unique label/number. That is how one can define the cardinality or the size of a finite set in general.

Definition 1.7.1. We say that a set A is **finite** if either $A = \emptyset$ or if there exists a natural number n such that the elements of A can be labeled/listed as a_1, \dots, a_n such that each element of A has a unique label/number between 1 and n . The **cardinality** or the **size** of A is defined as 0 if $A = \emptyset$ or n otherwise. We denote the cardinality of a finite set A by $|A|$.

If we have 5 apples and 4 oranges, then obviously we have $5 + 4 = 9$ fruits. This is the most basic idea of counting which states that when we put together disjoint collections of objects, the total number of objects will equal the sum of the number of objects in each collection. It is sometimes referred to as the **addition principle**.

Proposition 1.7.1. *Let k be a natural number. If A_1, \dots, A_k are pairwise disjoint finite sets, then*

$$|A_1 \cup \dots \cup A_k| = |A_1| + \dots + |A_k|.$$

Proof: We give a proof by induction on k . For the base case $k = 1$, the statement above is $|A_1| = |A_1|$ which is true. Let us give the proof of the case $k = 2$ as it contains the key idea of the general statement. Let A_1 and A_2 be two disjoint finite sets. If either set is empty, then clearly $|A_1 \cup A_2| = |A_1| + |A_2|$. Otherwise, if both sets are nonempty, then let $|A_1| = m$ and $|A_2| = n$. Label elements of A_1 as x_1, \dots, x_m and the elements of A_2 as y_1, \dots, y_n . Because A_1 and A_2 are disjoint, $x_i \neq y_j$ for any $1 \leq i \leq m$ and $1 \leq j \leq n$. Label the elements of $A_1 \cup A_2$ as follows:

$$z_k = \begin{cases} x_k & \text{if } 1 \leq k \leq m \\ y_{k-m} & \text{if } m+1 \leq k \leq m+n. \end{cases}$$

Then $A_1 \cup A_2 = \{x_1, \dots, x_m, y_1, \dots, y_n\} = \{z_1, \dots, z_{m+n}\}$ and $|A_1 \cup A_2| = m + n$.

Let $k \geq 3$ and assume that our statement above is true for any pairwise disjoint finite sets B_1, \dots, B_{k-1} . Let A_1, \dots, A_k be k pairwise disjoint finite sets. Consider the $k-1$ sets: $A_1, \dots, A_{k-2}, A_{k-1} \cup A_k$. Because A_1, \dots, A_k are pairwise disjoint, it is not too hard to show that $A_1, \dots, A_{k-2}, A_{k-1} \cup A_k$ are pairwise disjoint. We leave this as an exercise. Applying the induction hypothesis, we get that

$$\begin{aligned} |A_1 \cup \dots \cup A_{k-2} \cup (A_{k-1} \cup A_k)| &= |A_1| + \dots + |A_{k-2}| + |A_{k-1} \cup A_k| \\ &= |A_1| + \dots + |A_{k-2}| + |A_{k-1}| + |A_k|. \end{aligned}$$

This finishes our proof. ■

Let us apply this result.

Example 1.7.1. If we decide to participate in the end of the year student convocation, we will have to walk in line to enter the stadium. If there are 2 professors and 8 students, one might wonder in how many ways can we arrange these 10 people in a line? If there are no restrictions about having faculty in front of the line, then there are 10 choices for the person who will be in front of the line, 9 choices for the second in line, and so on with 2 choices for the 9th person in the line and one choice for the last person in line. This gives us $10 \cdot 9 \dots \cdot 1 = 3628800$ choices.

This simple example serves as a preamble for our next result which tells us the number of permutations of a finite set. A permutation of a set A is a bijective function from A to A . In particular, a permutation of order n is a rearrangement of the elements of $[n] = \{1, \dots, n\}$. For example, when $n = 2$, there are two permutations (12) and (21) and for $n = 3$, there are 6 permutations:

$$(123), (132), (213), (231), (312), (321).$$

We denote by S_n the set of all permutations of order n . In order to formulate the next result we recall the following definition.

Definition 1.7.2. Given a nonnegative integer n we define $n!$ (called n factorial) as follows

$$0! := 1 \quad \text{and} \quad n! := 1 \cdot 2 \cdot 3 \cdots (n-1) \cdot n, \quad n \geq 1.$$

Theorem 1.7.2. *The number of permutations of a set with $n \geq 1$ elements is $n!$.*

$$(|A| = n) \Rightarrow (|\{\text{All permutations on } A\}| = n!) \Rightarrow (|S_n| = n!)$$

Proof: One can give a direct proof of this result following the ideas of the paragraph at the beginning of this section. For a permutation of $\{1, \dots, n\}$, we have n choices for what entry will go on the first position. Once the first position is filled, we cannot use that entry anymore and therefore, we have $n - 1$ choices for what entry goes in the second position. By repeating this argument, we argue that there will be $n - 2$ choices for what is the third position and so on. Finally, after we filled out the first $n - 1$ entries of our permutation, we have one number left and one position left to fill, so we will have one choice for this to happen. Hence, the total number of permutations will be $n \cdot (n-1) \cdots 1 = n!$.

One can also give a proof by induction on n . For the base case $n = 1$, it is clear that there is one permutation of the set with one element. For the induction step, let $n \geq 1$ and assume that the number of permutations of the set $\{1, \dots, n\}$ equals $n!$. We will prove that the number of permutations of the set $\{1, \dots, n, n+1\}$. To count the permutations of $\{1, \dots, n, n+1\}$, we break them down according to the position of $n+1$. For $1 \leq j \leq n+1$, denote by A_j the set of permutations of $\{1, \dots, n, n+1\}$ that have $n+1$ in the position j . So permutations in A_1 have $n+1$ in the first position and those in A_{n+1} have $n+1$ in the $(n+1)-th$ position. When $n = 2$, for example, A_1 would consist of (312) and (321) , A_2 would have (132) and (231) and A_3 would be formed by (123) and (213) . For each $1 \leq j \leq n+1$, A_j consists of all permutations whose j -th position is fixed (it is $n+1$), but the remaining n positions can be filled with any permutation of $\{1, 2, \dots, n\}$. By the induction hypothesis, $|A_j| = n!$. For $1 \leq j \neq \ell \leq n+1$, A_j and A_ℓ are disjoint because a permutation cannot have $n+1$ in both j -th and ℓ -th position. It should also be clear that the family of permutations of $\{1, \dots, n, n+1\}$ equals $A_1 \cup \dots \cup A_{n+1}$. Using Proposition 1.7.1, the number of permutations of $\{1, \dots, n, n+1\}$ equals $|A_1 \cup \dots \cup A_{n+1}| = |A_1| + \dots + |A_{n+1}| = (n+1) \cdot n! = (n+1)!$. This finishes our induction proof. ■

It is also useful to see how fast $n!$ grows with n , at least for some small values of n . Note that for $n \geq 2$, $n! = n \cdot (n-1) \cdot \dots \cdot 1 = n \cdot (n-1)!$. It is convenient to define $0! = 1$ which extends the identity $n! = n \cdot (n-1)!$ to any natural number n . This identity may be used to calculate the values of $n!$ recursively. While $2^{10} = 1024$, $10! = 3628800$ and $12! = 479001600$. The previous identity also explains the rapid growth of $n!$ and the curious reader may wish to prove that $n!$ grows faster than any exponential of fixed basis (see assertions (5) and (6) in Proposition 5.2.9 below).

Another basic idea of counting is that if we have 3 pairs of pants and 6 shirts, then we have $3 \cdot 6 = 18$ combinations (pants, shirt). Going back to our example, what

happens if we wish to have faculty march in front, then we have 2 possible orders for the professors and $8! = 40320$ possible orderings of the 8 students behind them. Since we can pair up any permutations of the professors with any permutation of the students, we have $2 \cdot 40320 = 80640$ possible choices this time. The underlying general result used here is sometimes called the **multiplication principle**.

Proposition 1.7.3. *Let A_1, \dots, A_k be finite sets. Then*

$$|A_1 \times \dots \times A_k| = |A_1| \times \dots \times |A_k|.$$

Proof: As in Proposition 1.7.1 the proof is done by induction on k , the number of sets. For the base case $k = 1$, the statement is $|A_1| = |A_1|$ which is clearly true.

Although not necessary, we also give the proof for the case $k = 2$. So, let A_1 and A_2 be two finite sets. Our objective is to show that

$$(1.7.1) \quad |A_1 \times A_2| = |A_1| \times |A_2|.$$

If one of these two sets is empty, then so is $A_1 \times A_2$, and (1.7.1) is true by trivial reason.

Thus, we may assume that A_1 as well as A_2 are nonempty. Say,

$$A_1 = \{x_1, \dots, x_m\} \quad \text{and} \quad A_2 = \{y_1, \dots, y_n\}.$$

Then their Cartesian product consists of ordered pairs (x_i, y_j) with $1 \leq i \leq m$ and $1 \leq j \leq n$ or, if we enumerate these pairs according to their first coordinate,

$$A_1 \times A_2 = \left\{ \begin{array}{c} (x_1, y_1), \dots, (x_1, y_n) \\ (x_2, y_1), \dots, (x_2, y_n) \\ \vdots \quad \dots \quad \vdots \\ (x_m, y_1), \dots, (x_m, y_n) \end{array} \right\}.$$

Enumerating these elements line by line, it follows that there are $m \times n$ of them, hence by $|A_1| = m$ and $|A_2| = n$ equation (1.7.1) is true.

Let us now treat the general case. The induction hypothesis is that for any k finite sets we have

$$(1.7.2) \quad |A_1 \times \dots \times A_k| = |A_1| \times \dots \times |A_k|.$$

Our goal is to prove that then this is also valid in the case of $k + 1$ finite sets. So let A_1, \dots, A_{k+1} be given. Setting

$$A = A_1 \times \dots \times A_k$$

it follows (compare formula (A.2.3) in the Appendix) that

$$A_1 \times \dots \times A_{k+1} = A \times A_{k+1},$$

which by an application of (1.7.1) and of (1.7.2) implies

$$|A_1 \times \dots \times A_{k+1}| = |A \times A_{k+1}| = |A| \times |A_{k+1}| = |A_1| \times \dots \times |A_k| \times |A_{k+1}|.$$

Consequently, if (1.7.1) is true for k finite sets, then it is also satisfied in the case of $k + 1$ ones. This completes the proof by induction. ■

We give a simple and important application of this principle.

Proposition 1.7.4. *The number of subsets of a set with n elements equals 2^n .*

Proof: If $n = 0$, then our set is \emptyset and it has exactly $1 = 2^0$ subset, namely itself. Assume that $n \geq 1$. We will show that the number of subsets of $\{1, \dots, n\}$ equals 2^n . The argument will be the same for any other set with n elements. The key observation is a subset A of $\{1, \dots, n\}$ is precisely determined by the answers to the following n questions:

$$1 \in A? \text{ answer: yes or no}$$

...

$$n \in A? \text{ answer: yes or no.}$$

If we label *yes* by 1 and *no* by 0, then the answers for a set A create a vector or sequence of length n called the characteristic vector of A and denoted by $\mathbf{1}_A$. This is defined as follows:

$$\mathbf{1}_A(a) = \begin{cases} 1 & : a \in A \\ 0 & : a \notin A. \end{cases}$$

For example, when $n = 4$ and $A = \{1, 4\}$, then $\mathbf{1}_A = (1, 0, 0, 1)$. Note that $A \neq B$ if and only if $\mathbf{1}_A \neq \mathbf{1}_B$. Also, for every 0, 1-vector y of length n , there is a unique subset A such that $\mathbf{1}_A = y$. Hence, there is a bijective correspondence between the subsets of $\{1, \dots, n\}$ and the Cartesian product

$$\underbrace{\{0, 1\} \times \{0, 1\} \times \dots \times \{0, 1\}}_{n \text{ factors}}.$$

By the Proposition 1.7.3,

$$|\underbrace{\{0, 1\} \times \{0, 1\} \times \dots \times \{0, 1\}}_{n \text{ factors}}| = |\underbrace{\{0, 1\}| \cdot |\{0, 1\}| \cdot \dots \cdot |\{0, 1\}|}_{n \text{ factors}}| = 2^n.$$

■

Definition 1.7.3. For a nonnegative integer k and a set A , by a **k -subset** of A we mean a subset of A that has size k .

Definition 1.7.4. For integers n, k , let $\binom{n}{k}$ denote the number of k -subsets of $[n]$.

$$(|A| = n) \Rightarrow \left[\binom{n}{k} := |\{B \subseteq A : |B| = k\}| \right].$$

Remark 1.7.1. The number $\binom{n}{k}$ is an example of a *binomial coefficient*, and we can refer to it as n choose k . The name n choose k should be pretty clear since $\binom{n}{k}$ counts the number of ways of choosing k objects from a set with n elements. When $n < k$ or when $k < 0$, then $\binom{n}{k} = 0$. When $k = 0$, it is easy to see that $\binom{n}{0} = 1$ as \emptyset is the only subset of $[n]$ whose cardinality is 0. When $k = 1$, it is also easy to observe that $\binom{n}{1} = n$ as there are exactly n subsets of $[n]$ that have cardinality 1, namely $\{1\}, \dots, \{n\}$.

We give now a general formula for $\binom{n}{k}$.

Proposition 1.7.5. *Let $n \geq k \geq 0$ be integers. Then*

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

Proof: To select a k -subset of $[n]$, note that we have n choices for its first element, $n - 1$ choices for its 2nd, and so on, $n - k + 1$ choices for its k -th element. However, in the count of $n(n - 1)\dots(n - k + 1)$ every k -subset will appear exactly $k!$ times corresponding to all the possible $k!$ permutations of that subset. Thus, the number of k -subsets will equal

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

as promised. Since

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

the second formula follows immediately. ■

It is likely that many people have seen or heard of the Pascal's triangle. This is an arrangement of the numbers of the form $\binom{n}{k}$ constructed as follows: the first row is $\binom{0}{0} = 1$, the second row is $\binom{1}{0} = 1, \binom{1}{1} = 1$, the third row is $\binom{2}{0} = 1, \binom{2}{1} = 2, \binom{2}{2} = 1$ and so on with the $(n + 1)$ -th row having $n + 1$ entries:

$$\binom{n}{0}, \binom{n}{1}, \dots, \binom{n}{n-1}, \binom{n}{n}.$$

We compute below the first seven rows of Pascal's triangle:

| | | | | | | | |
|---------|---|--|--|--|--|--|--|
| $n = 0$ | 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 | | | | | | |
| $n = 6$ | 1 6 15 20 15 6 1 | | | | | | |

Blaise Pascal (1623–1662) was a French mathematician with important contributions in geometry and probability. Pascal's triangle was known before Pascal's time in the work of Indian mathematicians such as Pingala (around 2nd century BCE), Persian mathematicians such as Omar Khayyam (1048–1131) or Chinese mathematicians such as Yang Hui (1238–1298).

The term **binomial coefficient** is explained by the following result.

Theorem 1.7.6 (Binomial Theorem). *Let n be a natural number. If a and b are real numbers, then*

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

Proof: Expanding $(a + b)^n$ as

$$\underbrace{(a + b) \dots (a + b)}_{n \text{ factors}}$$

observe that the number of times the term $a^k b^{n-k}$ will appear in the expansion of $(a + b)^n$ equals the number of ways of choosing exactly k brackets/factors $a + b$ from which we select a (and leaving the remaining $n - k$ brackets/factors $a + b$ from which we pick b). This number of choices equals $\binom{n}{k}$ by definition. The possible values of k are in the range $0, 1, \dots, n$ and therefore $(a + b)^n = \sum_{k=0}^n \binom{n}{k} a^k b^{n-k}$. ■

Note that the binomial theorem holds over more general sets of numbers with essentially the same proof as above. The binomial coefficients have many beautiful properties. We list a few basic ones below.

Proposition 1.7.7. *If $n \geq k$ are two natural numbers, then*

$$(1.7.3) \quad \binom{n}{k} = \binom{n}{n-k} \quad \text{and} \quad \binom{n}{k} = \binom{n-1}{k} + \binom{n-1}{k-1}.$$

Proof: The number of k -subsets of $[n]$ equals the number of $(n - k)$ -subsets of $[n]$. To see this, consider the function f from the collection to k -subsets of $[n]$ to the family of $(n - k)$ -subsets of $[n]$ which maps each k -subset A to its complement $[n] \setminus A$. This function is bijective, and we give a short proof here. To show that f is injective, assume that A and B are k -subsets such that $f(A) = f(B)$. Using the definition of f , this means that $[n] \setminus A = [n] \setminus B$. Taking complements of both sides gives us that $A = B$. Hence, f is injective. To show that f is surjective, consider a $(n - k)$ -subset C of $[n]$. Define D to be the complement of C in $[n]$. Therefore, D is a k -subset of $[n]$ and $f(D) = C$. This shows that f is surjective. Because f is bijective, we deduce that the number of k -subsets of $[n]$ equals the number of $(n - k)$ -subsets of $[n]$ which is the same as $\binom{n}{k} = \binom{n}{n-k}$.

The second identity can be verified algebraically in a straightforward manner, and we leave this as an exercise. There is another and more insightful proof of this identity which can be obtained by showing that the two sides of the identity count the same quantity in different ways. Clearly, $\binom{n}{k}$ counts the number of k -subsets of $[n]$. The k -subsets of $[n]$ can be split into the k -subsets not containing n and the k -subsets containing n . Obviously, there is no overlap between these groups. A k -subset of $[n]$ that does not contain n is actually a k -subset of $[n - 1]$ and the number of such k -subsets is $\binom{n-1}{k}$. A k -subset of $[n]$ that contains n must be of the form $\{n\} \cup A$, where A is a $(k - 1)$ -subset of $[n - 1]$. Therefore, the number of such k -subsets is $\binom{n-1}{k-1}$ leading to the desired conclusion. ■

Remark 1.7.2. The first equation in (1.7.3) tells us that each row of the Pascal triangle is a palindrome. The second equation in (1.7.3) indicates that each binomial coefficient is the sum of the two coefficients directly above it in the Pascal triangle.

The binomial formula can be used to obtain results involving binomial coefficients.

Example 1.7.2. What is the coefficient of x^2 in $(3x^2 - 2x^{-1})^7$ where $x \neq 0$ is some variable? Applying the binomial formula with $a = 3x^2$, $b = -2x^{-1}$ and $n = 7$ gives that

$$\begin{aligned}(3x^2 - 2x^{-1})^7 &= \sum_{k=0}^7 \binom{7}{k} (3x^2)^k (-2x^{-1})^{7-k} \\&= \sum_{k=0}^7 \binom{7}{k} 3^k (-2)^{7-k} x^{2k+k-7} \\&= \sum_{k=0}^7 \binom{7}{k} 3^k (-2)^{7-k} x^{3k-7}.\end{aligned}$$

To find the coefficient of x^2 , we first have to identify when $x^{3k-7} = x^2$ happens. In this case, this occurs when $3k - 7 = 2$ meaning that $k = 3$. Therefore, the coefficient of x^2 will be $\binom{7}{3} \cdot 3^3 \cdot (-2)^{7-3} = 35 \cdot 27 \cdot 16 = 15120$.

Example 1.7.3. Consider the identity

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

which is true for any x . If we use the binomial formula on the right-hand side, we obtain that the coefficient of x^n is $\binom{2n}{n}$. On the other hand, the left-hand side can be expanded as follows:

$$(1+x)^n(1+x)^n = \sum_{k=0}^n \binom{n}{k} x^k \times \sum_{\ell=0}^n \binom{n}{\ell} x^\ell.$$

To get an x^n in the product above, an x^k from the first sum must meet an x^ℓ in the second sum, where $\ell = n-k$. This can happen for any k between 0 and n and therefore, the coefficient of x^n will be

$$\sum_{k=0}^n \binom{n}{k} \cdot \binom{n}{n-k} = \sum_{k=0}^n \binom{n}{k}^2,$$

since $\binom{n}{k} = \binom{n}{n-k}$ for any $0 \leq k \leq n$. On the other hand, the coefficient of x^n in $(1+x)^{2n}$ equals $\binom{2n}{n}$. Therefore,

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

Earlier in this section, we saw that if A_1, \dots, A_n are pairwise disjoint subsets, then

$$|A_1 \cup \dots \cup A_n| = |A_1| + \dots + |A_n|.$$

One may ask what is the relation between $|A_1 \cup \dots \cup A_n|$ and $|A_1| + \dots + |A_n|$ in general, when the sets A_1, \dots, A_n may intersect? A bit of experimentation with Venn diagrams

and a few sets leads us to deduce that for any finite sets A_1, A_2, A_3 :

$$\begin{aligned} |A_1 \cup A_2| &= |A_1| + |A_2| - |A_1 \cap A_2| \\ |A_1 \cup A_2 \cup A_3| &= |A_1| + |A_2| + |A_3| \\ &\quad - |A_1 \cap A_2| - |A_1 \cap A_3| - |A_2 \cap A_3| + |A_1 \cap A_2 \cap A_3|. \end{aligned}$$

The general results is stated below and is called the principle of inclusion and exclusion.

Theorem 1.7.8 (Principle of Inclusion and Exclusion). *Let A_1, \dots, A_n be finite subsets of a finite set A . For $J \subset [n]$, denote $A_J = \bigcap_{j \in J} A_j$ with $A_\emptyset = A$. Then*

$$|A \setminus (A_1 \cup \dots \cup A_n)| = \sum_{J \subset [n]} (-1)^{|J|} |A_J|$$

and equivalently,

$$(1.7.4) \quad |A_1 \cup \dots \cup A_n| = \sum_{\substack{J \subset [n] \\ J \neq \emptyset}} (-1)^{|J|-1} |A_J|.$$

Proof: If x is an element of A not contained in any of A_1, \dots, A_n , then the contribution of x to both sides of the first identity we are trying to prove is the same, namely 1. If x is an element of $A_1 \cup \dots \cup A_n$, then let us assume without any loss of generality that x is contained in A_1, \dots, A_k and not contained in A_{k+1}, \dots, A_n for some integer k with $1 \leq k \leq n$. This means that for any $J \subset [n]$, $x \in A_J$ if and only if $J \subset [k]$. This means that the contribution of x to the right-hand side equals $\sum_{J \subset [k]} (-1)^{|J|} = \sum_{j=0}^k (-1)^j \binom{k}{j} = (1-1)^k = 0$. This proves the first identity. The second identity follows from $|A_1 \cup \dots \cup A_n| = |A| - |A \setminus (A_1 \cup \dots \cup A_n)|$.

Alternative proof of (1.7.4): Recall that for a subset X of A , we can define its characteristic vector or function $\mathbf{1}_X$ by

$$\mathbf{1}_X(a) = \begin{cases} 1 & : a \in X \\ 0 & : a \notin X. \end{cases}$$

The basic properties of the characteristic function are that for any $X, Y \subset A$ and any $a \in A$,

$$(1.7.5) \quad \mathbf{1}_X(a) \cdot \mathbf{1}_Y(a) = \mathbf{1}_{X \cap Y}(a), \quad a \in A$$

and

$$\sum_{a \in A} \mathbf{1}_X(a) = |X|.$$

We claim now that

$$(1.7.6) \quad \mathbf{1}_{A_1 \cup \dots \cup A_n}(a) = 1 - \prod_{j=1}^n (1 - \mathbf{1}_{A_j}(a)), \quad a \in A.$$

The argument is similar to the one in the previous proof.

If $a \in A_1 \cup \dots \cup A_n$, then the left-hand side equals 1. Also, there is at least one ℓ between 1 and n such that $a \in A_\ell$ and therefore $1 - \mathbf{1}_{A_\ell}(a) = 0$, meaning that the right-hand side equals 1 as well. If $a \notin A_1 \cup \dots \cup A_n$, then the left-hand side equals 0. In

this case, $a \notin A_j$ for any $1 \leq j \leq n$ and therefore, $\prod_{j=1}^n (1 - \mathbf{1}_{A_j}(a)) = 1$. This means that the right-hand side equals 0, thus proves (1.7.6).

Let us evaluate the right-hand side of (1.7.6) differently. From (1.7.5), we can show by induction on $|J|$ that $\prod_{j \in J} \mathbf{1}_{A_j}(a) = \mathbf{1}_{A_J}(a), \forall a \in A$ (we leave the details as an exercise). Applying this identity, elementary calculations (use formula (1.7.10) in Exercise 1.7.12 with $\alpha_j = -\mathbf{1}_{A_j}(a)$) for any $a \in A$,

$$\begin{aligned} 1 - \prod_{j=1}^n (1 - \mathbf{1}_{A_j}(a)) &= 1 - [(1 - \mathbf{1}_{A_1}(a)) \cdots (1 - \mathbf{1}_{A_n}(a))] \\ &= 1 - \left(1 + \sum_{\substack{J \subset [n] \\ J \neq \emptyset}} \prod_{j \in J} (-\mathbf{1}_{A_j}(a))\right) = \sum_{\substack{J \subset [n] \\ J \neq \emptyset}} (-1)^{|J|-1} \mathbf{1}_{A_J}(a). \end{aligned}$$

Combining this with (1.7.6) we arrive at

$$(1.7.7) \quad \mathbf{1}_{A_1 \cup \dots \cup A_n}(a) = \sum_{\substack{J \subset [n] \\ J \neq \emptyset}} (-1)^{|J|-1} \mathbf{1}_{A_J}(a).$$

We may now sum up (1.7.7) over all $a \in A$ and obtain

$$\begin{aligned} |A_1 \cup \dots \cup A_n| &= \sum_{a \in A} \mathbf{1}_{A_1 \cup \dots \cup A_n}(a) = \sum_{a \in A} \sum_{\substack{J \subset [n] \\ J \neq \emptyset}} (-1)^{|J|-1} \mathbf{1}_{A_J}(a) \\ &= \sum_{\substack{J \subset [n] \\ J \neq \emptyset}} (-1)^{|J|-1} \sum_{a \in A_J} \mathbf{1}_{A_J}(a) = \sum_{\substack{J \subset [n] \\ J \neq \emptyset}} (-1)^{|J|-1} |A_J|, \end{aligned}$$

which completes the proof of (1.7.4). ■

Example 1.7.4. We want to know how many numbers in $\{1, \dots, 1000\}$ are not divisible by the numbers 3, 5 and 7. Those numbers are for example 1, 2, 4, 8, 11,

To answer this question we calculate the cardinality of the complementary set, which consists of those numbers which are divided by at least one of the three numbers 3, 5 and by 7. Thus, if

$$A_1 = \{n \leq 1000 : 3 \mid n\}, \quad A_2 = \{n \leq 1000 : 5 \mid n\}, \quad A_3 = \{n \leq 1000 : 7 \mid n\},$$

then the union of these three sets is exactly the complement of the set we are interested in.

Easy counting leads to $|A_1| = 333$, $|A_2| = 200$ and $|A_3| = 142$. Moreover, numbers in $A_1 \cap A_2$ are exactly those divisible by 15, those in $A_1 \cap A_3$ which are divisible by 21, those in $A_2 \cap A_3$ which are divisible by 35 and, finally, those in $A_1 \cap A_2 \cap A_3$ which are divisible by 105. This leads to

$$|A_1 \cap A_2| = 66, \quad |A_1 \cap A_3| = 47, \quad |A_2 \cap A_3| = 28 \text{ and } |A_1 \cap A_2 \cap A_3| = 9.$$

Hence,

$$|A_1 \cup A_2 \cup A_3| = 333 + 200 + 142 - 66 - 47 - 28 + 9 = 543,$$

which tells us that among the integers between 1 and 1000 there exist exactly $1000 - 543 = 457$ which are not divisible by 3, 5, and by 7.

Example 1.7.5. How many ways exist to order the numbers from 1 to n so that at least one number k is left at its original k -th place? More precisely, evaluate $|S_{\text{inv}}|$ where

$$(1.7.8) \quad S_{\text{inv}} := \{\pi \in S_n : \exists k \in \{1, \dots, n\}, \pi(k) = k\}.$$

To answer this question, we introduce sets

$$A_k := \{\pi \in S_n : \pi(k) = k\}, \quad 1 \leq k \leq n.$$

Then, by (1.7.4), it follows that

$$(1.7.9) \quad |S_{\text{inv}}| = |A_1 \cup \dots \cup A_n| = \sum_{J \subseteq [n]} (-1)^{|J|-1} |A_J| = \sum_{j=1}^n \sum_{|J|=j} (-1)^{j-1} |A_J|.$$

Observe that permutations belonging to A_J are exactly those which leave the numbers in J invariant. Thus, if $|J| = j$, there are exactly $(n-j)!$ those permutations. Moreover, since there are $\binom{n}{j}$ subsets $J \subseteq [n]$ with cardinality j , formula (1.7.9) implies

$$|S_{\text{inv}}| = \sum_{j=1}^n \sum_{|J|=j} (-1)^{j-1} (n-j)! = \sum_{j=1}^n (-1)^{j-1} \binom{n}{j} (n-j)! = n! \sum_{j=1}^n \frac{(-1)^{j-1}}{j!}.$$

It might be of interest that the proportion of the number of permutations leaving one number invariant within the set of all permutations equals

$$\frac{|S_{\text{inv}}|}{|S_n|} = \frac{n! \sum_{j=1}^n \frac{(-1)^{j-1}}{j!}}{n!} = \sum_{j=1}^n \frac{(-1)^{j-1}}{j!},$$

which is known to approximate $1 - e^{-1} \approx 0.63212$ when n grows. Here $e \approx 2.71828 \dots$ is the Euler number introduced in Definition 5.3.1.

Remark 1.7.3. Another equivalent formulation of the problem in Example 1.7.5 is as follows. At a Christmas party attend n guests. Each participant of the party brings a gift with him. Then the n parcels are distributed among the n participants. How many distributions of the parcels exist such that at least one participant gets his own gift? The result in Example 1.7.5 tells us that for large n this happens approximately 63.2% of the time.

We conclude with solving the problem discussed in Section 1.2 determining R_n which is the number of regions in which the chords drawn between n points on the circle divide the circle. Our assumption was also that no three chords intersect inside the circle.

Example 1.7.6. In the beginning, when there are no chords, there is one region. Drawing the first chord will create one new region as this chord divides the circle into two regions. Continue the process by considering the interaction of a new chord with the previously drawn ones. If a new chord intersects k old chords in the interior, it crosses $k+1$ regions of which those chords are edges and cuts each of them into two regions; if it intersects no old chords, it creates one new region; either way it adds $k+1$ to the number of regions. In the end, there will be one new region for any chord and one new region for any intersection of two chords.

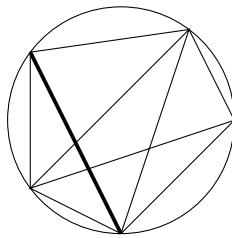


Figure 1.7.1. The bold chord creates three new regions.

As the number of chords is $\binom{n}{2}$ and the number of chord intersections is $\binom{n}{4}$ (any four points on the circle will produce one such point), this gives $1 + \binom{n}{2} + \binom{n}{4}$ regions¹².

Another solution can be obtained using Euler's formula. We create a planar graph from our configuration of n points on the circle. The vertices are the points on the circle and the intersection points of the diagonals. Two points are adjacent if they are on the same line corresponding to a side or a diagonal and there are no other points between them. Also, two points on the same side are also joined by an edge corresponding to the arc of the circle between them. An example of this graph is drawn in Figure 1.7.2. This graph is planar and has n vertices on the circle and $\binom{n}{4}$ vertices inside the circle corresponding to all possible intersections of two diagonals. Hence, the number of vertices is $n + \binom{n}{4}$. Given that each vertex on the circle has degree n and any other vertex has degree four, we deduce that there are $2\binom{n}{4} + \binom{n+1}{2}$ edges in this graph. The reader is invited to apply Euler's formula for planar graphs (Theorem 1.2.10) to get that $R_n = 1 + \binom{n}{2} + \binom{n}{4}$.

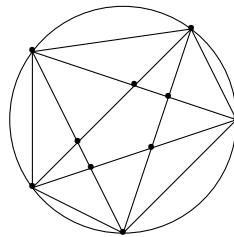


Figure 1.7.2. A planar graph created using five points on the circle.

Exercise 1.7.1. Expand Pascal's triangle by adding the rows 8 and 9. Which entry is the largest and which one is the smallest? What other patterns do you observe?

Exercise 1.7.2. Let n be a natural number.

- (1) If k is a natural number such that $k < n$, show that $\binom{2n}{k} < \binom{2n}{k+1}$.
- (2) If ℓ is a natural number such that $n \leq \ell < 2n$, prove that $\binom{2n}{\ell} > \binom{2n}{\ell+1}$.
- (3) If k is a natural number such that $k < n$, show that $\binom{2n+1}{k} < \binom{2n+1}{k+1}$.

¹²This short proof is due to Marc Noy in [26].

- (4) Prove that $\binom{2n+1}{n} = \binom{2n+1}{n+1}$.
(5) If ℓ is a natural number such that $n+1 \leq \ell < 2n+1$, show that $\binom{2n+1}{\ell} > \binom{2n+1}{\ell+1}$.

Exercise 1.7.3. Let n be a natural number. Prove that

$$\frac{2^{2n}}{2n+1} < \binom{2n}{n} < 2^{2n} \text{ and } \frac{2^{2n+1}}{2n+2} < \binom{2n+1}{n} = \binom{2n+1}{n+1} < 2^{2n+1}.$$

Exercise 1.7.4. Let $n \geq k$ be two natural numbers. Give algebraic proofs of the identities

$$\binom{n}{k} = \binom{n-1}{k} + \binom{n-1}{k-1} \text{ and } k \binom{n}{k} = n \binom{n-1}{k-1}.$$

Exercise 1.7.5. Let n be a natural number. Give two proofs for the identity:

$$\binom{n}{0} + \binom{n}{1} + \cdots + \binom{n}{n} = 2^n.$$

Exercise 1.7.6. Let n be a natural numbers. Give two proofs for the identity:

$$\binom{n}{0} - \binom{n}{1} + \cdots + (-1)^{n-1} \binom{n}{n} = 0.$$

Exercise 1.7.7. Let m and n be two natural numbers. Consider the walks between the point $(0, 0)$ and the point (m, n) , where each step is either an East step E: $(x, y) \rightarrow (x+1, y)$ or a North step N: $(x, y) \rightarrow (x, y+1)$. For example, when $(m, n) = (1, 1)$, there are two such walks: EN and NE. Show that the number of walks of this type from $(0, 0)$ to (m, n) equals $\binom{m+n}{m} = \binom{m+n}{n}$.

Exercise 1.7.8. Assume that one has m apples and n oranges. Choosing k fruits from these $m+n$ ones and counting in two ways, show that

$$\binom{m+n}{k} = \sum_{j=0}^k \binom{m}{j} \binom{n}{k-j}.$$

This is called the **Vandermonde's identity**¹³.

Exercise 1.7.9. Calculate the coefficient of x^k in two ways in the expansion

$$(1+x)^m(1+x)^n = (1+x)^{m+n},$$

and give another proof of Vandermonde's identity.

Exercise 1.7.10. The Fibonacci numbers $(F_n)_{n \geq 0}$ are defined recursively as follows: $F_0 = 0, F_1 = 1$ and $F_{n+1} = F_n + F_{n-1}$ for $n \geq 1$. Show that

$$F_{n+1} = \sum_{k=0}^{\lfloor n/2 \rfloor} \binom{n-k}{k},$$

for any $n \geq 1$.

Exercise 1.7.11. Let α and β be two numbers and $(x_n)_{n \geq 0}$ be a sequence of numbers defined recursively by $x_0 = \alpha, x_1 = \beta$ and $x_{n+1} = x_n + x_{n-1}$ for $n \geq 1$. Prove that $x_n = \alpha F_{n-1} + \beta F_n$, for every $n \geq 1$.

¹³This identity bears the name of the French mathematician Alexandre-Théophile Vandermonde (1735–1796).

Exercise 1.7.12. Suppose we are given n numbers $\alpha_1, \dots, \alpha_n$. Prove that

$$(1.7.10) \quad \prod_{j=1}^n (1 + \alpha_j) = 1 + \sum_{\substack{J \subset [n] \\ J \neq \emptyset}} \prod_{j \in J} \alpha_j.$$

For example, if $n = 3$, this means that

$$(1 + \alpha_1)(1 + \alpha_2)(1 + \alpha_3) = 1 + [\alpha_1 + \alpha_2 + \alpha_3] + [\alpha_1\alpha_2 + \alpha_1\alpha_3 + \alpha_2\alpha_3] + [\alpha_1\alpha_2\alpha_3].$$

Here the related nonvoid subsets J of $[3] = \{1, 2, 3\}$ are given as $\{1\}, \{2\}, \{3\}, \{1, 2\}, \{1, 3\}, \{2, 3\}$ and $\{1, 2, 3\}$.

How does formula (1.7.10) look like in the case $\alpha_1 = \dots = \alpha_n = \alpha$ and how is this formula related to the binomial formula in this case?

Use (1.7.10) to express $\prod_{j=1}^n (1 - \alpha_j)$ similarly.

Exercise 1.7.13. How many numbers in $\{1, \dots, 1000\}$ are not divisible by the numbers 7, 11, and 13?

Exercise 1.7.14. Find all permutations in S_3 and in S_4 belonging to S_{inv} defined in (1.7.8).

- Exercise 1.7.15.** (1) How many permutations π in S_n exist so that $\pi(k) \neq k$ for all $1 \leq k \leq n$?
(2) How many permutations π of order n leave exactly one number invariant? That is, the permutations π satisfy $\pi(k) = k$ for some $1 \leq k \leq n$ and $\pi(j) \neq j$ if $j \neq k$.

1.8. More Exercises

- Exercise 1.8.1.** (1) How many pairs $(q, d) \in \mathbb{N}_0 \times \mathbb{N}_0$ are there for which one has $25q + 10d \leq 100$? Compare this answer to Exercise 1.1.8.
(2) How many pairs $(q, d) \in \mathbb{N} \times \mathbb{N}$ are there such that $25q + 10d \leq 100$? Compare this answer to Exercise 1.1.9.

Exercise 1.8.2. Use strong induction to prove that in any graph, a closed odd walk contains an odd cycle.

Exercise 1.8.3. Show that a graph is bipartite if and only if it does not contain any odd cycles.

Exercise 1.8.4. Let $n \geq k$ be two natural numbers. Prove that $k!$ always divides $n(n-1)\dots(n-k+1)$.

Exercise 1.8.5. If $p \neq q$ are distinct primes and n is a natural number such that $p \mid n$ and $q \mid n$, then $pq \mid n$.

Exercise 1.8.6. If three primes greater than 10 form an arithmetic progression, show that the common difference is divisible by 6.

Exercise 1.8.7. The following is a problem from the Rhind Papyrus around 1550 BCE. How do you divide 100 loaves among 5 people so that the shares received are in an arithmetic progression such that one seventh of the sum of the largest three shares will equal the sum of the smallest two?

Exercise 1.8.8. Prove by induction that $2^n > 3n + 1$ for any integer $n \geq 4$.

Exercise 1.8.9. Let $N \geq 3$ be a fixed natural number. Consider the arithmetic progression $(a_n)_{n \geq 1}$ with initial term $a_1 = 1$ and common difference $d = N - 2$. The N -gonal numbers are given by the partial sums of this arithmetic progression:

$$s_n = a_1 + \dots + a_n.$$

- (1) Show that for $N = 3$, the 3-gonal numbers are the same as the triangular numbers appearing in Figure 1.2.2.
- (2) Show that for $N = 4$, the 4-gonal numbers are the same as the square numbers appearing in Figure 1.2.3.
- (3) When $N = 5$, what is the connection between 5-gonal numbers and pentagons?
- (4) Same question as above for $N = 6$.

Exercise 1.8.10. Let n be a natural number. What is the maximum number of regions formed by n circles in a plane? For $n = 1$, the answer is 2 (the inside and the outside of the circle), for $n = 2$, the answer is 4 and for $n = 3$ the answer is 8.

Exercise 1.8.11. Let $a > b$ be two natural numbers such that $\gcd(a, b) = 1$. Show that $\gcd(a+b, a-b) = 1$ or $\gcd(a+b, a-b) = 2$.

Exercise 1.8.12. Let p be a prime number. Show that for a natural number k with $1 \leq k \leq p-1$, p divides the binomial coefficient $\binom{p}{k}$.

Exercise 1.8.13. Show that $n^2 - n + 41$ is a prime number for any $1 \leq n \leq 40$, but not a prime when $n = 41$.

Exercise 1.8.14 (Problem A1, Putnam contest¹⁴ 1985). Determine, with proof, the number of ordered triples (A_1, A_2, A_3) of sets such that

$$A_1 \cup A_2 \cup A_3 = \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10\} \text{ and } A_1 \cap A_2 \cap A_3 = \emptyset.$$

Exercise 1.8.15. At a party, there are 10 people with white shirts and 8 people with red shirts. Four people have black shoes and white shirts. Moreover, there are 3 people wearing black shoes and red shirts. The total number of people with white or red shirts or black shoes is 21. How many people have black shoes?

Exercise 1.8.16. Recall the definition of the Fermat numbers. For $n \in \mathbb{N}_0$, denote $f_n = 2^{2^n} + 1$.

- (1) Prove that for any $n \in \mathbb{N}$, the following identity holds:

$$\prod_{k=0}^{n-1} f_k = f_n - 2.$$

- (2) Use the previous equation to show that any two distinct Fermat numbers are coprime.
- (3) Use part (2) to give another proof that there are infinitely many prime numbers.

¹⁴The William Lowell Putnam Mathematical Competition is the most famous mathematics competition for undergraduate college students in the United States and Canada. It takes place each year on the first Saturday of December and consists of two three-hour sessions, one in the morning and one in the afternoon. In each session, every student works individually on six problems.

Exercise 1.8.17. Let n be a natural number.

- (1) Prove that $(n+1)^4 - n^4 = 4n^3 + 6n^2 + 4n + 1$.
- (2) Using the previous statement, show that

$$1^3 + \dots + n^3 = \frac{n^2(n+1)^2}{4}.$$

Exercise 1.8.18. Let n be a natural number.

- (1) If $n \geq 3$, show that $\binom{3}{3} + \dots + \binom{n}{3} = \binom{n+1}{4}$.
- (2) Prove that $\binom{n}{3} = \frac{n^3}{6} - \frac{n^2}{2} + \frac{n}{3}$.
- (3) Using the previous statements, show that

$$1^3 + \dots + n^3 = \frac{n^2(n+1)^2}{4}.$$

Exercise 1.8.19. Let n be a natural number.

- (1) Prove that $(n+1)^5 - n^5 = 5n^4 + 10n^3 + 10n^2 + 5n + 1$.
- (2) Using the previous statement, show that

$$1^4 + \dots + n^4 = \frac{n(n+1)(2n+1)(3n^2+3n-1)}{30}.$$

Exercise 1.8.20. Let n be a natural number.

- (1) If $n \geq 4$, show that $\binom{4}{4} + \dots + \binom{n}{4} = \binom{n+1}{5}$.
- (2) Prove that $\binom{n}{4} = \frac{n^4}{24} - \frac{n^3}{4} + \frac{11n^2}{24} - \frac{n}{4}$.
- (3) Using the previous statements, show that

$$1^4 + \dots + n^4 = \frac{n(n+1)(2n+1)(3n^2+3n-1)}{30}.$$

Exercise 1.8.21. Let n be a natural number. Prove that

$$\frac{n}{2} < 1 + \frac{1}{2} + \frac{1}{3} + \dots + \frac{1}{2^n - 2} + \frac{1}{2^n - 1} < n.$$

Integer Numbers \mathbb{Z}

Beginning is easy. Continuing is hard.
Japanese proverb

2.1. Basic Properties

It took humanity a while to come to grips with zero and negative numbers. In ancient times, the equations considered were such that their coefficients would be natural numbers. For example, one would solve $2x + 14 = 16$ or $x^2 + 3 = 4x$ or $x^3 + 3x = 4$ instead of $2x - 2 = 0$ or $x^2 - 4x + 3 = 0$ or $x^3 + 3x - 4 = 0$. The ancient Greeks dismissed equations of the form $4x + 20 = 0$ as *absurd*. Negative numbers appeared around that time¹ in China in the book *Nine chapters on the Mathematical Art* (*Jiu zhang suan-shu*), which in its present form dates from the period of the Han Dynasty. These negative numbers were represented by black counting rods while positive numbers were represented by red counting rods. This is the opposite from what we use today where red numbers are debts and black numbers are income. Around the 7th century CE, negative numbers and rules for adding and multiplying them appear in the work of the Indian mathematician Brahmagupta with the same significance: negative numbers were called *debts* and positive numbers *fortunes*. It took until the 17th century for the negative numbers to be generally accepted by the mathematical community and not to be dismissed as *absurd*.

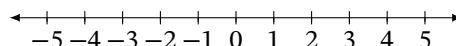


Figure 2.1.1. A graphic representation of some integer numbers.

¹It is a bit surprising that negative numbers occurred relatively late in human history, much later than fractions. Maybe merchants used them already much earlier, yet they were not considered as mathematical objects which can be added, multiplied, or ordered.

As in the previous chapter, we start with a less axiomatic description² of the set of integers as obtained by taking the union of the set of natural numbers \mathbb{N} , the number 0 and the set of negatives of natural numbers

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

The elements of these two last sets may be described as follows:

$$(\forall n \in \mathbb{N})(0 + n = n + 0 = n) \text{ and } (\forall n \in \mathbb{N})((-n) + n = n + (-n) = 0).$$

Definition 2.1.1. The set of integers \mathbb{Z} is defined as:

$$\mathbb{Z} = (-\mathbb{N}) \cup \{0\} \cup \mathbb{N}.$$

While using \mathbb{N} to denote the set of natural numbers seems obvious and natural (duh!), the curious reader may wonder why is \mathbb{Z} the notation for the set of integer numbers instead of \mathbb{I} for example. The notation \mathbb{Z} has its origins in German language where *ganze Zahl/ Ganze Zahlen* stands for *integer/ integers*. The notation seems to have been introduced³ by N. Bourbaki (pseudonym used by a group of mostly French mathematicians) around 1930 in their book *Algèbre, Chapter 1*.

Addition and multiplication are easily extended from \mathbb{N} to \mathbb{Z} . The basic properties of these operations such as associativity, commutativity as well as the distributive law remain valid. As already mentioned, these properties of \mathbb{Z} will be treated more thoroughly later in Section A.7.2. We describe them briefly below.

$$(\forall a, b \in \mathbb{Z})(a \cdot (-b) = (-a) \cdot b = -(a \cdot b) \text{ and } (-a) \cdot (-b) = a \cdot b)$$

Let us still mention three conventions. First, instead of $a + (-b)$ one may write $a - b$.

$$(\forall a, b \in \mathbb{Z})(a - b = a + (-b)).$$

In other words, $c = a - b$ is the unique integer satisfying $c + b = a$.

Second, if $a \in \mathbb{Z}$ and $m \geq 1$, define the m th power of a by

$$a^m = \underbrace{a \cdot \dots \cdot a}_{m \text{ times}}$$

and, moreover, we let $a^0 = 1$.

Third, for each $a \in \mathbb{Z}$ its absolute value is defined by

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

By this construction, $|a| \in \mathbb{N}$ for each nonzero integer a .

The natural order on \mathbb{N} is also extended to \mathbb{Z} .

$$\begin{aligned} (\forall a, b \in \mathbb{Z})[(a < b) &\Leftrightarrow (\exists n \in \mathbb{N})(b = n + a)] \quad \text{and} \\ (a \leq b) &\Leftrightarrow (a < b \text{ or } a = b). \end{aligned}$$

²For a (mathematical) exact introduction of integers we refer to Section A.7.2.

³Historically, as in the case of \mathbb{N} etc. true bold fonts have been used. But because bold font is hard to express on a blackboard or a sheet of paper, the bold font has typically been emulated by doubling one of the strokes to draw the letter, resulting in an “open bold” commonly referred to as “blackboard bold”. Some people not realizing this historical background and thinking the open bold is the correct format and some people preferring the distinction that the open bold actually provides, use \mathbb{N} , \mathbb{Z} , \mathbb{Q} , \mathbb{R} , and \mathbb{C} , respectively, instead for formal publication.

In this way

$$\dots < -3 < -2 < -1 < 0 < 1 < 2 < 3 < \dots.$$

This order is compatible with the algebraic operations in the following sense:

$$(\forall a, b, c \in \mathbb{Z})[(a \leq b) \Leftrightarrow (a + c \leq b + c)],$$

while

$$(\forall a, b, c \in \mathbb{Z}, c > 0)[(a \leq b) \Rightarrow (a \cdot c \leq b \cdot c)] \text{ and}$$

$$(\forall a, b, c \in \mathbb{Z}, c < 0)[(a \leq b) \Rightarrow (b \cdot c \leq a \cdot c)]$$

It seems that many concepts and properties extend from natural numbers to integer numbers without many complications. However, there is one important property that \mathbb{N} has, but \mathbb{Z} does not have anymore.

Proposition 2.1.1. *There is no least element in \mathbb{Z} .*

Proof: The proof is by contradiction. Assume that there exists an integer k that is the smallest in \mathbb{Z} . Then $k - 1$ is smaller than k and is also an integer. This contradicts our assumption that \mathbb{Z} has a smallest element and finishes our proof. ■

Informally, we can remember that any subset of natural numbers has a bottom lid (or a least element) while this is not the case for the set of integer numbers.

Exercise 2.1.1. Prove the following assertions about addition and multiplication of integers. Use the properties of these operations as stated in Section 2.1.

(1) For all $a \in \mathbb{Z}$ and $n \geq 1$ we have

$$\underbrace{a + a + \dots + a}_{n \text{ times}} = n \cdot a \quad \text{and} \quad \underbrace{-a - a - \dots - a}_{n \text{ times}} = -(n \cdot a).$$

(2) For all $a \in \mathbb{Z}$ and all $m, n \geq 0$ it follows that

$$(a^m)^n = (a^n)^m = a^{mn} \quad \text{and} \quad a^m \cdot a^n = a^{m+n}.$$

(3) Given $a, b \in \mathbb{Z}$ and $n \geq 0$, then

$$a^n \cdot b^n = (ab)^n.$$

Exercise 2.1.2. Find all integers x such that $|x + 2| = 3$.

Exercise 2.1.3. Find all integers y such that $|y + 2| < 3$.

Exercise 2.1.4. Prove the following multiplicative properties of the absolute value of integers.

(1) For any integers $a, b \in \mathbb{Z}$, the following identity holds:

$$|a \cdot b| = |a| \cdot |b|.$$

(2) More generally, for any natural number n and any integers a_1, \dots, a_n :

$$|a_1 \cdots a_n| = |a_1| \cdots |a_n|.$$

(3) For any natural number n and any integer a :

$$|a^n| = |a|^n.$$

Exercise 2.1.5. Prove the following properties of the absolute value of integers.

- (1) If $a, b \in \mathbb{Z}$, then

$$|a + b| \leq |a| + |b|.$$

When does equality happen? This is usually called the triangle inequality and is true for larger sets of numbers such as complex numbers.

- (2) If n is a natural number and a_1, \dots, a_n are integers, then

$$|a_1 + \dots + a_n| \leq |a_1| + \dots + |a_n|.$$

When does equality happen?

Exercise 2.1.6. An integer number n is called even if $n = 2k$ for some integer number k and is called odd, otherwise. For m and n integer numbers, prove the following:

- (1) The number n is odd if and only if $n = 2\ell - 1$ for some integer number ℓ .
- (2) If m and n are even, then $m + n$ is even.
- (3) If m and n are odd, then $m + n$ is even.
- (4) If $m + n$ is odd, then exactly one of m and n is odd.
- (5) If mn is odd, then m is odd and n is odd.

Exercise 2.1.7. How many pairs (q, d) of nonnegative integers are there such that $25q + 10d \leq 100$? Compare this answer to Exercise 1.8.1.

Exercise 2.1.8. For each of the following statements, decide if it is true or false. If true, prove it. If false, provide a counterexample.

- (1) For any two integers a and b , if $a < b$, then $a^2 < b^2$.
- (2) For any two integers a and b , if $a < b$, then $|a| < |b|$.

Exercise 2.1.9. Consider the function $g : \mathbb{N} \rightarrow \mathbb{Z}$ defined as $g(n) = n$ for any $n \in \mathbb{N}$. Prove that g is injective, but not surjective.

Exercise 2.1.10. Consider the function $f : \mathbb{N} \rightarrow \mathbb{Z}$ defined as

$$f(n) = \begin{cases} 0, & \text{if } n = 1, \\ n/2, & \text{if } n \geq 2 \text{ is even,} \\ (1-n)/2, & \text{if } n \geq 3 \text{ is odd.} \end{cases}$$

- (1) Determine the image of the set $\{1, 2, 3, 4, 5\}$ under the function f .
- (2) Find the pre-image of $\{-1, 0, 1\}$ and of \mathbb{N} .
- (3) Show that f is bijective.
- (4) If $f^{-1} : \mathbb{Z} \rightarrow \mathbb{N}$ denotes its inverse function, what are $f^{-1}(17)$ and $f^{-1}(-8)$?

2.2. Integer Division

We can add and multiply natural numbers and get other natural numbers. With the integers, we can add, multiply, or subtract them, and we get other integers. What happens when we divide integers? We will extend the notions of divisibility and division from natural numbers to integers below.

Definition 2.2.1. Given two integers a and b with $b \neq 0$, we say that b **divides** a , which we write as $b | a$ if there exists an integer k such that $a = b \cdot k$,

$$(b \text{ divides } a) \Leftrightarrow (\exists k \in \mathbb{Z})(a = b \cdot k) \Leftrightarrow (a/b \in \mathbb{Z}).$$

If b divides a , we also say that b is a divisor of a or that a is a multiple of b . Otherwise, we say that b does not divide a and we write $b \nmid a$. For example, $2 | -6$, but $-2 \nmid 7$ and $-6 | -24$, but $9 \nmid -24$.

Remark 2.2.1. The number 0 plays a special role. Since $0 = b \cdot 0$, any nonzero integer b divides zero, $b | 0$ for all integers $b \neq 0$.

Many divisibility properties extend easily from natural numbers to integers. For example, the following result is an extension of Proposition 1.6.2 and its proof is left as an exercise.

Proposition 2.2.1. Let $d \neq 0$ be an integer. If $k \in \mathbb{N}$ and $a_1, \dots, a_k \in \mathbb{Z}$ are such that d divides each number a_1, \dots, a_k , then d divides $\sum_{j=1}^k \lambda_j a_j$, for any $\lambda_1, \dots, \lambda_k \in \mathbb{Z}$.

One has to be careful with other statements such as Proposition 1.6.1. While division over integers is reflexive (a divides a for any integer) and transitive (if $a | b$ and $b | c$, then $a | c$ for any integers a, b, c), we note here that it is not anti-symmetric (5 divides -5 and -5 divides 5, but $5 \neq -5$).

Remark 2.2.2. For all $a, b \in \mathbb{Z}, b \neq 0$,

$$b | a \Leftrightarrow -b | a \Leftrightarrow b | -a \Leftrightarrow -b | -a,$$

so divisibility involving negative integers is equivalent to divisibility involving natural numbers.

As we saw, sometimes the quotient of two integers is again an integer, sometimes this is not so. But what happens if for two integers a and b we have $b \nmid a$? For example, how would one divide -17 by 5 or -17 by -5 ? In the previous chapter, we saw how such a process works when dividing one natural number by another. Extending this idea to integers is pretty straightforward and goes as follows when trying to divide -17 by 5. Geometrically, one can imagine the integer numbers as points on the real line, and we put some special fence on the multiples of 5 marking the numbers below: These fences break the real line into intervals of length 5 and the number -17 must live in exactly one of these intervals. It is easy to see that $-17 \in [-20, -15]$ which is the same as $(-4) \cdot 5 < -22 < (-3) \cdot 5$. We can easily calculate by how much -17 goes above the lower bound below: $-17 = (-4) \cdot 5 + 3$. We call -4 the **quotient** and 3 the **remainder** when dividing -17 by 5 using integer division. Another way to find this interval where

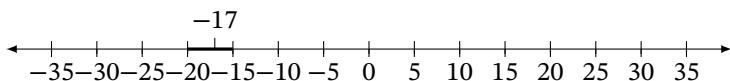


Figure 2.2.1. Integer division of -17 by 5.

-17 lives is by determining the corresponding interval for 17 , namely $17 \in [3 \cdot 5, 4 \cdot 5]$ and then multiply by -1 the inequalities $3 \cdot 5 < 17 < 4 \cdot 5$.

The curious reader may wonder how does one deal with the situation where we want to divide -17 by -5 ? It is a similar situation to the one above since the integer multiples of 5 are the same as the integer multiples of -5 . Hence, -17 will land in the same small yard as before $-17 \in [-20, -15]$. Since we are dividing -17 by -5 , we get that $-17 = 4 \cdot (-5) + 3$. In this case, the quotient will be 4 and the remainder will be 3 .

This simple example leads us to the next result called integer division which extends the similar result obtained in the previous chapter for natural numbers. Informally, it says you can always divide an integer by a nonzero integer, and you will get a *quotient* and a *small* (less than the number you divided by) and nonnegative leftover part called the *remainder*. These properties make the remainder unique.

Proposition 2.2.2 (Euclidean Division or Division with Remainder). *Let a and b be two integers with $b \neq 0$. Then there exist unique integers q and r such that*

$$(2.2.1) \quad a = bq + r, \quad 0 \leq r < |b|.$$

*The number q is called the **quotient** when a is divided by b while r is called the **remainder** under the integer division of a by b .*

$$(\forall a \in \mathbb{Z}, b \in \mathbb{Z} \setminus \{0\})(\exists! q, r \in \mathbb{Z})((a = bq + r) \text{ and } (0 \leq r < |b|))$$

Proof: The proof is similar to the one of Proposition 1.5.1 for natural numbers. However, we have different cases to consider depending on whether the numbers a and b are negative or not. We start our proof by splitting it into two cases: $b > 0$ and $b < 0$.

In the first case, suppose $b > 0$. Let $S = \{n \in \mathbb{N} : bn > |a|\}$. The set S is a subset of the set of natural numbers \mathbb{N} and it is nonempty since $|a|+1 \in S$. By the least element principle (see Section 1.4 if you need a recap), the set S has a smallest element which we will call $m+1$ where $m \geq 0$ is an integer. We claim now that

$$(2.2.2) \quad bm \leq |a| < b(m+1).$$

Indeed, since $m+1 \in S$, we get that $b(m+1) > |a|$. Also, because $m \notin S$, we deduce that $|a| \geq bm$. Recall that $|a| = -a$ for negative integers a , hence (2.2.2) says that

$$\begin{aligned} bm \leq a &< b(m+1), & \text{if } a \geq 0, \quad \text{or} \\ bm \leq -a &< b(m+1), & \text{if } a < 0. \end{aligned}$$

If $a \geq 0$, set $q = m$ and $r = a - bm = a - b q$. Then $a = b q + r$. Also, $bm \leq a < b(m+1)$ implies that $0 \leq r < b$ as asserted. If $a < 0$, then we have that $bm \leq -a < b(m+1)$ which implies that $b(-m-1) < a \leq b(-m)$. If $a = b(-m)$, we take $q = -m$ and $r = 0$ and we get a representation of a as stated in (2.2.1). Otherwise, if $a \neq b(-m)$, then $b(-m-1) < a < b(-m)$. Set $q = -m-1$ and $r = a - b q = a - b(-m-1)$. Then $a = b q + r$ and $0 < r < b$ as asserted. Hence, we proved the existence of representation (2.2.1) for any $a \in \mathbb{Z}$ and $b > 0$.

For the second case, assume $b < 0$. Then $-b > 0$, hence we can apply the results of our work from the first case to deduce that there are $q' \in \mathbb{Z}$ and $0 \leq r' < -b = |b|$ such that

$$a = (-b)q' + r'.$$

Letting $q = -q'$ and $r = r'$, we get

$$a = b q + r, \quad 0 \leq r < |b|.$$

This completes the proof of the existence of the representation.

It remains to show the uniqueness of representation (2.2.1). Suppose we have two representations

$$a = b q + r = b q' + r', \quad 0 \leq r, r' < |b|.$$

Therefore, we have that

$$r - r' = b(q' - q).$$

This implies that $|b|$ divides $r - r'$. Because $0 \leq r, r' < |b|$, we get that $-|b| < r - r' < |b|$. Since 0 is the only integer number strictly between $-|b|$ and $|b|$ divided by $|b|$, it follows that $r - r' = 0$, hence, $r = r'$. Of course, this implies $qb = q'b$, and since $b \neq 0$ we conclude that also $q = q'$. Consequently, q and r in representation (2.2.1) are unique, and this completes the proof. ■

Before describing some examples, let us remark that the proof above gives a quick recipe to determine the quotient and the remainder of an integer division. For any real number x , let $\lfloor x \rfloor$ denote the largest integer less than or equal to x . This is called the **floor function** and $\lfloor x \rfloor$ is called the floor of x . For example, $\lfloor 3.6 \rfloor = 3$, $\lfloor -3.6 \rfloor = -4$ and $\lfloor \sqrt{2} \rfloor = 1$ (since $1 < \sqrt{2} < 2$). Similarly, one can define $\lceil x \rceil$ as the smallest integer that is greater than or equal to x . This is the **ceiling function** and $\lceil x \rceil$ is called the ceiling of x . For example, $\lceil 3.6 \rceil = 4$, $\lceil -3.6 \rceil = -3$ and $\lceil \sqrt{2} \rceil = 2$.

With these notations, the proof of Proposition 2.2.2 tells us that given $a \in \mathbb{Z}$, for $b > 0$ the quotient q in representation (2.2.1) equals $q = \lfloor a/b \rfloor$ while the remainder r is evaluated as

$$r = a - bq.$$

For example, when $a = -22$ and $b = 5$, we obtain that $q = \lfloor -22/5 \rfloor = -5$ and $r = -22 - 5 \cdot (-5) = 3$.

How do we get q and r in the case $b < 0$? Recall that

$$a = q \cdot b + r$$

where $q = -\lfloor a/b \rfloor = \lceil a/b \rceil$ and $r = a - qb$. For example, if $a = 17$ and $b = -5$, then $\lceil 17/-5 \rceil = -3$, hence

$$17 = (-3) \cdot (-5) + 2 \quad \text{with } 0 \leq 2 < 5 = |-5|.$$

Another example is $a = -17$ and $b = -5$. Here $\lceil -17/-5 \rceil = \lceil 17/5 \rceil = 4$, thus

$$-17 = 4 \cdot (-5) + 3 \quad \text{with } 0 \leq 3 < 5 = |-5|.$$

We conclude with a simple, but useful result that will be used many times in this chapter.

Proposition 2.2.3. *Let d, m , and n be three integers such that d is nonzero. The numbers m and n have the same remainder when divided by d if and only if $d \mid m - n$.*

Proof: Assume that m and n have the same remainder when divided by d . Then $m = dq_1 + r, n = dq_2 + r$ for some integers q_1, q_2 , and r with $0 \leq r < |d|$. Subtracting one equation from the other, we get that $m - n = d(q_1 - q_2)$ which implies that $d \mid m - n$.

For the converse, assume that $d \mid m - n$. Let r_1 be the remainder of m when divided by d . Therefore, $m = dq_1 + r_1$ for some integer q_1 and $0 \leq r_1 < |d|$. Let r_2 be the remainder of n when divided by d . Hence, $n = dq_2 + r_2$ for some integer q_2 and $0 \leq r_2 < |d|$. Therefore, $m - n = d(q_1 - q_2) + r_1 - r_2$ which implies that $r_1 - r_2 = (m - n) - d(q_1 - q_2)$. Because $d \mid m - n$ and $d \mid d(q_1 - q_2)$, we deduce that $d \mid r_1 - r_2$. Since $0 \leq r_1, r_2 < |d|$, the only way this can happen is if $r_1 - r_2 = 0$ which means that $r_1 = r_2$ and finishes our proof. ■

Exercise 2.2.1. Apply the integer division with remainder to the following pairs of integers: $(89, -55)$, $(-143, 12)$ and $(-1001, -324)$.

Exercise 2.2.2. Let a_1 and a_2 be two integers and $d \neq 0$ a nonzero integer. Assume that the remainder of the integer division of a_j by d is r_j for $1 \leq j \leq 2$. Is it true that the remainder of the integer division of $a_1 + a_2$ by d is $r_1 + r_2$? Prove it or give counterexamples.

Exercise 2.2.3. Assume the same hypothesis of Exercise 2.2.2. If also $r_1 \geq r_2$, then is it true that the remainder of the integer division of $a_1 - a_2$ by d equals $r_1 - r_2$? Prove it or give a counterexample.

Exercise 2.2.4. Let a be an integer and d a natural number. Denote by q and r the quotient and the remainder of the integer division of a by d . Denote by q' and r' the quotient and the remainder of the integer division of a by $-d$. Is there a relation between q and q' ? What about r and r' ?

Exercise 2.2.5. Let x be a real number. Prove that $|x| \leq x < |x| + 1$ as well as $|x| - 1 < x \leq |x|$.

Exercise 2.2.6. Let d and $n \neq 0$ be two integers such that $d \mid n$. Prove that $d \leq |n|$.

Exercise 2.2.7. Prove that $n^2 + 3n + 5$ is odd for any integer n .

Exercise 2.2.8. Let x be a real number. Which number is larger: $2x$ or $|x| + |x|$?

Exercise 2.2.9. Let x and y be real numbers. Show that $|x+y| \geq |x| + |y|$. Generalize this inequality to n real numbers x_1, \dots, x_n , where $n \geq 3$ is a natural number. When does equality hold in the previous inequalities?

Exercise 2.2.10. Let x and y be real numbers. Prove that $|x| + |y| \leq |x+y|$. Generalize this inequality to n real numbers x_1, \dots, x_n , where $n \geq 3$ is a natural number. When does equality hold in the previous inequalities?

2.3. Euclidean Algorithm Revisited

In Section 1.6, we introduced the greatest common divisor of nonnegative integers. This concept can be easily extended to integers as follows.

Definition 2.3.1. Let a and b be two integers, not both equal to 0. The **greatest common divisor** $\gcd(a, b)$ of a and b is defined as the largest integer that divides both a and b .

Similar to natural numbers, this definition makes sense for integers. The set of common divisors for a and b is nonempty since $1 \mid a$ and $1 \mid b$. If $a = 0$, then $|b|$ will be the largest integer dividing a and b and $\gcd(a, b) = |b|$ in this case. Similarly, if $b = 0$, then $\gcd(a, b) = |a|$. If both numbers are nonzero, then any common divisor of a and b is less than the minimum of $|a|$ and $|b|$ (see also Exercise 2.2.6). This argument also implies that $\gcd(a, b)$ is a natural number for any integers a and b that are not both 0.

The following proposition shows that the computation of the greatest common divisor of two integers can always be reduced to the computation of the greatest common divisor of nonnegative integer numbers as in Section 1.6.

Proposition 2.3.1. *If a and b are two integers, not both equal to 0, then*

$$\gcd(a, b) = \gcd(|a|, |b|).$$

Proof: The set of integer divisors of a is the same as the set of integer divisors of $|a|$. The same is true for b and $|b|$. The result then follows from the definition above. ■

In Section 1.6, we describe the Euclidean algorithm for computing $\gcd(a, b)$ when $a, b \in \mathbb{N}_0$ are not both 0. The Euclidean algorithm is a succession of integer divisions with decreasing remainders that ends when the remainder is 0. The last nonzero remainder is the greatest common divisor of a and b . This algorithm can be extended with essentially the same proof to pairs of integers (see Exercise 2.3.1) or can be used in conjunction with the proposition above to calculate $\gcd(a, b)$ for any integers a and b that are not both 0.

Example 2.3.1. Let $a = -23$ and $b = 17$. First, we use the Euclidean algorithm to compute $\gcd(|-23|, 17) = \gcd(23, 17)$. We have the following integer divisions:

$$\begin{aligned} 23 &= 17 \cdot 1 + 6 \\ 17 &= 6 \cdot 2 + 5 \\ 6 &= 5 \cdot 1 + 1 \\ 5 &= 5 \cdot 1 + 0. \end{aligned}$$

Therefore, $\gcd(23, 17) = 1$. The previous proposition implies that $\gcd(-23, 17) = 1$.

We can also run the Euclidean algorithm for the numbers -23 and 17 as follows:

$$\begin{aligned} -23 &= 17 \cdot (-2) + 11 \\ 17 &= 11 \cdot 1 + 6 \\ 11 &= 6 \cdot 1 + 5 \\ 6 &= 5 \cdot 1 + 1 \\ 5 &= 5 \cdot 1 + 0. \end{aligned}$$

Interestingly enough, the number of steps to get a zero remainder was different from the previous case, but the greatest common divisor was the same.

However, we can add one interesting twist to the Euclidean algorithm and run it *in reverse* as follows.

Example 2.3.2. Let us go back to our example with $\gcd(23, 17) = 1$. We retrace our steps as above and in addition, for each integer division, we write down the remainder in terms of the rest of the numbers involved as follows:

$$(2.3.1) \quad 23 = 17 \cdot 1 + 6 \Leftrightarrow 6 = 23 - 17 \cdot 1$$

$$(2.3.2) \quad 17 = 6 \cdot 2 + 5 \Leftrightarrow 5 = 17 - 6 \cdot 2$$

$$6 = 5 \cdot 1 + 1 \Leftrightarrow 1 = 6 - 5 \cdot 1$$

$$5 = 5 \cdot 1 + 0.$$

We start from the equation at the bottom of our list $1 = 6 - 5 \cdot 1$ and work our way *up* as follows:

$$1 = 6 - 5 \cdot 1 \Rightarrow 1 = 6 - (17 - 6 \cdot 2) \cdot 1 \text{ using (2.3.2)}$$

$$1 = 6 \cdot 3 - 17 \cdot 1 \text{ just rearranging the above}$$

$$1 = (23 - 17) \cdot 3 - 17 \cdot 1 \text{ using (2.3.1)}$$

$$1 = 23 \cdot 3 - 17 \cdot 4 \text{ just rearranging the above}$$

$$1 = 3 \cdot 23 + (-4) \cdot 17.$$

We have managed to write $1 = \gcd(23, 17)$ as a linear combination with integer coefficients of 23 and 17.

It turns out that this procedure will work in general and for any two integers a and b that are not both zero, their greatest common divisor can be written as an integer linear combination of a and b . This result is called Bézout's lemma or Bézout's identity in honor of the French mathematician Etienne Bézout (1730–1783) who proved a similar result for polynomials.

Proposition 2.3.2 (Bézout's Lemma). *Let a and b be two integers that are not both zero. Then there are integers s and t such that*

$$(2.3.3) \quad \gcd(a, b) = s \cdot a + t \cdot b.$$

Proof: We give two proofs of this result. One is constructive using the Euclidean algorithm and the other is nonconstructive. For both proofs, we may assume that $a, b \in \mathbb{N}_0$. Otherwise, replace a by $-a$ and/or b by $-b$.

For the proof using the Euclidean algorithm, we use strong induction on the minimum $\min(a, b)$ between a and b . Without loss of generality, we assume that $a \geq b$ so $\min(a, b) = b$. For the base case, we deal with the situation when we have $b = \min(a, b) = 0$. In this case, $\gcd(a, b) = \gcd(a, 0) = a$ and we have that

$$a = \gcd(a, b) = \gcd(a, 0) = 1 \cdot a + 1 \cdot 0 = 1 \cdot a + 0 \cdot b.$$

This implies that the assertion (2.3.3) is true with $s = 0$ and $t = 1$. Let us examine the next case when $b = \min(a, b) = 1$ as well. In this case, $\gcd(a, b) = \gcd(a, 1) = 1$ and we can write

$$1 = \gcd(a, 1) = 1 \cdot a + (-a + 1) \cdot 1 = 1 \cdot a + (-a + 1) \cdot b.$$

It is clear that our assertion (2.3.3) is satisfied with $s = 1$ and $t = -a + 1$ and this case is also true.

For the induction step, assume that $\min(a, b) \geq 2$ and assume that our statement is true for any pairs of nonnegative integers k and ℓ for which we have $\min(k, \ell) < \min(a, b) = b$. We perform the first step of the Euclidean algorithm for a and b and we obtain that:

$$a = bq + r, \quad 0 \leq r < b.$$

If we consider the pair of nonnegative integers b and r , it follows the estimate $\min(b, r) = r < b = \min(a, b)$. By the induction hypothesis, we deduce that there exist integers s' and t' such that

$$\gcd(b, r) = s'b + t'r.$$

We know that $\gcd(a, b) = \gcd(b, r)$ (see Section 1.6 or Exercise 2.3.1) and $r = a - bq$ and therefore,

$$\gcd(a, b) = \gcd(b, r) = s'b + t'r = s'b + t'(a - bq) = t'a + (s' - t'q)b.$$

This equation implies that our equation (2.3.3) is satisfied for a and b with $s = t'$ and $t = s' - t'q$. This finishes our first proof.

For the second proof, we assume that both a and b are natural numbers. The case when one of them is zero can be dealt with as in the base case of the previous proof. Let A be the set of all natural numbers that may be represented as $m \cdot a + n \cdot b$ for some integers m and n . This set is nonempty since, for example, $a + b$ belongs to A . The least element principle implies the existence of a least element d in A . We will prove that $d = \gcd(a, b)$ which will imply that $\gcd(a, b)$ admits a representation (2.3.3).

Because $d = m \cdot a + n \cdot b$ for some integers m and n , $\gcd(a, b) \mid a$, $\gcd(a, b) \mid b$ and Proposition 2.2.1 imply that $\gcd(a, b) \mid d$. Hence, $\gcd(a, b) \leq d$.

Now we prove that $d \leq \gcd(a, b)$ which implies $d = \gcd(a, b)$ and completes the proof. To this end, we use Proposition 2.2.2. Consequently, there exist a quotient q and a remainder $0 \leq r < d$ such that

$$(2.3.4) \quad a = q \cdot d + r.$$

Substituting $d = m \cdot a + n \cdot b$ and rearranging the previous equation, we get that

$$r = a - qd = a - q(m \cdot a + n \cdot b) = (1 - q \cdot m) \cdot a + (-q \cdot n) \cdot b.$$

What does this tell us? It says that the remainder r has representation (2.3.3), hence belongs to A . But $0 \leq r < d$, and since d was taken as the least element in A , the remainder r cannot be a natural number, thus, has to be zero. But this implies $d \mid a$ by equation (2.3.4). By exactly the same argument it follows that $d \mid b$, hence d is a common divisor of a and b , and $d \leq \gcd(a, b)$. As already observed, this completes the proof. ■

If for the numbers 23 and 17, there is a possibility of being able to guess numbers s and t such that $1 = \gcd(23, 17) = s \cdot 23 + t \cdot 17$, the next example shows that such a strategy would not be useful for larger numbers in general.

Example 2.3.3. We describe obtaining a representation (2.3.3) for $a = 7596$ and $b = 3242$. Reversing the calculations in Example 1.6.2 we obtain the following Bézout representation of $2 = \gcd(a, b)$.

$$\begin{aligned} & \left. \begin{array}{rcl} 2 & = & 16 - 14 \\ 14 & = & 78 - 4 \cdot 16 \end{array} \right\} \Rightarrow 2 = 5 \cdot 16 - 78 \\ & \left. \begin{array}{rcl} 2 & = & 5 \cdot 16 - 78 \\ 16 & = & 94 - 78 \end{array} \right\} \Rightarrow 2 = 5 \cdot 94 - 6 \cdot 78 \\ & \left. \begin{array}{rcl} 2 & = & 5 \cdot 94 - 6 \cdot 78 \\ 78 & = & 1018 - 10 \cdot 94 \end{array} \right\} \Rightarrow 2 = 65 \cdot 94 - 6 \cdot 1018 \\ & \left. \begin{array}{rcl} 2 & = & 65 \cdot 94 - 6 \cdot 1018 \\ 94 & = & 1112 - 1018 \end{array} \right\} \Rightarrow 2 = 65 \cdot 1112 - 71 \cdot 1018 \\ & \left. \begin{array}{rcl} 2 & = & 65 \cdot 1112 - 71 \cdot 1018 \\ 1018 & = & 3242 - 2 \cdot 1112 \end{array} \right\} \Rightarrow 2 = 207 \cdot 1112 - 71 \cdot 3242 \\ & \left. \begin{array}{rcl} 2 & = & 207 \cdot 1112 - 71 \cdot 3242 \\ 1112 & = & 7596 - 2 \cdot 3242 \end{array} \right\} \Rightarrow 2 = 207 \cdot 7596 - 485 \cdot 3242. \end{aligned}$$

Therefore, one possible Bézout representation of $\gcd(a, b)$ with $a = 7596$ and with $b = 3242$ is given by

$$2 = 207 \cdot 7596 + (-485) \cdot 3242.$$

Remark 2.3.1. The Bézout representation (2.3.3) of $d = \gcd(a, b)$ is not unique. We first describe the idea for one of our previous examples, $a = 23$ and $b = 17$, where we proved that

$$1 = \gcd(23, 17) = 3 \cdot 23 + (-4) \cdot 17.$$

The idea to obtain other Bézout representations for 23 and 17 is to tweak the coefficients 3 and -4 in a clever way. If we increase 3 to $3 + 17$, then this operation changes $3 \cdot 23$ to

$$(2.3.5) \quad (3 + 17) \cdot 23 = 3 \cdot 23 + 17 \cdot 23.$$

At the same time, if we decrease -4 to $-4 - 23$, then we change the value $(-4) \cdot 17$ to

$$(2.3.6) \quad (-4 - 23) \cdot 17 = (-4) \cdot 17 - 23 \cdot 17.$$

Adding up the equations (2.3.5) and (2.3.6), we obtain that

$$\begin{aligned} (3 + 17) \cdot 23 + (-4 - 23) \cdot 17 &= 3 \cdot 23 + 17 \cdot 23 + (-4) \cdot 17 - 23 \cdot 17 \\ &= 3 \cdot 23 + (-4) \cdot 17 \\ &= 1. \end{aligned}$$

The idea was to modify the coefficients of 23 and 17 such that the new contributions from each term would cancel out. Of course, we can do a more general procedure as follows. Let k be an integer. Then

$$\begin{aligned} (3 + 17k) \cdot 23 + (-4 - 23k) \cdot 17 &= 3 \cdot 23 + 17 \cdot 23 \cdot k + (-4) \cdot 17 - 23 \cdot 17 \cdot k \\ &= 3 \cdot 23 + (-4) \cdot 17 \\ &= 1. \end{aligned}$$

What we did so far is to show that the set

$$(2.3.7) \quad A = \{(3 + 17k, -4 - 23k) : k \in \mathbb{Z}\}$$

is a *subset* of the set of integer solutions of the equation

$$(2.3.8) \quad 23x + 17y = 1.$$

Are these all the solutions? Consider an arbitrary pair of integers (x, y) such that $23x + 17y = 1$. We know that

$$\begin{aligned} 23x + 17y &= 1 \\ 23 \cdot 3 + 17 \cdot (-4) &= 1, \end{aligned}$$

and therefore,

$$23x + 17y = 23 \cdot 3 + 17 \cdot (-4)$$

which further implies that

$$23(x - 3) = 17(-y - 4).$$

Because 17 divides $17(-y - 4)$, it must divide $23(x - 3)$. Now $\gcd(23, 17) = 1$ and Bézout's lemma will tell us that $17 \mid x - 3$. Hence, there exists an integer ℓ such that $x - 3 = 17\ell$. So $x = 17\ell + 3$. Also, the equation $23(x - 3) = 17(-y - 4)$ will imply that $-y - 4 = 23\ell$ and therefore, $y = -23\ell - 4$. We just proved that if (x, y) is an integer solution of the equation $23x + 17y = 1$, then (x, y) is an element of A . Thus, the set A in (2.3.7) is the set of solutions of the equation (2.3.8).

We will see later in this section how to determine all the integer solutions (x, y) of an equation of the form

$$ax + by = c,$$

where a, b , and c are integers.

Definition 2.3.2. Two nonzero integers a and b are said to be **coprime** (or **relatively prime**) provided that

$$\gcd(a, b) = 1.$$

In other words, 1 is the only natural number dividing both a and b .

For example, 9 and -20 are coprime, -15 and -21 are not. The next result shows that *peeling off* the greatest common divisor from a pair of integers yields two coprime integers. It is an extension of Proposition 1.6.8 and we leave its proof as an exercise.

Proposition 2.3.3. *Let a and b be two integers, not both equal to zero. If we have $d = \gcd(a, b)$, then*

$$\gcd(a/d, b/d) = 1.$$

We present other applications of Bézout's lemma below. The first one is a generalization of Euclid's lemma (Corollary 1.6.6 in Section 1.6).

Proposition 2.3.4. *Let $d, a, b \in \mathbb{Z}$. If $d \mid ab$ and $\gcd(d, a) = 1$, then $d \mid b$.*

Proof: Proposition 2.3.2 and $\gcd(d, a) = 1$ imply that there are integers s and t such that $1 = sa + td$. We multiply this equation by b and obtain

$$(2.3.9) \quad b = sab + tdb.$$

By our hypothesis, $d \mid sab$ and, of course, $d \mid tdb$. Using (2.3.9), we get that $d \mid sab + tdb = b$. This finishes the proof. \blacksquare

We explain now how the previous result is a generalization of Proposition 1.6.6.

Corollary 2.3.5 (Euclid's Lemma for Integers). *Let p be a prime. If a and b are integers such that $p \mid ab$, then $p \mid a$ or $p \mid b$.*

Proof: If $p \mid a$, then we are done. Otherwise, if $p \nmid a$, then $\gcd(a, p) \neq p$. From definition, $\gcd(a, p) \mid p$ and since p is prime, $\gcd(a, p)$ can only equal 1 or p . Hence, $\gcd(a, p) = 1$. Now we apply Proposition 2.3.4 with $d = p$ and get $p \mid b$. \blacksquare

Remark 2.3.2. Note that if $p \mid ab$, we may have both that $p \mid a$ and $p \mid b$ as easy examples show. Only the case $p \nmid a$ and $p \nmid b$ cannot occur.

The previous result can be generalized to products of more than two numbers. The proof of the next result is left as an exercise.

Corollary 2.3.6. *Let p be a prime and n a natural number. If a_1, \dots, a_n are integers such that $p \mid a_1 \cdots a_n$, then there is a $j = 1, \dots, n$ such that p divides a_j .*

We now give another application of Bézout's lemma in solving equations of the form

$$(2.3.10) \quad ax + by = c,$$

where a, b, c are given integers and the unknowns are the integers x and y . Such equations are examples of **Diophantine equations** which are polynomial equations in two or more variables whose coefficients are integers and whose solutions must be integers. The name comes from the Greek mathematician Diophantus of Alexandria who lived around the 3rd century CE and whose book *Arithmetica* was extremely influential in the development of number theory and algebra. The equation (2.3.10) is the simplest example of a Diophantine equation. Before proceeding with the general result, let us do some more examples.

Example 2.3.4. Earlier in the section, we solved the equation $23x + 17y = 1$. Let us try solving a similar equation:

$$(2.3.11) \quad 23x + 17y = 4.$$

From earlier work, we know that for any integer k , the pair $(3 + 17k, -4 - 23k)$ is a solution of (2.3.8):

$$23(3 + 17k) + 17(-4 - 23k) = 1.$$

Multiplying both sides by 4, we get that

$$23(12 + 68k) + 17(-16 - 92k) = 4.$$

Hence, the set $B = \{(12 + 68k, -16 - 92k) : k \in \mathbb{Z}\}$ is a subset of the set of solutions for the equation (2.3.11). To see if these are all the solutions, we proceed as in the case of

(2.3.8) and consider an arbitrary solution (x, y) of (2.3.11) along with a specific solution of the same equation:

$$\begin{aligned} 23x + 17y &= 4 \\ 23 \cdot 12 + 17 \cdot (-16) &= 4. \end{aligned}$$

Subtracting these equations, we get that $23(x - 12) = 17(-y - 16)$. This lets us conclude that $17 \mid 23(x - 12)$ which combined with $\gcd(23, 17) = 1$ and Proposition 2.3.4, gives us that $17 \mid x - 12$. Thus, there exists $\ell \in \mathbb{Z}$ such that $x - 12 = 17\ell$ and therefore, $x = 12 + 17\ell$. Plugging this back into the previous equation will give us that $y = -16 - 92\ell$. Consequently, (x, y) is an element of B and this shows that B is the set of solutions of (2.3.11).

Example 2.3.5. Let us solve a different equation over integers:

$$(2.3.12) \quad 24x + 16y = 10.$$

Before trying the Euclidean algorithm, let us pause and notice that the coefficients 24 and 16 are no longer coprime as before. Actually, $\gcd(24, 16) = 4$ can be checked easily. This implies that for any integers x and y , 4 must divide $24x + 16y$. Hence, if (x, y) was an integer solution of (2.3.12), then $4 \mid 24x + 16y = 10$ which is a contradiction. This means that the equation (2.3.12) has no integer solutions.

The examples above illustrate the possible situations when solving a Diophantine linear equation in two unknowns. We give the details in the next proposition.

Proposition 2.3.7. *Let a , b , and c be integers such that both a and b are nonzero. The equation (2.3.10) has integer solutions if and only if $\gcd(a, b) \mid c$. Moreover, if $\gcd(a, b) \mid c$ and (x_0, y_0) is an integer pair satisfying $ax_0 + by_0 = c$, then the set of integer solutions (x, y) of the equation (2.3.10) equals*

$$\{(x_0 + b_1 k, y_0 - a_1 k) : k \in \mathbb{Z}\}.$$

where $a_1 = a/\gcd(a, b)$ and $b_1 = b/\gcd(a, b)$.

Proof: Assume that the equation $ax + by = c$ has an integer solution (x, y) . Because $\gcd(a, b) \mid a$ and $\gcd(a, b) \mid b$, it follows that $\gcd(a, b) \mid ax + by = c$ which finishes the proof of one implication.

Assume now that $\gcd(a, b) \mid c$. Let $d = \gcd(a, b)$ and $c = dc_1$ with $c_1 \in \mathbb{Z}$. Bézout's lemma implies that there exist integers s and t such that $as + bt = d$. Multiplying both sides by c_1 , we get that $a(sc_1) + b(tc_1) = dc_1 = c$. Hence, the pair (sc_1, tc_1) is an integer solution for the equation (2.3.10). Note that such pairs can be found using the Euclidean algorithm in reverse as described in the first proof of Bézout's lemma (Proposition 2.3.2).

Let (x_0, y_0) be a given integer solution of (2.3.10). If (x, y) is another integer solution of (2.3.10), then we have that $ax + by = ax_0 + by_0$. Then we get that $a(x - x_0) = b(y_0 - y)$. Because $a = da_1$ and $b = db_1$, dividing both sides of the previous equation by d , we get that $a_1(x - x_0) = b_1(y_0 - y)$. Now Proposition 2.3.3 implies that $\gcd(a_1, b_1) = 1$. Because $b_1 \mid b_1(y_0 - y) = a_1(x - x_0)$ and $\gcd(a_1, b_1) = 1$, Proposition 2.3.4 implies that $b_1 \mid x - x_0$. Hence, $x - x_0 = b_1 k$ for some integer k . Therefore,

$x = x_0 + b_1 k$ and $a_1 b_1 k = a_1(x - x_0) = b_1(y_0 - y)$. Because $b_1 = b/d \neq 0$, dividing both sides by b_1 gives that $y = y_0 - a_1 k$. This shows that the set of solutions (x, y) of the equation (2.3.10) is contained in the set $\{(x_0 + b_1 k, y_0 - a_1 k) : k \in \mathbb{Z}\}$. To show that these two sets are actually equal, we have to check that any pair of the form $(x_0 + b_1 \ell, y_0 - a_1 \ell)$ with $\ell \in \mathbb{Z}$ satisfies equation (2.3.10). A simple calculation and the fact that $a_1 b = ab_1 = da_1 b_1$ imply that

$$a(x_0 + b_1 \ell) + b(y_0 - a_1 \ell) = ax_0 + by_0 + (ab_1 - a_1 b)\ell = ax_0 + by_0 = c.$$

This finishes the proof. ■

Example 2.3.6. Consider the following equation over integers:

$$(2.3.13) \quad 20x + 16y = 8.$$

In this case, $a = 20$, $b = 16$ and $\gcd(a, b) = 4 \mid 8$. Therefore, the equation above is equivalent to the one obtained after dividing both sides by 4:

$$5x + 4y = 2.$$

To find all the solutions of this equation, we must first find one solution. It is not too hard to see that $(2, -2)$ is a solution since $5 \cdot 2 + 4 \cdot (-2) = 10 - 8 = 2$. By the previous Proposition, the solution set of the equation (2.3.13) is $\{(2 + 4k, 2 - 5k) : k \in \mathbb{Z}\}$.

Our results related to solving the equation (2.3.10) imply that if a and b are natural numbers such that $\gcd(a, b) = 1$, then any integer c can be written as $ax + by$ for some integers x and y .

Example 2.3.7. If $a = 3$ and $b = 5$, then any integer c can be written as $3x + 5y$ with x and y integers. For example,

$$\begin{aligned} 3 \cdot 2 + 5 \cdot (-1) &= 1 \\ 3 \cdot (-3) + 5 \cdot 2 &= 1 \\ 3 \cdot (2c) + 5 \cdot (-c) &= c \\ 3 \cdot (-3c) + 5 \cdot (2c) &= c. \end{aligned}$$

Let us slightly change the previous problem.

Problem 2.3.8. Given two natural numbers a and b , what natural numbers can be written as $ax + by$ for x and y **nonnegative integer numbers**?

This problem was studied by James Joseph Sylvester (1814–1897), one of the greatest English mathematicians, who made important contributions in algebra, number theory, and combinatorics. Sylvester came up with the words *graph* (as the ones we study in Section 1.2) and *matrix* among other accomplishments.

Definition 2.3.3. Let a and b be two natural numbers. A natural number n is called **(a, b) -achievable** if there exist nonnegative integers x and y such that

$$n = ax + by.$$

Example 2.3.8. For example, when $a = 3$ and $b = 5$, since $x, y \geq 0$ now, then it is clear we cannot write 1 as $3x + 5y$. This is because if $x \geq 1$, then $3x + 5y \geq 3$ and if $y \geq 1$, then $3x + 5y \geq 5$. A simple case analysis shows that also 2, 4 and 7 cannot be written as $3x + 5y$ for any $x, y \in \mathbb{N}_0$. The attentive reader must recall a similar result in Section 1.4, namely Proposition 1.4.6 where we proved that every natural number $n \geq 8$ can be written as $3x + 5y$ for some $x, y \in \mathbb{N}_0$. Now we can uncover the key idea in the proof of Proposition 1.4.6 which is that 8, 9, 10 is the first sequence of three consecutive numbers that are $(3, 5)$ -achievable. Strong induction can be used to prove Proposition 1.4.6. We summarize the situation for each natural number n in the table below (*yes* means n is $(3, 5)$ -achievable and *no* the opposite). Hence, for $a = 3$ and $b = 5$, any natural number greater than or equal to 8 is $(3, 5)$ -achievable and there are exactly 4 numbers $n \leq 7$ that are not $(3, 5)$ -achievable.

Table 2.3.1. Numbers up to 10 split between $(3, 5)$ -achievable and the rest.

| | |
|-----|--------------|
| yes | 3,5,6,8,9,10 |
| no | 1,2,4,7 |

Example 2.3.9. Let us try two different numbers, say $a = 4$ and $b = 7$. It is fairly easy to see that 1, 2, 3, 5, 6 cannot be written as $4x + 7y$ for any $x, y \in \mathbb{N}_0$. With a bit more effort which we leave as an exercise to the reader, one can extend this work and compile the following table for small values of n : As before, *yes* means that the numbers in that row are $(4, 7)$ -achievable and *no* means the opposite. By a similar argument as in the previous example, note that 18, 19, 20, 21 is the first sequence of four consecutive numbers that are $(4, 7)$ -achievable. We will leave it as an exercise for the reader to use strong induction as in Proposition 1.4.6 and show that any natural number $n \geq 18$ is $(4, 7)$ -achievable. Hence, for $a = 4$ and $b = 7$, any natural number greater than or equal to 18 is $(4, 7)$ -achievable and there are exactly 8 numbers $n \leq 17$ that are $(4, 7)$ -achievable.

Table 2.3.2. Numbers up to 21 split between $(4, 7)$ -achievable and the rest.

| | |
|-----|----------------------------------|
| yes | 4,7,8,11,12,14,15,16,18,19,20,21 |
| no | 1,2,3,5,6,9,10,13,17 |

Example 2.3.10. We do one more short example before describing the general solution of Sylvester's problem. Consider $a = 4$ and $b = 6$. It is easy to see that no odd natural number can be written as $4x + 6y$ for x, y nonnegative integers. To see what even natural numbers can be written as $4x + 6y = 2(2x + 3y)$, we would first solve the problem of determining the natural numbers that are $(2, 3)$ -achievable and take the double of each such number.

The purpose of the last example is that we can restrict ourselves to the case $\gcd(a, b) = 1$ when studying Problem 2.3.8. The following result shows that $ab - a - b$ is the largest natural number that is not (a, b) -achievable and determines how many numbers less than $ab - a - b$ are (a, b) -achievable.

Theorem 2.3.9. Let a and b be two natural numbers such that $\gcd(a, b) = 1$.

- (1) The number $ab - a - b$ is not (a, b) -achievable.
- (2) Any natural number $n \geq ab - a - b + 1$ is (a, b) -achievable.

Proof: For the first part, we use proof by contradiction. Assume that $ab - a - b$ is (a, b) -achievable. It means that there exist nonnegative integers m and n such that $ab - a - b = ma + nb$. This further implies that $ab = (m+1)a + (n+1)b$. Because $a \mid ab$ and $a \mid (m+1)a$, we deduce that $a \mid (n+1)b$. Because $\gcd(a, b) = 1$, we can apply Proposition 2.3.4 and deduce that $a \mid n+1$. Since $n+1$ and a are natural numbers, it follows that $n+1 \geq a$. By a similar argument which we leave as an exercise, we can deduce that $m+1 \geq b$. Therefore,

$$ab = (m+1)a + (n+1)b \geq b \cdot a + a \cdot b = 2ab,$$

which means that $ab \leq 0$, in contradiction with a and b being both natural numbers.

For the second part, we will need some simple lemmas.

Lemma 2.3.10. Let $ab - a - b, ab - a - b + 1, \dots, ab - a - b + a - 1$ be the sequence of a consecutive integers starting at $ab - a - b$. The set of their remainders when divided by a is the set $\{0, \dots, a - 1\}$.

Proof: Since every remainder of an integer division by a is an element of the set $\{0, \dots, a - 1\}$, it is clear that the set of remainders of the a consecutive numbers above is a subset of $\{0, \dots, a - 1\}$. To show that equality happens, it is sufficient to show that no two remainders of numbers in the consecutive list are the same. We use proof by contradiction and assume that there exist $0 \leq s_1 < s_2 \leq a - 1$ such that $ab - a - b + s_1$ and $ab - a - b + s_2$ have the same remainder when divided by a . It follows that a must divide the difference $(ab - a - b + s_2) - (ab - a - b + s_1) = s_2 - s_1$. This is impossible since $0 < s_2 - s_1 < a$ and finishes our proof of Lemma 2.3.10. ■

Lemma 2.3.11. Let a and b be two natural numbers such that $\gcd(a, b) = 1$. The set of the remainders of the numbers $\{0 \cdot b, 1 \cdot b, \dots, (a-1) \cdot b\}$ when divided by a equals the set $\{0, 1, \dots, a-1\}$.

Proof: The proof is similar to the previous one. It is clear that the set of remainders of $\{0 \cdot b, 1 \cdot b, \dots, (a-1) \cdot b\}$ when divided by a is a subset of $\{0, \dots, a-1\}$. To show that equality happens, we prove that no two numbers in the set $\{0 \cdot b, 1 \cdot b, \dots, (a-1) \cdot b\}$ have the same remainder when divided by a . We use proof by contradiction and assume that there exist $0 \leq t_1 < t_2 \leq a-1$ such that $t_1 b$ and $t_2 b$ have the same remainders when divided by a . This means that $a \mid (t_2 - t_1)b$. Because $\gcd(a, b) = 1$, Proposition 2.3.4 implies that $a \mid t_2 - t_1$. This is impossible since $0 < t_2 - t_1 < a$ and finishes our proof. ■

From the previous two lemmas, we deduce that the set of remainders (when divided by a) of the numbers

$$ab - a - b, ab - a - b + 1, \dots, ab - a - b + a - 1$$

is the same as the set of remainders of the numbers

$$0 \cdot b, 1 \cdot b, \dots, (a-1) \cdot b.$$

Furthermore, notice that $(a-1)b - (ab - a - b) = a$ which implies that the remainders of $ab - a - b$ and $(a-1)b$ are the same when divided by a . This allows us to make the following argument. Let r be a natural number between 1 and $a-1$. There exists (exactly one) integer y_r between 0 and $a-2$ such that the remainders of the numbers $ab - a - b + r$ and $y_r \cdot b$ are the same. Hence,

$$a \mid (ab - a - b + r) - y_r \cdot b,$$

which implies that there exists an integer x_r such that

$$ab - a - b + r = x_r \cdot a + y_r \cdot b.$$

Note that $0 \leq y_r \leq a-2$ and that

$$\begin{aligned} x_r \cdot a &= (ab - a - b + r) - y_r \cdot b \geq (ab - a - b + r) - (a-2)b \\ &= b - a + r > 0. \end{aligned}$$

This implies that x_r is a natural number. Hence, $ab - a - b + r$ is (a, b) -achievable for any $1 \leq r \leq a-1$. Note also that $ab - a - b + a = ab - b = (a-1)b$ is also (a, b) -achievable. We have a list of a consecutive (a, b) -achievable numbers:

$$ab - a - b + 1, \dots, ab - a - b + a.$$

Using strong induction as in the previous examples will show that any natural number $n \geq ab - a - b + 1$ is (a, b) -achievable. We leave the details to the reader. ■

Note that one can figure out how many numbers less than $ab - a - b$ are (a, b) -achievable (see Exercise 2.3.7).

Our last application of Bézout's lemma and its consequences in this section deals with solving the Diophantine equation involving Pythagorean triples, namely determining all integers x, y and z such that

$$(2.3.14) \quad x^2 + y^2 = z^2.$$

Note that integer triples where one of the numbers x and y is 0 and the other two numbers have the same absolute value are solutions for the equation above. We call such solutions trivial, and we will try now to find the nontrivial solutions to this equation. Perhaps the reader knows some solutions already such as $(3, 4, 5)$ or $(4, 6, 8)$ or $(5, 12, 13)$ and we will try to explain how all such Pythagorean triples are obtained. We first prove the following useful lemma.

Lemma 2.3.12. *Let a, b , and c be integers. If $\gcd(a, b) = 1$ and $ab = c^2$, then both a and b are squares of integers, namely there exist integers c_1 and c_2 such that $a = c_1^2$, $b = c_2^2$, and $c = c_1 c_2$.*

Proof: We show that a is a square by proving that for every prime p that divides a , the exponent of p in the prime factorization of a given by Theorem 1.6.5 is even. Let p be a prime that divides a and assume that the power of p in the prime decomposition of a equals e . If e is even, then we are done. Otherwise, if e is odd, then we want to obtain a contradiction. Because $p \mid a$ and $ab = c^2$, we get that $p \mid c^2$. By Corollary 2.3.5, we

get that $p \mid c$. Let f denote the exponent of p in the prime decomposition of c . The exponent of p in the prime decomposition of c^2 must be $2f$. Since $ab = c^2$, we deduce that the exponent of p in the prime decomposition of b must equal $2f - e$. Because we assumed that e is odd, this implies that $2f - e$ is odd. Hence, p divides b . Therefore, we have obtained that $p \mid a$ and $p \mid b$ which is a contradiction with $\gcd(a, b) = 1$. ■

Before moving on to the Pythagorean triples, note that Lemma 2.3.12 can be generalized to higher powers (see Exercise 2.3.6). Also, note that the result would not hold in the absence of the hypothesis $\gcd(a, b) = 1$. For example, $2 \cdot 18 = 6^2$, but neither 2 nor 18 is a square of an integer.

As one may guess, the notion of the greatest common divisor of two integers can be extended to more than two integers as follows.

Definition 2.3.4. Let n be a natural number. If a_1, \dots, a_n are integers not all equal to zero, the **greatest common divisor of a_1, \dots, a_n** is defined as the largest integer that divides each of a_1, \dots, a_n . It is denoted by $\gcd(a_1, \dots, a_n)$.

It is not hard to see that $\gcd(a_1, \dots, a_n)$ is well-defined and is a natural number. The reason this notion is relevant for problem (2.3.14) is that if x, y , and z are integers such that $x^2 + y^2 = z^2$ and $d = \gcd(x, y, z)$, then $x/d, y/d$, and z/d are integers satisfying $(x/d)^2 + (y/d)^2 = (z/d)^2$. On the other hand, if one has a triple of integers a, b , and c such that $a^2 + b^2 = c^2$, then for any integer k , the integers ka, kb , and kc satisfy (2.3.14): $(ka)^2 + (kb)^2 = (kc)^2$. Also, since $a^2 = (-a)^2$, we can also restrict ourselves to solving (2.3.14) in natural numbers. Hence, it is sufficient to restrict ourselves to natural numbers satisfying $\gcd(x, y, z) = 1$ when solving the Pythagorean equation $x^2 + y^2 = z^2$.

Theorem 2.3.13. Let $x, y, z \in \mathbb{N}$ such that $\gcd(x, y, z) = 1$ and

$$x^2 + y^2 = z^2.$$

Then the following are true:

- (1) The number z is odd and exactly one of the numbers x and y is odd.
- (2) If x is even and y is odd, then there exist coprime natural numbers $m > n$ such that

$$x = 2mn$$

$$y = m^2 - n^2$$

$$z = m^2 + n^2.$$

Conversely, for any natural numbers m and n ,

$$(2mn)^2 + (m^2 - n^2)^2 = (m^2 + n^2)^2.$$

Proof: Suppose (x, y, z) is a Pythagorean triple of natural numbers satisfying $\gcd(x, y, z) = 1$. Because $\gcd(x, y, z) = 1$, at least one of the numbers x, y , and z is odd. If exactly one of the numbers x, y , and z is odd, then $x^2 + y^2 = z^2$ implies that either the sum of two even numbers is odd or that the sum of an odd number and an even number is even. Both situations are clearly impossible and therefore, we must have that at least two of the numbers x, y , and z are odd. If all of them

are odd and $x = 2x_1 + 1, y = 2y_1 + 1, z = 2z_1 + 1$, for some integers x_1, y_1 , and z_1 , then $x^2 + y^2 = z^2$ implies that $4x_1^2 + 4x_1 + 1 + 4y_1^2 + 4y_1 + 1 = 4z_1^2 + 4z_1 + 1$ which implies that $1 = 4(z_1^2 + z_1 - x_1^2 - x_1 - y_1^2 - y_1)$. This means that $4 \mid 1$ which is impossible. Hence, this situation cannot happen and exactly two of the numbers x, y , and z are odd. To prove the first part, we have to show that z is odd. Again, by contradiction assume that z is even and x and y are odd. As before, we write $x = 2x_1 + 1, y = 2y_1 + 1, z = 2z_1$ for some integers x_1, y_1 , and z_1 . The equation $x^2 + y^2 = z^2$ is the same as $4x_1^2 + 4x_1 + 1 + 4y_1^2 + 4y_1 + 1 = 4z_1^2$ which implies that $1 = 2(z_1^2 - x_1^2 - x_1 - y_1^2 - y_1)$. This means that 1 is an even number which is false. Hence, our assumption was false, and we conclude that z must be odd and exactly one of the numbers x and y is odd.

Assume that x is even and y and z are odd. The equation $x^2 + y^2 = z^2$ can be rewritten as

$$(2.3.15) \quad x^2 = z^2 - y^2 \Leftrightarrow \left(\frac{x}{2}\right)^2 = \frac{z+y}{2} \cdot \frac{z-y}{2}.$$

Note that $\frac{x}{2}, \frac{z+y}{2}$ and $\frac{z-y}{2}$ are all natural numbers because x is even, y and z are odd and $z > y$. We are now in a situation similar to the one in Lemma 2.3.12 except that we do not know what is $\gcd\left(\frac{z+y}{2}, \frac{z-y}{2}\right)$. We want to prove that we have $\gcd\left(\frac{z+y}{2}, \frac{z-y}{2}\right) = 1$. Assume by contradiction that this is not the case. Consider a prime p that divides both $\frac{z+y}{2}$ and $\frac{z-y}{2}$. Therefore, $p \mid \frac{z+y}{2} - \frac{z-y}{2} = y$ and $p \mid \frac{z+y}{2} + \frac{z-y}{2} = z$. Hence, p divides both y and z and must divide $z^2 - y^2 = x^2$. Because p is prime and divides x^2 , it must divide x by Corollary 2.3.5. Thus, p divides x, y , and z which contradicts our assumption that $\gcd(x, y, z) = 1$.

Now we can apply Lemma 2.3.12 in (2.3.15) and deduce that there exist natural numbers m and n such that

$$\begin{aligned} \frac{z+y}{2} &= m^2 \\ \frac{z-y}{2} &= n^2 \\ \frac{x}{2} &= mn. \end{aligned}$$

Note that m and n are coprime since $\frac{z+y}{2}$ and $\frac{z-y}{2}$ are coprime. Solving for x, y , and z , we get that

$$\begin{aligned} x &= 2mn \\ y &= m^2 - n^2 \\ z &= m^2 + n^2, \end{aligned}$$

which finishes this part of the proof.

The last part follows from the following calculation:

$$\begin{aligned} (m^2 - n^2)^2 + (2mn)^2 &= (m^2)^2 + (n^2)^2 - 2m^2n^2 + 4m^2n^2 \\ &= (m^2)^2 + (n^2)^2 + 2m^2n^2 \\ &= (m^2 + n^2)^2. \end{aligned}$$



We illustrate this theorem with examples of Pythagorean triples obtained for some special choices of m and n . As we see below, just because $\gcd(m, n) = 1$, that does not mean that $\gcd(x, y, z) = 1$ necessarily. Actually, one can prove that the solution $(x, y, z) = (2mn, m^2 - n^2, m^2 + n^2)$ given by Theorem 2.3.13 has the property that $\gcd(x, y, z) = 1$ if and only if m and n are coprime and not both odd. If m and n are both odd, then all the numbers $m^2 - n^2, 2mn$ and $m^2 + n^2$ are even. Dividing each of them by 2 yields a Pythagorean triple (x, y, z) with $\gcd(x, y, z) = 1$ when m and n are coprime. We leave these statements as exercises for the reader.

Table 2.3.3. Some small Pythagorean triples.

| m | n | x | y | z |
|-----|-----|-----|-----|-----|
| 2 | 1 | 4 | 3 | 5 |
| 3 | 1 | 6 | 8 | 10 |
| 4 | 1 | 8 | 15 | 17 |
| 3 | 2 | 12 | 5 | 13 |
| 5 | 1 | 10 | 24 | 26 |
| 4 | 3 | 24 | 7 | 25 |
| 6 | 1 | 12 | 35 | 37 |

A natural question is what happens when the exponent 2 from the Pythagorean equation $x^2 + y^2 = z^2$ is replaced by a larger natural number. Do we have a similar situation as above with infinitely many solutions? If so, then how would we obtain such solutions? For example, can we determine all the triples of integers (x, y, z) such that

$$x^3 + y^3 = z^3.$$

Note that the triples $(0, t, t)$ and $(t, 0, t)$ are solutions for any integer t . Also, because the power is odd, the triple $(t, -t, 0)$ is also a solution for any integer t . Such solutions in which one of our variables x, y or z is zero, are called **trivial** solutions. Are there any nontrivial solutions of the equation above? In general, let $n \geq 3$ be a natural number. We are interested in finding nontrivial solutions of the equation:

$$(2.3.16) \quad x^n + y^n = z^n.$$

This is one of the most famous equations in mathematics. It was settled in mid-1990s by Andrew Wiles, an English mathematician and professor at Princeton University who proved the following result.

Theorem 2.3.14 (Fermat's Last Theorem). *Let $n \geq 3$ be a natural number. The equation $x^n + y^n = z^n$ has no nontrivial integer solutions.*

The word “last” refers to the fact that it is the last conjecture or statement of Fermat to be proved or disproved. Around 1637, Pierre de Fermat (1607–1665) wrote a note in

Latin in his copy of Diophantus's *Arithmetica* stating that

*Cubum autem in duos cubos, aut quadrato quadratum in duos quadrato quadratos,
& generaliter nullam in infinitum ultra quadratum potestatem in duos ejusdem nominis
fas est divide re; cuius rei demonstrationem mirabilem sane detexi. Hanc marginis exigu-
itas non caperet.*

In case your Latin is a bit rusty as is ours, the English translation is

*It is impossible to separate a cube into two cubes, or a fourth power into two fourth
powers, or in general, any power higher than the second, into two like powers. I have
discovered a truly marvelous proof of this, which this margin is too narrow to contain.*

Fermat actually proved that the equation (2.3.16) has no nontrivial integer solutions when $n = 4$. He used what is now called the method of infinite descent or Fermat's method of descent which is a proof by contradiction method which shows that if a statement holds for a number, then it will be true for a smaller number, which would lead to an infinite descent and contradiction.

However, few mathematicians today believe that he actually had a proof for general $n \geq 3$. Nevertheless, Fermat's paragraph would set up an amazing chapter in the history of mathematics with contributions from hundreds of mathematicians over more than 350 years. Several important mathematicians handled various small values of n in the two centuries following Fermat's conjecture. Sophie Germain (1776–1831) was a French mathematician who made important contributions towards solving Fermat's conjecture. She corresponded with Gauss on topics related to number theory and Fermat's conjecture. Because of the prejudices of that era, Sophie Germain used a pseudonym *Monsieur Le Blanc* in her correspondence with Gauss and when she submitted her work for publication. Among other results, Sophie Germain proved that if p is a prime such that $2p + 1$ is also a prime, then Fermat's conjecture is true for p . Such primes p are called these days Sophie Germain primes. For example, 2, 3, 5, 11, and 23 are Sophie Germain primes. However, we do not know if there are infinitely many such primes.

Exercise 2.3.1. Let a and b be two integers with $b \neq 0$. If q and r are integers such that $a = bq + r$, then prove that $\gcd(a, b) = \gcd(b, r)$. This exercise shows that running the Euclidean algorithm for any two integers will produce the greatest common divisor similar to Section 1.6.

Exercise 2.3.2. Write down the steps of the Euclidean algorithm for calculating $\gcd(-23, -17)$, $\gcd(23, -17)$ and $\gcd(7596, -3242)$.

Exercise 2.3.3. A point (a, b) in the Cartesian plane is called a **lattice point** if both a and b are integers. How many lattice points are on the segment whose endpoints have coordinates $(0, 0)$ and $(12, -15)$? How many lattice points are on the segment whose endpoints have coordinates $(4, 7)$ and $(12, 15)$?

Exercise 2.3.4. Let (a, b) be a lattice point. What is the number of lattice points on the segment between $(0, 0)$ and (a, b) ? What is the number of lattice points on the segment between two lattice points of coordinates (a, b) and (c, d) , respectively?

Exercise 2.3.5. Solve the following Diophantine linear equations:

- (1) $23x + 17y = 9$.
- (2) $12x - 30y = 17$.
- (3) $12x - 30y = -8$.

Exercise 2.3.6. Let a , b , and c be integers such that $ab = c^k$ for some natural number k . If $\gcd(a, b) = 1$, prove that both a and b are k -th powers of integers, namely $a = c_1^k$, $b = c_2^k$ for some integers c_1 and c_2 such that $c = c_1c_2$.

Exercise 2.3.7. Let a and b be two coprime natural numbers. Show that there are $\frac{(a-1)(b-1)}{2}$ natural numbers between 1 and $ab - a - b$ that are not (a, b) -achievable.

Exercise 2.3.8. Show that any Pythagorean triple contains an integer divisible by the number 3.

Exercise 2.3.9. Prove that any Pythagorean triple contains an integer which is divisible by 5.

Exercise 2.3.10. Prove that the solution $(x, y, z) = (2mn, m^2 - n^2, m^2 + n^2)$ given by Theorem 2.3.13 has the property that $\gcd(x, y, z) = 1$ if and only if m and n are coprime and not both odd.

2.4. Congruences and Modular Arithmetic

The best way to understand congruences and modular arithmetic is by looking at a clock with 12 replaced by 0. If the clock is showing 9 in the morning now (yes, math reading often happens in the morning), then what time will it be 22 hours later? Clearly $9 + 22 = 31$, but we do not have 31 on the clock. However, the *clock arithmetic* is special, and we know that $12 = 0$ so in $9 + 22 = 9 + 12 + 3 + 7 = ((9 + 12) + 3) + 7$ meaning that $9 + 12$ will be 9 in the evening, $(9 + 12) + 3$ will be midnight and finally, $((9 + 12) + 3) + 7$ will be 7 in the morning the next day. Similarly, what time was it 22 hours ago? By a similar argument, the answer is $((9 - 9) - 12) - 1$ which can be calculated by noting that 9 hours ago, it was $9 - 9 = 0$ which is midnight, 21 hours ago, it was $(9 - 9) - 12$ which is noon and finally, 22 hours ago, it was 11 in the morning.

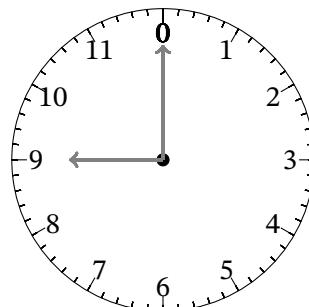


Figure 2.4.1. It is 9 in the morning, time to do some math!

So what is so special about *clock arithmetic*? We observe that $12 = 0$ in this realm and this observation leads us to identify 9 with 21, 33, -3 . But what is the common property of 9, -3 , 21 and 33? All the numbers 9, -3 , 21 and 33 have, when divided by 12, the remainder 9.

Let us make these facts more precise.

Definition 2.4.1. Let $n \in \mathbb{N}$. We say that two integers a and b are **congruent modulo n** which we write as $a \equiv b \pmod{n}$ if n divides $a - b$. The number $n \geq 1$ is called the **modulus of the congruence**. If $n \nmid a - b$, then we say that a and b are **not congruent modulo n** which we denote as $a \not\equiv b \pmod{n}$.

$$(a \equiv b \pmod{n}) \Leftrightarrow (n \mid a - b)$$

Example 2.4.1. As a simple illustration of the definition, we have that

$$\begin{aligned} 31 &\equiv 1 \pmod{10} & \text{because } 10 \mid 31 - 1, \\ 43 &\equiv 22 \pmod{7} & \text{because } 7 \mid 43 - 22, \\ -8 &\not\equiv 8 \pmod{3} & \text{because } 3 \nmid -8 - 8, \\ 91 &\not\equiv 18 \pmod{6} & \text{because } 6 \nmid 91 - 18. \end{aligned}$$

Another useful reformulation of the previous definition involves integer division. This is essentially the same result as Proposition 2.2.3 stated now using congruences. The proof is already given in Section 2.2 and will not be repeated here.

Proposition 2.4.1. Let n be a natural number. If a and b are integers, then $a \equiv b \pmod{n}$ if and only if a and b have the same remainder when divided by n .

$$\begin{aligned} &(\forall a, b \in \mathbb{Z})[(a \equiv b \pmod{n}) \\ &\quad \Leftrightarrow (\exists 0 \leq r < n, \exists k, \ell \in \mathbb{Z})(a = kn + r, b = \ell n + r)] \end{aligned}$$

The following result is an important consequence of Proposition 2.4.1.

Corollary 2.4.2. Let n be a natural number. For each $a \in \mathbb{Z}$ there exists exactly one $r \in \{0, 1, \dots, n - 1\}$ such that $a \equiv r \pmod{n}$.

$$(\forall a \in \mathbb{Z})(\exists! r \in \mathbb{Z}_n)(a \equiv r \pmod{n})$$

Proof: Take r as the remainder of the integer division of a by n . ■

Definition 2.4.2. Fix $n \geq 1$. For each $a \in \mathbb{Z}$ the (unique) $r \in \{0, 1, \dots, n - 1\}$ with $a \equiv r \pmod{n}$ is called the **modulo- n residue** of a . We denote by

$$\mathbb{Z}_n := \{0, 1, \dots, n - 1\}$$

the collection of all possible modulo- n residues. This is also called the **set of integers modulo n** or the **ring of integers modulo n** .

For the clock arithmetic in Figure 2.4.1, we do operations in the set of integers \mathbb{Z}_{12} modulo 12. The binary relation of congruence modulo n satisfies the following important properties.

Proposition 2.4.3. Let n be a natural number. If a, b, c are integers, then

- (1) $a \equiv a \pmod{n}$ [Reflexivity]
- (2) $a \equiv b \pmod{n}$ if and only if $b \equiv a \pmod{n}$ [Symmetry]
- (3) $a \equiv b \pmod{n}$ and $b \equiv c \pmod{n}$ imply $a \equiv c \pmod{n}$ [Transitivity]

Proof: Reflexivity can be proved by noticing that $n \mid a - a = 0$. For symmetry, note that if $a \equiv b \pmod{n}$, then n divides $a - b$. This implies that n divides $b - a = (-1) \cdot (a - b)$ and therefore $b \equiv a \pmod{n}$. Transitivity can be deduced by observing that $a \equiv b \pmod{n}$ and $b \equiv c \pmod{n}$ imply that $n \mid a - b$ and $n \mid b - c$. Hence, $n \mid (a - b) + (b - c) = a - c$ meaning that $a \equiv c \pmod{n}$. ■

The properties above imply that the relation of congruence modulo n is an **equivalence relation** on the set of integers. An equivalence relation is a binary relation on a given set that has these three properties: reflexive, symmetric and transitive. It is an important concept in mathematics. In the case of integers modulo n , each integer a gives rise to its equivalence class $\hat{a} = \{b \in \mathbb{Z} : b \equiv a \pmod{n}\}$ which consists of all the integers that are congruent to a modulo n . Because of the three properties mentioned above, the equivalence classes $\hat{0}, \dots, \hat{n-1}$ partition the set of integers (they are pairwise disjoint, and their union is \mathbb{Z}). For example, when $n = 2$, there are two equivalence classes $\hat{0}$ (consisting of all even integers) and $\hat{1}$ (consisting of all odd integers). We refer to Definition A.5.8 and the subsequent considerations for the general theory of equivalence relations and the generated classes.

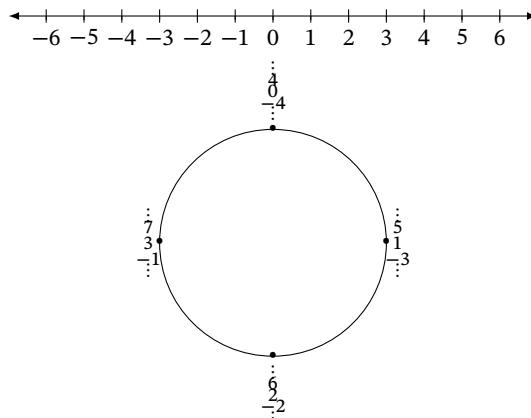


Figure 2.4.2. A representation of the integers mod 4.

Example 2.4.2. Let us consider again the question about adding or subtracting hours on a clock. Here we may use $n = 12$ and figure out that in 22 hours from 9 in the morning, the time will be $9 + 22 = 9 + 22 - 24 \equiv 7 \pmod{12}$ which is 7 (in the morning). Or which time will it be 14 hours after 9 in the morning? A possible slightly different way to answer this would be to use the 24 hours clock in which case there is no need for modular arithmetic for our example as $9 + 14 = 23$ and thus, the time

will be 23 which is 11 in the evening. For another example, when using a twelve hours clock and $n = 12$, we get that 15 hours ago (from 9 in the morning), the time was $9 - 15 = -6 \equiv 6 \pmod{12}$ which is 6 (in the evening). With a twenty-four hours clock and $n = 24$, the time was $9 - 15 = -6 \equiv 18 \pmod{24}$ which is 18, the same as 6 in the evening.

These examples show the importance not only of the integers a and b , but also of the natural number n . The same two numbers can be equal modulo n , but will not necessarily be equal modulo a different natural number m for example. Also, the examples above show that one may use addition and subtraction together with the modulo relation. This is actually true in general as the next result shows.

Proposition 2.4.4. *Let n be a natural number. If $a_1 \equiv a_2 \pmod{n}$ and $b_1 \equiv b_2 \pmod{n}$, then*

- (1) $a_1 + b_1 \equiv a_2 + b_2 \pmod{n}$ and $a_1 - b_1 \equiv a_2 - b_2 \pmod{n}$.
- (2) $a_1 b_1 \equiv a_2 b_2 \pmod{n}$.

Proof: If $a_1 \equiv a_2 \pmod{n}$ and $b_1 \equiv b_2 \pmod{n}$, then $n \mid a_1 - a_2$ and $n \mid b_1 - b_2$. Therefore, n divides $(a_1 - a_2) + (b_1 - b_2) = (a_1 + b_1) - (a_2 + b_2)$ and by the same argument n divides $(a_1 - a_2) - (b_1 - b_2) = (a_1 - b_1) - (a_2 - b_2)$. Hence, $a_1 + b_1 \equiv a_2 + b_2 \pmod{n}$ and $a_1 - b_1 \equiv a_2 - b_2 \pmod{n}$.

For the second part, because $n \mid a_1 - a_2$, we write $a_1 - a_2 = nr$ for some integer r . Similarly, because $n \mid b_1 - b_2$, we write $b_1 - b_2 = ns$ for some integer s . Thus, $a_1 = a_2 + nr$ and $b_1 = b_2 + ns$. Multiplying these two equations, we obtain that

$$a_1 b_1 = (a_2 + nr)(b_2 + ns) = a_2 b_2 + n(a_2 s + r b_2 + nrs)$$

and this implies $a_1 b_1 - a_2 b_2 = n(a_2 s + r b_2 + nrs)$. Hence, n divides $a_1 b_1 - a_2 b_2$ and $a_1 b_1 \equiv a_2 b_2 \pmod{n}$. ■

The following example shows that one benefit of congruences is reducing the calculations from larger numbers to small numbers.

Example 2.4.3. What are the last two digits of $32769 \cdot 4574$? To answer this question we use $n = 100$ as modulus of congruence. Then we get

$$32769 \cdot 4574 \equiv 69 \cdot 74 = 5106 \equiv 06 \pmod{100}.$$

Consequently, the last two digits of this product are 0 and 6.

Let us summarize some properties of modular addition and multiplication.

Proposition 2.4.5. *Let $n \geq 1$ be a fixed modulus of congruence. The following statements are true.*

- (1) *Modular addition and multiplication are associative.*

$$(\forall a, b, c \in \mathbb{Z})[(a + b) + c \equiv a + (b + c) \text{ and } a \cdot (b \cdot c) \equiv (a \cdot b) \cdot c \pmod{n}].$$

- (2) *Modular addition and multiplication are commutative.*

$$(\forall a, b \in \mathbb{Z})[a + b \equiv b + a \text{ and } a \cdot b \equiv b \cdot a \pmod{n}].$$

(3) *The cancellation law for modular addition is true.*

$$(\forall a, b, c \in \mathbb{Z})[(a + c \equiv b + c \pmod{n}) \Rightarrow ((a \equiv b \pmod{n}) \wedge (c \equiv 0 \pmod{n}))].$$

(4) *The distributive law is valid for modular addition and multiplication.*

$$(\forall a, b, c \in \mathbb{Z})[(a + b) \cdot c \equiv a \cdot c + b \cdot c \pmod{n}].$$

$$(5) \quad (\forall a \in \mathbb{Z})[a \cdot 0 \equiv 0 \text{ and } a \cdot 1 \equiv a \pmod{n}],$$

$$(6) \quad (\forall a, b \in \mathbb{Z}, k \geq 1)[(a \equiv b \pmod{n}) \Rightarrow (a^k \equiv b^k \pmod{n})].$$

Proof: For completeness, we include the proof of (6). The proofs of the remaining statements are left as exercises to the reader. We use induction on k . The result is obviously true for $k = 1$. Suppose now we already know $a^k \equiv b^k \pmod{n}$ provided that a and b are congruent. Write $a^{k+1} = a \cdot a^k$ and $b^{k+1} = b \cdot b^k$. Since we know that property (6) is valid for k , it follows that $a^k \equiv b^k \pmod{n}$. Now, by property (2) of Proposition 2.4.4 we conclude that also $a \cdot a^k \equiv b \cdot b^k \pmod{n}$. Thus, we took the step from k to $k + 1$ which shows that (6) is true for all $k \geq 1$. ■

Remark 2.4.1. Let n be a natural number. The addition properties above mean that the set $\mathbb{Z}_n = \{0, 1, \dots, n - 1\}$ of integers modulo n is a mathematical structure called a **group**. A group consists of a set (not necessarily finite as in the \mathbb{Z}_n case) with a binary operation that is associative, has a neutral element for that operation and such that every element has an inverse with respect to that operation. For example, the set of integers \mathbb{Z} with the binary operation of addition is a group, but the set of natural numbers \mathbb{N} with the same operation is not a group as not every natural number has an additive inverse.

Note that for any element $a \in \mathbb{Z}_n$, $\underbrace{a + \dots + a}_{n \text{ times}} \equiv 0 \pmod{n}$. This observation makes the following notion well-defined by the least element principle of \mathbb{N} .

Definition 2.4.3. Let n be a natural number. The **additive order in \mathbb{Z}_n** (or **modulo n**) of an element $a \in \mathbb{Z}_n$ is the smallest natural number k such that

$$\underbrace{a + \dots + a}_{k \text{ times}} \equiv 0 \pmod{n}.$$

Example 2.4.4. The additive order modulo n of 0 is 1, for any $n \in \mathbb{N}$.

Example 2.4.5. The additive order modulo n of 1 is n , for any $n \in \mathbb{N}$.

Example 2.4.6. Let $n = 6$. The additive order modulo 6 of 2 equals 3.

We give another simple and useful application of congruences below.

Proposition 2.4.6. Suppose a natural number m is written in base 10 as

$$m = x_k 10^k + x_{k-1} 10^{k-1} + \dots + x_1 10 + x_0,$$

for certain integers x_0, \dots, x_k in $\{0, \dots, 9\}$ and with $x_k \neq 0$. Then $3 \mid m$ if and only if $3 \mid x_0 + \dots + x_k$.

Proof: Since $10 \equiv 1 \pmod{3}$, for all $\ell \geq 1$ we also get $10^\ell \equiv 1 \pmod{3}$. Consequently,

$$m = x_k 10^k + x_{k-1} 10^{k-1} + \cdots + x_1 10 + x_0 \equiv x_k \cdot 1 + \cdots + x_1 \cdot 1 + x_0 \pmod{3},$$

and, hence,

$$3 \mid m \Leftrightarrow m \equiv 0 \pmod{3} \Leftrightarrow x_0 + \cdots + x_k \equiv 0 \pmod{3} \Leftrightarrow 3 \mid (x_0 + \cdots + x_k).$$

This completes the proof. ■

Example 2.4.7. The number $x = 3453$ is divisible by 3 because $3 + 5 + 4 + 3 = 15$ is divisible by 3. On the contrary, 3 does not divide 4672 because $2 + 7 + 6 + 4 = 19$ is not divisible by 3. But the proof of Proposition 2.4.6 gives some additional information also in this case. Namely, since $19 \equiv 1 \pmod{3}$ it also follows $4672 \equiv 1 \pmod{3}$.

Example 2.4.8. What is the last digit of 2^{2020} ? This problem can be rephrased in the language of congruences as finding r between 0 and 9 such that

$$2^{2020} \equiv r \pmod{10}.$$

We can start looking for patterns by examining the behavior of $2^k \pmod{10}$ for small values of k .

$$\begin{array}{lll} 2^0 \equiv 1 \pmod{10}, & 2^1 \equiv 2 \pmod{10}, & 2^2 \equiv 4 \pmod{10}, \\ 2^3 \equiv 8 \pmod{10}, & 2^4 \equiv 6 \pmod{10}, & 2^5 \equiv 2 \pmod{10}, \\ 2^6 \equiv 4 \pmod{10}, & 2^7 \equiv 8 \pmod{10}, & 2^8 \equiv 6 \pmod{10}. \end{array}$$

Our data seem to indicate that the positive powers of 2 modulo 10 repeat with a period of 4, namely $2^k \equiv 2^{k+4} \pmod{10}$ for $1 \leq k \leq 4$. A natural question is whether this pattern extends for any k greater than 4? The answer is yes and to see this, take k to be any natural number and note that $2^{k+4} - 2^k = 2^k(2^4 - 1) = 2^k \cdot 15$. Because k is positive, the number $2^k \cdot 15$ is divisible by 10 and therefore, $2^k \equiv 2^{k+4} \equiv 10 \pmod{10}$ for any positive k . This implies that $2^{k+4n} \equiv 2^k \pmod{10}$ for any natural numbers k and n . We can write $2020 = 4 \cdot 504 + 4$ and therefore, $2^{2020} \equiv 2^4 \equiv 6 \pmod{10}$.

Example 2.4.9. What are the last 3 digits of 3^{100} ? We can try to replicate the idea from the previous example and calculate the last three digits of small powers of 3 and look for patterns. As we will see later in this chapter, there is a pattern, but we may have to wait awhile before the three digits start repeating. An alternative approach is using the following fast powering algorithm. The first step is to write 100 in base 2 which can be done fairly easily using the methods from Section 1.5:

$$100 = 64 + 32 + 4 = 2^6 + 2^5 + 2^2.$$

Therefore,

$$(2.4.1) \quad 3^{100} = 3^{2^6+2^5+2^2} = 3^{2^2} \cdot 3^{2^5} \cdot 3^{2^6}.$$

We can compute each factor above modulo 1000 by repeated squaring and using the fact that

$$(3^{2^k})^2 = 3^{2^k \cdot 2} = 3^{2^{k+1}}.$$

We collect our computations in Table 2.4.1. We computed $3^{2^3} \pmod{1000}$ by first squaring $3^{2^2} = 81$ (which gives us $81^2 = 6561$) and then reducing the result modulo

Table 2.4.1. The fast powering algorithm computations.

| k | 0 | 1 | 2 | 3 | 4 | 5 | 6 |
|-----------------------|---|---|----|-----|-----|-----|-----|
| $3^{2^k} \pmod{1000}$ | 3 | 9 | 81 | 561 | 721 | 841 | 281 |

1000 to get 561. We then compute $3^{2^4} \pmod{1000}$ by squaring 561 (which gives us $561^2 = 314721$) and then reducing the result modulo 1000 to get 721. We leave it to the reader to verify the remaining two entries of the above table. To compute $3^{100} \pmod{1000}$, we plug in some values from Table 2.4.1 into equation (2.4.1) to get that

$$3^{100} = 3^{2^2} \cdot 3^{2^5} \cdot 3^{2^6} \equiv 81 \cdot 841 \cdot 281 = 19142001 = 001 \pmod{1000}.$$

The attentive reader has noticed that one operation is missing from the previous proposition, namely division. When using ordinary operations over integers or real numbers for example, we can divide by a number as long as that number is not 0. Also, when we perform division of a number a by a number b , we are actually multiplying a by $b^{-1} = \frac{1}{b}$ (the multiplicative inverse of b). Since multiplication in modular arithmetic works without any restrictions as seen in Proposition 2.4.4, we only have to figure out how to deal with the multiplicative inverse of an integer in modular arithmetic. Now the multiplicative inverse of a real number a is the real number b such that $a b = 1$. We can use a similar definition for modular arithmetic.

Definition 2.4.4. Let n be a natural number and a an integer. We say that a has a **multiplicative inverse** modulo n if there exists an integer b such that $a \cdot b \equiv 1 \pmod{n}$. A number $b \in \mathbb{Z}$ satisfying $a \cdot b \equiv 1 \pmod{n}$ is said to be a **(multiplicative) inverse** of $a \in \mathbb{Z}$.

It is easy to see that for any natural number $n \geq 2$, 0 does not have a multiplicative inverse modulo n while 1 has a multiplicative inverse modulo n . This is because $0 \cdot k \equiv 0 \not\equiv 1 \pmod{n}$ for any integer k and $1 \cdot 1 \equiv 1 \pmod{n}$. We do some examples with small values of n to figure out which other numbers in the set $\{0, 1, \dots, n-1\}$ of residues modulo n have an inverse modulo n .

Example 2.4.10. Let $n = 5$. Observe that $2 \cdot 3 \equiv 3 \cdot 2 = 6 \equiv 1 \pmod{5}$ and $4 \cdot 4 = 16 \equiv 1 \pmod{5}$. Every number 1, 2, 3, and 4 has a multiplicative inverse modulo 5.

Example 2.4.11. Consider $n = 12$. What integers in the set $\{0, 1, \dots, 11\}$ have a multiplicative inverse modulo 12? Assume that 2 has a multiplicative inverse modulo 12 and $2 \cdot k \equiv 1 \pmod{12}$ for some integer k . Then $12 \mid 2k - 1$. Since $2 \mid 12$, we deduce that 2 divides $2k - 1$ which would mean $2 \mid 1$, impossible. Hence, 2 does not have a multiplicative inverse modulo 12. We leave it as an exercise for the reader to apply similar arguments and compile the Table 2.4.2, where *no* means the corresponding number does not have multiplicative inverse modulo 12, otherwise the entry is a multiplicative inverse of the corresponding number.

Remark 2.4.2. The following is obvious:

$$(\forall a, b \in \mathbb{Z})[(b \text{ is inverse to } a \pmod{n}) \Leftrightarrow (a \text{ is so to } b \pmod{n})].$$

Table 2.4.2. The elements of \mathbb{Z}_{12} with a multiplicative inverse.

| | | | | | | | | | | | | |
|----------------------|----|---|----|----|----|---|----|---|----|----|----|----|
| a | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 |
| $a^{-1} \pmod{12}$? | no | 1 | no | no | no | 5 | no | 7 | no | no | no | 11 |

Why do some numbers possess a multiplicative inverse and other ones do not? The answer will be given in the next important result.

Proposition 2.4.7. *Let n be a natural number. An integer a has a multiplicative inverse modulo n if and only if a and n are coprime, that is $\gcd(a, n) = 1$.*

Proof: We first assume that a and n are coprime. By Proposition 2.3.2 (Bézout's lemma) there exist integers s and t such that $1 = s \cdot a + t \cdot n$. Because $n \equiv 0 \pmod{n}$, we get that

$$1 = s \cdot a + s \cdot 0 \equiv s \cdot a \pmod{n}.$$

Hence, s is a multiplicative inverse of a .

To prove the converse suppose that there is an integer $b \in \mathbb{Z}$ for which $ab \equiv 1 \pmod{n}$. Take some $d > 0$ with $d \mid a$ and $d \mid n$. Our aim is to show that then necessarily $d = 1$. By the choice of d , there are two nonzero integers k, ℓ such that $a = d k$ and $n = d \ell$. Consequently, in view of $ab \equiv 1 \pmod{n}$ it follows that

$$1 \equiv ab \equiv d k b \pmod{n},$$

which implies

$$\ell \equiv \ell a b \equiv \ell d k b \equiv n k b \equiv 0 \pmod{n}.$$

Therefore, $n \mid \ell$. Hence, $\ell \mid n$ lets us conclude that $n = \ell$. From $n = d \ell$ we obtain $d = 1$, which shows $\gcd(a, n) = 1$ as asserted. ■

We apply the previous result to congruences modulo a prime number, and we obtain the following important result.

Corollary 2.4.8. *If $p \geq 2$ is prime, then any $a \in \mathbb{Z}$ with $a \not\equiv 0 \pmod{p}$ possesses a multiplicative inverse modulo p .*

Proof: If $p \geq 2$ is prime, then for any $a \in \mathbb{Z}$ either $\gcd(a, p) = 1$ or $\gcd(a, p) = p$. The latter happens if and only if $p \mid a$ or, equivalently, if and only if $a \equiv 0 \pmod{p}$. Since we assumed $a \not\equiv 0 \pmod{p}$, it follows $\gcd(a, p) = 1$, hence by Proposition 2.4.7 a multiplicative inverse of a exists. ■

Another consequence of Proposition 2.4.7 is that it allows us to cancel factors that have a multiplicative inverse.

Corollary 2.4.9 (Cancellation Rule for Multiplication). *Let n be a natural number and a an integer such that $\gcd(a, n) = 1$. If b and c are integers such that $ab \equiv ac \pmod{n}$, then $b \equiv c \pmod{n}$.*

$$(\forall a, b, c \in \mathbb{Z})[(\gcd(a, n) = 1 \text{ and } ab \equiv ac \pmod{n}) \Rightarrow (b \equiv c \pmod{n})].$$

Proof: Because $\gcd(a, n) = 1$, a must have a multiplicative inverse a' modulo n . Multiplying both sides of the equation $ab \equiv ac \pmod{n}$ by a' , we deduce that $b \equiv c \pmod{n}$ as desired. ■

Remark 2.4.3. The proof of Proposition 2.4.7 shows us how we can determine efficiently a multiplicative inverse element of an integer a if it exists. We can use the Euclidean Algorithm to compute $\gcd(a, n)$. If it is not 1, then a has no multiplicative inverse modulo n . If $\gcd(a, n) = 1$, then the Euclidean Algorithm reversed will produce integers s and t such that $1 = s \cdot a + t \cdot n$. Thus, s is the multiplicative inverse of a modulo n .

Example 2.4.12. Suppose $n = 20$ and $a = 13$. Since 13 is prime, it follows that $\gcd(13, 20) = 1$. From $1 = (-3) \cdot 13 + 2 \cdot 20$, we get that -3 is inverse to 13 modulo 20. But $-3 \equiv 17 \pmod{20}$, hence 17 (as well as -3) is the multiplicative inverse to 13 modulo 20.

Example 2.4.13. Take $n = 72$ and $a = 43$. Check that $\gcd(43, 72) = 1$. Because $1 = (-5) \cdot 43 + 3 \cdot 72$, -5 is multiplicative inverse to 43 modulo 72. But $-5 \equiv 67 \pmod{72}$, hence another positive inverse of 43 is given by 67. This also follows from $43 \cdot 67 = 2881 = 40 \cdot 72 + 1$.

In the previous example we already used a property that we have not proved yet, namely that congruent integers possess congruent multiplicative inverses. We are going to verify this now.

Proposition 2.4.10. Let n be a natural number. Suppose that a_1 and a_2 are integers such that $a_1 \equiv a_2 \pmod{n}$. If a_1 has a multiplicative inverse modulo n , then so does a_2 and in that case, any multiplicative inverse of a_1 and any multiplicative inverse of a_2 are congruent modulo n .

Proof: By assumption $n \mid (a_2 - a_1)$, hence there is some $k \in \mathbb{Z}$ such that $a_2 = a_1 + k \cdot n$. Thus, $\gcd(a_1, n) = \gcd(a_2, n)$. In particular, a_1 and n are coprime if and only if this is so for a_2 and n . Proposition 2.4.7 lets us conclude that a_1 possesses a multiplicative inverse if and only if this is so for a_2 .

Suppose now that b_1 and b_2 are multiplicative inverse integers to a_1 and a_2 , respectively. Because of $a_1 \equiv a_2 \pmod{n}$ it follows that

$$b_1 = 1 \cdot b_1 \equiv b_2 \cdot a_2 \cdot b_1 \equiv b_2 \cdot a_1 \cdot b_1 \equiv b_2 \cdot 1 = b_2 \pmod{n}.$$

Hence, $b_1 \equiv b_2 \pmod{n}$ as asserted. ■

Remark 2.4.4. Let n be a natural number such that $n \geq 2$. Given an integer $a \in \mathbb{Z}$ that has a multiplicative inverse element b modulo n , it follows that

$$a \cdot b' \equiv 1 \pmod{n},$$

for all $b' \in \mathbb{Z}$ with $b' \equiv b \pmod{n}$. Thus, technically there is not only one inverse, but infinitely many integers b' whose multiplication by a gives 1 modulo n . For example, if $n = 9$, 7 is a multiplicative inverse to 4 modulo 9 since $7 \cdot 4 = 28 \equiv 1 \pmod{9}$, but many other numbers such as $\dots, -20, -11, -2, 7, 16, 25, \dots$ satisfy the property of being a multiplicative inverse to 4 modulo 9. However, any such number is congruent to 7 modulo 9 so 7 is the unique element in the set $\{0, 1, \dots, 8\}$ that is the multiplicative inverse of 4. Sometimes, this is written as $4^{-1} \equiv 7 \pmod{9}$. This notation is important in the higher level abstract algebra context where one studies structures as groups, rings, and fields.

Example 2.4.14. The set \mathbb{Z}_n of integers modulo n with the operations of addition and multiplication modulo n is an example of a **ring**. Table (2.4.3) contains the addition and multiplication tables for \mathbb{Z}_6 , the ring of integers modulo 6. The binary addition operation modulo 6 is associative, commutative with 0 as a neutral element ($0 + a = a + 0 \equiv a \pmod{6}$ for any $a \in \mathbb{Z}_6$) and with every element in \mathbb{Z}_6 possessing an additive inverse (for any $a \in \mathbb{Z}_6$, there is $b \in \mathbb{Z}_6$ such that $a + b = 0$). The binary multiplication operation modulo 6 is also associative, commutative with 1 as a neutral element ($a \cdot 1 = 1 \cdot a \equiv a \pmod{6}$ for any $a \in \mathbb{Z}_6$). There are distributivity laws as in Proposition 2.4.5. However, not every nonzero element has a multiplicative inverse (for example, there is no element $b \in \mathbb{Z}_6$ such that $2 \cdot b \equiv 1 \pmod{6}$).

Table 2.4.3. The addition and multiplication tables for \mathbb{Z}_6 .

| + | 0 | 1 | 2 | 3 | 4 | 5 |
|----------|----------|----------|----------|----------|----------|----------|
| 0 | 0 | 1 | 2 | 3 | 4 | 5 |
| 1 | 1 | 2 | 3 | 4 | 5 | 0 |
| 2 | 2 | 3 | 4 | 5 | 0 | 1 |
| 3 | 3 | 4 | 5 | 0 | 1 | 2 |
| 4 | 4 | 5 | 0 | 1 | 2 | 3 |
| 5 | 5 | 0 | 1 | 2 | 3 | 4 |

| · | 0 | 1 | 2 | 3 | 4 | 5 |
|----------|----------|----------|----------|----------|----------|----------|
| 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 1 | 0 | 1 | 2 | 3 | 4 | 5 |
| 2 | 0 | 2 | 4 | 0 | 2 | 4 |
| 3 | 0 | 3 | 0 | 3 | 0 | 3 |
| 4 | 0 | 4 | 2 | 0 | 4 | 2 |
| 5 | 0 | 5 | 4 | 3 | 2 | 1 |

Example 2.4.15. For any prime p , the set $\mathbb{Z}_p^* = \{1, \dots, p - 1\}$ of nonzero integers modulo p is also an example of a **group** with the operation of multiplication modulo p . We have seen this notion in Remark 2.4.1. The set \mathbb{Z}_p with the operations of addition and multiplication modulo p is an example of a special kind of ring called **field**. Table 2.4.4 contains the addition and multiplication tables for \mathbb{Z}_7 , the field of integers modulo 7.

Table 2.4.4. The addition and multiplication tables of \mathbb{Z}_7 .

| + | 0 | 1 | 2 | 3 | 4 | 5 | 6 |
|----------|----------|----------|----------|----------|----------|----------|----------|
| 0 | 0 | 1 | 2 | 3 | 4 | 5 | 6 |
| 1 | 1 | 2 | 3 | 4 | 5 | 6 | 0 |
| 2 | 2 | 3 | 4 | 5 | 6 | 0 | 1 |
| 3 | 3 | 4 | 5 | 6 | 0 | 1 | 2 |
| 4 | 4 | 5 | 6 | 0 | 1 | 2 | 3 |
| 5 | 5 | 6 | 0 | 1 | 2 | 3 | 4 |
| 6 | 6 | 0 | 1 | 2 | 3 | 4 | 5 |

| · | 0 | 1 | 2 | 3 | 4 | 5 | 6 |
|----------|----------|----------|----------|----------|----------|----------|----------|
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 1 | 0 | 1 | 2 | 3 | 4 | 5 | 6 |
| 2 | 0 | 2 | 4 | 6 | 1 | 3 | 5 |
| 3 | 0 | 3 | 6 | 2 | 5 | 1 | 4 |
| 4 | 0 | 4 | 1 | 5 | 2 | 6 | 3 |
| 5 | 0 | 5 | 3 | 1 | 6 | 4 | 2 |
| 6 | 0 | 6 | 5 | 4 | 3 | 2 | 1 |

We proved earlier in Theorem 1.6.4 that there are infinitely many primes. Except for 2, every prime number p must be odd and therefore, either $p \equiv 1 \pmod{4}$ or $p \equiv 3 \pmod{4}$. In the next result, we refine the proof of Theorem 1.6.4 to show the following result.

Theorem 2.4.11. *There are infinitely many primes p such that $p \equiv 3 \pmod{4}$.*

Proof: Assume that there exists a natural number n such that the list of primes congruent to 3 modulo 4 is: $3, p_1, \dots, p_n$. Consider the number $N := 4p_1 \dots p_n + 3$. Because each of p_1, \dots, p_n are greater than 3, $p_j \nmid N$ for any $1 \leq j \leq n$. Also, since $3 \nmid p_j$ for $1 \leq j \leq n$, we deduce that $3 \nmid N$. Hence, N is not divisible by any of our primes $3, p_1, \dots, p_n$. Hence, any prime that divides N must be congruent to 1 mod 4. Therefore, $N \equiv 1 \pmod{4}$. However, $N \equiv 3 \pmod{4}$ which gives us a contradiction. ■

It turns out that there are infinitely many primes congruent to 1 modulo 4 (see Exercise 2.8.10). However, the proof above can be adapted to other situations such as showing that there are infinitely many primes congruent to 5 modulo 6. Much more is true.

Theorem 2.4.12. *For any coprime natural numbers r and n , there are infinitely many primes p such that*

$$p \equiv r \pmod{n}.$$

This famous theorem was proved by the German mathematician Peter Gustav Lejeune Dirichlet (1805–1859) using complex analysis which is calculus with complex numbers.

We present now an application of modular arithmetic and congruences related to books and an area of mathematics called coding theory.

Example 2.4.16. Until 2007, each book was assigned an International Standard Book Number or ISBN for short. Such a code may look like 0–521–00601–5 for example. It represents a 10 digit word 0521006015. In general, an ISBN is a 10 digit word $x_1 \dots x_{10}$, where the digits x_1, \dots, x_{10} are the nonzero elements of \mathbb{Z}_{11} , the set of integers modulo 11, with the convention that 10 is represented by an X. The first 9 digits are assigned according to some rules related to the language and the publishing house. The last digit is a so-called check digit, and it depends on the digits x_1, \dots, x_9 by satisfying the following equation:

$$x_1 + 2x_2 + \dots + 9x_9 + 10x_{10} \equiv 0 \pmod{11}.$$

Note that 11 being prime implies that, given x_1, \dots, x_9 , there is a unique $x_{10} \in \mathbb{Z}_{11}$ satisfying the above equation.

The ISBN is designed to

- (1) detect any single error,
- (2) detect any double-error created by transposing two digits.

The process of detecting these errors is simple. For detecting a single error, assume that the word $y = y_1 \dots y_{10}$ is the same as the word $x = x_1 \dots x_{10}$ in every position except one. This means that $y_j = x_j$ for all j between 1 and 10 except for index ℓ , where $y_\ell \neq x_\ell$. Because $x_1 \dots x_{10}$ is a correct ISBN, we must have that

$$\sum_{k=1}^{10} kx_k \equiv 0 \pmod{11}.$$

Because $y_\ell \neq x_\ell$, we have that $y_\ell = x_\ell + a$ some integer a such that $a \not\equiv 0 \pmod{11}$. Then the calculation of the weighted sum of the digits of the word $y_1 \dots y_{10}$ gives that

$$\sum_{k=1}^{10} ky_k = \sum_{j:j \neq \ell} jx_j + \ell(x_\ell + a) = \sum_{k=1}^{10} kx_k + \ell \cdot a \equiv \ell \cdot a \pmod{11}.$$

As $\ell \not\equiv 0 \pmod{11}$ and $a \not\equiv 0 \pmod{11}$, we deduce that $\ell \cdot a \not\equiv 0 \pmod{11}$ and therefore, $y_1 \dots y_{10}$ is not a valid ISBN.

The process of detecting an interchange of digits is similar. Assume now that $x_1 \dots x_{10}$ is a valid ISBN and $y_1 \dots y_{10}$ is the same as it in all but two positions $j < \ell$, where $x_j = y_\ell$ and $x_\ell = y_j$. Clearly, if $x_j = x_\ell$, the word y is the same as the word x so we will work under the assumption that $x_j \neq x_\ell$. Because $x_1 \dots x_{10}$ is a valid ISBN, we have that

$$\sum_{k=1}^{10} kx_k \equiv 0 \pmod{11}.$$

Therefore,

$$\begin{aligned} \sum_{k=1}^{10} ky_k &\equiv \sum_{k=1}^{10} ky_k - \sum_{k=1}^{10} kx_k \\ &= jy_j + \ell y_\ell - jx_j - \ell x_\ell \\ &= jx_\ell + \ell x_j - jx_j - \ell x_\ell \\ &= (j - \ell)(x_\ell - x_j) \\ &\not\equiv 0 \pmod{11}, \end{aligned}$$

where in the last step we used that the product of nonzero elements of \mathbb{Z}_{11} is also nonzero modulo 11.

Since 2017, a 13 digit ISBN has been introduced. The word $x_1 \dots x_{13}$ has its digits in \mathbb{Z}_{11} with the digits x_1, \dots, x_{12} assigned according to various rules related to language, publishing house and so on. The last digit x_{13} is the unique number in \mathbb{Z}_{11} satisfying the equation:

$$(2.4.2) \quad \begin{aligned} x_1 + 3x_2 + x_3 + 3x_4 + x_5 + 3x_6 + x_7 + 3x_8 + x_9 \\ + 3x_{10} + x_{11} + 3x_{12} + x_{13} \equiv 0 \pmod{11}. \end{aligned}$$

A 10 digit ISBN is converted to a 13 digit ISBN by adding 978 at the beginning and recalculating the check digit x_{13} using equation (2.4.2). The ISBNs are particular examples of codes. A code is a finite set of words over some alphabet such that any two distinct codewords differ in several positions. The idea is that one can use these codewords to reliably transmit information over a communication channel that has noise (meaning that errors can appear due to the distance or other factors). A simple example is a repetition code of the form $\{00000, 11111\}$. If we wish to transmit a string of 0, 1 bits over some channel, then we can encode 0 as 00000 and 1 as 11111. If our message is 101, then we send 111110000011111. If at the other end of the channel, the receiver gets the message 001110010011001, then by analyzing the first bits 00111, the receiver will deduce that 11111 and therefore, 1 was sent as the first letter of the message. By a similar argument, the receiver deduces that 0 and 1 are the other letters

of the message. The take-home message from this example is that in order to prevent errors from distorting our message, we introduce redundancy (repeating each bit five times) in our message.

Definition 2.4.5. Let n be a natural number and \mathcal{A} be a finite alphabet. The **distance** $d(x, y)$ between two ordered n -tuples or words $x = (x_1, \dots, x_n)$ as well as $y = (y_1, \dots, y_n)$ whose entries/letters are from \mathcal{A} is defined as the number of positions where x and y are different:

$$d(x, y) = |\{j : x_j \neq y_j\}|.$$

Equivalently, $d(x, y)$ is the smallest number of changes one has to make to transform x into y (or y into x).

Definition 2.4.6. A q -ary (n, M, d) -code is a collection \mathcal{C} of M words of the same length n over some alphabet of length q such that

$$d = \min_{x \neq y \in \mathcal{C}} d(x, y).$$

When $q = 2$, such a code is commonly called binary code. The **length** of \mathcal{C} is n and the **size** of \mathcal{C} is M . The parameter d is called **the minimum distance** of \mathcal{C} .

Example 2.4.17. The code $\{00000, 11111\}$ is a binary $(5, 2, 5)$ -code.

In coding theory, it is of great interest to construct codes with large size and large minimum distance. One of the parameters of importance in coding theory is $A_q(n, d)$ which is defined as the maximum M such that there exists a q -ary (n, M, d) -code.

We now describe another application of modular congruences to coding theory via Latin squares.

Definition 2.4.7. Let n be a natural number. A **Latin square** L of order n is a $n \times n$ matrix whose entries come from a set with n elements, say $\{0, \dots, n - 1\}$, such that each row and each column is a permutation of these entries.

For the sake of convenience, we also label our rows and columns with the integers $0, \dots, n - 1$ so when discussing Latin squares L , we refer to their entry in row j and column k as $\ell_{j,k}$ for $0 \leq j, k \leq n - 1$. An equivalent characterization of a Latin square is that any element from our set of entries $\{0, \dots, n - 1\}$ appears exactly once in each row and each column. It should be fairly straightforward to see that we can construct a Latin square of order n for any natural number n . Actually, we have seen such examples earlier in the addition tables from Table 2.4.3 and Table 2.4.4. Actually, for any natural number n , the addition table of the integers modulo n will form a Latin square of order n . Obviously, these are not the only examples, and we leave it to the reader to construct other Latin squares.

We can construct a code from a Latin square L with entries $(\ell_{j,k})_{0 \leq j, k \leq n - 1}$ as follows. Write down all the ordered triples

$$(j, k, \ell_{j,k}), 0 \leq j, k \leq n - 1.$$

For example, when $n = 2$ and $L = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$, then we obtain the following: This is a

Table 2.4.5. A code from a 2×2 Latin square.

$$\begin{array}{l} 000, 011 \\ 101, 110 \end{array}$$

binary code of length 3 with minimum distance 2. Actually, it has the largest number of codewords one can construct with these parameters (see Exercise 2.8.7) and shows that $A_2(3, 2) = 4$.

So far so good, but let us try to push our luck a bit more.

Definition 2.4.8. Let n be a natural number and assume that $A = (a_{j,k})_{0 \leq j, k \leq n-1}$ and $B = (b_{j,k})_{0 \leq j, k \leq n-1}$ are two Latin squares of order n with the same entries $\{0, \dots, n-1\}$. We say that A and B are **orthogonal** if the set of ordered pairs

$$\{(a_{j,k}, b_{j,k}) \mid 0 \leq j, k \leq n-1\}$$

equals the set of all possible ordered pairs

$$\{0, \dots, n-1\} \times \{0, \dots, n-1\} = \{(u, v) \mid u, v \in \{0, \dots, n-1\}\}.$$

Informally, one can interpret the set of ordered pairs $\{(a_{j,k}, b_{j,k}) \mid 0 \leq j, k \leq n-1\}$ as placing the Latin square B on the Latin square A and gathering all the n^2 ordered pairs obtained in such way.

Example 2.4.18. When $n = 2$, we have two Latin squares:

$$A = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}, \quad B = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}.$$

The set of ordered pairs $\{(a_{j,k}, b_{j,k}) \mid 0 \leq j, k \leq 1\}$ is given by:

$$\begin{bmatrix} (0, 1) & (1, 0) \\ (1, 0) & (0, 1) \end{bmatrix}.$$

Since the ordered pair $(0, 0)$ or $(1, 1)$ does not appear on this list, we deduce that A and B are not orthogonal.

Example 2.4.19. Let $n = 3$ and consider the following Latin squares:

$$C = \begin{bmatrix} 0 & 1 & 2 \\ 1 & 2 & 0 \\ 2 & 0 & 1 \end{bmatrix}, \quad D = \begin{bmatrix} 0 & 1 & 2 \\ 2 & 0 & 1 \\ 1 & 2 & 0 \end{bmatrix}.$$

The set of ordered pairs $\{(c_{j,k}, d_{j,k}) \mid 0 \leq j, k \leq 2\}$ is the following:

$$\begin{bmatrix} (0, 0) & (1, 1) & (2, 2) \\ (1, 2) & (2, 0) & (0, 1) \\ (2, 1) & (0, 2) & (1, 0) \end{bmatrix}.$$

Since the collection of ordered pairs above is the same as $\{0, 1, 2\} \times \{0, 1, 2\}$, we deduce that C and D are orthogonal.

Before going into details about other orthogonal Latin squares, let us describe how to construct another code from the orthogonal Latin squares C and D above. Consider the code \mathcal{C} over the alphabet $\{0, 1, 2\}$ consisting of the following 9 words of length 4:

$$(j, k, c_{j,k}, d_{j,k}) : 0 \leq j, k \leq 2.$$

More explicitly, the codewords are the following (we omit the brackets and commas):

Table 2.4.6. A code from two orthogonal 3×3 Latin squares.

$$\begin{aligned} & 0000, 0111, 0222 \\ & 1012, 1120, 1201 \\ & 2021, 2102, 2210 \end{aligned}$$

One can check that any two codewords differ in at least 3 positions and that these 9 words form a 3-ary $(4, 9, 3)$ -code. Similar to before, one can prove that this code has the largest number of codewords among all 3-ary codes of length 4 and minimum distance 3 (see Exercise 2.8.8) which implies that $A_3(4, 3) = 9$.

We generalize the previous construction of two orthogonal Latin squares as follows.

Example 2.4.20. Let $n \geq 3$ be an odd natural number. Let $X = (x_{j,k})_{0 \leq j, k \leq n-1}$ be the $n \times n$ Latin square which is the addition table of the integers modulo n . Hence,

$$x_{j,k} \equiv (j + k) \pmod{n}, \forall 0 \leq j, k \leq n - 1.$$

Let $Y = (y_{j,k})_{0 \leq j, k \leq n-1}$ be the $n \times n$ matrix defined as:

$$y_{j,k} \equiv (2j + k) \pmod{n}, \forall 0 \leq j, k \leq n - 1.$$

We leave to the reader to figure out that when $n = 3$, the matrices X and Y are the same as the matrices C and D in Example 2.4.19. Below you can see these matrices when $n = 9$,

$$X = \begin{bmatrix} 0 & 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 \\ 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 0 \\ 2 & 3 & 4 & 5 & 6 & 7 & 8 & 0 & 1 \\ 3 & 4 & 5 & 6 & 7 & 8 & 0 & 1 & 2 \\ 4 & 5 & 6 & 7 & 8 & 0 & 1 & 2 & 1 \\ 5 & 6 & 7 & 8 & 0 & 1 & 2 & 3 & 4 \\ 6 & 7 & 8 & 0 & 1 & 2 & 3 & 4 & 5 \\ 7 & 8 & 0 & 1 & 2 & 3 & 4 & 5 & 6 \\ 8 & 0 & 1 & 2 & 3 & 4 & 5 & 6 & 7 \end{bmatrix}, \quad Y = \begin{bmatrix} 0 & 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 \\ 2 & 3 & 4 & 5 & 6 & 7 & 8 & 0 & 1 \\ 4 & 5 & 6 & 7 & 8 & 0 & 1 & 2 & 3 \\ 6 & 7 & 8 & 0 & 1 & 2 & 3 & 4 & 5 \\ 8 & 0 & 1 & 2 & 3 & 4 & 5 & 6 & 7 \\ 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 0 \\ 3 & 4 & 5 & 6 & 7 & 8 & 0 & 1 & 2 \\ 5 & 6 & 7 & 8 & 0 & 1 & 2 & 3 & 4 \\ 7 & 8 & 0 & 1 & 2 & 3 & 4 & 5 & 6 \end{bmatrix}.$$

It seems plausible that each of the matrices above is a Latin square. Exercise 2.4.10 should take care of dealing with X being a Latin square.

We prove now that Y is a Latin square for any odd n . Pick a row j between 0 and $n-1$. We show that the j -th row is a permutation of the elements $0, \dots, n-1$ of the set of integers modulo n . The entries of the j -th row are:

$$2j + k, 0 \leq k \leq n - 1,$$

modulo n of course. It is pretty straightforward to see that no two entries are equal since $2j + k \equiv 2j + k' \pmod{n}$ implies that $k \equiv k' \pmod{n}$. Since $0 \leq k \leq n - 1$

and no two entries of the form $2j + k$ are equal, it means that on the row j , we will see n distinct values from $0, \dots, n - 1$. Hence, the row j is a permutation of the entries $0, \dots, n - 1$. Note that our proof is essentially showing that for any $j \in \mathbb{Z}_n$, the function $f : \mathbb{Z}_n \rightarrow \mathbb{Z}_n, f(x) = x + 2j$ is a bijection.

Pick now a column k between 0 and $n - 1$. We prove that the k -th column is a permutation of the elements of \mathbb{Z}_n . The entries of the k -th column are:

$$2j + k, 0 \leq j \leq n - 1,$$

modulo n . It is less obvious that these entries should be distinct. Let us assume that $2j + k \equiv 2j' + k \pmod{n}$ for some $j, j' \in \mathbb{Z}_n$. Subtracting k from both sides, we get that $2j \equiv 2j' \pmod{n}$. Now we use the fact that n is odd as it means that $n = 2t - 1$ for some natural number t . This implies that t is the multiplicative inverse of 2 modulo n as $2t - 1 = n \equiv 0 \pmod{n}$. From $2j \equiv 2j' \pmod{n}$, multiplying both sides by t , we get that $j \equiv j' \pmod{n}$. This proves that the entries of the k -th column are distinct. As before, this means that the k -th column is also a permutation of the elements of \mathbb{Z}_n . This completes the proof that Y is a Latin square.

Even for a small number like $n = 9$, attempting to do a proof that X and Y are orthogonal by superimposing the entries of Y on the entries of X seems like a tedious task. Perhaps the reader will be convinced again about the power of mathematical proofs going over the following argument. We will prove that the ordered pairs

$$(x_{j,k}, y_{j,k}) : 0 \leq j, k \leq n - 1,$$

are all distinct. Let us assume that there are two pairs of coordinates (a, b) and (s, t) such that

$$(x_{a,b}, y_{a,b}) = (x_{s,t}, y_{s,t}).$$

Using the definitions for X and Y , this means that

$$a + b \equiv s + t \pmod{n}$$

$$2a + b \equiv 2s + t \pmod{n}.$$

Subtracting the first equation from the second equation, we get that $a \equiv s \pmod{n}$. Combining this with our first equation above, we also get that $b = t$. Hence, $(a, b) = (s, t)$ which proves our assertion.

An alternative proof for the fact that X and Y are orthogonal is the following. We will show that for any ordered pair $(\alpha, \beta) \in \{0, \dots, n - 1\} \times \{0, \dots, n - 1\}$, there exists a unique ordered pair $(j, k) \in \{0, \dots, n - 1\} \times \{0, \dots, n - 1\}$ such that

$$(x_{j,k}, y_{j,k}) = (\alpha, \beta).$$

Let $(\alpha, \beta) \in \{0, \dots, n - 1\} \times \{0, \dots, n - 1\}$. Finding (j, k) as above is equivalent to solving the system:

$$j + k \equiv x_{j,k} \equiv \alpha \pmod{n}$$

$$2j + k \equiv y_{j,k} \equiv \beta \pmod{n}.$$

Subtracting the first equation from the second one, we get that $j \equiv \beta - \alpha \pmod{n}$. From the first equation above, we get that $k \equiv \alpha - j \equiv 2\alpha - \beta \pmod{n}$. This proves our assertion and shows that X and Y are orthogonal.

The notion of orthogonal Latin squares can be extended as follows.

Definition 2.4.9. Let n be a natural number. For $k \geq 1$, a collection A_1, \dots, A_k of $n \times n$ Latin squares is said to be **mutually orthogonal** if any two different Latin squares in our collection are orthogonal.

Define $N(n)$ as the maximal k for which a collection of k mutually orthogonal $n \times n$ Latin squares exists. The discussion in Example 2.4.18 implies that $N(2) = 1$. Also, when $n = 3$, Example 2.4.19 gives us that $N(3) \geq 2$. Actually, this is the best possible.

Theorem 2.4.13. For any natural number n , $N(n) \leq n - 1$.

Proof: Note that for any $n \times n$ Latin square L whose entries are $0, \dots, n - 1$, we can permute the entries $0, \dots, n - 1$ such that the first row is the ordered n -tuple $(0, \dots, n - 1)$ and the resulting $n \times n$ is also a Latin square. For example, when $L = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$, switching

0 with 1 gives $\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$. When $L = \begin{bmatrix} 2 & 0 & 1 \\ 0 & 1 & 2 \\ 1 & 2 & 0 \end{bmatrix}$ applying the permutation $\begin{pmatrix} 0 & 1 & 2 \\ 1 & 2 & 0 \end{pmatrix}$ to

its entries we get the matrix $\begin{bmatrix} 0 & 1 & 2 \\ 1 & 2 & 0 \\ 2 & 0 & 1 \end{bmatrix}$ which is a Latin square.

Let $k = N(n)$ and A_1, \dots, A_k be a collection of mutually orthogonal $n \times n$ Latin squares. For each A_j , we permute its entries such that the first row is $(0, \dots, n - 1)$. Note that this actually preserves the orthogonality property between any two Latin squares, so we obtain another sequence B_1, \dots, B_k of mutually orthogonal Latin squares. The $(0, 0)$ -th entry for each of them is 0. Remove the first row and the first column of B_1, \dots, B_k . For each $1 \leq j \leq k$, in the remaining $(n - 1) \times (n - 1)$ -subsquare of B_j , there are exactly $n - 1$ entries equal to 0. For $j \neq \ell$, no 0 entry from B_j can have the same coordinates as a 0 entry from B_ℓ because B_j and B_ℓ are orthogonal. Thus, we have $n - 1$ entries from the $(n - 1) \times (n - 1)$ subsquares of B_1, \dots, B_k and therefore $k(n - 1)$ in total such that no two occupy the same coordinates. Since there are $(n - 1)^2 = (n - 1) \times (n - 1)$ coordinates, we get that $k(n - 1) \leq (n - 1)^2$ and therefore, $k \leq n - 1$, finishing our proof.

Another proof may be obtained similarly as follows. If A_1, \dots, A_k are a collection of mutually orthogonal $n \times n$ Latin squares whose first rows are $(0, 1, \dots, n - 1)$ as above, then consider the entries in the second row and first column in each of A_1, \dots, A_k . These entries have to be distinct (why?) and cannot equal 0 (why?). Therefore, $k \leq n - 1$. ■

Example 2.4.20 implies that $N(n) \geq 2$ for any odd n . Actually, $N(n) \geq 2$ for any $n \neq 2, 6$. When $n = 6$, the fact that $N(6) = 1$ was first observed by Euler, who in 1779, studied the *36 officers problem*:

Six different regiments have six officers, each one having a different rank. Can these 36 officers be arranged in a square formation so that each row and column contains one officer of each rank and one of each regiment?

Euler believed that $N(6) = 1$ and convinced himself of this fact, but a proof did not appear until 1900 when the French mathematician Gaston Tarry (1843–1913) analyzed all possible cases and showed that $N(6) = 1$, meaning that there are no two orthogonal Latin squares of order 6. The story gets better. We saw earlier how one can construct two orthogonal Latin squares of order n when n is odd. Euler also figured out how to construct two orthogonal Latin squares of any order n when n is divisible by 4. The remaining congruence class was $n \equiv 2 \pmod{4}$. Based on the fact that $N(2) = N(6) = 1$, Euler conjectured that there are no orthogonal Latin squares of order n whenever $n \equiv 2 \pmod{4}$. In 1960, Bose, Parker, and Shrikhande (cf. [2]) disproved this conjecture and constructed orthogonal Latin squares of order n for any $n \equiv 2 \pmod{4}$, $n \geq 10$. Hence, the only values of n where $N(n) = 1$ are 2 and 6. Even strong mathematicians such as Euler can be sometimes wrong so don't beat yourself up for not solving a problem the first time you try it or for getting a wrong answer sometimes. We conclude this part with the following important result.

Theorem 2.4.14. *For any prime p , $N(p) = p - 1$.*

Proof: We have seen that $N(n) \leq n - 1$ for any n . We use a method similar to the one from Example 2.4.20 to construct $p - 1$ mutually orthogonal Latin squares A_1, \dots, A_{p-1} . This will imply that $N(p) = p - 1$. For any $t \in \{1, \dots, p - 1\}$, define A_t as the $p \times p$ matrix whose rows and columns are indexed by the elements of \mathbb{Z}_p whose entries are defined as:

$$A_t(j, k) \equiv tj + k \pmod{p}, 0 \leq j, k \leq p - 1.$$

As before, A_1 is just the addition table of \mathbb{Z}_p and when $p \geq 3$, A_2 is the same construction as the matrix Y from Example 2.4.20. When $p = 5$, we list the Latin squares obtained by this construction below:

$$A_1 = \begin{bmatrix} 0 & 1 & 2 & 3 & 4 \\ 1 & 2 & 3 & 4 & 0 \\ 2 & 3 & 4 & 0 & 1 \\ 3 & 4 & 0 & 1 & 2 \\ 4 & 0 & 1 & 2 & 3 \end{bmatrix}, \quad A_2 = \begin{bmatrix} 0 & 1 & 2 & 3 & 4 \\ 2 & 3 & 4 & 0 & 1 \\ 4 & 0 & 1 & 2 & 3 \\ 1 & 2 & 3 & 4 & 0 \\ 3 & 4 & 0 & 1 & 2 \end{bmatrix},$$

and

$$A_3 = \begin{bmatrix} 0 & 1 & 2 & 3 & 4 \\ 3 & 4 & 0 & 1 & 2 \\ 1 & 2 & 3 & 4 & 0 \\ 4 & 0 & 1 & 2 & 3 \\ 2 & 3 & 4 & 0 & 1 \end{bmatrix}, \quad A_4 = \begin{bmatrix} 0 & 1 & 2 & 3 & 4 \\ 4 & 0 & 1 & 2 & 3 \\ 3 & 4 & 0 & 1 & 2 \\ 2 & 3 & 4 & 0 & 1 \\ 1 & 2 & 3 & 4 & 0 \end{bmatrix}.$$

Again, it may be a tedious task to check any two Latin squares above to make sure they are orthogonal. Instead, we can give a proof as follows. First, each of the matrices A_1, \dots, A_{p-1} is a Latin square (see Exercise 2.8.16). Second, if $p \geq 3$ and $1 \leq t \neq t' \leq p - 1$, then

$$\begin{aligned} A_t(j, k) &\equiv tj + k \pmod{p} \\ A_{t'}(j, k) &\equiv t'j + k \pmod{p}. \end{aligned}$$

If there are two coordinate pairs (a, b) and (r, s) such that $A_t(a, b) = A_t(r, s)$ and $A_{t'}(a, b) = A_{t'}(r, s)$, then

$$\begin{aligned} ta + b &\equiv tr + s \pmod{p} \\ t'a + b &\equiv t'r + s \pmod{p}. \end{aligned}$$

Subtracting the first equation from the second one, we get that

$$(t' - t)a \equiv (t - t')r \pmod{p}.$$

Because $1 \leq t \neq t' \leq p - 1$, $t - t'$ is a nonzero element of \mathbb{Z}_p . Since p is prime, this implies that $t - t'$ has a multiplicative inverse x modulo p . Multiplying both sides of the above equation by x , we deduce that $a \equiv r \pmod{p}$. Combining this fact with $ta + b \equiv tr + s \pmod{p}$, we obtain that $b \equiv s \pmod{p}$. This finishes our proof.

An alternative proof of the orthogonality of A_t and $A_{t'}$ may be given using the last part of Example 2.4.20. We leave the details for the reader. ■

Note that with a bit more work involving using finite fields, one can show that $N(q) = q - 1$ for any prime power q . We will not prove this statement here, but we encourage the reader to figure out why $N(4) = 3$ (see Exercise 2.8.17). Determining when $N(n) = n - 1$ or how large can $N(n)$ be is an important research problem lying at the intersection of algebra, geometry and combinatorics.

Another important application of modular arithmetic is in cryptography. A typical scenario in this area is the following. Alice is trying to send a plain text message to Bob via a communication channel where Eve can eavesdrop and intercept the message. Alice will encrypt the plain text message into a cipher text with the hope that such cipher text cannot be decrypted by Eve, but can be decrypted by Bob. To connect letters with numbers, we can associate the 26 letters of the English alphabet: a,b,...,z with the numbers $0, 1, \dots, 25$ from the set of integers modulo 26. This enables us to perform arithmetic operations on the letters via the operations allowed in \mathbb{Z}_{26} .

Table 2.4.7. Letters as elements of \mathbb{Z}_{26} .

| a | b | c | d | e | f | g | h | i | j | k | l | m |
|----|----|----|----|----|----|----|----|----|----|----|----|----|
| 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
| n | o | p | q | r | s | t | u | v | w | x | y | z |
| 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 | 21 | 22 | 23 | 24 | 25 |

Example 2.4.21. The famous Caesar cipher from antiquity created a cipher text letter by shifting the plaintext letter by 3 which is the same as adding 3 to the element from \mathbb{Z}_{26} corresponding to the plaintext letter. We summarize the correspondence below. A plaintext message such as come here would be encrypted as the ciphertext FRPH.

Table 2.4.8. The Caesar cipher.

| a | b | c | d | e | f | g | h | i | j | k | l | m |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| D | E | F | G | H | I | J | K | L | M | N | O | P |
| n | o | p | q | r | s | t | u | v | w | x | y | z |
| Q | R | S | T | U | V | W | X | Y | Z | A | B | C |

KHUH. The plaintext message `work harder` would be encrypted as ZRUN KDUGHU. That seems pretty straightforward. How hard is it to decrypt this cipher? For example, if Eve sees PRYH OHIW, how could she decrypt it? The solution would be to shift each letter backward by 3 positions and get `move left`.

Of course, one can create similar encryption schemes by using other numbers instead of 3. However, in each situation, the system would be easy to decrypt by Eve as she could take an intercepted encrypted message and try all possible 25 shifts until she would get a valid message.

Exercise 2.4.1. For each expression below, find the integer between 0 and $n - 1$ that is equal to it, where n is the modulus of congruence:

- (1) $1743 + 2341 \pmod{2021}$
- (2) $7224 - 9242 \pmod{5001}$
- (3) $523 \cdot 543 \pmod{217}$
- (4) $1351 \cdot 2315 \cdot 1582 \cdot 3051 \cdot 5310 \pmod{2021}$
- (5) $123^2 \pmod{200}$
- (6) $459^7 \pmod{501}$

Exercise 2.4.2. Using congruences, prove that a natural number is divisible by 9 if and only if the sum of its digits in its decimal expansion is divisible by 9.

Exercise 2.4.3. Using congruences, prove that a natural number is divisible by 4 if and only if the number formed by the last two digits of its decimal expansion is divisible by 4.

Exercise 2.4.4. Determine the last three digits of k^{2022} for any $2 \leq k \leq 9$.

Exercise 2.4.5. Write down the addition as well as the multiplication tables for \mathbb{Z}_{12} and \mathbb{Z}_{13} . Determine which elements have additive inverses and which ones have multiplicative inverses.

Exercise 2.4.6. Find the inverse of every nonzero element in \mathbb{Z}_{17} . Which elements in \mathbb{Z}_{18} have a multiplicative inverse?

Exercise 2.4.7. If (x, y, z) is a Pythagorean triple with $\gcd(x, y, z) = 1$, prove that z is odd and exactly one of x and y is odd.

Exercise 2.4.8. Show that $x_1 \dots x_{10}$ is a 10 digit ISBN if and only if

$$10x_1 + 9x_2 + \dots + 2x_9 + x_{10} \equiv 0 \pmod{11}.$$

Exercise 2.4.9. Verify that 0–521–00601–5, 0–486–60255–9, 81–85931–93–2 are valid ISBNs.

Exercise 2.4.10. For any natural number n , show that the addition table of the integers modulo n is a Latin square. Construct another Latin square of order n .

2.5. Modular Equations

In the realm of integers, we are familiar with solving linear equations. The aim of this section is to investigate linear equations involving congruences. What is meant by this? Fix $n \geq 1$ as modulus of congruence.

Problem 2.5.1. Given integers a and b with $a \not\equiv 0 \pmod{n}$, find integers $x \in \mathbb{Z}$ for which

$$(2.5.1) \quad a \cdot x \equiv b \pmod{n}.$$

We give some examples below.

Example 2.5.1. Consider first the equation $2x = 5$ over integers. This equation has no integer solution. Now let us change this problem and consider the equation: find all integers x such that

$$2x \equiv 5 \pmod{11}.$$

The idea in solving this equation is the same as the one we use in solving usual equations, namely trying to make the 2 in front of the x change to 1. To do that, we can try multiplying both sides of the equation by a suitable number, the multiplicative inverse of 2 modulo 11 (if it exists). In our situation, $\gcd(2, 11) = 1$, and therefore 2 has a multiplicative inverse modulo 11. It is not hard to see that $2 \cdot 6 = 6 \cdot 2 \equiv 1 \pmod{11}$. Multiplying the equation above by 6, we obtain that

$$x = 1 \cdot x \equiv (6 \cdot 2) \cdot x \equiv 6 \cdot (2 \cdot x) \equiv 6 \cdot 5 = 30 \equiv 8 \pmod{11}.$$

So we easily get

$$2x \equiv 5 \pmod{11} \Leftrightarrow x \equiv 8 \pmod{11},$$

and therefore the set of solutions of our equation consists of all integers congruent to 8 modulo 11.

Example 2.5.2. Let us give another example: find all integers x such that

$$2x \equiv 5 \pmod{6}.$$

In this case, $\gcd(2, 6) = 2$ and 2 does not have a multiplicative inverse modulo 6. We interpret the congruence above using divisibility: $6 \mid 2x - 5$. There exists an integer k such that $2x - 5 = 6k$ which means that $2x - 6k = 5$. This leads to a contradiction as the left-hand side is an even integer and the right-hand side is an odd integer. Hence, the equation $2x \equiv 5 \pmod{6}$ has no integer solutions.

Example 2.5.3. Find all integers x such that

$$2x \equiv 4 \pmod{6}.$$

Before solving this equation, we note the obvious fact that the equation $2x = 4$ has exactly one integer solution, namely $x = 2$. For the modular equation above, we remark first 2 does not have a multiplicative inverse modulo 6 and we cannot make it disappear by multiplication. We rewrite the equation as $6 \mid 2x - 4$ which means that $2x - 4 = 6\ell$ for some integer ℓ . Therefore, $x - 2 = 3\ell$ which is the same as $x \equiv 2 \pmod{3}$. It is not too difficult to show that

$$2x \equiv 4 \pmod{6} \Leftrightarrow x \equiv 2 \pmod{3},$$

meaning that the set of solutions of our equation consists of all integers congruent to 2 modulo 3.

The basic questions about the equation (2.5.1) are:

- (1) Are there integers $x \in \mathbb{Z}$ for which (2.5.1) is satisfied?
- (2) If so, how do we obtain **all** possible integer solutions?

The next proposition answers the first question and gives a partial answer to the second one. We emphasize that understanding the examples and methods described above is more important than memorizing the next result.

Proposition 2.5.2. *Let n be a natural number. Given integers $a \not\equiv 0 \pmod{n}$ and b , the equation*

$$(2.5.2) \quad a \cdot x \equiv b \pmod{n}$$

has solutions if a and n are coprime. In this case all solutions of (2.5.2) are given by

$$x \equiv bc \pmod{n}, \quad \text{where} \quad c \cdot a \equiv 1 \pmod{n}.$$

If a and n are not coprime, i.e., $d = \gcd(a, n) > 1$, then equation (2.5.2) has solutions if and only if $d \mid b$.

Proof: The case $\gcd(a, n) = 1$ is easy to handle. In this case, a has a multiplicative inverse c such that $a \cdot c = c \cdot a \equiv 1 \pmod{n}$. As in Example 2.5.1, multiplying both sides of the equation $ax \equiv b \pmod{n}$ by c will make a disappear:

$$a \cdot x \equiv b \Leftrightarrow x = 1 \cdot x \equiv c \cdot a \cdot x = c \cdot b \Leftrightarrow x \equiv bc \pmod{n}.$$

This implies that x satisfies (2.5.2) if and only if $x \equiv bc \pmod{n}$.

Suppose now $d = \gcd(a, n) > 1$ and $d \nmid b$. The aim is to show that then (2.5.2) has no solutions. We use a similar method as in Example 2.5.2. We use proof by contradiction and assume the contrary, namely that there exists a $x \in \mathbb{Z}$ for which $a \cdot x \equiv b \pmod{n}$. Then $n \mid ax - b$ which means that $ax - b = kn$ for some $k \in \mathbb{Z}$. Therefore,

$$b = ax - kn.$$

Since $d \mid a$ and $d \mid n$, it follows $d \mid (a \cdot x - k \cdot n)$, hence $d \mid b$ which contradicts our assumption. Hence, a solution $x \in \mathbb{Z}$ cannot exist.

Finally, let us assume that $d > 1$ and $d \mid b$. In this case, we follow the solution from Example 2.5.3. Let $a' = a/d$, $b' = b/d$ and $n' = n/d$. By assumption, a' and b' are integers and n' is a natural number. Moreover, (check this) $\gcd(a', n') = 1$. An integer $x \in \mathbb{Z}$ solves (2.5.2) if and only if there is an integer $k \in \mathbb{Z}$ for which

$$a \cdot x = b + k \cdot n$$

which, by dividing by d , is equivalent to

$$a' \cdot x = b' + k \cdot n'.$$

Consequently, x solves (2.5.2) if and only if it is a solution of

$$a' \cdot x \equiv b' \pmod{n'}.$$

But $\gcd(a', n') = 1$, hence by the first case this equation has solutions. This completes the proof. \blacksquare

Corollary 2.5.3. *If $d = \gcd(a, n) > 1$ and $d \mid b$, then all solutions of (2.5.2) are given by*

$$x \equiv c \cdot b' \quad \text{where} \quad c \cdot a' \equiv 1 \pmod{n'}$$

and $a' = a/d$, $b' = b/d$ and $n' = n/d$.

One can also solve systems of linear equations in modular arithmetic.

Example 2.5.4. Find all integers x and y such that

$$2x + 5y \equiv 3 \pmod{6} \text{ and } 3x + 4y \equiv 2 \pmod{6}.$$

When solving such systems in the usual arithmetic, we often use the substitution method. This uses one equation to express one variable in terms of the others and then substitutes such a formula in the other equations and reduces the number of variables. We can try to do the same approach here, but one has to be a bit careful. Say we start with $3x + 4y \equiv 2 \pmod{6}$, then neither 3 nor 4 has a multiplicative inverse modulo 6. We are therefore forced to dig a bit deeper and rewrite this equation as $3x + 4y = 6k + 2$ for some integer k . Because $3x = 6k + 2 - 4y$, we deduce that x must be divisible by 2 and therefore $x = 2x_1$ for some integer x_1 . Hence, $6x_1 + 4y = 3x + 4y = 6k + 2$ which means that $3x_1 + 2y = 3k + 2$. From here we obtain that $3 \mid 2y - 2$ which implies that $y - 1 = 3y_1$ for some integer y_1 . Hence, $x = 2x_1$ and $y = 3y_1 - 1$. Plugging this information back into the other equation, we get that $6x_1 + 5(3y_1 - 1) = 2x + 5y \equiv 3 \pmod{6}$ which implies that $21y_1 - 4 \equiv 3 \pmod{6}$. Hence, $3y_1 \equiv 1 \pmod{6}$. However, there are no solutions of this equation. Therefore, the system has no solutions.

A general method for solving linear systems is taught in linear algebra courses. The readers familiar with such methods can try them on other linear systems of equations in modular arithmetic.

Modular equations arise naturally in cryptography. We scratched the surface of this large important subject in the previous section. We will add more examples of some simple cryptosystems.

Example 2.5.5. The Caesar cipher from the previous section is a particular case of an affine cipher. Take two nonzero numbers a and b in \mathbb{Z}_{26} such that a has a multiplicative inverse (we will see in a second why this matters). The affine cipher with parameters a and b will encrypt any plain letter x (seen as an element of \mathbb{Z}_{26}) as the cipher letter C by the following rule:

$$C \equiv ax + b \pmod{26}.$$

When $a = 1$ and $b = 3$, this is the same as the Caesar cipher.

Let us take $a = 7$ and $b = 10$ for example. The letter a (corresponding to 0 in \mathbb{Z}_{26}) will be encrypted as $7 \cdot 0 + 10 \equiv 10 \pmod{26}$ (which is K using Table 2.4.7). The letter b (which is 1 $\in \mathbb{Z}_{26}$) will be encrypted as $7 \cdot 1 + 10 \equiv 17 \pmod{26}$ (which is R using Table 2.4.7). Similarly, the letter c will be encrypted as Y, d as G and so on. We think of the encryption as a function $f : \mathbb{Z}_{26} \rightarrow \mathbb{Z}_{26}$, $f(x) = 7x + 10 \pmod{26}$. To decrypt a given letter, say S, we have to determine what element $x \in \mathbb{Z}_{26}$ has the property

that $7x + 10 \equiv 18 \pmod{26}$. Using the previous work in this section, we can get that $7x \equiv 8 \pmod{26}$. To get x , we need to find the multiplicative inverse of 7 modulo 26. We perform the Euclidean algorithm forward and in reverse to get the following:

$$26 = 3 \cdot 7 + 5, \quad 7 = 1 \cdot 5 + 2, \quad 5 = 2 \cdot 2 + 1,$$

and

$$1 = 5 - 2 \cdot 2 = 5 - 2 \cdot (7 - 5) = 3 \cdot 5 - 2 \cdot 7 = 3 \cdot (26 - 3 \cdot 7) - 2 \cdot 7 = 3 \cdot 26 - 11 \cdot 7.$$

Hence, $1 \equiv (-11) \cdot 7 \equiv 15 \cdot 7 \pmod{26}$. From $7x \equiv 8 \pmod{26}$, multiplying both sides by 15 (or -11 if you prefer), then we get that $x \equiv 8 \cdot 15 = 120 \equiv 16 \pmod{26}$. Hence, S is decrypted as q.

Since $\gcd(7, 26) = 1$, the function $f : \mathbb{Z}_{26} \rightarrow \mathbb{Z}_{26}$, $f(x) = 7x + 10 \pmod{26}$ is bijective. To see this, we first prove that f is injective. If $f(x_1) = f(x_2)$, then $7x_1 + 10 \equiv 7x_2 + 10 \pmod{26}$ which gives $x_1 \equiv x_2 \pmod{26}$. In some sense, we have almost done the proof that f is surjective in the previous paragraph. Going over that proof, we can see that for any $y \in \mathbb{Z}_{26}$, there exists $x \in \mathbb{Z}_{26}$ (take $x \equiv 15(y - 10) \pmod{26}$) such that $f(x) \equiv y \pmod{26}$. What we just did is finding the inverse function of f which is $f^{-1} : \mathbb{Z}_{26} \rightarrow \mathbb{Z}_{26}$, $f^{-1}(y) \equiv 15(y - 10) \pmod{26}$.

It seems that the pattern of encryption is not so obvious as for the Caesar cipher, so perhaps this is a more difficult cipher to decrypt.

Before we give the answer to this question, let us analyze some possible situations. Assume that a plaintext message was encrypted via an affine cipher and that Eve managed to figure out that d is encrypted as U and m is encrypted as V. How can Eve figure the whole affine cipher? Using Table 2.4.7, this problem boils down to finding a and b in \mathbb{Z}_{26} such that

$$3a + b \equiv 20 \pmod{26} \text{ and } 12a + b \equiv 21 \pmod{26}.$$

The quick way to solve this problem is to subtract the first equation from the second one to obtain $9a \equiv 1 \pmod{26}$. Because $\gcd(9, 26) = 1$, 9 has a multiplicative inverse modulo 26. We can quickly use the Euclidean algorithm to figure it out:

$$26 = 2 \cdot 9 + 8, \quad 9 = 1 \cdot 8 + 1,$$

and the reverse

$$1 = 9 - 8 = 9 - (26 - 2 \cdot 9) = 3 \cdot 9 - 26.$$

Hence, $1 \equiv 3 \cdot 9 \pmod{26}$. Therefore, $a = 3$ and $b \equiv 20 - 3a = 20 - 9 = 11 \pmod{26}$. This completely determines the affine cipher. In conclusion, it seems that if Eve can figure the encryptions of two plaintext letters, then she can figure out the whole affine cipher and decrypt it as shown above. In practice, such a guess can be done efficiently using letter frequencies from the English alphabet. Moreover, such a guess can be skipped entirely and one can try all possible affine ciphers and their inverses to see which one produces a valid plaintext message. The number of affine ciphers is $26 \cdot 12 = 312$ as there are 26 choices for b and there are 12 choices of a since $\gcd(a, 26) = 1$. In conclusion, the affine cipher has a more theoretical and pedagogical value than a practical one.

The curious reader may ask the natural question about how to solve quadratic equations in modular arithmetic. We will not give the full details on how to solve such equations, but we will describe some simple examples.

Example 2.5.6. Consider the quadratic equation $x^2 = 0$ over the integers. It is clear that this equation has only one solution, namely $x = 0$. If we consider a similar equation in modular arithmetic, things will sometimes be similar and sometimes will be different. Consider the equation $x^2 \equiv 0 \pmod{5}$. This means that $5 \mid x^2$. Since 5 is a prime, we deduce that $5 \mid x$ and therefore, $x \equiv 0 \pmod{5}$. Hence, the only solution of our equation in \mathbb{Z}_5 is 0 which is the same as saying that the set of integer solutions of the equation $x^2 \equiv 0 \pmod{5}$ consists of all integers that are divisible by 5. The similarity to the integer equation $x^2 = 0$ is that the equation $x^2 \equiv 0 \pmod{5}$ has exactly one solution in \mathbb{Z}_5 . If phrased in terms of integer solutions, things are different and the set of integer solutions of the equation $x^2 \equiv 0 \pmod{5}$ is the set of integers $\{5k : k \in \mathbb{Z}\}$. Therefore, this modular equation is quite simple, but one has to be a bit careful when writing things down.

Example 2.5.7. Let us investigate the equation $x^2 \equiv 0 \pmod{9}$ in \mathbb{Z}_9 . As before, we can see that $9 \mid x^2$. However, 9 is not prime, and we cannot deduce that $9 \mid x$. Since $9 = 3^2$, one can prove that $9 \mid x^2$ implies that $3 \mid x$. Therefore, x can be 0, 3 or 6. One can easily check that each of these numbers satisfies $x^2 \equiv 0 \pmod{9}$. We have obtained quite a curious result, namely that a quadratic equation in the realm of modular arithmetic has 3 solutions. This is quite different from our integer or real numbers realm where a quadratic equation can have 0, 1, or 2 solutions.

Example 2.5.8. Let us try a similar equation $x^2 = 1$ over integers. One way to solve it is by moving everything to one side and getting that $(x - 1)(x + 1) = 0$. This implies that $x - 1 = 0$ or $x + 1 = 0$ and thus, 1 and -1 are the solutions. When we solve the modular equation $x^2 \equiv 1 \pmod{5}$, we can use a similar argument. We have that $5 \mid x^2 - 1 = (x - 1)(x + 1)$. Since 5 is a prime, it follows that $5 \mid x - 1$ or $5 \mid x + 1$. Hence, $x \equiv 1 \pmod{5}$ or $x \equiv -1 \pmod{5}$. Thus, there are two solutions in modular arithmetic as well.

Consider the equation $x^2 \equiv 1 \pmod{24}$ for example. This means, we have $24 \mid x^2 - 1 = (x - 1)(x + 1)$. Because 24 is not a prime, then we cannot deduce that $24 \mid x - 1$ or $24 \mid x + 1$.

Another approach here would be to try each element of \mathbb{Z}_{24} .

Example 2.5.9. Consider the equation $x^2 = -1$ over integers. Since $a^2 \geq 0$ for any real number a , our equation has no solutions. If we study this equation over integers modulo 5, $x^2 \equiv -1 \pmod{5}$, we get that $x^2 \equiv -1 \equiv 4 \pmod{5}$. This means that 5 divides $x^2 - 4 = (x + 2)(x - 2)$ and therefore, $5 \mid x + 2$ or $5 \mid x - 2$. In the first case, we get that $x \equiv -2 \equiv 3 \pmod{5}$ and in the second case, we get that $x \equiv 2 \pmod{5}$. Hence, the equation $x^2 \equiv -1 \pmod{5}$ has two solutions modulo 5: 2 and 3. What happens if we change the modulus of congruence? Let us study the equation $x^2 \equiv -1 \pmod{7}$. Our previous argument does not work as -1 is not congruent to a

square modulo 7. The number 7 is fairly small, and we can calculate the square of each element in \mathbb{Z}_7 :

Table 2.5.1. The squares of the elements of \mathbb{Z}_7 .

| | | | | | | | |
|-------|---|---|---|---|---|---|---|
| a | 0 | 1 | 2 | 3 | 4 | 5 | 6 |
| a^2 | 0 | 1 | 4 | 2 | 2 | 4 | 1 |

Table 2.5.1 shows that $a^2 \in \{0, 1, 2, 4\}$ for any $a \in \mathbb{Z}_7$. Since none of the elements 0, 1, 2 or 4 equals $-1 \equiv 6 \pmod{7}$, we deduce that the equation $x^2 \equiv -1 \pmod{7}$ has no solutions.

Exercise 2.5.1. Prove that for any natural number $n \geq 2$, $(n-1)^2 \equiv 1 \pmod{n}$. If $k \in \{1, \dots, n-1\}$, then show that $(n-1)(n-k) \equiv k \pmod{n}$.

Exercise 2.5.2. Solve the following modular equations:

- $8x \equiv 3 \pmod{13}$.
- $10x \equiv 5 \pmod{4}$.
- $9x \equiv 6 \pmod{12}$.
- $24x \equiv 12 \pmod{36}$.

Exercise 2.5.3. Find all the integers x between 0 and 100 such that

$$3x \equiv 2 \pmod{5}.$$

Exercise 2.5.4. Find all the integers x and y such that

$$2x + 3y \equiv 4 \pmod{5},$$

$$3x + 4y \equiv 1 \pmod{5}.$$

Exercise 2.5.5. Find all the integers x, y , and z such that

$$x + y + z \equiv 3 \pmod{7},$$

$$x + 2y + 2z \equiv 4 \pmod{7},$$

$$x + 3y + 4z \equiv 6 \pmod{7}.$$

Exercise 2.5.6. Find a nonidentity affine cipher for which there exists a letter of the alphabet that is encrypted as itself.

Exercise 2.5.7. Give an example of a modular quadratic equation that has exactly 10 solutions.

Exercise 2.5.8. How many integers x between -200 and 200 are there such that $x^2 \equiv 0 \pmod{12}$?

Exercise 2.5.9. Let k be an integer. If $n \in \{3, 4\}$, prove that $k^2 \equiv 0 \pmod{n}$ or $k^2 \equiv 1 \pmod{n}$. Is the statement true for $n \in \{5, 6, 7, 8, 9\}$?

Exercise 2.5.10. Give proofs of Exercise 2.2.7 and Exercise 2.2.10 using congruences.

2.6. The Chinese Remainder Theorem

The starting point of the problem is the following question.⁴ A Chinese King wanted to count the number of a group of soldiers. Thus, he asked them to stand together in groups of three. Doing so, one soldier remained. After that, the soldiers stood together in groups of five. Then two soldiers remained. And, finally, he asked the soldiers to stand in groups of seven. Here four soldiers were not in a group of seven. The question is, what is the minimal number of soldiers such that this happens?

Transforming this problem into a mathematical question, it reads as follows.

Example 2.6.1. Find all integers x (or maybe the smallest positive one) satisfying

$$x \equiv 1 \pmod{3}, \quad x \equiv 2 \pmod{5}, \quad \text{and} \quad x \equiv 4 \pmod{7}.$$

Let us try to find a solution here. We will give the general solution later in this section, but the method we describe now is simple and useful. The main idea is to use the information from the first equation to the second equation and then use the information from the first and second equation to the last equation. More precisely, from $x \equiv 1 \pmod{3}$, we deduce that $x = 3k + 1$ for some integer k . We then substitute this formula into the second equation $x \equiv 2 \pmod{5}$ and obtain that $3k + 1 \equiv 2 \pmod{5}$ which is the same as $3k \equiv 1 \pmod{5}$. To find out more about k , we get rid of the 3 by multiplying both sides of the previous equation by 2 which is the multiplicative inverse of 3 modulo 5. Therefore, $k \equiv 6k = 2(3k) \equiv 2 \pmod{5}$. This means that $k = 5\ell + 2$ for some integer ℓ . Combining $x = 3k + 1$ and $k = 5\ell + 2$, we get that $x = 3(5\ell + 2) + 1 = 15\ell + 7$. We now plug this formula into the last equation $x \equiv 4 \pmod{7}$ to obtain that $15\ell + 7 \equiv 4 \pmod{7}$. This is the same as $\ell \equiv 4 \pmod{7}$ meaning that $\ell = 7m + 4$ for some integer m . Combining $x = 15\ell + 7$ with $\ell = 7m + 4$, we get that $x = 15(7m + 4) + 7 = 105m + 67$. We have proved that the set of integer solutions of our system above is contained in the set $\{105n + 67 : n \in \mathbb{Z}\}$. To show that actually equality happens, note that any number of the form $105n + 67$ will have the property that

$$\begin{aligned} 105n + 67 &\equiv 0 + 1 = 1 \pmod{3} \\ 105n + 67 &\equiv 0 + 2 = 2 \pmod{5} \\ 105n + 67 &\equiv 0 + 4 = 4 \pmod{7}. \end{aligned}$$

If we are interested in the smallest natural number satisfying our congruences, then 67 would be the answer. Thus, coming back to the original question, the Chinese King watches at least 67, but maybe also 172 or 277 or, in general, $67 + 105m$, $m \geq 0$, soldiers.

The general formulation of the problem is as follows.

⁴The earliest known statement of the theorem is by the Chinese mathematician Sun-tzu in the Sun-tzu Suan-ching in the 3rd century CE.

Problem 2.6.1. Given k pairwise coprime natural numbers n_1, \dots, n_k and integers a_1, \dots, a_k , find $x \in \mathbb{Z}$ for which all following modular equations are valid:

$$(2.6.1) \quad \begin{aligned} x &\equiv a_1 \pmod{n_1} \\ x &\equiv a_2 \pmod{n_2} \\ &\vdots \\ x &\equiv a_k \pmod{n_k} \end{aligned}$$

The question is whether there are any $x \in \mathbb{Z}$ satisfying the k equations in (2.6.1). If this is so, how do we get all solutions from a special one? The next theorem answers these questions.

Theorem 2.6.2 (Chinese Remainder Theorem). *Let a_1, \dots, a_k and n_1, \dots, n_k be as above. Set $N := n_1 \cdot n_2 \cdots n_k$. If n_1, \dots, n_k are pairwise coprime, then there exists exactly one solution x_0 of (2.6.1) satisfying $0 \leq x_0 < N$. Any other solution of (2.6.1) is congruent to x_0 modulo N .*

Proof: We first prove the result for $k = 2$. Thus, suppose we have two coprime integers n_1 and n_2 and we want to find an integer x for which

$$(2.6.2) \quad x \equiv a_1 \pmod{n_1} \quad \text{and} \quad x \equiv a_2 \pmod{n_2},$$

for given $a_1, a_2 \in \mathbb{Z}$. Because n_1 and n_2 are coprime, there exist inverse elements N_1 and N_2 of n_1 and n_2 modulo n_2 and modulo n_1 , respectively. That is

$$(2.6.3) \quad 1 \equiv n_1 N_1 \pmod{n_2} \quad \text{and} \quad 1 \equiv n_2 N_2 \pmod{n_1}.$$

Define $x \in \mathbb{Z}$ by

$$(2.6.4) \quad x = a_2 n_1 N_1 + a_1 n_2 N_2.$$

We claim now that x solves (2.6.2). Observe that $a_2 n_1 N_1 \equiv 0 \pmod{n_1}$ and $a_1 n_2 N_2 \equiv 0 \pmod{n_2}$. Hence, the equation (2.6.3) implies $x \equiv a_1 n_2 N_2 \equiv a_1 \cdot 1 \equiv a_1 \pmod{n_1}$ as well as $x \equiv a_2 n_1 N_1 \equiv a_2 \pmod{n_2}$. Thus, x is a solution of the two equations in (2.6.2).

It remains to show that x is unique modulo $n_1 n_2$. Given any other solution $x' \in \mathbb{Z}$ of (2.6.2), we have that

$$x' - x \equiv a_1 - a_1 \equiv 0 \pmod{n_1} \quad \text{as well as} \quad x' - x \equiv a_2 - a_2 \equiv 0 \pmod{n_2}.$$

That is, $n_1 | x' - x$ and $n_2 | x' - x$. Because n_1 and n_2 are coprime, we obtain $n_1 n_2 | x' - x$ (cf. Exercise 1.6.7), hence $x' - x \equiv 0 \pmod{n_1 n_2}$ and this completes the proof.

Let $k \geq 3$ now. Recall that $N = n_1 \dots n_k$ and for $1 \leq j \leq k$, define $m_j = N/n_j$. Because n_1, \dots, n_k are pairwise coprime, we get that $\gcd(n_j, m_j) = 1$ for each $1 \leq j \leq k$. This implies that for each $1 \leq j \leq k$, m_j has a multiplicative inverse N_j modulo n_j , that is $m_j N_j \equiv 1 \pmod{n_j}$. Define $x = a_1 m_1 N_1 + \dots + a_k m_k N_k$. From our definition of N_1, \dots, N_k , we deduce that $a_j m_j N_j \equiv a_j \pmod{n_j}$ for any $1 \leq j \leq k$. Because $n_j | N_\ell$ for any $1 \leq j \neq \ell \leq k$, we deduce that $a_\ell m_\ell N_\ell \equiv 0 \pmod{n_j}$ for any $1 \leq j \neq \ell \leq k$. Putting these things together, we get that for any $1 \leq j \leq k$, $x \equiv a_j \pmod{n_j}$. The uniqueness of the solution modulo N can be proved similarly to the case $k = 2$ and is left as an exercise. ■

Example 2.6.2. Let us use the general methods above for the example considered in the beginning of this section:

$$x \equiv 1 \pmod{3}, x \equiv 2 \pmod{5}, \text{ and } x \equiv 4 \pmod{7}.$$

In this case $k = 3$, $n_1 = 3$, $n_2 = 5$, $n_3 = 7$, $N = 105$, $m_1 = 35$, $m_2 = 21$, $m_3 = 15$. Now we need to find N_1 such that $m_1 N_1 \equiv 1 \pmod{n_1}$ which is the same as $35N_1 \equiv 1 \pmod{3}$ or $2N_1 \equiv 1 \pmod{3}$. Multiplying both sides by 2, we get that $N_1 \equiv 2 \pmod{3}$ so we can take $N_1 = 2$. We need to find N_2 such that $m_2 N_2 \equiv 1 \pmod{n_2}$ which is the same as $21N_2 \equiv 1 \pmod{5}$ or $N_2 \equiv 1 \pmod{5}$. We got lucky here with less work, so we can take $N_2 = 1$. Finally, we have to find N_3 such that $m_3 N_3 \equiv 1 \pmod{n_3}$ which is the same as $15N_3 \equiv 1 \pmod{7}$ or $N_3 \equiv 1 \pmod{7}$. Lucky here again, we can take $N_3 = 1$. By the previous theorem, our solution is

$$\begin{aligned} x &= a_1 m_1 N_1 + a_2 m_2 N_2 + a_3 m_3 N_3 \\ &= 1 \cdot 35 \cdot 2 + 2 \cdot 21 \cdot 1 + 4 \cdot 15 \cdot 1 \\ &= 70 + 42 + 60 \\ &= 172. \end{aligned}$$

Note that 67 was the solution that we obtained at the beginning of this section and that $172 \equiv 67 \pmod{105}$. Thus, our work here is correct even though the result we obtained is not between 0 and $N - 1 = 104$.

Corollary 2.6.3. *Let n_1 and n_2 be two natural numbers. If $\gcd(n_1, n_2) = 1$, then for any integers $a_1 \in \{0, \dots, n_1 - 1\}$ and $a_2 \in \{0, \dots, n_2 - 1\}$ there exists a unique $x \in \{0, \dots, n_1 n_2 - 1\}$ such that*

$$(2.6.5) \quad x \equiv a_1 \pmod{n_1} \quad \text{and} \quad x \equiv a_2 \pmod{n_2}.$$

Conversely, given any integer $x \in \{0, \dots, n_1 n_2 - 1\}$, there exists a unique pair of integers (a_1, a_2) , with $a_1 \in \{0, \dots, n_1 - 1\}$ and $a_2 \in \{0, \dots, n_2 - 1\}$ for which (2.6.5) is satisfied.

Proof: Given numbers a_1 and a_2 the existence of x follows by Theorem 2.6.2. Take as x the solution of the equations in (2.6.2). On the other hand, if we are given an integer $x \in \{0, \dots, n_1 n_2 - 1\}$, let a_1 and a_2 be the remainders of the division of x by n_1 and n_2 , respectively. Solve (2.6.2) with these two numbers a_1 and a_2 . Then, by uniqueness this solution has to be x , which completes the proof. ■

Example 2.6.3. Let $n_1 = 3$ and $n_2 = 5$. Then $x = 12 \in \mathbb{Z}_{15}$ corresponds to the pair $(a_1, a_2) = (0, 2) \in \mathbb{Z}_3 \times \mathbb{Z}_5$. That is, the solution of $x \equiv 0 \pmod{3}$ and $x \equiv 2 \pmod{5}$ is $x \equiv 12 \pmod{15}$. We summarize the correspondence between $\mathbb{Z}_{15} = \{0, 1, \dots, 14\}$ and $\mathbb{Z}_3 \times \mathbb{Z}_5 = \{0, 1, 2\} \times \{0, 1, 2, 3, 4\}$ in Table 2.6.1. The rows are indexed by $\{0, 1, 2\}$, the columns by $\{0, 1, 2, 3, 4\}$, and the (r, c) -th entry is the unique $x \in \{0, 1, \dots, 14\}$ such that $x \equiv r \pmod{3}$ and $x \equiv c \pmod{5}$.

Table 2.6.1. A bijective correspondence between \mathbb{Z}_{15} and $\mathbb{Z}_3 \times \mathbb{Z}_5$.

| $\mathbb{Z}_3 \times \mathbb{Z}_5$ | 0 | 1 | 2 | 3 | 4 |
|------------------------------------|----|----|----|----|----|
| 0 | 0 | 6 | 12 | 3 | 9 |
| 1 | 10 | 1 | 7 | 13 | 4 |
| 2 | 5 | 11 | 2 | 8 | 14 |

Remark 2.6.1. The condition $\gcd(n_i, n_j) = 1, 1 \leq i \neq j \leq k$ in Theorem 2.6.2 is important. Without it, the conclusion of Theorem 2.6.2 will not hold. To see this, take for example $k = 2, n_1 = 6, n_2 = 9$. Assume that we want to solve:

$$x \equiv 4 \pmod{6} \text{ and } x \equiv 6 \pmod{9}.$$

If a solution $x \in \mathbb{Z}$ existed, then this would imply that for some $k, \ell \in \mathbb{Z}$

$$x = 4 + 6k = 6 + 9\ell,$$

and therefore, $3\ell - 2k = -2/3$. This is impossible as $k, \ell \in \mathbb{Z}$ and therefore, the equation above has no solutions.

On the other hand, say we are interested in solving the system:

$$x \equiv 1 \pmod{6} \text{ and } x \equiv 4 \pmod{9}.$$

In this situation, each of 13 and 31 and 49 is a solution of the system, so the uniqueness of the solution fails in this case.

We finish this section with one more example comparing our two methods for solving the Chinese remainder theorem.

Example 2.6.4. Find all $x \in \mathbb{Z}$ satisfying

$$x \equiv 3 \pmod{5} \text{ and } x \equiv 7 \pmod{11}.$$

First Solution: The first equation tells us that $x = 3 + 5k$ for a certain $k \in \mathbb{Z}$. Plugging this into the second equation and using that the number 9 is the multiplicative inverse of 5 modulo 11, we get that

$$3 + 5k \equiv 7 \pmod{11} \Leftrightarrow 5k \equiv 4 \pmod{11} \Leftrightarrow k \equiv 36 \equiv 3 \pmod{11}.$$

Consequently, $k = 3 + 11\ell$ for some $\ell \in \mathbb{Z}$, which implies that

$$x = 3 + 5k = 3 + 5(3 + 11\ell) = 18 + 55\ell, \quad \ell \in \mathbb{Z}.$$

In particular, the smallest positive solution is $x = 18$.

Second Solution: We use now the approach presented in the proof of Theorem 2.6.2. In the notation of this result we have $k = 2, n_1 = 5, n_2 = 11, a_1 = 3$ and $a_2 = 7$. We also have that $N_1 = 9$ as $5 \cdot 9 \equiv 1 \pmod{11}$ and $N_2 = 1$ as $11 \cdot 1 \equiv 1 \pmod{5}$. Plugging these numbers into (2.6.4) we get the solution

$$x = a_2 n_1 N_1 + a_1 n_2 N_2 = 7 \cdot 5 \cdot 9 + 3 \cdot 11 \cdot 1 = 348 \equiv 18 \pmod{55}.$$

Thus, as above the general solution is $x = 18 + 55\ell$ for some $\ell \in \mathbb{Z}$.

We now give an example showing how the Chinese remainder theorem can be used to find last digits of large numbers.

Example 2.6.5. We would like to find out the last 2 digits of a very large number N . The only information that we have is that $N \equiv 3 \pmod{4}$ and $N \equiv 17 \pmod{25}$. The Chinese remainder theorem will help us find the answer here. As before, from $N \equiv 3 \pmod{4}$, we get that $N = 4k + 3$ for some natural number k . Plugging this value into the second congruence $N \equiv 17 \pmod{25}$, we deduce that $4k + 3 \equiv 17 \pmod{25}$ which means that $4k \equiv 14 \pmod{25}$. A short calculation using Euclid's algorithm will tell us that 19 is the multiplicative inverse of 4 modulo 25. Therefore, multiplying both sides $4k \equiv 14 \pmod{25}$ by 19, we deduce that $k \equiv 14 \cdot 19 = 266 \equiv 16 \pmod{25}$. Hence, $k = 25\ell + 16$ for some nonnegative integer ℓ . Thus,

$$n = 4k + 3 = 4(25\ell + 16) + 3 = 100\ell + 67.$$

This shows that the last two digits of n are 67.

Exercise 2.6.1. Find all integers x such that

$$x \equiv 1 \pmod{2} \text{ and } x \equiv 2 \pmod{5}.$$

Exercise 2.6.2. Determine all the integers x such that

$$2x \equiv 5 \pmod{7} \text{ and } 3x \equiv 8 \pmod{11}.$$

Exercise 2.6.3. Find all the integers x such that

$$x \equiv 2 \pmod{3}, \quad x \equiv 3 \pmod{5}, \quad \text{and} \quad x \equiv 2 \pmod{7}.$$

Exercise 2.6.4. Construct a table similar to Example 2.6.3 for

$$(n_1, n_2) \in \{(2, 5), (3, 4), (4, 5), (3, 7)\}.$$

Exercise 2.6.5. Determine the additive orders of the elements of \mathbb{Z}_4 .

Exercise 2.6.6. Define a binary operation on $\mathbb{Z}_2 \times \mathbb{Z}_2$ as follows: for any pairs (a, c) , $(b, d) \in \mathbb{Z}_2 \times \mathbb{Z}_2$, define their sum as:

$$(a, c) + (b, d) = ((a + b) \pmod{2}, (c + d) \pmod{2}).$$

- (1) Prove that this operation is associative, commutative, and $(a, c) + (0, 0) = (a, c)$ for any $(a, c) \in \mathbb{Z}_2 \times \mathbb{Z}_2$.
- (2) Show that for any $(a, c) \in \mathbb{Z}_2 \times \mathbb{Z}_2$, there exists a unique $(b, d) \in \mathbb{Z}_2 \times \mathbb{Z}_2$ such that $(a, c) + (b, d) = (0, 0)$.
- (3) For each $(a, c) \in \mathbb{Z}_2 \times \mathbb{Z}_2$, determine the smallest k such that

$$\underbrace{(a, c) + \dots + (a, c)}_{k \text{ times}} = (0, 0).$$

- (4) Compare⁵ the values above with the additive orders of the elements in \mathbb{Z}_4 .

Exercise 2.6.7. Write down the addition table of \mathbb{Z}_{15} .

Exercise 2.6.8. Define a binary operation on $\mathbb{Z}_3 \times \mathbb{Z}_5$ as follows: for any pairs (a, c) , $(b, d) \in \mathbb{Z}_3 \times \mathbb{Z}_5$, define their sum as

$$(a, c) + (b, d) = ((a + b) \pmod{3}, (c + d) \pmod{5}).$$

⁵This exercise shows that $\mathbb{Z}_2 \times \mathbb{Z}_2$ is a group with 4 elements that is different (formally, nonisomorphic) from the group $(\mathbb{Z}_4, +)$. See also Exercise A.8.14 for a general approach to the product of groups and compare Exercise A.8.13 about Klein's 4-group \mathbb{K}_4 .

- (1) Prove that this operation is associative, commutative, and that for any pair $(a, c) \in \mathbb{Z}_3 \times \mathbb{Z}_5$, $(a, c) + (0, 0) = (a, c)$.
- (2) Show that for any $(a, c) \in \mathbb{Z}_3 \times \mathbb{Z}_5$, there exists a unique $(b, d) \in \mathbb{Z}_3 \times \mathbb{Z}_5$ such that $(a, c) + (b, d) = (0, 0)$.
- (3) For each $(a, c) \in \mathbb{Z}_3 \times \mathbb{Z}_5$, determine the smallest k such that

$$\underbrace{(a, c) + \dots + (a, c)}_{k \text{ times}} = (0, 0).$$

Exercise 2.6.9. Using the correspondence in Table 2.6.3, compare the addition table of \mathbb{Z}_{15} from Exercise 2.6.7 with the addition table of $\mathbb{Z}_3 \times \mathbb{Z}_5$ from Exercise 2.6.8 above⁶.

Exercise 2.6.10. Consider the correspondence between the integers n between 0 and 29 and the triples (r_2, r_3, r_5) , where r_j is the remainder of the division of n by j for $j \in \{2, 3, 5\}$. Show that this is a bijective function.

2.7. Fermat and Euler Theorems

The aim of this section is to prove several important properties of prime numbers. Many results in mathematics start with concrete examples and calculations from which patterns often emerge. We start with some simple concrete examples that will lead us to some famous results of Fermat, Euler and others.

Example 2.7.1. We tabulate the powers of the nonzero elements of \mathbb{Z}_5 , \mathbb{Z}_7 and of \mathbb{Z}_{11} .

Table 2.7.1. The powers of the elements of \mathbb{Z}_5 .

| a | 1 | 2 | 3 | 4 |
|-------|----------|----------|----------|----------|
| a^2 | 1 | 4 | 4 | 1 |
| a^3 | 1 | 3 | 2 | 4 |
| a^4 | 1 | 1 | 1 | 1 |
| a^5 | 1 | 2 | 3 | 4 |

Table 2.7.2. The powers of the elements of \mathbb{Z}_7 .

| a | 1 | 2 | 3 | 4 | 5 | 6 |
|-------|----------|----------|----------|----------|----------|----------|
| a^2 | 1 | 4 | 2 | 2 | 4 | 1 |
| a^3 | 1 | 1 | 6 | 1 | 6 | 6 |
| a^4 | 1 | 2 | 4 | 4 | 2 | 1 |
| a^5 | 1 | 4 | 5 | 2 | 3 | 6 |
| a^6 | 1 | 1 | 1 | 1 | 1 | 1 |
| a^7 | 1 | 2 | 3 | 4 | 5 | 6 |

⁶The purpose of these exercises is to convince the reader that $(\mathbb{Z}_{15}, +)$ is the same group as $\mathbb{Z}_3 \times \mathbb{Z}_5$ with the binary operation above.

Table 2.7.3. The powers of the elements of \mathbb{Z}_{11} .

| a | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|-----------|
| a^2 | 1 | 4 | 9 | 5 | 3 | 3 | 5 | 9 | 4 | 1 |
| a^3 | 1 | 8 | 5 | 9 | 4 | 7 | 2 | 6 | 3 | 10 |
| a^4 | 1 | 5 | 4 | 3 | 9 | 9 | 3 | 4 | 5 | 1 |
| a^5 | 1 | 10 | 1 | 1 | 1 | 10 | 10 | 10 | 1 | 10 |
| a^6 | 1 | 9 | 3 | 4 | 5 | 5 | 4 | 3 | 9 | 1 |
| a^7 | 1 | 7 | 9 | 5 | 3 | 8 | 6 | 2 | 4 | 10 |
| a^8 | 1 | 3 | 5 | 9 | 4 | 4 | 9 | 5 | 3 | 1 |
| a^9 | 1 | 6 | 4 | 3 | 9 | 2 | 8 | 7 | 5 | 10 |
| a^{10} | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| a^{11} | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |

There are several patterns one may observe starting at these tables. We will focus on the rows and observe that in each case the top row equals the bottom row. This is also true in the case when $a = 0$. The following result is one of the most famous theorems in number theory. It is called is **Fermat's little theorem**⁷.

Theorem 2.7.1 (Fermat's Little Theorem). *Let $p \geq 2$ be a prime number. If $a \in \mathbb{Z}$, then*

$$(2.7.1) \quad a^p \equiv a \pmod{p}.$$

Proof: The result is true for $p = 2$ because $2 | a^2 - a = a(a - 1)$ for any integer a .

We assume $p \geq 3$ for the rest of proof. First, we observe that it suffices to verify (2.7.1) for nonnegative integers a . To see this, let us assume for a moment that we already proved (2.7.1) for $a \geq 0$. If $p \geq 3$, then p is odd, and since $-a > 0$ we get

$$(-a)^p - (-a) = -(a^p - a) \equiv 0 \pmod{p}.$$

Hence, if (2.7.1) is true for nonnegative integers, then it is true for all integers.

We will use induction on a to prove equation (2.7.1) for any nonnegative integer a . The base case of induction is $a = 0$ and the equation (2.7.1) is satisfied by trivial reason. Now suppose that the assertion is valid for some integer $a \geq 0$, namely that

$$(2.7.2) \quad a^p \equiv a \pmod{p}.$$

We want to prove that it is also true for $a + 1$ meaning that

$$(a + 1)^p \equiv a + 1 \pmod{p}.$$

To this end we use the binomial formula from Theorem 1.7.6:

$$(2.7.3) \quad (a + 1)^p = \sum_{k=0}^p \binom{p}{k} a^k = 1 + \binom{p}{1} a^1 + \cdots + \binom{p}{p-1} a^{p-1} + a^p.$$

Consider the occurring binomial coefficients $\binom{p}{k}$. If $1 \leq k \leq p - 1$, then

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

⁷The name was chosen in order to distinguish it from the famous *Fermat's last theorem*.

Therefore, $k(p) = p \binom{p-1}{k-1}$ which implies that $p \mid k(p)$. Now the conditions $1 \leq k \leq p-1$ and p prime come into play as together they imply that $\gcd(p, k) = 1$. Hence, by Corollary 2.3.5, we deduce that $p \mid \binom{p}{k}$. Hence, $\binom{p}{k} \equiv 0 \pmod{p}$. Equation (2.7.3) implies that

$$(a+1)^p = 1 + \binom{p}{1}a^1 + \cdots + \binom{p}{p-1}a^{p-1} + a^p \equiv 1 + a^p \pmod{p}.$$

By our induction hypothesis (2.7.2), we know that $a^p \equiv a \pmod{p}$. Therefore, combining these last equations, we get that $(a+1)^p \equiv a+1 \pmod{p}$ which is our desired conclusion. ■

Example 2.7.2. To see a quick benefit of this theorem, use your favorite primes p (of at least two digits let's say) and your favorite nonzero integers a to calculate a^p first and then determine its remainder when divided by p . For example, try calculating $12^{23} \pmod{23}$ by first computing 12^{23} and then dividing by 23.

The following is an equivalent formulation of Fermat's little theorem.

Theorem 2.7.2. *Let p be a prime. If a is an integer such that $p \nmid a$, then*

$$(2.7.4) \quad a^{p-1} \equiv 1 \pmod{p}.$$

Proof: To see how this result follows from Theorem 2.7.1, observe that $p \nmid a$ implies that $\gcd(a, p) = 1$. Hence, a has a multiplicative inverse b modulo p : $ab \equiv 1 \pmod{p}$. From $a^p \equiv a \pmod{p}$, multiplying both sides by b , we get that $a^{p-1} \equiv 1 \pmod{p}$ which is our desired conclusion. On the other hand, multiplying both sides of $a^{p-1} \equiv 1 \pmod{p}$ by a implies that $a^p \equiv a \pmod{p}$.

We use this occasion to give another proof of (2.7.4) without using the induction method. Let a be an integer such that $\gcd(a, p) = 1$. Denote

$$\mathbb{Z}_p^* = \mathbb{Z}_p \setminus \{0\} = \{1, \dots, p-1\}$$

the set of nonzero residues modulo p . Define the function $f : \mathbb{Z}_p^* \rightarrow \mathbb{Z}_p^*$ by $f(x) \equiv ax \pmod{p}$. We claim that this function is one-to-one and onto. To see that f is one-to-one, let $x_1, x_2 \in \mathbb{Z}_p^*$ be such that $f(x_1) = f(x_2)$. This means that $ax_1 \equiv ax_2 \pmod{p}$. Multiplying both sides by the multiplicative inverse b of a modulo p , we get that $x_1 \equiv x_2 \pmod{p}$ which shows that f is one-to-one. To see that f is onto, take $y \in \mathbb{Z}_p^*$. Define $x \equiv by \pmod{p}$. Because $y \not\equiv 0 \pmod{p}$ and $b \not\equiv 0 \pmod{p}$, we deduce that $x \not\equiv 0 \pmod{p}$. Hence, $x \in \mathbb{Z}_p^*$. Also, $f(x) \equiv ax \equiv aby \equiv y \pmod{p}$. This proves that f is onto.

Because f is one-to-one and onto, it means that the set

$$\{f(1), \dots, f(p-1)\}$$

is the same as $\mathbb{Z}_p^* = \{1, \dots, p-1\}$. This implies that the product of the elements in one set is the same as the product of the elements in the other set (modulo p of course). Hence,

$$f(1) \cdot \dots \cdot f(p-1) \equiv 1 \cdot \dots \cdot (p-1) \pmod{p}$$

which means that

$$(a \cdot 1) \cdot \dots \cdot (a \cdot (p-1)) \equiv 1 \cdot \dots \cdot (p-1) \pmod{p}.$$

Rearranging the left-hand side and using the factorial notation

$$(p-1)! = 1 \cdot \dots \cdot (p-1),$$

we get that

$$(2.7.5) \quad a^{p-1} \cdot (p-1)! \equiv (p-1)! \pmod{p}.$$

Because each of the elements $1, \dots, p-1$ is nonzero modulo p and p is a prime, the product $(p-1)! = 1 \cdot \dots \cdot (p-1)$ is nonzero modulo p . Hence, it has a multiplicative inverse c modulo p . Multiplying both sides of equation (2.7.5) by c , we deduce that $a^{p-1} \equiv 1 \pmod{p}$. ■

Another equivalent formulation of Fermat's little theorem is that if a is a nonzero element of \mathbb{Z}_p , then

$$a^{p-1} \equiv 1 \pmod{p}.$$

When we examine the previous tables, another pattern may become apparent. For example, it seems that the *middle* row seems to always consist of two entries: 1 and -1 (modulo the corresponding prime). It turns out that this is not a coincidence. The next result is a further refinement of Theorem 2.7.2 due to Euler.

Corollary 2.7.3 (Euler's Corollary). *Let $p \geq 3$ be prime. If a is an integer such that $p \nmid a$, then*

$$(2.7.6) \quad a^{\frac{p-1}{2}} \equiv 1 \pmod{p} \quad \text{or} \quad a^{\frac{p-1}{2}} \equiv -1 \pmod{p}.$$

Proof: By assumption $a \not\equiv 0 \pmod{p}$, hence equation (2.7.1) may be written as it is formulated in (2.7.4). From this we derive that

$$(2.7.7) \quad \left(a^{\frac{p-1}{2}} - 1\right) \cdot \left(a^{\frac{p-1}{2}} + 1\right) = a^{p-1} - 1 \equiv 0 \pmod{p}.$$

Recall that p is prime, hence by Corollary 2.4.8 every integer different from zero modulo p possesses an inverse number. Due to this fact, from (2.7.7) it follows that at least one of the two factors has to be zero modulo p . Otherwise, both numbers would possess inverse elements and their product could never be equal to zero modulo p . Summing up, either

$$a^{\frac{p-1}{2}} - 1 \equiv 0 \pmod{p} \quad \text{or} \quad a^{\frac{p-1}{2}} + 1 \equiv 0 \pmod{p}.$$

Of course, this implies (2.7.6) and completes the proof. ■

Example 2.7.3. If $p = 5$ and $a = 3$, then

$$a^{\frac{p-1}{2}} = 3^{\frac{5-1}{2}} = 9, \quad \text{hence here we get} \quad a^{\frac{p-1}{2}} \equiv -1 \pmod{5}.$$

Next let $p = 13$ and $a = 12$. Then we obtain

$$a^{\frac{p-1}{2}} = 12^{\frac{13-1}{2}} = 2,985,984.$$

Since

$$2,985,984 - 1 = 13 \cdot 229,691 \quad \text{it follows that} \quad 12^{\frac{13-1}{2}} \equiv +1 \pmod{13}.$$

So we see that in (2.7.6) both cases (-1 and $+1$) may occur.

We mentioned in Section 1.6 the story of Fermat numbers. These are numbers of the form $2^{2^n} + 1$ which Fermat conjectured to be primes for any $n \geq 0$. He was certainly correct for small values of n , namely when $1 \leq n \leq 4$, where the numbers are:

$$3, 5, 17, 257, 65537.$$

We leave checking the primality of these numbers as an exercise. When $n = 5$, one observes that

$$2^{2^5+1} = 2^{32} + 1 = 4294967297.$$

Euler proved that this number is not a prime by showing that

$$4294967297 = 641 \cdot 6700417.$$

How did he come up with such a factorization? Here are his tricks. Euler started by analyzing the possible values of a prime divisor p of a number of the form $a^{2^n} + 1$.

Proposition 2.7.4. *Let a be an even integer and p be a prime.*

- (1) *If $p \mid a^{2^0} + 1 = a + 1$, then $p \equiv 1 \pmod{2}$.*
- (2) *If $p \mid a^{2^1} + 1 = a^2 + 1$, then $p \equiv 1 \pmod{2^2}$.*
- (3) *If $p \mid a^{2^2} + 1 = a^4 + 1$, then $p \equiv 1 \pmod{2^3}$.*
- (4) *If $k \geq 1$ and $p \mid a^{2^k} + 1$, then $p \equiv 1 \pmod{2^{k+1}}$.*

Proof: The first part is easy. If a is even, then $a + 1$ is odd. Therefore, any divisor of $a + 1$ must be odd. This means that if $p \mid a + 1$, p must be odd which is the same as $p \equiv 1 \pmod{2}$.

Let us move to the second part. If a is even, then $a^{2^1} + 1 = a^2 + 1$ is odd. From the first part, $p \mid a^2 + 1$ implies that p is odd. Therefore, $p \equiv 1 \pmod{4}$ or $p \equiv 3 \pmod{4}$. We show that the case $p \equiv 3 \pmod{4}$ cannot happen using a proof by contradiction. Suppose that $p \equiv 3 \pmod{4}$. This means that $p = 4t + 3$ for some nonnegative integer t . Because $p \mid a^2 + 1$, we deduce that p cannot divide a . Otherwise, we would get that $p \mid 1$ implying $p = 1$. Hence, $p \nmid a$ which means that $\gcd(a, p) = 1$. From Fermat's little theorem, we obtain that

$$p \mid a^{p-1} - 1 = a^{4t+2} - 1.$$

Using (1.3.3) from Section 1.3, we get that

$$a^{4t+2} + 1 = (a^2)^{2t+1} + 1 = (a^2 + 1)[(a^2)^{2t} - (a^2)^{2t-1} + \cdots - (a^2)^1 + (a^2)^0].$$

Because $p \mid a^2 + 1$, it follows that $p \mid a^{4t+2} + 1$. Combining with $p \mid a^{4t+2} - 1$, we deduce that $p \mid (a^{4t+2} + 1) - (a^{4t+2} - 1) = 2$. This means that $p = 2$, in contradiction with p being odd. Hence $p \equiv 1 \pmod{4}$.

The third part builds on our previous argument and can be proved similarly. Because a is even, then a^2 is even. Since $p \mid a^4 + 1 = (a^2)^2 + 1$, from the second part we get that $p \equiv 1 \pmod{4}$. Therefore, $p \equiv 1 \pmod{8}$ or $p \equiv 5 \pmod{8}$. We show that

$p \equiv 5 \pmod{8}$ cannot happen using proof by contradiction. If $p \equiv 5 \pmod{8}$, then $p = 8u + 5$ for some nonnegative integer u . As before, we use Fermat's little theorem and obtain that

$$p \mid a^{p-1} - 1 = a^{8u+4} - 1.$$

Using (1.3.3) from Section 1.3, we get that

$$a^{8u+4} + 1 = (a^4)^{2u+1} + 1 = (a^4 + 1)[(a^4)^{2u} - (a^4)^{2u-1} + \cdots - (a^4)^1 + 1].$$

Because $p \mid a^4 + 1$, the identity above implies that $p \mid a^{8u+4} + 1$. Therefore, $p \mid (a^{8u+4} + 1) - (a^{8u+4} - 1) = 2$ which means that $p = 2$, contradiction. Hence, $p \equiv 1 \pmod{8}$.

The proof of the last part follows the same pattern and can be done using induction on k . We leave it as an exercise for the reader to complete it. ■

Using the results stated above, Euler argued that if a prime p divides $2^{2^5} + 1$, then $p \equiv 1 \pmod{64}$, meaning that p is of the form $p = 64k + 1$. He then rolled up his sleeves and tried successive values of k .

- (1) If $k = 1$, then $p = 64 + 1 = 65 = 5 \cdot 13$, not a prime so move to the next k .
- (2) If $k = 2$, then $p = 64 \cdot 2 + 1 = 129 = 3 \cdot 43$, not a prime.
- (3) If $k = 3$, then $p = 64 \cdot 3 + 1 = 193$. This number is a prime, but it does not divide $2^{32} + 1$. The reader should check this.
- (4) If $k = 4$, then $p = 64 \cdot 4 + 1 = 257$. This is a prime which is not a divisor of $2^{32} + 1$.
- (5) If $k = 5$, then $p = 64 \cdot 5 + 1 = 321 = 3 \cdot 107$, not a prime.
- (6) If $k = 6$, then $p = 64 \cdot 6 + 1 = 385 = 5 \cdot 77 = 5 \cdot 7 \cdot 11$, not a prime.
- (7) If $k = 7$, then $p = 64 \cdot 7 + 1 = 449$. This number is a prime, but it does not divide $2^{32} + 1$.
- (8) If $k = 8$, then $p = 64 \cdot 8 + 1 = 513 = 3 \cdot 171 = 3 \cdot 3 \cdot 57 = 3 \cdot 3 \cdot 3 \cdot 19$, not a prime.
- (9) If $k = 9$, then $p = 64 \cdot 9 + 1 = 577$. This is a prime, but does not divide $2^{32} + 1$.
- (10) If $k = 10$, then $p = 64 \cdot 10 + 1 = 641$. This is a prime which divides $2^{32} + 1$.

One takeaway message from this story is that even geniuses like Euler were not afraid of tedious case analysis in search for mathematical proofs. As mentioned in Section 1.6, at present time, the only Fermat primes known are the ones Fermat mentioned: $2^{2^n} + 1$ for $0 \leq n \leq 4$. As of 2014, it is known that $2^{2^n} + 1$ is composite for any $5 \leq n \leq 32$, although explicit factorizations of these numbers are only known for $5 \leq n \leq 11$. There are many open questions about these numbers. For example, it is not known whether $2^{2^n} + 1$ is composite for any $n \geq 5$, if there are infinitely many composite numbers of the form $2^{2^n} + 1$ or if there are infinitely many prime numbers of the form $2^{2^n} + 1$.

Another amazing fact about the Fermat primes was proved by Gauss and Wantzel. A **constructible polygon** is a regular polygon that can be constructed with compass and straightedge. It is not too hard to show that the equilateral triangle and the square are constructible. The Greeks knew that, and also that pentagons are constructible. They could also prove that if a regular n -gon is constructible, then a regular $2n$ -gon is constructible. Perhaps the reader has seen some of these constructions in a geometry

class. In 1796, Gauss obtained a breakthrough result and proved that the regular 17-gon is constructible. Gauss wanted the regular 17-gon to be engraved on his gravestone, but unfortunately that did not happen. However, the famous Mathematical Sciences Research Institute MSRI located in Berkeley, California has as its address *17 Gauss Way*.

Later on, Gauss obtained a sufficient condition for a regular n -gon to be constructible and stated that the condition is also necessary, but without providing a proof. The proof of necessity was given in 1837 by the French mathematician Pierre Wantzel (1814–1848). Thus, the following result is commonly known as the Gauss–Wantzel theorem.

Theorem 2.7.5 (Gauss–Wantzel). *A regular n -gon is constructible if and only if $n = 2^k f_1 \dots f_t$, where $k \geq 0$ is an integer and f_1, \dots, f_t are distinct Fermat primes.*

We will not give the proof here, but we will mention that its proof uses Euler's totient function that we discuss next. Euler extended Fermat's little theorem to composite numbers. To understand Euler's result, we need to introduce a new function called Euler's totient function. The name *totient* was given by James Joseph Sylvester (1814–1897) in 1879. It comes from the Latin *tot* meaning *so many*.

Definition 2.7.1. Let n be a natural number. The **Euler totient function** or **Euler's ϕ -function** $\phi(n)$ is defined as the number of integers k with $1 \leq k \leq n$ that are **coprime** with n .

In Table 2.7.4 we list some values of $\phi(n)$ for small n . To determine $\phi(n)$ in general, we start with some simple observations.

Table 2.7.4. Euler's totient function $\phi(n)$ for $1 \leq n \leq 15$.

| | | | | | | | | | | | | | | | |
|-----------|---|---|---|---|---|---|---|---|---|----|----|----|----|----|----|
| n | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 |
| $\phi(n)$ | 1 | 1 | 2 | 2 | 4 | 2 | 6 | 4 | 6 | 4 | 10 | 4 | 12 | 6 | 8 |

Proposition 2.7.6. *If p is a prime, then $\phi(p^k) = p^k - p^{k-1} = p^k \left(1 - \frac{1}{p}\right)$ for any natural number k .*

Proof: We start with proving $\phi(p) = p - 1$. Because p is a prime, it has no factor other than 1 and p . This means that the only number between 1 and p that is not coprime with p must be p itself. Hence, $\phi(p)$ must equal $p - 1$.

Let $k \geq 2$. Because p is prime, any number that is not coprime with p^k must have p in its prime factorization. This means that the numbers between 1 and p^k that are not coprime with p^k are the multiples of p . There are $p^k/p = p^{k-1}$ such multiples and therefore, $\phi(p^k) = p^k - p^{k-1}$. ■

Since we have determined Euler's totient function for prime powers, with the Prime Factorization theorem in mind, it seems natural to try to understand $\phi(n_1 n_2)$ when n_1 and n_2 are coprime. If we could do that, then we could determine Euler's

totient function for any product of powers of distinct primes and therefore, for any natural number. We can examine the table above for some possible patterns. In particular,

$$\begin{aligned}\phi(6) &= 2 = \phi(2)\phi(3), \quad \phi(10) = 4 = \phi(2)\phi(5), \quad \phi(12) = 4 = \phi(3)\phi(4), \\ \phi(14) &= 6 = \phi(2)\phi(7) \quad \text{and} \quad \phi(15) = 8 = \phi(3)\phi(5)\end{aligned}$$

point to a connection between $\phi(n_1 n_2)$ and $\phi(n_1)\phi(n_2)$ when $\gcd(n_1, n_2) = 1$.

Proposition 2.7.7. *Let n_1 and n_2 be two natural numbers. If $\gcd(n_1, n_2) = 1$, then*

$$\phi(n_1 n_2) = \phi(n_1)\phi(n_2).$$

Proof: Let

$$\begin{aligned}A &= \{k : 1 \leq k \leq n_1 n_2, \gcd(k, n_1 n_2) = 1\}, \\ A_1 &= \{\ell : 1 \leq \ell \leq n_1, \gcd(\ell, n_1) = 1\}, \\ A_2 &= \{m : 1 \leq m \leq n_2, \gcd(m, n_2) = 1\}.\end{aligned}$$

Clearly, $\phi(n_1 n_2) = |A|$, $\phi(n_1) = |A_1|$ and $\phi(n_2) = |A_2|$. We will prove the identity $|A| = |A_1||A_2|$ which is equivalent to our desired conclusion. To prove this statement, we will show that there is a bijective function between A and $A_1 \times A_2$.

Define $f : A \rightarrow A_1 \times A_2$ as follows. For $x \in A$, let a_j be the remainder of the division of x by n_j for $1 \leq j \leq 2$ and define $f(x) = (a_1, a_2)$. To prove that f is well-defined, we have to check that actually $a_1 \in A_1$ and $a_2 \in A_2$. Let $d_1 = \gcd(a_1, n_1)$. Because $x \equiv a_1 \pmod{n_1}$, we get that $n_1 \mid x - a_1$ which implies $x - a_1 = n_1 s$ for some integer s . Since $d_1 \mid a_1$ and $d_1 \mid n_1$, we get that $d_1 \mid x$. Now we have that $d_1 \mid x$ and $d_1 \mid n_1 n_2$ (since $d_1 \mid n_1$) which imply that $d_1 = 1$ as $\gcd(x, n_1 n_2) = 1$. Hence, $\gcd(a_1, n_1) = 1$. By a similar argument, $\gcd(a_2, n_2) = 1$. Therefore, $f : A \rightarrow A_1 \times A_2$ is well-defined.

To prove that f is a bijective function, we will use the Chinese remainder theorem. To see that f is onto, we use Corollary 2.6.3. Let $(a_1, a_2) \in A_1 \times A_2$. Corollary 2.6.3 implies that there exists $x \in \{0, \dots, n_1 n_2 - 1\}$ such that $x \equiv a_1 \pmod{n_1}$ and $x \equiv a_2 \pmod{n_2}$. The only thing left to prove is that we have $\gcd(x, n_1 n_2) = 1$. Assume that there exists a prime p that divides both x and $n_1 n_2$. Because $p \mid n_1 n_2$ and $\gcd(n_1, n_2) = 1$, we get that $p \mid n_1$ or $p \mid n_2$. Without loss of generality, assume that $p \mid n_1$. From $x \equiv a_1 \pmod{n_1}$, we get that $x = a_1 + n_1 t$ for some integer t . Since $p \mid x$ and $p \mid n_1$, we deduce that $p \mid a_1$. Therefore, $p \mid n_1$ and $p \mid a_1$ which is a contradiction with $\gcd(n_1, a_1) = 1$. Therefore, $\gcd(x, n_1 n_2) = 1$ which implies that $x \in A$ and the function f is onto. The fact that f is one-to-one follows again from Corollary 2.6.3. ■

We can generalize this result to larger products.

Proposition 2.7.8. *Let $t \geq 2$ be a natural number and n_1, \dots, n_t be pairwise coprime natural numbers. Then*

$$\phi(n_1 \dots n_t) = \phi(n_1) \dots \phi(n_t).$$

Proof: We use induction on t . The base case $t = 2$ is true and follows from the previous proposition. For the induction step, let $t \geq 3$ and assume that the statement is true for $t - 1$. Let n_1, \dots, n_{t-1}, n_t be t natural numbers that are pairwise coprime. We first

show that $n_1 \dots n_{t-1}$ and n_t are coprime. Assume by contradiction that there exists a prime p that divides both $n_1 \dots n_{t-1}$ and n_t . Because $p \mid n_1 \dots n_{t-1}$, it follows that p must divide n_j for some j between 1 and $t - 1$. Then $p \mid n_j$ and $p \mid n_t$ which gives a contradiction with $\gcd(n_j, n_t) = 1$. Hence, $n_1 \dots n_{t-1}$ and n_t are coprime. By the previous proposition, we deduce that $\phi(n_1 \dots n_{t-1} n_t) = \phi(n_1 \dots n_{t-1})\phi(n_t)$. Now the induction hypothesis gives us that $\phi(n_1 \dots n_{t-1}) = \phi(n_1) \dots \phi(n_{t-1})$. Putting these two last equations together, we deduce that $\phi(n_1 \dots n_t) = \phi(n_1) \dots \phi(n_t)$ which finishes our proof. ■

We give now a precise formula for $\phi(n)$ for any natural number $n \geq 2$ in terms of the unique prime factorization of n .

Theorem 2.7.9. *Let $n \geq 2$ be a natural number. If $n = p_1^{e_1} \dots p_t^{e_t}$ is the prime factorization of n , where p_1, \dots, p_t are distinct primes and e_1, \dots, e_t are natural numbers, then*

$$\phi(n) = n \left(1 - \frac{1}{p_1}\right) \dots \left(1 - \frac{1}{p_t}\right).$$

Proof: We give two proofs of this result. The first proof builds on our work in Proposition 2.7.6 and Proposition 2.7.8. From Proposition 2.7.8, we get that $\phi(n) = \phi(p_1^{e_1}) \dots \phi(p_t^{e_t})$. Using Proposition 2.7.6, we obtain that

$$\begin{aligned} \phi(n) &= \phi(p_1^{e_1}) \dots \phi(p_t^{e_t}) = p_1^{e_1} \left(1 - \frac{1}{p_1}\right) \cdot \dots \cdot p_t^{e_t} \left(1 - \frac{1}{p_t}\right) \\ &= n \left(1 - \frac{1}{p_1}\right) \cdot \dots \cdot \left(1 - \frac{1}{p_t}\right), \end{aligned}$$

which is the required result.

The second proof uses the principle of inclusion and exclusion from Section 1.7 (see Theorem 1.7.8). Let A be the set of natural numbers from 1 to n . We want to remove from A those *bad* numbers that are not coprime with n . We keep track of these numbers in the following sets. For $1 \leq j \leq t$, we define sets A_j by $A_j = \{k : 1 \leq k \leq n, p_j \mid k\}$. Then $A_1 \cup \dots \cup A_t$ contains all these *bad* numbers and its complement $A \setminus (A_1 \cup \dots \cup A_t)$ contains precisely the numbers between 1 and t that are coprime with n . The fact that p_1, \dots, p_t are distinct primes comes into play here as we will see now. Recall our notation: for $J \subseteq \{1, \dots, t\}$, A_J denotes the intersection $\cap_{j \in J} A_j$, when $J \neq \emptyset$ and $A_J = A$ when $J = \emptyset$. Consider a nonempty subset $J \subseteq \{1, \dots, t\}$. Let $k \in A_J$. This means that $p_j \mid k$ for any $j \in J$. Because p_1, \dots, p_t are distinct primes, it follows that $\prod_{j \in J} p_j \mid k$. Conversely, if $\prod_{j \in J} p_j \mid k$, then $p_j \mid k$ for any $j \in J$. Hence,

$$\begin{aligned} A_J &= \bigcap_{j \in J} A_j = \{k : 1 \leq k \leq n, p_j \mid k, \forall j \in J\} \\ &= \{k : 1 \leq k \leq n, \prod_{j \in J} p_j \mid k\}. \end{aligned}$$

Thus, any element in A_J is of the form $\prod_{j \in J} p_j \cdot s$ with $1 \leq s \leq \frac{n}{\prod_{j \in J} p_j}$. Therefore,

$$|A_J| = \frac{n}{\prod_{j \in J} p_j}.$$

The principle of inclusion and exclusion (Theorem 1.7.8) implies that

$$\begin{aligned}\phi(n) &= |A \setminus (A_1 \cup \dots \cup A_t)| = \sum_{J \subseteq [t]} (-1)^{|J|} |A_J| = \sum_{J \subseteq [t]} (-1)^{|J|} \frac{n}{\prod_{j \in J} p_j} \\ &= n - \sum_{j=1}^t \frac{n}{p_j} + \sum_{1 \leq j < \ell \leq t} \frac{n}{p_j p_\ell} - \dots + (-1)^t \frac{n}{p_1 \cdot \dots \cdot p_t} \\ &= n \left(1 - \frac{1}{p_1}\right) \cdot \dots \cdot \left(1 - \frac{1}{p_t}\right),\end{aligned}$$

which is the claimed formula. ■

Example 2.7.4. Let $n = 8$. Because $8 = 2^3$, our previous formula gives us that $\phi(8) = 8 \cdot \left(1 - \frac{1}{2}\right) = 4$. This is reassuring since we can observe that the set $\{1, 3, 5, 7\}$ is the set of $\phi(8) = 4$ natural numbers between 1 and 8 that are coprime with 8. We describe below the multiplication table for these elements (modulo 8). Note that if we pick any element $a \in \{1, 3, 5, 7\}$ and we multiply each element of $\{1, 3, 5, 7\}$ by it (modulo 8), we obtain a permutation of $\{1, 3, 5, 7\}$.

Table 2.7.5. The multiplication table for the elements of \mathbb{Z}_8 coprime with 8.

| . | 1 | 3 | 5 | 7 |
|---|---|---|---|---|
| 1 | 1 | 3 | 5 | 7 |
| 3 | 3 | 1 | 7 | 5 |
| 5 | 5 | 7 | 1 | 3 |
| 7 | 7 | 5 | 3 | 1 |

Example 2.7.5. For $n = 15$ we have $\phi(15) = 8$ and the set $\{1, 2, 4, 7, 8, 11, 13, 14\}$ consists of all the numbers between 1 and 15 that are coprime with 15. We now examine the multiplication table of these elements (modulo 15). As in the previous example, note that every row of the multiplication table above is a permutation of $\{1, 2, 4, 7, 8, 11, 13, 14\}$. This will be useful in proving Euler's theorem next.

Table 2.7.6. The multiplication table of the invertible elements of \mathbb{Z}_{15} .

| . | 1 | 2 | 4 | 7 | 8 | 11 | 13 | 14 |
|----|----|----|----|----|----|----|----|----|
| 1 | 1 | 2 | 4 | 7 | 8 | 11 | 13 | 14 |
| 2 | 2 | 4 | 8 | 14 | 1 | 7 | 11 | 13 |
| 4 | 4 | 8 | 1 | 13 | 2 | 14 | 7 | 11 |
| 7 | 7 | 14 | 13 | 4 | 11 | 2 | 1 | 8 |
| 8 | 8 | 1 | 2 | 11 | 4 | 13 | 14 | 7 |
| 11 | 11 | 7 | 14 | 2 | 13 | 1 | 8 | 4 |
| 13 | 13 | 11 | 7 | 1 | 14 | 8 | 4 | 2 |
| 14 | 14 | 13 | 11 | 8 | 7 | 4 | 2 | 1 |

Euler's extension of Fermat's little theorem is the following result. To see why it implies Fermat's little theorem, note that when p is prime, $\phi(p) = p - 1$.

Theorem 2.7.10 (Euler). *Let $n \geq 2$ be a natural number. If a is an integer coprime with n , then*

$$a^{\phi(n)} \equiv 1 \pmod{n}.$$

Proof: Let $U_n = \{k : 1 \leq k \leq n, \gcd(k, n) = 1\}$ be the set of natural numbers between 1 and n that are coprime with n . Let a be an integer coprime with n . Because $\gcd(a, n) = 1$, it is sufficient for us to prove that $a^{\phi(n)} \equiv 1 \pmod{n}$ when $a \in U_n$. This is because $a \equiv r \pmod{n}$, where r is the remainder of integer division when a is divided by n and $\gcd(r, n) = \gcd(a, n) = 1$ (so $r \in U_n$).

Assume $a \in U_n$. Consider the effect that a has on the elements of U_n when they are multiplied by a . More formally, define a function $f : U_n \rightarrow U_n$ by $f(x)$ equal the remainder of the division of ax by n . We claim that f is well-defined and moreover, is a bijective function which permutes the elements of U_n . To see that f is well-defined, we have to argue that if $x \in U_n$ (meaning that $1 \leq x \leq n$ and $\gcd(x, n) = 1$), then the remainder y of ax when divided by n is also between 1 and n and is coprime with n . Because $\gcd(a, n) = \gcd(x, n) = 1$, we get that $\gcd(ax, n) = 1$. This implies that the remainder y of the integer division of ax by n cannot be 0. Hence, $1 \leq y \leq n$. We also have that $\gcd(ax, n) = \gcd(y, n)$ which implies that $\gcd(y, n) = 1$. Hence, $y \in U_n$. To see why f is injective, take $x_1, x_2 \in U_n$ such that $f(x_1) = f(x_2)$. This implies that $n \mid ax_1 - ax_2 = a(x_1 - x_2)$. Since $\gcd(a, n) = 1$, we deduce that $n \mid x_1 - x_2$. As $1 \leq x_1, x_2 \leq n$, the only possibility is that $x_1 = x_2$. Hence, f is injective. To see why f is surjective, take $y \in U_n$. Because $\gcd(a, n) = 1$, there exists $b \in U_n$ such that $ab \equiv 1 \pmod{n}$. Take x to be the remainder of the integer division of by by n . Because $\gcd(b, n) = \gcd(y, n) = 1$, we get that $\gcd(by, n) = 1$. This implies that $x \in U_n$. Also, $f(x) \equiv ax \equiv aby \equiv y \pmod{n}$. Since $1 \leq f(x), y \leq n$, we deduce that $f(x) = y$. Thus, f is surjective.

Because f is bijective, this means that the two following sets are equal:

$$\{x : x \in U_n\} = \{f(x) : x \in U_n\}.$$

Therefore, the product of the elements in one set equals the product of the elements in the other set:

$$\prod_{x \in U_n} x = \prod_{x \in U_n} f(x) \equiv \prod_{x \in U_n} ax \equiv a^{\phi(n)} \prod_{x \in U_n} x \pmod{n}.$$

Let $M = \prod_{x \in U_n} x$. Because $\gcd(x, n) = 1$ for any $x \in U_n$, we must have that $\gcd(M, n) = 1$. Hence, M has a multiplicative inverse modulo n . Multiplying both sides of the previous equation by the multiplicative inverse of M , we get that

$$1 \equiv a^{\phi(n)} \pmod{n}.$$

This finishes our proof. ■

We give some applications of Euler's theorem.

Example 2.7.6. What are the last 3 digits of 7^{2020} ? Denote $M = 7^{2020}$. Because $\gcd(7, 1000) = 1$, we can apply Euler's theorem to deduce that $7^{\phi(1000)} \equiv 1 \pmod{1000}$.

Because $1000 = 2^3 \cdot 5^3$, we have that $\phi(1000) = 1000 \left(1 - \frac{1}{2}\right) \left(1 - \frac{1}{5}\right) = 400$. Therefore, $7^{400} \equiv 1 \pmod{1000}$ and consequently,

$$7^{2020} = (7^{400})^5 \cdot 7^{20} \equiv 1 \cdot 7^{20} \equiv 7^{20} \pmod{1000}.$$

To calculate $7^{20} \pmod{1000}$, it is not too hard. We start with $7^2 = 49$ and

$$7^4 = 49^2 = (50 - 1)^2 = 2500 - 100 + 1 = 2401 \equiv 401 \pmod{1000}.$$

Therefore, $7^{20} \equiv 401^5 \pmod{1000}$. Also, $401^2 = (400 + 1)^2 = 16000 + 800 + 1 \equiv 801 \pmod{1000}$ and

$$401^4 \equiv 801^2 = (800 + 1)^2 = 64000 + 1600 + 1 \equiv 601 \pmod{1000}.$$

Hence,

$$7^{20} = 401^5 = 401^4 \cdot 401 \equiv 601 \cdot 401 = 24000 + 1000 + 1 \equiv 001 \pmod{1000},$$

which implies that the last 3 digits of 7^{2020} are 001.

Example 2.7.7. What are the last 3 digits of 2^{2020} ? Denote $N = 2^{2020}$. The question is asking to figure out $N \pmod{1000}$. Because of $1000 = 2^3 \cdot 5^3$ and due to $\gcd(N, 1000) = 8 > 1$, we cannot apply Euler's theorem to N and 1000. However, we can observe first that $N \equiv 0 \pmod{8}$, determine $N \pmod{125}$ and apply the Chinese remainder theorem to get $N \pmod{1000}$. Note that $\gcd(2, 125) = 1$ and we can apply Euler's theorem to deduce that $2^{\phi(125)} \equiv 1 \pmod{125}$. Because $\phi(125) = 100$, we have that $2^{100} \equiv 1 \pmod{125}$. This is helpful as it implies that $2^{2000} \equiv (2^{100})^{20} \equiv 1 \pmod{125}$. Hence, $2^{2020} \equiv 2^{20} \pmod{125}$. It is not too hard to see that $2^{10} = 1024 \equiv 24 \pmod{125}$ and that $2^{20} \equiv 24^2 = 576 \equiv 76 \pmod{125}$. Thus, we know now that $N \equiv 0 \pmod{8}$ and $N \equiv 76 \pmod{125}$. The first equation gives us that $N = 8k$ for some natural number k . Hence $8k \equiv 76 \pmod{125}$. We need to find the multiplicative inverse of 8 modulo 125. Applying the Euclidean algorithm, we get that

$$\begin{aligned} 125 &= 8 \cdot 15 + 5 \\ 8 &= 5 \cdot 1 + 3 \\ 5 &= 3 \cdot 1 + 2 \\ 3 &= 2 \cdot 1 + 1, \end{aligned}$$

and therefore,

$$\begin{aligned} 1 &= 3 - 2 = 2 \cdot 3 - 5 = 2 \cdot (8 - 5) - 5 = 2 \cdot 8 - 3 \cdot 5 \\ &= 2 \cdot 8 - 3 \cdot (125 - 15 \cdot 8) = 47 \cdot 8 - 3 \cdot 125. \end{aligned}$$

The multiplicative inverse of 8 modulo 125 is 47. From $8k \equiv 76 \pmod{125}$, multiplying both sides by 47, we get that $k \equiv 47 \cdot 76 = 3572 \equiv 72 \pmod{125}$. Hence, $k = 125\ell + 72$ for some natural number ℓ . Combining with $N = 8k$, we get that $N = 8(125\ell + 72) = 1000\ell + 608$. Therefore, the last three digits of 2^{2020} are 608.

The proof we gave for Euler's theorem is very similar to the second proof we gave for Fermat's little theorem. In the case of a prime, any number between 1 and $p - 1$ has a multiplicative inverse and the product M from the end of the proof of Euler's theorem is exactly $(p - 1)!$, the product of all these numbers. It turns out that one can figure out $(p - 1)! \pmod{p}$ for any prime p . The following result is usually called

Wilson's theorem. It was first published without a proof by the English mathematician Edward Waring (1734–1798) in 1770. Waring, a Lucasian Professor at Cambridge University, attributed it to John Wilson (1741–1793), another English mathematician who was Waring's student.

Theorem 2.7.11 (Wilson 1770). *If p is a prime, then*

$$(p - 1)! \equiv -1 \pmod{p}.$$

Proof: Let p be a prime. If $p = 2$, then $(2 - 1)! = 1 \equiv -1 \pmod{2}$. If $p = 3$, then $(3 - 1)! = 2 \equiv -1 \pmod{3}$. Assume that $p \geq 5$ for the rest of the proof. Because p is prime, we know that for any natural number a between 1 and $p - 1$, there exists a natural number b such that $ab \equiv 1 \pmod{p}$. First, we claim that the only such a with the property that $b = a$ are 1 and $p - 1$. This is because if $a^2 \equiv 1 \pmod{p}$, then $p \mid a^2 - 1 = (a+1)(a-1)$ which implies that $p \mid a+1$ or $p \mid a-1$. Because $1 \leq a \leq p-1$, we deduce that $a = p - 1$ or $a = -1$.

Our previous argument implies that for any a between 2 and $p - 2$, there exists $b \neq a$ such that $2 \leq b \leq p - 2$ and $ab \equiv 1 \pmod{p}$. This means that the set $\{2, \dots, p-2\}$ of $p-3$ natural numbers can be partitioned into $\frac{p-3}{2}$ pairs of the form $\{a, b\}$ with $ab \equiv 1 \pmod{p}$. Therefore, the product $\prod_{k=2}^{p-2} k$ is congruent to 1 modulo p . This implies that

$$(p - 1)! = (p - 1) \prod_{k=2}^{p-2} k \equiv p - 1 \equiv -1 \pmod{p},$$

which is our desired conclusion. ■

Example 2.7.8. To see an illustration of the previous proof on a concrete example, take $p = 13$ for example. Then the set $\{2, \dots, 11\}$ of $p - 3 = 10$ numbers can be split into the pairs

$$\{2, 7\}, \{3, 9\}, \{4, 10\}, \{5, 8\}, \{6, 11\}$$

such that

$$\begin{aligned} 2 \cdot 7 &= 14 \equiv 1 \pmod{13} \\ 3 \cdot 9 &= 27 \equiv 1 \pmod{13} \\ 4 \cdot 10 &= 40 \equiv 1 \pmod{13} \\ 5 \cdot 8 &= 40 \equiv 1 \pmod{13} \\ 6 \cdot 11 &= 66 \equiv 1 \pmod{13}. \end{aligned}$$

Therefore, $2 \cdot 3 \cdot \dots \cdot 11 \equiv 1 \pmod{13}$ which implies that

$$12! = 12 \cdot (2 \cdot 3 \cdot \dots \cdot 11) \equiv 12 \equiv -1 \pmod{13}.$$

By the way, $12! = 479001600$. The reader is invited to divide it by 13.

The first proof of Wilson's theorem appeared in 1771 and is due to Joseph Louis Lagrange or Giuseppe Luigi Lagrangia (1736–1813) in 1773. Lagrange was an Italian-born French mathematician who made important contributions in number theory, analysis, algebra, and probability. Lagrange also proved the converse to Wilson's theorem.

Theorem 2.7.12. Let $n \geq 2$ be a natural number. If $(n - 1)! \equiv -1 \pmod{n}$, then n is a prime.

Proof: The proof is by contradiction. Assume that $n \geq 2$ is a natural number such that $(n - 1)! \equiv -1 \pmod{n}$ and n is not a prime. There exists a natural number d such that $2 \leq d \leq n - 1$ and $d \mid n$. Because $n \mid (n - 1)! + 1$, we deduce that $d \mid (n - 1)! + 1$. Also, since $2 \leq d \leq n - 1$, d must appear as a factor in the product $(n - 1)!$. Thus, $d \mid (n - 1)!$. Because $d \mid (n - 1)! + 1$, we deduce that $d \mid (n - 1)! + 1 - (n - 1)! = 1$. Hence, $d = 1$, contradiction with our assumption $2 \leq d \leq n - 1$. ■

While this is a useful theoretical result recognizing when a natural number is a prime, the converse to Wilson's theorem is not useful in practice. Imagine calculating $1002! \pmod{1003}$ just to try to verify if 1003 is a prime or not.

As stated in Fermat's little theorem, $a^{p-1} \equiv 1 \pmod{p}$ for any prime p and nonzero element $a \in \mathbb{Z}_p$. When examining the tables in Example 2.7.1, other patterns are perhaps apparent. For some $a \neq 1$ and p , there is a smaller exponent $e < p - 1$ such that $a^e \equiv 1 \pmod{p}$. In our examples with $p \in \{5, 7, 11\}$, this happens when

$$(p, a) \in \{(5, 4), (7, 2), (7, 4), (11, 3), (11, 4), (11, 5)\}$$

and we invite the curious reader to double-check our computations.

Definition 2.7.2. Let p be a prime and $a \in \mathbb{Z}_p \setminus \{0\}$. The **multiplicative order of a modulo p** is defined as the smallest natural number e such that

$$a^e \equiv 1 \pmod{p}.$$

We denote the multiplicative order of a modulo p by $\text{ord}_p(a)$.

Proposition 2.7.13. Let p be a prime and $a \in \mathbb{Z}_p \setminus \{0\}$. The multiplicative order $\text{ord}_p(a)$ of a modulo p divides $p - 1$.

Proof: Let $e = \text{ord}_p(a)$. Since $a^{p-1} \equiv 1 \pmod{p}$, clearly $1 \leq e \leq p - 1$. Dividing $p - 1$ by e we get that $p - 1 = qe + r$, $0 \leq r < e$. Therefore, $a^{p-1} \equiv a^{qe+r} \equiv (a^e)^q \cdot a^r \pmod{p}$. Because $a^{p-1} \equiv a^e \equiv 1 \pmod{p}$, we deduce that $a^r \equiv 1 \pmod{p}$. Because e is the smallest natural number such that $a^e \equiv 1 \pmod{p}$, we obtain that r must be 0. Therefore, $e \mid p - 1$ and finishes our proof. ■

When $p = 5$, all the powers a^1, \dots, a^{p-1} are distinct when $a \in \{2, 3\}$. The same phenomenon happens for $p = 7$ and $a \in \{3, 5\}$ and when $p = 11$ and $a \in \{2, 6, 7, 8\}$. This leads to the following definition.

Definition 2.7.3. Let p be a prime. A nonzero element $g \in \mathbb{Z}_p$ is called a **primitive root modulo p** if

$$\mathbb{Z}_p \setminus \{0\} = \{g^1, \dots, g^{p-1}\}.$$

It is clear that g is a primitive root modulo p if and only if $\text{ord}_p(g) = p - 1$.

Theorem 2.7.14. Let p be a prime. There are exactly $\phi(p - 1)$ primitive roots modulo p in \mathbb{Z}_p .

Note that $\phi(5-1) = 2$, $\phi(7-1) = 2$ and $\phi(11-1) = 4$ which confirms our previous observations. We will not give the proof here, but we will mention that while it tells us that there are many primitive roots modulo a prime number, it does not give a recipe for finding such roots. The following conjecture due to the Austrian mathematician Emil Artin (1898–1962) is still open.

Conjecture 2.7.15. *There are infinitely many primes p such that 2 is a primitive root modulo p .*

We describe now an application of primitive roots to the construction of Costas arrays that have practical use in radar and sonar technologies.

Definition 2.7.4. A **permutation matrix** is an $n \times n$ array with 0, 1 entries having exactly one 1 in each row and in each column.

Permutation matrices are essentially equivalent to permutations. Given a permutation $\sigma : [n] \rightarrow [n]$, we can construct a permutation matrix P_σ as follows:

$$P_\sigma(j, k) = \begin{cases} 1 & \text{if } \sigma(j) = k \\ 0 & \text{otherwise.} \end{cases}$$

Example 2.7.9. For $n = 3$ and the identity permutation $\epsilon = \begin{pmatrix} 1 & 2 & 3 \\ 1 & 2 & 3 \end{pmatrix}$, the corresponding permutation matrix P_ϵ is the identity matrix $\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$. Conversely, the matrix $M = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{bmatrix}$ corresponds to the permutation $\sigma = \begin{pmatrix} 1 & 2 & 3 \\ 2 & 3 & 1 \end{pmatrix}$.

This is a bijective correspondence between the set of all permutations with n elements and the collection of $n \times n$ permutation matrices. Hence, there are $n!$ permutation matrices with n rows and columns. A permutation matrix can also be described by the *coordinates* or the ordered pairs (column, row) of its nonzero entries. For example, for the matrix M from the previous example, these coordinates are $(1, 3), (2, 1)$, and $(3, 2)$. Note that these coordinates are not the same as the usual coordinates in a Cartesian representation of a Euclidean plane.

Definition 2.7.5. A **Costas array**⁸ is an $n \times n$ permutation matrix with the property that among the $\binom{n}{2}$ segments whose endpoints are the coordinates of the entries 1, there are no two segments with the same length and the same slope.

Before we give a general construction of Costas arrays using primitive roots, we mention some examples and nonexamples.

Example 2.7.10. The identity matrix I_n is the $n \times n$ matrix that has 0 everywhere except 1 on the main diagonal. The *coordinates* of its nonzero entries are $(1, 1), \dots, (n, n)$.

⁸These matrices are named after the American electrical engineer John Peter Costas (1923–2008) who studied them in 1965. Around the same time, this concept was studied by the American mathematician Edgar Gilbert (1923–2013).

When $n = 1$, there are no segments and I_1 is a Costas array. When $n = 2$, there is only one segment and I_2 is also a Costas array. For $n \geq 3$, any segment with endpoints among the coordinates has slope -1 and one can find two such segments with the same length and slope; for example, $(1, 1)$ to $(2, 2)$ and $(2, 2)$ to $(3, 3)$ each have length $\sqrt{2}$ and slope -1 . Hence, for $n \geq 3$, I_n is not a Costas array.

Example 2.7.11. For the matrix M mentioned earlier, the three segments with endpoints in the coordinates $(1, 3), (2, 1)$, and $(3, 2)$, have slopes $-2, 1$ and $-1/2$ and lengths $\sqrt{5}, \sqrt{2}$, and $\sqrt{5}$, respectively. While there is one repeat in the lengths, the corresponding slopes are different and therefore, M is a Costas array.

We leave the complete verification of the claims in the next example as an exercise.

Example 2.7.12. The matrix $\begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix}$ is a Costas array. Out of the $3! = 6$ permutation matrices of order 3, only $\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$ and $\begin{bmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{bmatrix}$ are not Costas arrays.

For applications, one is interested in finding larger Costas arrays and that is where primitive roots prove to be useful as the next result shows.

Proposition 2.7.16. Let p be a prime and g a primitive root modulo p . Define the $(p - 1) \times (p - 1)$ matrix $A_{p,g}$ whose rows and columns are indexed by the elements of $\mathbb{Z}_p \setminus \{0\}$ that has all entries equal to 0 except for the entries in the positions (j, g^j) for $1 \leq j \leq p - 1$ that are equal to 1. The matrix $A_{p,g}$ is a $(p - 1) \times (p - 1)$ Costas array.

Proof: Before we give the proof, we write down the matrices constructed by the above procedure for several pairs (p, g) :

$$A_{5,2} = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 \end{bmatrix}, \quad A_{5,3} = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{bmatrix}$$

and

$$A_{7,3} = \begin{bmatrix} 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}, \quad A_{7,5} = \begin{bmatrix} 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}.$$

Assume that there exist two segments S and T of same length and slope whose endpoints are the coordinates of 1s in the matrix $A_{p,g}$. Let us denote the coordinates of the endpoints of S by (s_1, g^{s_1}) and (s_2, g^{s_2}) and those of T by (t_1, g^{t_1}) and (t_2, g^{t_2}) . The condition that S and T have the same length and slope means that

$$\begin{aligned} s_2 - s_1 &= t_2 - t_1 \\ g^{s_2} - g^{s_1} &= g^{t_2} - g^{t_1}. \end{aligned}$$

The second equation gives us that

$$g^{s_1}(g^{s_2-s_1} - 1) = g^{t_1}(g^{t_2-t_1} - 1).$$

As $s_2 - s_1 = t_2 - t_1 \neq 0$, we get that $g^{s_2-s_1} - 1 = g^{t_2-t_1} - 1 \neq 0$ and therefore, $g^{s_1} = g^{t_1}$. This implies that $s_1 = t_1$. Combining with $s_2 - s_1 = t_2 - t_1$, we get that $s_2 = t_2$. Hence, the segments S and T must be the same. ■

The readers interested in learning more about Costas arrays may wish to consult the sequence A008404 <https://oeis.org/A008404> in the *Online Encyclopedia of Integer Sequences* which counts the number of Costas arrays of order n .

Primitive roots are also useful in other circumstances, such as deciding which elements are squares in modular arithmetic.

Definition 2.7.6. Let p be a prime number. An element $a \in \mathbb{Z}_p$ is called a **quadratic residue modulo p** if

$$a \equiv x^2 \pmod{p}$$

for some nonzero element $x \in \mathbb{Z}_p$.

Example 2.7.13. The table 2.7.1 shows that the 1 and 4 are the quadratic residues in \mathbb{Z}_5 . Table 2.7.2 tells us that 1, 2 and 4 are the quadratic residues modulo 7. We also have that 1, 3, 4, 5, 9 are the quadratic residues modulo 11 from Table 2.7.3.

Proposition 2.7.17. Let p be a prime and g a primitive root modulo p . If a is a nonzero element of \mathbb{Z}_p , then a is a quadratic residue modulo p if and only if $a = g^{2k}$ for some nonnegative integer k .

Proof: If $a \equiv g^{2k} \pmod{p}$, then $a = (g^k)^2$. Because g is a primitive root modulo p , g^k is nonzero and therefore, a is a quadratic residue modulo p . Conversely, assume that a is a quadratic residue modulo p . Then $a \equiv x^2 \pmod{p}$ for some nonzero element $x \in \mathbb{Z}_p$. Because g is a primitive root modulo p , it follows that $x \equiv g^e \pmod{p}$ for some exponent e . Hence, $a \equiv x^2 \equiv g^{2e} \pmod{p}$ which finishes our proof. ■

The following notation is useful when studying quadratic residues modulo primes.

Definition 2.7.7. Let p be a prime and a a nonzero element of \mathbb{Z}_p . The **Legendre symbol** $\left(\frac{a}{p}\right)$ is defined as 1 if a is a quadratic residue modulo p and -1 if a is not a quadratic residue modulo p .

Proposition 2.7.18. If $p \geq 3$ is a prime and a is a nonzero element of \mathbb{Z}_p , then

$$a^{\frac{p-1}{2}} \equiv \left(\frac{a}{p}\right) \pmod{p}.$$

Proof: The result can be deduced from the previous Proposition and Corollary 2.7.3. ■

As a consequence, we can figure out when -1 is a quadratic residue modulo a prime.

Corollary 2.7.19. If $p \geq 3$ is a prime, then -1 is a quadratic residue modulo p if and only if $p \equiv 1 \pmod{4}$.

Proof: The conclusion follows from Proposition 2.7.17 and the previous result. ■

One of the most celebrated results in number theory is the quadratic reciprocity result obtained by Gauss.

Theorem 2.7.20 (Quadratic Reciprocity). *If p and q are two primes, then*

$$\left(\frac{p}{q}\right) \cdot \left(\frac{q}{p}\right) = (-1)^{\frac{p-1}{2} \cdot \frac{q-1}{2}}.$$

We conclude this section and chapter with an application of Euler's theorem and congruences to cryptography by describing the basics of the RSA encryption algorithm.

This was developed by Ron Rivest, Ron Shamir, and Leonard Adleman in 1977 and was named using their initials⁹. We briefly describe it here. The scenario is as before: Alice would like to send an encrypted message to Bob without Eve being able to decrypt it. Because of our previously described correspondence between letters and numbers, transmitting words over any alphabet is equivalent to transmitting numbers. We can use Table 2.4.7 or Table 2.4.8 for example.

Example 2.7.14. For simplicity, let us use Table 2.4.7. Say Alice wants to send Bob the message:

Do your homework.

In the first step, this is converted into a number using Table 2.4.7 and becomes

| | | | | | | | | | | | | | |
|---|----|----|----|----|----|---|----|----|---|----|----|----|----|
| d | o | y | o | u | r | h | o | m | e | w | o | r | k |
| 3 | 14 | 24 | 14 | 20 | 17 | 7 | 14 | 12 | 4 | 22 | 14 | 17 | 10 |

which becomes the following number after removing the spaces in between:

3142414201771412422141710.

To get this number encrypted, Bob chooses two *large*¹⁰ primes: $p = 317$ and $q = 709$. Bob will keep these numbers secret. Bob will calculate $pq = 224753$. Because pq has 6 digits, we will break our original message/number into blocks/numbers of 5 consecutive digits and encrypt each number. These numbers will be

31424 14201 77141 24221 41710.

In order to see how these numbers will be encrypted, we need the following result.

Proposition 2.7.21. *Let p and q be two distinct primes. Let e be a natural number such that $\gcd(e, (p-1)(q-1)) = 1$ and denote by d the multiplicative inverse of e modulo $(p-1)(q-1)$. For any integer M , the equation*

$$(2.7.8) \quad x^e \equiv M \pmod{pq}$$

has a unique solution (modulo pq), namely $x \equiv M^d \pmod{pq}$.

⁹It appears that this work was known before in some intelligence agencies of the UK.

¹⁰These primes are large when doing computations by hand, not when doing them by computer.

Proof: We first deal with the case $\gcd(M, pq) = 1$. Because

$$de \equiv 1 \pmod{(p-1)(q-1)},$$

there exists an integer s such that $de = 1 + s(p-1)(q-1)$. We show that M^d satisfies our equation $x^e \equiv M \pmod{pq}$. We use Euler's theorem, $\phi(pq) = (p-1)(q-1)$ and $\gcd(M, pq) = 1$ to deduce that $M^{(p-1)(q-1)} \equiv 1 \pmod{pq}$. Therefore,

$$\begin{aligned} (M^d)^e &= M^{de} = M^{1+s(p-1)(q-1)} \\ &= M \cdot (M^{(p-1)(q-1)})^s \\ &\equiv M \cdot 1^s \pmod{pq} \\ &\equiv M \pmod{pq}. \end{aligned}$$

To see why M^d is the unique solution (modulo pq) of the equation $x^e \equiv M \pmod{pq}$, assume that y is a solution of our equation (2.7.8). From $y^e \equiv M \pmod{pq}$, by raising both sides to the power d , we get that $y^{de} \equiv M^e \pmod{pq}$. Also, because $pq \mid M - y^e$ and $\gcd(M, pq) = 1$, we deduce that $\gcd(y^e, pq) = 1$ and therefore, $\gcd(y, pq) = 1$. Hence, by Euler's theorem $y^{(p-1)(q-1)} \equiv 1 \pmod{pq}$. Now since $de = 1+s(p-1)(q-1)$, we obtain that

$$\begin{aligned} y &= y^1 = y^{de-s(p-1)(q-1)} \equiv y^{de} \cdot (y^{(p-1)(q-1)})^{-s} \pmod{pq} \\ &\equiv M^d \cdot 1^{-s} \pmod{pq} \\ &\equiv M^d \pmod{pq}. \end{aligned}$$

This proves our assertion.

If $\gcd(M, pq) \neq 1$, then because p and q are distinct primes, we must have that $p \mid M$ or $q \mid M$. If both $p \mid M$ and $q \mid M$, then $pq \mid M$ and therefore, $M \equiv 0 \pmod{pq}$. Our equation (2.7.8) becomes $x^e \equiv 0 \pmod{pq}$. Therefore, $pq \mid x^e$. Since $p \neq q$, this means that $p \mid x^e$ and $q \mid x^e$. Euclid's lemma implies that $p \mid x$ and $q \mid x$. Hence, $pq \mid x$ and $x \equiv 0 \pmod{pq}$. Thus, $x \equiv 0 \equiv M^d \pmod{pq}$ is the unique solution of our equation. If primes p and q do not both divide M , then we may assume without loss of generality that $p \mid M$ and $q \nmid M$. Hence, $M = kp$, for some integer k such that $\gcd(k, q) = 1$. Because $p \mid M$, we get that $p \mid M^{de} - M$. Also, $M^{de-1} - 1 = M^{s(p-1)(q-1)} - 1 = (M^{(p-1)})^{q-1} - 1 \equiv 0 \pmod{q}$ because $\gcd(M, q) = 1$ and Fermat's little theorem. Hence, $q \mid M^{de} - M$. Since p and q are distinct primes, we obtain that $pq \mid M^{de} - M$. Therefore, $(M^d)^e \equiv M^{de} \equiv M \pmod{pq}$ showing that $x \equiv M^d \pmod{pq}$ is a solution of our equation in this case as well. ■

We continue the description of our RSA toy-example by explaining the encryption of the five numbers in Example 2.7.14. The RSA cryptosystem is an example of public key cryptography. This means that Bob will make public certain information, so that everyone can send him encrypted information. At the same time, Bob will keep some crucial information secret.

Example 2.7.15. Suppose Bob uses the primes $p = 317$ and $q = 709$. In our situation, we can find such primes using lists such as the ones from

<https://primes.utm.edu/>

or generate numbers at random and checking their primality using SageMath for example. Bob then calculates $n := pq = 224753$ and $m := (p - 1)(q - 1) = 223728$. Bob needs an exponent e such that $\gcd(e, (p - 1)(q - 1)) = 1$. This can be found using a SageMath code:

```
e=ZZ.random_element(m)
while gcd(e, m) != 1:
    e = ZZ.random_element(m)
print(e, gcd(e,m))
```

The output (when we ran the code) was:

```
31956 12
222316 4
46888 8
157738 2
91855 1.
```

Bob takes $e = 91855$. He also needs the inverse d of e modulo m . One can find it by hand using the Euclidean algorithm or one can use the SageMath command `e.inverse_mod(m)` and get that $d = 32767$.

Bob will make public the pair of numbers $(n, e) = (224753, 91855)$ and keep secret the numbers $(p, q, m, d) = (317, 709, 223728, 32767)$. Anyone who wishes to send Bob a number of at most 5 digits x , will do so by sending Bob the encrypted number

$$x^e \pmod{n}.$$

In our situation from Example 2.7.14, the first number to be sent by Alice equals $x = 31424$. Alice will encrypt it as

$$M = 31414^{91855} \pmod{224753}.$$

This calculation can be done similar to the fast powering algorithm explained in Example 2.4.9 and is also available in SageMath as:

```
x=31424
power_mod(x, e, n)
```

producing $31424^{91855} \equiv 50567 \pmod{224753}$. Hence, 31424 will be encrypted as 50567. Similarly, we can get the following correspondence

| | | | | | |
|----------------|-------|--------|--------|-------|--------|
| x | 31424 | 14201 | 77141 | 24221 | 41170 |
| $x^e \pmod{n}$ | 50567 | 112888 | 184532 | 49319 | 118659 |

Alice will send Bob the following sequence of numbers

$$50567, 112888, 184532, 49319, 118659.$$

To decrypt these numbers, Bob will raise each of these numbers to the power d modulo n . Proposition 2.7.21 guarantees that raising each of these numbers M to power d

modulo n will give back x (from $x^e \equiv M \pmod{n}$):

$$\begin{aligned} 50567^{32767} &\equiv 31424 \pmod{224753} \\ 112888^{32767} &\equiv 14201 \pmod{224753} \\ 184532^{32767} &\equiv 77141 \pmod{224753} \\ 49319^{32767} &\equiv 24221 \pmod{224753} \\ 118659^{32767} &\equiv 41170 \pmod{224753}. \end{aligned}$$

In practice, one has to use much larger numbers (hundreds of digits) than the ones we use here. To recap, Bob makes n and e available to everyone, but keeps p, q, d and $(p-1)(q-1)$ secret. If Eve knew p , then she would easily find out $q = n/p, (p-1)(q-1)$ and use them to find out d . The key idea here is that while it is easy to multiply two numbers, it is much harder to factor a number (especially a very large one).

Exercise 2.7.1. Let $N = 2^{32} + 1 = 4294967297$. What are the remainders of the division of N by 193, 257, 449 and 577?

Exercise 2.7.2. Is 6700417 a prime number? Justify your answer.

Exercise 2.7.3. Prove that $2^{2^6} + 1 = 2^{64} + 1$ is not a prime by finding a prime divisor of it of the form $128k + 1$.

Exercise 2.7.4. Pick two of your favorite prime numbers $p \geq 13$ and find all the primitive roots modulo p .

Exercise 2.7.5. Pick two of your favorite prime numbers $p \geq 13$ and find all the quadratic residues modulo p .

Exercise 2.7.6. Show that there are infinitely many primes that are of the form $3k + 1$. Write down the smallest ten of them.

Exercise 2.7.7. Let $p \geq 3$ be a prime. If a and b are integers, prove that

$$\left(\frac{ab}{p}\right) = \left(\frac{a}{p}\right) \cdot \left(\frac{b}{p}\right).$$

Exercise 2.7.8. Let $p \geq 3$ be a prime number. Show that there are exactly $\frac{p-1}{2}$ quadratic residues modulo p .

Exercise 2.7.9. Using primitive roots, construct Costas arrays of order 11 and of order 13.

Exercise 2.7.10. Pick two of your favorite prime numbers $p \neq q$ with four digits. Find a natural number e with four digits such that $\gcd(e, (p-1)(q-1)) = 1$ and determine its multiplicative inverse d modulo $(p-1)(q-1)$. Using the RSA algorithm, encrypt the message

mathematics is beautiful

with the public key (pq, e) and decrypt the result using the private key equal to $(p, q, (p-1)(q-1), d)$.

2.8. More Exercises

Exercise 2.8.1. Let n be a natural number possessing 3 digits, say $n = abc$ for some $0 \leq a, b, c \leq 9$. Build a new integer m with 6 digits by setting $m = abcabc$. For example, if $n = 398$, then $m = 398398$. Why is m always divisible by 7? Does this remain valid if a and/or a and b equal 0?

Exercise 2.8.2. Prove the following divisibility rules. A natural number n is divisible by

- (1) 7 if and only if when one subtracts twice the last digit from the remaining part, the result is a multiple of 7. For example, 4347 is divisible by 7 because the number $434 - 2 \cdot 7 = 420$ is divisible by 7,
- (2) 8 if and only if the last three digits are a multiple of 8,
- (3) 9 if and only if the sum of the digits of n is a multiple of 9,
- (4) 11 if and only if the difference of the alternating sum is a multiple of 11. For example, 1,054,482 is divisible by 11 because $1 - 0 + 5 - 4 + 4 - 8 + 2 = 0$ is a multiple of 11.

Exercise 2.8.3. Let m and n be two natural numbers. Consider the Cartesian product $\mathbb{Z}_m \times \mathbb{Z}_n$ with the binary operation:

$$(a, c) + (b, d) = ((a + c) \pmod{m}, (b + d) \pmod{n}).$$

- (1) Prove that this operation is associative, commutative with $(a, c) + (0, 0) = (a, c)$ for any $(a, c) \in \mathbb{Z}_m \times \mathbb{Z}_n$.
- (2) Show that for any $(a, c) \in \mathbb{Z}_m \times \mathbb{Z}_n$, there exists $(b, d) \in \mathbb{Z}_m \times \mathbb{Z}_n$ such that

$$(a, c) + (b, d) = (0, 0).$$

- (3) Pick your favorite coprime numbers m and n and compare the addition table of \mathbb{Z}_{mn} to the addition table of $\mathbb{Z}_m \times \mathbb{Z}_n$ ¹¹.
- (4) Pick your favorite noncoprime numbers m and n and determine the additive orders of the elements in \mathbb{Z}_{mn} and in $\mathbb{Z}_m \times \mathbb{Z}_n$ with the previous operation¹².

Exercise 2.8.4. The Pythagorean triples $(3, 4, 5), (5, 12, 13), (7, 24, 25)$, contain two consecutive natural numbers. Do all Pythagorean triples have this property?

Exercise 2.8.5. If n is an integer, then what are the possible remainders of n^2 when divided by 3, 4, 5, 6, or 7?

Exercise 2.8.6. Show that there are infinitely many primes that are of the form $3k + 2$. Write down the smallest ten of them.

Exercise 2.8.7. Prove that a binary code of length 3 and minimum distance 2 must have at most 4 codewords.

¹¹This exercise is meant to illustrate that when $\gcd(m, n) = 1$, the group $(\mathbb{Z}_{mn}, +)$ and the group $\mathbb{Z}_m \times \mathbb{Z}_n$ with the above operation are the same (isomorphic).

¹²This part is meant to indicate that when m and n are not coprime, the group $(\mathbb{Z}_{mn}, +)$ and the group $\mathbb{Z}_m \times \mathbb{Z}_n$ with the above operation are not the same (are nonisomorphic).

Exercise 2.8.8. Prove that a 3-ary code of length 4 and minimum distance 3 must have at most 9 codewords.

Exercise 2.8.9. Consider the set $\mathbb{Z}_2^3 = \mathbb{Z}_2 \times \mathbb{Z}_2 \times \mathbb{Z}_2$ of ordered triples (a, b, c) , where $a, b, c \in \mathbb{Z}_2$. Define the following binary operation:

$$(a, b, c) + (d, e, f) = (a + d, b + e, c + f),$$

where each entry is calculated modulo 2.

- (1) Prove that this operation is associative, commutative and that

$$(a, b, c) + (0, 0, 0) = (a, b, c),$$

for any $a, b, c \in \mathbb{Z}_2$.

- (2) Show that for any $(a, b, c) \in \mathbb{Z}_2^3$, there exists $(d, e, f) \in \mathbb{Z}_2^3$ such that

$$(a, b, c) + (d, e, f) = (0, 0, 0).$$

- (3) Determine the additive orders of the elements in \mathbb{Z}_8 , $\mathbb{Z}_4 \times \mathbb{Z}_2$ and \mathbb{Z}_2^3 .¹³

Exercise 2.8.10. Show that there are infinitely many primes that are of the form $4k+1$. Write down the smallest ten of them. Hint: assume there are finitely many such primes p_1, \dots, p_n and consider the number $N := (2p_1 \dots p_n)^2 + 1$.

Exercise 2.8.11. Show that there are infinitely many primes that are of the form $6k+5$. Write down the smallest ten of them.

Exercise 2.8.12. If $m > n$ are odd and coprime natural numbers, prove that the triple

$$(x, y, z) = \left(mn, \frac{m^2 - n^2}{2}, \frac{m^2 + n^2}{2} \right)$$

is a Pythagorean triple and satisfies $\gcd(x, y, z) = 1$.

Exercise 2.8.13. Are there integers x, y and z such that $101x - 47y + 71z = 1$?

Exercise 2.8.14. Find all the Sophie Germain primes less than 200.

Exercise 2.8.15. The ISBN $0 - 8018 - 9?99 - 4$ has a smudge where the ? is. What is the missing digit?

Exercise 2.8.16. Show that each of the matrices A_1, \dots, A_{p-1} constructed in the proof of Theorem 2.4.14 is a Latin square.

Exercise 2.8.17. Construct three mutually orthogonal Latin squares of order 4.

Exercise 2.8.18. Construct a Costas array of order 6.

¹³This is meant to indicate that these groups of order 8 are different (nonisomorphic).

Exercise 2.8.19. A **Golomb ruler**¹⁴ is a set of integers $\{a_1, \dots, a_m\}$ for some $m \geq 2$ such that $a_1 < \dots < a_m$ and all differences $a_j - a_i$ for $1 \leq i < j \leq m$ are distinct. The order is m and the length is $a_m - a_1$. A Golomb ruler of order m is called **optimal** if its length is the smallest among all rulers of order m .

- (1) Prove that $\{0, 1\}$ is an optimal Golomb ruler of order 2.
- (2) Prove that $\{0, 1, 3\}$ is an optimal Golomb ruler of order 3.
- (3) Prove that $\{0, 1, 4, 6\}$ is an optimal Golomb ruler of order 4.

Exercise 2.8.20. A **finite Sidon set**¹⁵ is a finite set $\{b_1, \dots, b_m\}$ of natural numbers such that all the pairwise sums $b_i + b_j$, for $1 \leq i \leq j \leq m$, are different. Prove that every finite Sidon set is a Golomb ruler and vice versa.

¹⁴This notion was named after the American mathematician Solomon Golomb (1932–2016).

¹⁵This object was introduced by the Hungarian mathematician Simon Sidon (1897–1941).

Rational Numbers \mathbb{Q}

Those who have one foot in the canoe, and one foot in the boat, are going to fall into the river.

Tuscarora Proverb

3.1. Basic Properties

Similar to the introduction of zero and negative integers, practical and theoretical considerations led to the development and study of rational numbers. In many situations, one is required to measure or divide certain quantities in equal parts (for example, a cookie in half, a dollar in four equal parts, a pizza in six slices, or a plot of land in some given number of parts). Splitting a cookie in half is simple enough and goes back to Thales, just draw a straight line through the center.

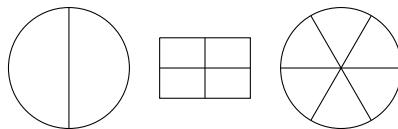


Figure 3.1.1. Dividing a cookie, a rectangle, and a pizza.

Dividing a dollar in four parts requires introducing a smaller unit (say a cent, such that one dollar consists of 100 cents) which allows the division of the dollar into four parts of 25 cents each. Of course, one can use fewer coins if we divide our dollar into four quarters. We can divide a disk into six equal parts using a ruler and a compass although such an operation may be messy when applied to a pizza.

What happens if we wish to divide our one dollar among three people? Our previous solution of creating *subunits* of cents or quarters to help us divide the quantity evenly does not seem to work. An alternative is to come up with another subunit of measurement so that when dividing one dollar into three parts using these subunits,

the parts will be equal. We denote this subunit part by $1/3$ or $\frac{1}{3}$. This is an example of a rational number or fraction or ratio. In general, for any natural number n , we may divide our unit of measurement into n equal parts. The length or size of each sub-unit is *one divided by n* which we denote by $\frac{1}{n}$ or $1/n$. A number that equals exactly m such subunits will be denoted by $\frac{m}{n}$ or m/n . In this situation, the problem of measurement/division has been reduced to the one of counting the number of subunits.

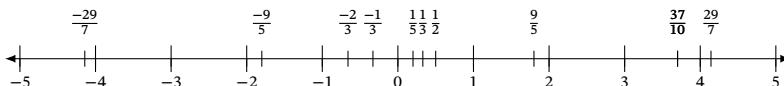


Figure 3.1.2. A representation of some rational numbers.

We call m the **numerator** and n the **denominator** of the fraction m/n . Note that while one third of a dollar is not a commonly used subunit, we use units and subunits all the time: 1 meter consists of 100 centimeters, 1 tonne is 1000 kilograms, 1 kilogram is 1000 grams, 1 foot is 12 inches, 1 pound is 16 ounces, 1 hour is 60 minutes, 1 minute is 60 seconds and so on.

Definition 3.1.1. The set of **rational numbers** \mathbb{Q} is

$$\mathbb{Q} = \{m/n : m, n \in \mathbb{Z}, n \neq 0\}.$$

Geometrically, we can think about rational numbers as staking out even more places on the real number line. In previous chapters, we observed that natural numbers can be seen as stakes starting at 1 that go on forever to the right one unit apart. The integers extended this collection to the left again going forever. The rational numbers will contain all these marks on the real line, but will have many more marks in the gaps between any consecutive integers.

The introduction of integer numbers allowed us to expand the collection of equations that we can solve from linear equations of the form $x - 3 = 0$ to those like $x + 7 = 0$ or $2x + 9 = 3$. The rational numbers allow us to further extend our reach and solve any linear equation of the form

$$(3.1.1) \quad ax = b, \text{ where } a \neq 0 \text{ and } b \text{ are integers.}$$

In the realm of natural numbers, this equation has a solution if and only if a divides b and both a and b are natural numbers. When solving the same equation in integer numbers, we will have a solution if and only if a divides b . Over rational numbers, the equation (3.1.1) always has a unique solution, namely b/a or $\frac{b}{a}$. Thus, rational numbers enable us to enlarge our horizons by making division as well as solving more equations possible. We will see this phenomenon again in the next chapters when extending the realm of numbers from rationals to reals and from reals to complex numbers.

Another interpretation of rational numbers can be given using the Cartesian plane system. A rational number m/n with $m, n \in \mathbb{Z}, n \neq 0$, can be *identified* with the line $mx = ny$. Thus, the number $\frac{2}{1}$ would be seen as the line $2x = y$, the number $\frac{1}{1}$ as

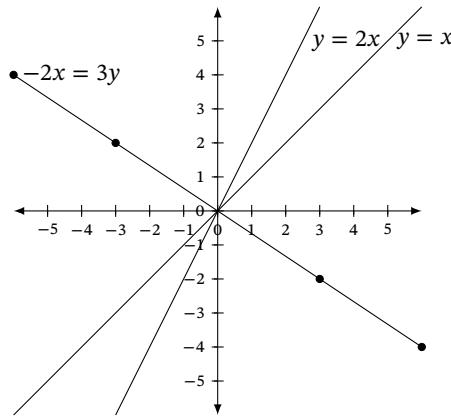


Figure 3.1.3. Interpretation of rational numbers as slopes of lines.

the line $x = y$ and the number $\frac{-2}{3}$ as the line $-2x = 3y$. This way, $\frac{-4}{6} = \frac{-2}{3} = \frac{2}{-3} = \frac{4}{-6}$ as each fraction corresponds to the same line $-2x = 3y$. Note that this line interpretation is a geometric way of viewing rational numbers as equivalence classes of pairs of integers. If \mathbb{Z}^* denotes the set of nonzero integers, then we can define a binary relation on the set

$$\mathbb{Z} \times \mathbb{Z}^* = \{(m, n) : m \in \mathbb{Z}, n \in \mathbb{Z}^*\}$$

as follows:

$$(3.1.2) \quad (m, n) \sim (p, q) \Leftrightarrow mq = np.$$

Proposition 3.1.1. *The binary relation \sim is reflexive, symmetric, and transitive.*

Proof: To prove reflexivity, we show that $(m, n) \sim (m, n)$ for any $(m, n) \in \mathbb{Z} \times \mathbb{Z}^*$. This is true since $mn = nm$.

For symmetry, let $(m, n), (p, q) \in \mathbb{Z} \times \mathbb{Z}^*$ such that $(m, n) \sim (p, q)$. We will prove that $(p, q) \sim (m, n)$. From $(m, n) \sim (p, q)$, it follows that $mq = np$, which is the same as $pn = qm$ implying that $(p, q) \sim (m, n)$.

To show transitivity, let $(m, n), (p, q), (r, s) \in \mathbb{Z} \times \mathbb{Z}^*$ such that $(m, n) \sim (p, q)$ and $(p, q) \sim (r, s)$. We want to prove that $(m, n) \sim (r, s)$. From $(m, n) \sim (p, q)$, we get that $mq = np$. From $(p, q) \sim (r, s)$, we have that $ps = qr$. Multiplying the equation $mq = np$ by s , we get that $mqs = nps$. Substituting ps by qr on the right-hand side, we deduce that $mqs = n(ps) = n(qr)$. Because $q \neq 0$, we can divide both sides by q and obtain that $ms = nr$. This means that $(m, n) \sim (r, s)$ which finishes our proof. ■

The rational numbers are the equivalence classes corresponding to this binary relation. The rules for adding, subtracting, multiplying and dividing rational numbers are taught to us since young ages and everyone is familiar with them. For any integers

m, n, p, q with $n \neq 0, q \neq 0$, we have that

$$(3.1.3) \quad \begin{aligned} \frac{m}{n} + \frac{p}{q} &= \frac{mq + np}{nq}, & \frac{m}{n} \cdot \frac{p}{q} &= \frac{mp}{nq} \\ \frac{m}{n} - \frac{p}{q} &= \frac{mq - np}{nq}, & \frac{\frac{m}{n}}{\frac{p}{q}} &= \frac{mq}{np} \text{ when } p \neq 0. \end{aligned}$$

These operations are well-defined, meaning $(m, n) \sim (m', n')$ and $(p, q) \sim (p', q')$ yield

$$\begin{aligned} \frac{m}{n} + \frac{p}{q} &= \frac{m'}{n'} + \frac{p'}{q'}, & \frac{m}{n} \cdot \frac{p}{q} &= \frac{m'}{n'} \cdot \frac{p'}{q'}, \\ \frac{m}{n} - \frac{p}{q} &= \frac{m'}{n'} - \frac{p'}{q'}, & \frac{\frac{m}{n}}{\frac{p}{q}} &= \frac{\frac{m'}{n'}}{\frac{p'}{q'}} \text{ when } pp' \neq 0. \end{aligned}$$

For example,

$$\begin{aligned} \frac{3}{4} + \frac{-2}{3} &= \frac{3 \cdot 3 + 4 \cdot (-2)}{4 \cdot 3} = \frac{9 - 8}{12} = \frac{1}{12}, & \frac{3}{4} \cdot \frac{-2}{3} &= \frac{3 \cdot (-2)}{4 \cdot 3} = \frac{-6}{12} = \frac{-1}{2}, \\ \frac{3}{4} - \frac{-2}{3} &= \frac{3 \cdot 3 - 4 \cdot (-2)}{4 \cdot 3} = \frac{9 + 8}{12} = \frac{17}{12}, & \frac{\frac{3}{4}}{\frac{-2}{3}} &= \frac{\frac{3}{4} \cdot 3}{4 \cdot (-2)} = \frac{9}{-8} = \frac{-9}{8}. \end{aligned}$$

If we identify the integer $m \in \mathbb{Z}$ with the rational number $\frac{m}{1}$, we may consider \mathbb{Z} as a subset of \mathbb{Q} . In particular, $\frac{1}{1} = 1$ and $\frac{0}{n} = 0$, for any nonzero integer n . Moreover, the algebraic operations on \mathbb{Q} introduced above are extensions of the corresponding operations for natural and integer numbers and satisfy the same properties (addition and multiplication are commutative and associative and addition is distributive with respect to multiplication). Similar to before, we have the special number 0 such that any rational number added to it does not change. Also, for any rational number m/n , there exists a unique rational number $-m/n$ such that $m/n + (-m/n) = 0$. The number $-m/n$ is sometimes called the additive inverse of m/n . The above property did not hold for natural numbers, but it did for integers. A new feature of operations over rational numbers involves division, as one may expect. We have another special number 1 such that the product of any rational number by it does not change the value of that rational number. This happened before for both natural and integer numbers. However, now for any nonzero rational number m/n , there exists a unique nonzero rational number n/m such that $m/n \cdot n/m = 1$. The number n/m is the multiplicative inverse of m/n and also denoted by $(m/n)^{-1}$. We can divide 1 by m/n and get n/m . This leads us to be able to divide any rational number by another nonzero rational number. In other words, if $m, n \in \mathbb{Z}^*$, then

$$\left(\frac{m}{n}\right)^{-1} = \frac{1}{\frac{m}{n}} = \frac{n}{m}.$$

This feature was not always possible for natural and integer numbers.

The set of rational numbers \mathbb{Q} with the operations of addition and multiplication (and implicitly, subtraction and division) is an example of a mathematical object called a **field**. It is an infinite field as there are infinitely many rational numbers. In the previous chapter, we saw examples of finite fields, the sets of integers \mathbb{Z}_p modulo p , where p is a prime number. Fields are fundamental mathematical objects and their

study is an important part of algebra. Other basic (infinite) fields are those of real and of complex numbers which we will treat in the following chapters.

The rational numbers have another important feature, namely, there exists an order on \mathbb{Q} which is compatible with the algebraic operations. Fields having such an order are usually called **ordered fields**. There exist many ordered fields different from \mathbb{Q} . The most important among them is \mathbb{R} , the field of real numbers. Another one is that of algebraic numbers as introduced in Remark 4.7.5, and there are many more of interest. In contrast to that, the field \mathbb{Z}_p cannot be ordered in a way such that the sum and the product of positive elements are positive again. Similarly, the field \mathbb{C} does not possess an order compatible with its algebraic operations.

The best way to introduce the order on \mathbb{Q} is by describing its positive elements.

Definition 3.1.2. A fraction m/n in \mathbb{Q} is said to be **positive** if either $m, n > 0$ or $m, n < 0$. Let us write $m/n > 0$ in this case.

$$\frac{m}{n} > 0 \Leftrightarrow m, n > 0 \text{ or } m, n < 0 \Leftrightarrow m \cdot n > 0.$$

Set

$$\mathbb{Q}_+ := \left\{ \frac{m}{n} \in \mathbb{Q} : \frac{m}{n} > 0 \right\} = \left\{ \frac{m}{n} \in \mathbb{Q} : m \cdot n > 0 \right\}.$$

Remark 3.1.1. Recall that \mathbb{Q} is in fact a set of equivalence classes with respect to the relation \sim introduced in (3.1.2). Therefore, in order to make the previous definition precise, one has to prove the following fact:

$$(3.1.4) \quad (m, n) \sim (p, q) \Rightarrow \left(\frac{m}{n} > 0 \Leftrightarrow \frac{p}{q} > 0 \right).$$

So, for example, we have to know that $4/8 > 0$ if and only if $8/16 > 0$ or if and only if $(-1)/(-2) > 0$. We leave its verification of (3.1.4) as an exercise (see Exercise 3.1.3).

We summarize now some properties of the set \mathbb{Q}_+ of positive rational numbers.

Proposition 3.1.2. *The set \mathbb{Q}_+ possesses the following properties:*

$$\begin{aligned} (1) \quad \frac{m}{n}, \frac{p}{q} \in \mathbb{Q}_+ &\Rightarrow \frac{m}{n} + \frac{p}{q} \in \mathbb{Q}_+ \\ (2) \quad \frac{m}{n}, \frac{p}{q} \in \mathbb{Q}_+ &\Rightarrow \frac{m}{n} \cdot \frac{p}{q} \in \mathbb{Q}_+ \end{aligned}$$

(3) *For any fraction $m/n \in \mathbb{Q}$ we have either $m/n \in \mathbb{Q}_+$ or $-m/n \in \mathbb{Q}_+$ or $m/n = 0$.*

Proof: To prove (1) and (2) choose two fractions m/n and (p, q) in \mathbb{Q}_+ . Since

$$(m, n) \sim (-m, -n) \quad \text{and} \quad (p, q) \sim (-p, -q),$$

without losing generality we may assume that all integers m, n, p and q are positive. Recall that neither addition nor multiplication of fractions depend on the special representation of the chosen fractions.

Of course, since all integers are positive, we also have

$$mq + np > 0 \quad \text{as well as} \quad nq > 0.$$

Hence, by definition

$$\frac{m}{n} + \frac{p}{q} = \frac{mq + np}{nq} > 0 \quad \Rightarrow \quad \frac{m}{n} + \frac{p}{q} \in \mathbb{Q}_+.$$

This completes the proof of (1).

The proof of (2) is even easier. Here we get

$$\frac{m}{n} \cdot \frac{p}{q} = \frac{mp}{nq} > 0$$

by $mp > 0$ and $nq > 0$.

Finally, if m/n is an arbitrary fraction, then either

$$mn > 0 \quad \text{or} \quad mn < 0 \quad \text{or} \quad mn = 0.$$

The first case corresponds to $m/n > 0$, the second one to

$$(-m)n > 0 \quad \Leftrightarrow \quad -\frac{m}{n} = -\frac{m}{n} > 0.$$

One has $mn = 0$ if and only if $m/n = 0$. Thus, the proposition is proven. ■

Remark 3.1.2. If we define the subset $\mathbb{Q}_- \subseteq \mathbb{Q}$ by

$$\mathbb{Q}_- := \left\{ \frac{m}{n} \in \mathbb{Q} : -\frac{m}{n} > 0 \right\},$$

then property (3) of Proposition 3.1.2 says that \mathbb{Q} is the disjoint union

$$\mathbb{Q} = \mathbb{Q}_+ \cup \mathbb{Q}_- \cup \{0\}.$$

Fractions in \mathbb{Q}_- are called **negative**.

Definition 3.1.3. Given two fractions m/n and p/q we write

$$(3.1.5) \quad \frac{m}{n} < \frac{p}{q} \quad :\Leftrightarrow \quad \frac{p}{q} - \frac{m}{n} > 0 \quad :\Leftrightarrow \quad \frac{np - mq}{nq} > 0.$$

In particular,

$$nq > 0 \quad \Rightarrow \quad \left(\frac{m}{n} < \frac{p}{q} \Leftrightarrow mq < pn \right).$$

Finally, we define a relation \leq on \mathbb{Q} by

$$\frac{m}{n} \leq \frac{p}{q} \quad \text{if either} \quad \frac{m}{n} < \frac{p}{q} \quad \text{or} \quad \frac{m}{n} = \frac{p}{q}.$$

Example 3.1.1.

$$\frac{3}{7} < \frac{5}{9} \quad \text{because of } 7 \cdot 9 > 0 \text{ and } 3 \cdot 9 < 5 \cdot 7.$$

$$\frac{5}{-9} < \frac{3}{-7} \quad \text{because of } (-7) \cdot (-9) > 0 \text{ and } 5 \cdot (-7) < 3 \cdot (-9).$$

The binary relation \leq is a (total) order on \mathbb{Q} . We refer to Section A.5 for the properties of an order relation. Furthermore, this relation is compatible with the algebraic operations. That is

$$(3.1.6) \quad \left(\forall \frac{s}{t} \in \mathbb{Q} \right) \left(\frac{m}{n} \leq \frac{p}{q} \Rightarrow \frac{m}{n} + \frac{s}{t} \leq \frac{p}{q} + \frac{s}{t} \right)$$

$$(3.1.7) \quad \left(\forall \frac{s}{t} > 0 \right) \left(\frac{m}{n} \leq \frac{p}{q} \Rightarrow \frac{m}{n} \cdot \frac{s}{t} \leq \frac{p}{q} \cdot \frac{s}{t} \right).$$

The proof of these properties is easy and follows directly from the definition of the order and by Proposition 3.1.2. Furthermore, we will investigate those problems in a more general context in Section 4.3 below.

Let us finally mention that not only are the algebraic operations on \mathbb{Q} natural extensions of those on \mathbb{Z} but also its order. Since we identify $m \in \mathbb{Z}$ with the fraction $m/1 \in \mathbb{Q}$ this easily follows by

$$\begin{aligned} \frac{m}{1} + \frac{p}{1} &= \frac{m \cdot 1 + p \cdot 1}{1 \cdot 1} = \frac{m + p}{1}, \\ \frac{m}{1} \cdot \frac{p}{1} &= \frac{m \cdot p}{1 \cdot 1} = \frac{m \cdot p}{1} \quad \text{and by} \\ \frac{m}{1} \leq \frac{p}{1} &\Leftrightarrow \frac{p \cdot 1 - m \cdot 1}{1 \cdot 1} > 0 \Leftrightarrow m \leq p. \end{aligned}$$

Exercise 3.1.1. Show that addition and multiplication are associative operations connected by the distributive law. In other words, verify the following: Given fractions m/n , p/q and s/t with $n, q, t \neq 0$, then

$$\begin{aligned} (i) \quad & \left(\frac{m}{n} + \frac{p}{q} \right) + \frac{s}{t} = \frac{m}{n} + \left(\frac{p}{q} + \frac{s}{t} \right) \\ (ii) \quad & \left(\frac{m}{n} \cdot \frac{p}{q} \right) \cdot \frac{s}{t} = \frac{m}{n} \cdot \left(\frac{p}{q} \cdot \frac{s}{t} \right) \\ (iii) \quad & \frac{m}{n} \cdot \left(\frac{p}{q} + \frac{s}{t} \right) = \frac{m}{n} \cdot \frac{p}{q} + \frac{m}{n} \cdot \frac{s}{t} \end{aligned}$$

Exercise 3.1.2. Order the following fractions by their size without using a pocket calculator.

$$\frac{3}{8}, \quad \frac{13}{17}, \quad \frac{2 + \frac{1}{2}}{3 + \frac{1}{3}}, \quad \frac{-2}{-3}, \quad \frac{1}{3} + \frac{1}{4}.$$

Exercise 3.1.3. Prove the statement (3.1.4) which asserts that the fractions of equivalent pairs of integers are either both positive or both are negative.

Exercise 3.1.4. Show that the order on \mathbb{Q} is compatible with the algebraic operations as it is stated in (3.1.6) and in (3.1.7).

Exercise 3.1.5. Prove the following properties of the order on \mathbb{Q} . Hereby assume that all appearing fractions are well-defined.

- (1) $\left(\forall \frac{m}{n}, \frac{p}{q}, \frac{s}{t}, \frac{u}{v}\right) \left[\left(\frac{m}{n} \leq \frac{p}{q}, \frac{s}{t} \leq \frac{u}{v} \right) \Rightarrow \left(\frac{m}{n} + \frac{s}{t} \leq \frac{p}{q} + \frac{u}{v} \right) \right].$
- (2) $\left(\forall \frac{s}{t} < 0\right) \left[\left(\frac{m}{n} \leq \frac{p}{q} \right) \Rightarrow \left(\frac{p}{q} \cdot \frac{s}{t} \leq \frac{m}{n} \cdot \frac{s}{t} \right) \right].$
- (3) $\left(\forall \frac{m}{n}, \frac{p}{q} > 0\right) \left[\left(\frac{m}{n} \leq \frac{p}{q} \right) \Leftrightarrow \left(\frac{q}{p} \leq \frac{n}{m} \right) \right].$
- (4) $\left(\forall \frac{m}{n}, \frac{p}{q}\right) \left[\left(\frac{m}{n} \leq \frac{p}{q} \right) \Leftrightarrow \left(-\frac{p}{q} \leq -\frac{m}{n} \right) \right].$
- (5) $\left(\forall m, n \in \mathbb{N}\right) \left[\left(m \geq n \right) \Rightarrow \left(\frac{m}{n} \geq 1 \right) \right].$
- (6) $\left(\forall m, n, p, q \in \mathbb{N}\right) \left[\left(m \leq p, q \leq n \right) \Rightarrow \left(\frac{m}{n} \leq \frac{p}{q} \right) \right].$

Exercise 3.1.6. For each $n \in \mathbb{N}$ define the rational number x_n by $x_n = \frac{n}{n+1}$. Show that

$$\frac{1}{2} = x_1 < x_2 < x_3 < \dots < x_n < x_{n+1} < \dots < 1.$$

Prove that for any $k \in \mathbb{N}$, there is a natural number N (that depends on k) such that

$$|x_n - 1| < 1/k, \forall n \geq N.$$

Here, as in the case of integers, for any $q \in \mathbb{Q}$ its absolute value is defined by

$$|q| = \begin{cases} q & : q \geq 0 \\ -q & : q < 0. \end{cases}$$

Exercise 3.1.7. Write down all the fractions $\frac{m}{n}$ in the interval $[0, 1]$ with integers $m > 0$ and $n \leq 5$.

Exercise 3.1.8. Find the fraction m/n that is the closest, but not equal to $1/3$, where $0 < m, n \leq 10$.

Exercise 3.1.9. Show that the set \mathbb{Q} of rational numbers satisfies the Archimedean property, meaning that for each rational number $q > 0$ there is some $n \in \mathbb{N}$ such that $1/n < q$.

Exercise 3.1.10. Let q be a nonnegative rational number. If $q < 1/n$ for all $n \in \mathbb{N}$, prove that $q = 0$.

3.2. Not Everything Is Rational

The aim of this section is to show that there exist numbers that are not rational¹. Our simplest example of such a number comes from geometry. In the Figure 3.2.1, consider the square whose opposite corners have coordinates $(0, 0)$ and $(1, 1)$. By Pythagorean theorem, if d is the length of the diagonal of our square (which is the segment joining $(0, 0)$ and $(1, 1)$), then d is a positive and $d^2 = 1^2 + 1^2 = 2$. Thus, $d^2 = 2$ and we give d the name $\sqrt{2}$.

¹This was already observed by the Pythagoreans, but the first proof was published in Euclid's *Elements*.

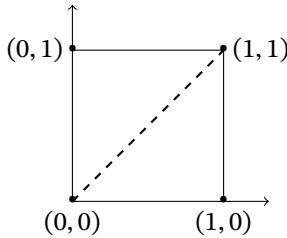


Figure 3.2.1. The diagonal of the square has length $\sqrt{2}$.

A *calculus* interpretation of $\sqrt{2}$ can be found by considering the graph of the function $y = x^2$ for $0 \leq x \leq 2$ in Figure 3.2.2. The function is increasing with x on the interval $[0, 2]$ and its graph is a continuous curve that connects $(1, 1)$ to $(2, 4)$. Thus, at some point this graph will intersect the horizontal line $y = 2$ and the x -coordinate of the intersection point will be $\sqrt{2}$.

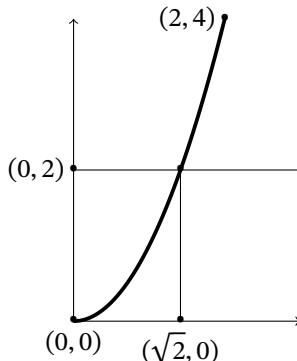


Figure 3.2.2. The graph of $y = x^2$ for $0 \leq x \leq 2$.

Another appearance of $\sqrt{2}$ is in the realm of paper. The sizes (height,width) of the paper sizes A_0, A_1, \dots, A_8 are all proportional, meaning that if (h_j, w_j) denotes the height and the width of the A_j paper size, then the ratio h_j/w_j is the same for every $0 \leq j \leq 8$. Let us denote it by r . In addition, for any $0 \leq j \leq 7$, folding an A_j size sheet in half on its larger side (which is the height) creates two halves that have the dimensions of the next smaller A_{j+1} size. This means that

$$h_{j+1} = w_j, \quad w_{j+1} = h_j/2.$$

Since $r = h_{j+1}/w_{j+1} = h_j/w_j$, dividing the first equation by the second one we get that

$$r = h_{j+1}/w_{j+1} = 2w_j/h_j = 2/r,$$

which implies that $r^2 = 2$. Since r is positive, $r = \sqrt{2}$.

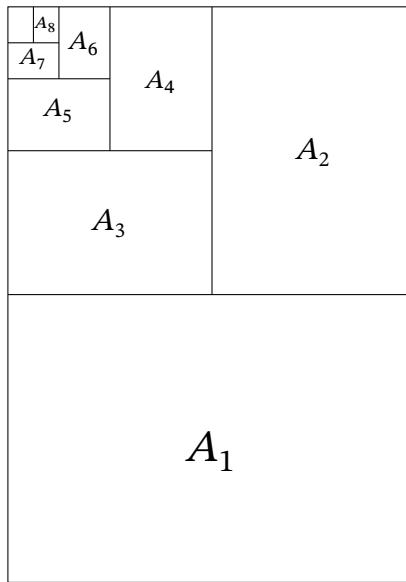


Figure 3.2.3. An illustration of various A paper sizes.

We will need the following definition for our next result.

Definition 3.2.1. A rational number or fraction a/b with a and $b \neq 0$ integers is said to be in **lowest terms** or **reduced** or **irreducible** if $\gcd(a, b) = 1$.

Example 3.2.1. The fractions $\frac{1}{2}, \frac{7}{3}, \frac{-2}{5}, \frac{8}{-11}, \frac{-9}{-22}$ are reduced.

The fractions $\frac{2}{4}, \frac{21}{9}, \frac{6}{-15}, \frac{-32}{44}, \frac{-45}{-110}$ are not reduced.

For any rational number a/b , there exists a reduced fraction a'/b' such that $a/b = a'/b'$. Simply, take $a' = a/d$ and $b' = b/d$, where $d = \gcd(a, b)$. Actually, there are precisely two such reduced fractions: one with positive denominator and one with negative denominator.

Definition 3.2.2. A nonrational number x is called **irrational**.

From this definition it follows that if x is irrational, then for any integers a and $b \neq 0$, $x \neq a/b$. Note that any proof of irrationality of a given number will be a proof by contradiction. This is because we know how a rational number looks like, but there is no such description for an irrational number.

Theorem 3.2.1. *The number $\sqrt{2}$ is irrational. That is, there are no integers a and b satisfying $(a/b)^2 = 2$.*

Proof: We will give several proofs below.

1st proof. This proof uses the unique factorization of natural numbers as product of primes (see Theorem 1.6.5). Assume that $\sqrt{2} = \frac{a}{b}$ for some integers a and $b \neq 0$. It follows that $a^2 = 2b^2$. By Theorem 1.6.5, the exponent of 2 must be the same in

both sides. However, the exponent of 2 in a^2 is even (as it equals twice the exponent of 2 in the prime factorization of a) while the exponent of 2 in $2b^2$ is odd (by a similar argument), a contradiction. This proof can be generalized (see Exercise 3.2.3).

2nd proof. Assume that $\sqrt{2} = \frac{a}{b}$ for some integers a and $b \neq 0$. Without loss of generality, we may assume that $\frac{a}{b}$ is in lowest terms, meaning that $\gcd(a, b) = 1$, and that a and b are positive. Squaring both sides, we obtain that $2 = (\sqrt{2})^2 = \left(\frac{a}{b}\right)^2 = \frac{a^2}{b^2}$. Therefore, $2b^2 = a^2$. From here on, we can get a contradiction in two different ways. The first approach is as follows. Because $\gcd(a, b) = 1$, we get that $\gcd(a^2, b^2) = 1$. Combining this fact with $b^2 \mid 2b^2 = a^2$ and Proposition 1.6.6, we deduce that $b^2 = 1$ and therefore, $\frac{a}{b}$ must be a natural number. However, the square of no natural number equals 2 and this gives us the desired contradiction. We will generalize this method later in Proposition 3.2.3.

3rd proof.² Another way of getting a contradiction is called the method of infinite descent. Start as in the previous proof. From $a^2 = 2b^2$, we deduce that $2 \mid a^2$. Therefore, $2 \mid a$. Hence, $a = 2a_1$ for some integer a_1 . Plugging this back into the equation $2b^2 = a^2$, we obtain that $2b^2 = a^2 = (2a_1)^2 = 4a_1^2$. Dividing both sides by 2, we have that $b^2 = 2a_1^2$. Hence, $2 \mid b^2$ implying that $2 \mid b$. Thus, $2 \mid a$ and $2 \mid b$ which is a contradiction with $\gcd(a, b) = 1$.

4th proof.³ We start with the observation that $\sqrt{2} = \frac{a}{b}$ for some natural numbers a and b is equivalent to $a^2 - 2b^2 = 0$. It is therefore natural to investigate the minimum value of $|a^2 - 2b^2|$ as a and b range over natural numbers. Clearly, this minimum is at most 1 since $|3^2 - 2 \cdot 2^2| = 1$. One can search for other pairs (a, b) such that $|a^2 - 2b^2| \leq 1$. This can be done by hand, but the easiest way is by using a computer. The short code below written in SageMath searches for all pairs (a, b) so that $1 \leq a, b \leq 300$ and $|a^2 - 2b^2| \leq 1$.

```
min=1;
for a in [1..300]:
    for b in [1..300]:
        if (abs(a*a-2*b*b)<=min):
            min=abs(a*a-2*b*b); print(a,b);
```

It produces the following pairs of numbers:

$$(3.2.1) \quad (1, 1), (3, 2), (7, 5), (17, 12), (41, 29), (99, 70), (239, 169).$$

Examining these pairs, we may notice a pattern that if (m, n) is a pair in the list, then the next pair is $(m + 2n, m + n)$. This leads us to note that

$$\begin{aligned} (m + 2n)^2 - 2(m + n)^2 &= m^2 + 4mn + 4n^2 - 2m^2 - 4mn - 2n^2 \\ &= 2n^2 - m^2, \end{aligned}$$

which implies that

$$(3.2.2) \quad |m^2 - 2n^2| = |(m + 2n)^2 - 2(m + n)^2|.$$

²This is Euclid's original proof.

³This proof is based on Doron Zeilberger's article *Two Motivated Concrete Proofs (much better than the usual one) that the Square-root of 2 is Irrational*. The article is available at <https://arxiv.org/abs/1410.2304>.

At this point, observe that equation (3.2.2) is true for any integers (m, n) and not just for any integers from our list above. If

$$s = m + 2n, t = m + n,$$

we can deduce that

$$m = 2t - s, n = s - t.$$

Hence, we have obtained another identity

$$(3.2.3) \quad |s^2 - 2t^2| = |(2t - s)^2 - 2(s - t)^2|,$$

for any integers s and t . If s and t are natural numbers, the sum of the pair $(2t - s, s - t)$ equals $2t - s + s - t = t$ and is strictly smaller than the sum of the pair (s, t) .

After all this work, we can now prove that $\sqrt{2}$ is irrational. Assume by contradiction that $\sqrt{2} = \frac{s}{t}$ for some natural number s and t . By the least element principle, choose such a pair (s, t) such that their sum is minimal among all such pairs. Since $\sqrt{2} > 1$, we have that $s > t$. Also, since $\sqrt{2} < 2$, we get that $s < 2t$. From the equation (3.2.3), we deduce that

$$0 = |s^2 - 2t^2| = |(2t - s)^2 - 2(s - t)^2|.$$

By our inequalities above the numbers $2t - s$ and $s - t$ are natural numbers and therefore

$$\sqrt{2} = \frac{2t - s}{s - t}.$$

Thus, we have found a pair of natural numbers whose ratio is $\sqrt{2}$. However, their sum is t which is strictly less than $s + t$. This contradicts the definition of (s, t) .

5th proof. We give now a geometric proof which is equivalent to our previous proof. It is due to the American mathematician Stanley Tennenbaum (1927–2005) in the 1950s. Assume as above that $\sqrt{2}$ is rational and that $\sqrt{2} = \frac{s}{t}$ for some natural numbers s and t . We can choose s and t such that $s + t$ is smallest. Because $s^2 = 2t^2$, it means that the area of a $s \times s$ square equals twice the area of a $t \times t$ square. This is described in Figure 3.2.4.

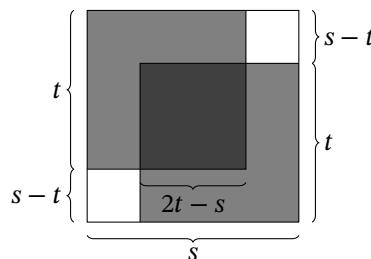


Figure 3.2.4. Geometric proof $\sqrt{2}$ is irrational.

Subtracting from the area of the $s \times s$ the area of one of the gray $t \times t$ squares, we deduce that the area of the dark gray $(2t - s) \times (2t - s)$ square equals twice the area of the white $(s - t) \times (s - t)$ square. This is essentially the geometric interpretation of the identity (3.2.3). This proof ends as the previous one with the contradiction that we

have a new pair $(2t - s, s - t)$ of natural numbers such that $(2t - s)^2 = 2(s - t)^2$ with a sum $2t - s + s - t = t$ which is smaller than the minimum sum $s + t$ of the pair (s, t) . This geometric proof can be generalized (compare [21]). ■

The following result gives a lower bound on the distance between $\sqrt{2}$ and any rational number.

Proposition 3.2.2. *If $\frac{a}{b}$ is a positive rational number, then*

$$\left| \sqrt{2} - \frac{a}{b} \right| \geq \frac{1}{3b^2}.$$

Proof: From the previous proposition, we know that $|a^2 - 2b^2| \geq 1$. Using the identity $(a + b\sqrt{2})(a - b\sqrt{2}) = a^2 - 2b^2$, we can write that

$$\left| \sqrt{2} - \frac{a}{b} \right| = \frac{|a^2 - 2b^2|}{b(a + b\sqrt{2})} \geq \frac{1}{b(a + b\sqrt{2})} = \frac{1}{b^2(\sqrt{2} + \frac{a}{b})}.$$

If $\frac{a}{b} \leq 3 - \sqrt{2}$, then the previous inequality implies that

$$\left| \sqrt{2} - \frac{a}{b} \right| \geq \frac{1}{b^2(\sqrt{2} + \frac{a}{b})} \geq \frac{1}{3b^2},$$

and we are done. Otherwise, if $\frac{a}{b} > 3 - \sqrt{2} > \sqrt{2}$, then

$$\left| \sqrt{2} - \frac{a}{b} \right| = \frac{a}{b} - \sqrt{2} > 3 - 2\sqrt{2}.$$

If $b \geq 2$, then $3 - 2\sqrt{2} > 1/12 \geq 1/(3b^2)$. The first inequality $3 - 2\sqrt{2} > 1/12$ is true as it is equivalent with $35/12 > 2\sqrt{2}$ or $35 > 24\sqrt{2}$ or $35^2 = 1225 > 1154 = (24\sqrt{2})^2$. If $b = 1$, then

$$\left| \sqrt{2} - \frac{a}{b} \right| = |\sqrt{2} - a| \geq \sqrt{2} - 1 > \frac{1}{3} = \frac{1}{3b^2}.$$

Here, we used the fact that the closest integer to $\sqrt{2}$ is 1. This is true as $1 < \sqrt{2} < 2$ and $2 - \sqrt{2} > \sqrt{2} - 1$. We also used the inequality $\sqrt{2} - 1 > 1/3$ which is the same as $3\sqrt{2} > 4$. This is true as $(3\sqrt{2})^2 = 18 > 16 = 4^2$, and it finishes our proof. ■

The number $\sqrt{2}$ is a solution of the equation $x^2 - 2 = 0$. The next result can be used to prove that other numbers are irrational.

Proposition 3.2.3. *Let $n \in \mathbb{N}$ and $c_0, \dots, c_n \in \mathbb{Z}$ such that $c_n \neq 0$. If a/b is a reduced rational number that is a solution of the equation:*

$$(3.2.4) \quad c_n x^n + \dots + c_1 x + c_0 = 0,$$

then $a | c_0$ and $b | c_n$.

Proof: Because a/b is a solution of the equation (3.2.4), we have that

$$c_n \left(\frac{a}{b} \right)^n + \dots + c_1 \left(\frac{a}{b} \right) + c_0 = 0.$$

Bringing the left-hand side to the common denominator b^n , we obtain that

$$c_n a^n + \dots + c_1 a b^{n-1} + c_0 b^n = 0.$$

Hence, a and b must divide the left-hand side above. This implies that $a \mid c_0 b^n$ and $b \mid c_n a^n$. Since $\gcd(a, b) = 1$, Lemma 1.6.6 implies that $a \mid c_0$ and $b \mid c_n$. ■

Example 3.2.2. We apply this result to prove that $\sqrt{2}$ is irrational. Assume that $\sqrt{2} = a/b$ with a/b a reduced rational number. Applying Proposition 3.2.3 with $n = 2$, $c_2 = 1$, $c_1 = 0$ and $c_0 = -2$, we deduce that $a \mid c_0 = -2$ and $b \mid c_2 = 1$. Therefore, $b \in \{-1, 1\}$ and a/b must be an integer. However, there is no integer whose square is 2, contradiction. The attentive reader will have noticed that this was our first proof of Theorem 3.2.1 and is invited to prove that $\sqrt{3}$ and $\sqrt{5}$ are irrational using this method.

Example 3.2.3. The numbers

$$\sqrt{3} + \sqrt{5}, \sqrt{3} - \sqrt{5}, -\sqrt{3} + \sqrt{5}, -\sqrt{3} - \sqrt{5}$$

are solutions of the equation

$$(x^2 + 2)^2 - 12x^2 = 0.$$

The reader is invited to verify this statement and to use Proposition 3.2.3 to show that the four numbers above are irrational.

Proposition 3.2.3 can be used to prove that certain trigonometric numbers are irrational. It may be used in combination with the trigonometric identities:

(3.2.5) $\cos(\alpha + \beta) = \cos \alpha \cos \beta - \sin \alpha \sin \beta$ and $\sin(\alpha + \beta) = \sin \alpha \cos \beta + \cos \alpha \sin \beta$, for our next few examples.

Example 3.2.4. From geometry, we know that $\sin 30^\circ = \frac{1}{2}$ and $\cos 30^\circ = \frac{\sqrt{3}}{2}$ so the first number is rational and the second is irrational. How about $\sin 15^\circ$ or $\cos 15^\circ$? From (3.2.5), $\cos(2\alpha) = 2\cos^2 \alpha - 1$ for any α . Taking $x = \cos 15^\circ$, then $\frac{\sqrt{3}}{2} = \cos 30^\circ = 2x^2 - 1$ giving that $\sqrt{3} = 4x^2 - 2$. If x is rational, then the right-hand side would be a rational number and therefore, $\sqrt{3}$ is rational, contradiction. Hence, $\cos 15^\circ$ is irrational. We leave it as an exercise for the reader to show that $\sin 15^\circ$ is also irrational.

Example 3.2.5. Is $\cos 10^\circ$ rational or irrational? Applying (3.2.5) twice gives that

$$\cos(3\alpha) = \cos(2\alpha)\cos(\alpha) - \sin(2\alpha)\sin(\alpha) = 4\cos^3 \alpha - 3\cos \alpha.$$

For $x = \cos 10^\circ$, we deduce that $\frac{1}{2} = \cos 30^\circ = 4x^3 - 3x$ and therefore, $8x^3 - 6x - 1 = 0$. If x is rational and equals $\frac{a}{b}$ in lowest terms, Proposition 3.2.3 implies that $a \mid 1$ and $b \mid 8$. Therefore, $x = \frac{a}{b}$ equals $1, \frac{1}{2}, \frac{1}{4}, \frac{1}{8}$ (why can we restrict to positive numbers?). However, none of these numbers satisfy the equation $8x^3 - 6x - 1 = 0$, a fact that the reader should check. Hence, $\cos 10^\circ$ is irrational.

Another method for finding irrational numbers involves logarithms.

Example 3.2.6. Consider the number $\log_3(10)$. This is the unique exponent e such that $3^e = 10$ (this can be proved formally using the continuity of the function $f : [2, 3] \rightarrow \mathbb{R}, f(x) = 3^x$, but we can take it for granted for now). Since we have $3^2 = 9 < 10 < 27 = 3^3$, the number $\log_3(10)$ is between 2 and 3. Assume that $\log_3(10)$ is a rational number. Therefore, $\log_3(10) = a/b$ for some natural numbers a and b . This means

that $3^{a/b} = 10$ which implies that $3^a = 10^b$. This gives us a natural number with two different prime factorizations 3^a and $2^b \cdot 5^b$, contradicting Theorem 1.6.5.

Exercise 3.2.1. A point (a, b) in the Cartesian plane is called a **lattice point** if both a and b are integers. Prove that a line in the plane may pass through exactly zero, one or infinitely many lattice points and give examples for each case.

Exercise 3.2.2. Let r be a rational number and x be an irrational number. Prove that $r + x$ is an irrational number. How about rx , x^2 , $\frac{r}{x}$?

Exercise 3.2.3. Let n be a natural number. Prove that \sqrt{n} is rational if and only if $n = m^2$ for some natural number m .

Exercise 3.2.4. Prove that the golden ratio $\varphi = \frac{1+\sqrt{5}}{2}$ is irrational.

Exercise 3.2.5. Prove that $\sqrt{2} + \sqrt{3}$ is an irrational number. Prove that the number $\sqrt{2} + \sqrt{3} + \sqrt{5}$ is irrational.

Exercise 3.2.6. The number $\sqrt[3]{2}$ is defined as the unique number $a > 0$ such that $a^3 = 2$. It can be interpreted as the length of the side of a cube whose volume is 2. Prove that $\sqrt[3]{2}$ is an irrational number.

Exercise 3.2.7. Is the number $\sqrt{2} + \sqrt[3]{2}$ rational or irrational? How about the number $\sqrt{3} + \sqrt[3]{2}$?

Exercise 3.2.8. Prove that $\log_6(12)$ is irrational.

Exercise 3.2.9. Let α be such that $\cos(2\alpha)$ is irrational. Prove that $\sin \alpha, \cos \alpha$ and $\tan \alpha$ are irrational.

Exercise 3.2.10. Let a and b be two natural numbers. Show that

$$\left| \sqrt{3} - \frac{a}{b} \right| \geq \frac{1}{4b^2}.$$

3.3. Fractions and Decimal Representations

In the previous section, we mentioned the interpretation of rational numbers as stakes/marks on the number line that contain and refine the marks of natural and integer numbers. For those numbers, the consecutive stakes were one unit apart. Clearly, such uniformity cannot be possible for the rational numbers since we can make our subunits of various size: $1/1, 1/2, 1/3, 1/4, \dots$ and so on. In some situations, one needs to choose a certain set of subunits and a most common one is the collection involving the powers of 10, namely $\frac{1}{10^0}, \frac{1}{10}, \frac{1}{10^2}, \frac{1}{10^3}, \dots$

Definition 3.3.1. Let n be a nonnegative integer. Let a, a_1, \dots, a_n be integers such that $0 \leq a_j \leq 9$ for any $1 \leq j \leq n$. The number

$$a + \frac{a_1}{10^1} + \cdots + \frac{a_n}{10^n}$$

is called a **finite decimal fraction** and is denoted by $a.a_1 \dots a_n$.

If we bring the expression above to the common denominator 10^n , we get that

$$a.a_1 \dots a_n = \frac{a \cdot 10^n + a_1 \cdot 10^{n-1} + \dots + a_n}{10^n}.$$

This shows that any finite decimal fraction is a rational number or fraction whose denominator is a power of 10.

Example 3.3.1. Take the number $1/4$ for example. How do we write it as a finite decimal fraction? The previous observation is useful, and we look to write $1/4$ as a fraction whose numerator is a power of 10. It is not too hard to see that

$$1/4 = \frac{25}{100} = \frac{0 \cdot 10^2 + 2 \cdot 10 + 5}{100} = 0 + \frac{2}{10} + \frac{5}{10^2} = 0.25.$$

Of course, we can always tag along a bunch of zeroes at the end of this decimal representation so 0.25 is the same as 0.250 or 0.2500. For this reason, we will assume that the last digit a_n is always nonzero.

Example 3.3.2. How about writing the number $1/3$ as a finite decimal fraction? If this were possible, then $\frac{1}{3}$ would equal $\frac{m}{10^n}$ for some natural numbers m and n . From $\frac{1}{3} = \frac{m}{10^n}$, we get that $3m = 10^n$. Thus, $3 \mid 10^n$. By Euclid's lemma, we deduce that $3 \mid 10$ which is impossible. Hence, $1/3$ is a rational number that cannot be written as a finite decimal fraction.

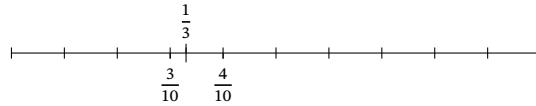


Figure 3.3.1. Dividing the interval $[0, 1]$ into $1/10$ subunits.

The natural question is how do we proceed here? We describe the situation in Figure 3.3.1. It is not too hard to figure out that $1/3$ lands between the consecutive stakes/numbers $\frac{3}{10}$ and $\frac{4}{10}$ as:

$$0.3 = 3/10 < 1/3 < 4/10 = 0.4.$$

This situation leads to use finer subunits of size $\frac{1}{100}$ and *squeeze* $1/3$ even more. Zooming in into the interval $[0.3, 0.4]$ in Figure 3.3.2, we observe that

$$0.3 = 30/100 < 33/100 = 0.33 < 1/3 < 0.34 = 34/100 < 40/100 = 0.4.$$

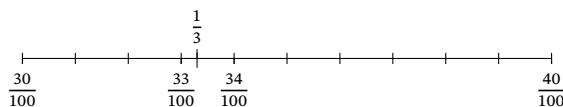


Figure 3.3.2. Dividing the interval $[0.3, 0.4]$ into $1/100$ subunits.

We can continue this process as follows:

$$\begin{aligned} 0.333 &< 1/3 < 0.334 \\ 0.3333 &< 1/3 < 0.3334 \\ &\dots \\ \underbrace{0.33\dots 3}_{n \text{ digits}} &< 1/3 < \underbrace{0.33\dots 4}_{n \text{ digits}}, \end{aligned}$$

for any natural number n .

Geometrically, we have found a sequence of *nested* intervals $(I_n)_{n \geq 1}$:

$$I_n = [\underbrace{0.33\dots 3}_{n \text{ digits}}, \underbrace{0.33\dots 4}_{n \text{ digits}})$$

such that⁴

$$1/3 \in I_n,$$

for any n . The endpoints of I_n are finite decimal fractions with denominator 10^n and the length of each I_n is $\frac{1}{10^n}$. The name *nested* means that

$$\dots I_n \subset I_{n-1} \subset \dots \subset I_2 \subset I_1$$

and

$$1/3 \in I_n$$

for each n . These intervals get closer to $1/3$ as n gets larger like the walls of a villain trap in a James Bond movie. By our analysis, the endpoints of the interval I_n will get close to each other, but will never reach $1/3$.

This example gives the intuition behind attributing to $1/3$ an **infinite decimal fraction** representation:

$$1/3 = 0.33\dots,$$

corresponding to the left endpoints of the intervals I_n . Another way of interpreting a number such as $0.33\dots$ is an infinite sum of the form:

$$\frac{3}{10} + \frac{3}{10^2} + \dots$$

If this infinite sum causes you problems, then just consider the finite decimal number

$$0.\underbrace{3\dots 3}_{n \text{ digits}} = \frac{3}{10} + \dots + \frac{3}{10^n}.$$

Note that this is the sum of the first n terms of a geometric progression with first term $3/10$ and ratio $1/10$ and therefore,

$$\frac{3}{10} + \dots + \frac{3}{10^n} = \frac{3}{10} \cdot \frac{1 - \frac{1}{10^{n+1}}}{1 - \frac{1}{10}} = \frac{1}{3} \cdot \left(1 - \frac{1}{10^{n+1}}\right).$$

⁴In the notation of (3.3.2) the interval I_n coincides with $I_{3,\dots,3}$. Note that in the decimal case there are at level n exactly 10^n disjoint intervals of length $1/10^n$, defined by

$$I_{a_1,\dots,a_n} = \left[\sum_{j=0}^n \frac{a_j}{10^j}, \sum_{j=0}^n \frac{a_j}{10^j} + \frac{1}{10^n} \right) = [0.a_1 \dots a_n, 0.a_1 \dots a_n + 10^{-n}], \quad 0 \leq a_j \leq 9.$$

We will not go into the formalism of taking limits now, but we argue that it seems intuitive that as n grows large, the number $\frac{1}{10^{n+1}}$ becomes smaller and gets closer to 0. Therefore, the finite decimal number $0.\underbrace{3\dots3}_{n \text{ digits}}$ will tend to $1/3$.

Definition 3.3.2. Let a be an integer and $(a_n)_{n \geq 1}$ be a sequence of integers such that $0 \leq a_n \leq 9$ for any $n \geq 1$. A number α is said to equal the **infinite decimal fraction** $a.a_1a_2\dots$ if for any $n \geq 1$,

$$a.a_1\dots a_n \leq \alpha < a.a_1\dots a_n + \frac{1}{10^n}.$$

The notation $0.33\dots$ is sometimes referred to as $0\bar{3}$. This is a special kind of infinite decimal representation that is called periodic since we have a certain collection of consecutive digits (in this case, just one digit 3) that repeats forever.

Similar to our example of $1/3$, one may figure the infinite decimal representation for numbers like $2/3$ or $1/9$ (see Exercise 3.3.3).

Example 3.3.3. How about the decimal representation of $1/7$? Since $0 < 1/7 < 1$, the integer part must be 0. To find out the first digit a_1 , we are searching for a number a_1 between 0 and 9 such that

$$0.a_1 < 1/7 < 0.a_1 + \frac{1}{10} \text{ which is the same as } \frac{a_1}{10} < \frac{1}{7} < \frac{a_1 + 1}{10}.$$

These inequalities lead to $3/7 < a_1 < 10/7$ and our only choice is $a_1 = 1$. To get the next digit a_2 , we must look for a number a_2 between 0 and 9 such that

$$0.1a_2 < 1/7 < 0.1a_2 + \frac{1}{10^2} \text{ or } \frac{10 + a_2}{100} < \frac{1}{7} < \frac{11 + a_2}{100}.$$

Therefore, $a_2 < 100/7 - 10 = 30/7$ and $a_2 > 100/7 - 11 = 23/7$. Our only choice is $a_2 = 4$. To get a_3 , we must solve

$$0.14a_3 < 1/7 < 0.14a_3 + \frac{1}{10^3},$$

which leads to

$$\frac{140 + a_3}{1000} < 7 < \frac{141 + a_3}{1000}.$$

Therefore, $a_3 < \frac{1000}{7} - 140 = \frac{1000-980}{7} = \frac{20}{7}$ and $a_3 > \frac{1000}{7} - 141 = \frac{1000-987}{7} = \frac{13}{7}$. Thus, $a_3 = 2$. The procedure goes on in similar fashion and one can show that the first six digits in the decimal representation of $1/7$ are 1, 4, 2, 8, 5, 7 which will repeat from then on in this order (see Exercise 3.3.2):

$$1/7 = 0.142857142857148257\dots = 0.\overline{142857}.$$

This is another periodic infinite decimal representation.

Example 3.3.4. Let us try an example in reverse. Say we have an infinite decimal representation of the form

$$2.\bar{47}.$$

What number does it correspond to? We can attempt to reverse our previous argument and start with

$$\begin{aligned} 2.4\overbrace{7\dots7}^n \text{ digits} &= 2.4 + \frac{7}{10^2} + \dots + \frac{7}{10^{n+1}} = 2.4 + \frac{7}{10^2} \cdot \left(1 + \frac{1}{10} + \dots + \frac{1}{10^{n-1}}\right) \\ &= 2.4 + \frac{7}{10^2} \cdot \frac{1 - \frac{1}{10^n}}{1 - \frac{1}{10}} = 2.4 + \frac{7\left(1 - \frac{1}{10^n}\right)}{90} = 2.4 + \frac{7}{90} - \frac{7/9}{10^{n+1}}. \end{aligned}$$

As n gets large, $\frac{7/9}{10^{n+1}}$ approaches 0 and therefore, $2.4\overbrace{7\dots7}^n$ will tend to

$$2.4 + \frac{7}{90} = \frac{2.4 \cdot 90 + 7}{90} = \frac{223}{90}.$$

In order to avoid the use of limits, one can rephrase the previous proofs as follows. Let α be the number whose decimal representation is $2.\overline{47}$. Then for any natural number n ,

$$\begin{aligned} 2.4\overbrace{7\dots77}^n < \alpha &< 2.4\overbrace{7\dots78}^n, \\ 2.4 + \frac{7}{90} - \frac{7/9}{10^{n+1}} < \alpha &< 2.4 + \frac{7}{90} - \frac{7/9}{10^{n+1}} + \frac{1}{10^{n+1}}, \\ 2.4 + \frac{7}{90} - \frac{7/9}{10^{n+1}} < \alpha &< 2.4 + \frac{7}{90} + \frac{2/9}{10^{n+1}}. \end{aligned}$$

We claim that the only number α that satisfies these inequalities for any n will be $2.4 + \frac{7}{90} = \frac{223}{90}$. To see this, note first that $2.4 + \frac{7}{90}$ satisfies the above inequalities.

To prove that $2.4 + \frac{7}{90}$ is the only number with this property, we use proof by contradiction. Assume that there were two numbers $\alpha' < \alpha''$ such that

$$2.4 + \frac{7}{90} - \frac{7/9}{10^{n+1}} < \alpha' < \alpha'' < 2.4 + \frac{7}{90} + \frac{2/9}{10^{n+1}},$$

Therefore,

$$0 < \alpha'' - \alpha' < \frac{1}{10^{n+1}},$$

for any $n \geq 1$. This means that

$$10^{n+1} < \frac{1}{\alpha'' - \alpha'},$$

for any $n \geq 1$. Let M be a natural number that is larger than $\frac{1}{\alpha'' - \alpha'}$. By our previous equation, $10^M < M$. However, the binomial formula implies that

$$10^M = (1 + 9)^M = 1 + \binom{M}{1}9^1 + \dots + \binom{M}{M}9^M > 1 + 9M > M,$$

contradiction. This proves that $2.4 + \frac{7}{90} = 2.\overline{47}$.

Example 3.3.5. How about the decimal representation of $\sqrt{2}$? Since $1 < \sqrt{2} < 2$, we know that it must start with 1. To find the first decimal digit, we need to figure out a number a_1 such that

$$1 + a_1/10 < \sqrt{2} < 1 + (a_1 + 1)/10.$$

Since $1.4^2 = 1.96 < 2 < 2.25 = 1.5^2$, we can deduce that $a_1 = 4$. The next digit a_2 must satisfy the inequality

$$1.4 + a_2/100 < \sqrt{2} < 1.4 + (a_2 + 1)/100.$$

Because $1.41^2 = 1.9881 < 2 < 1.42^2 = 2.0164$, we get that $a_2 = 1$. This recursive procedure may continue forever and one may determine the digits of $\sqrt{2}$ for its decimal representation recursively one at a time⁵. We encourage the reader to continue this argument and verify that $1.41421 < \sqrt{2} < 1.41422$.

After investigating *decimal* fractional representations of rational numbers, let us treat now the *general case*. The integer 10 as base of the representation was likely chosen because humans have 10 fingers, but many cultures used different bases as well. With the appearance of computers other bases became important, as for example the bases $b = 2$ and $b = 16$.

Our objective is to represent any rational number α by (positive and negative) powers of a given base $b \geq 2$. If $\alpha \in \mathbb{Z}$, this was done in Theorem 1.5.2. Therefore, it suffices to investigate rational numbers in the interval $(0, 1)$. If $\alpha \in \mathbb{Q} \setminus \mathbb{Z}$, then $\alpha = \lfloor \alpha \rfloor + \alpha'$, where $\alpha' := \alpha - \lfloor \alpha \rfloor \in (0, 1)$ denotes the fractional part of α . Since $\lfloor \alpha \rfloor \in \mathbb{Z}$, representing α in base b is equivalent to representing α' in base b .

When we ask for fractional representations with respect to an arbitrary base $b \geq 2$, as in the decimal case two different possibilities occur: finite and infinite expansions.

Definition 3.3.3. Fix a base $b \geq 2$. A number $\alpha \in (0, 1)$ admits a **finite expansion** as b -fraction if there exists a natural number n and integers a_1, \dots, a_n in $\{0, \dots, b-1\}$ such that

$$\alpha = \frac{a_1}{b^1} + \dots + \frac{a_n}{b^n}.$$

As usual we write

$$\alpha =_b 0.a_1 \dots a_n.$$

To avoid unnecessary zeroes we always assume that $a_n \neq 0$. As usual, if $b = 10$ we do not add the subscript $=_{10}$.

Example 3.3.6. In base $b = 2$ the number 7.75 admits the representation

$$7.75 = 7 + 0.75 = 2^2 + 2^1 + 2^0 + 2^{-1} + 2^{-2} = 111.11_2.$$

It is common to separate the integer part and the fractional part by a dot.

Example 3.3.7. If $b = 5$, then

$$43.02_5 = 4 \cdot 5^1 + 3 \cdot 5^0 + 0 \cdot 5^{-1} + 2 \cdot 5^{-2},$$

which equals 23.008 in decimal system.

Of course, each number α presentable as finite b -fraction is rational. However, we saw earlier in the case $b = 10$ that not every rational number can be written as a finite b -fraction. The next result characterizes those rational numbers which are presentable in this way.

⁵There are better ways of doing this work, and we will see some in Section 3.4.

Proposition 3.3.1. A number $\alpha \in (0, 1)$ can be written as finite b -fraction if and only if there are natural numbers n and k such that $1 \leq k < b^n$ and

$$\alpha = \frac{k}{b^n}.$$

Proof: The result follows from the equivalence

$$\alpha =_b 0.a_1 \dots a_n \Leftrightarrow \alpha = \frac{a_1 b^{n-1} + a_2 b^{n-2} + \dots + a_n}{b^n}$$

and Theorem 1.5.2 which asserts that every nonnegative integer $k < b^n$ can be represented uniquely as $k = a_1 b^{n-1} + \dots + a_n$, where $a_1, \dots, a_n \in \{0, \dots, b-1\}$. ■

Similar to integers, the expansion of a rational number as a finite b -fraction is unique.

Definition 3.3.4. A number $\alpha \in (0, 1)$ is said to be **b -rational** provided there exists $n \in \mathbb{N}$ and an integer k with $1 \leq k < b^n$ such that

$$\alpha = \frac{k}{b^n}.$$

Note that we always may assume that $b \nmid k$. If $b = 2$, then 2-rational numbers are also called **dyadic rational numbers**. These are the fractions

$$\frac{1}{2}, \frac{1}{4}, \frac{3}{4}, \frac{1}{8}, \frac{3}{8}, \frac{5}{8}, \frac{7}{8}, \frac{1}{16}, \dots, \frac{1}{2^n}, \dots, \frac{2^n - 1}{2^n}, \dots$$

Proposition 3.3.1 states that a number $\alpha \in (0, 1)$ can be represented as a finite b -fraction if and only if it is b -rational.

We turn now to the case of infinite representations as b -fractions.

Definition 3.3.5. Fix a base $b \geq 2$. Let α be a rational number in $(0, 1)$. The number α admits a representation as **b -fraction**⁶ if there are integers a_1, a_2, \dots with $0 \leq a_j < b$ such that for all $n \geq 1$ the following holds:

$$(3.3.1) \quad \frac{a_1}{b} + \frac{a_2}{b^2} + \dots + \frac{a_n}{b^n} \leq \alpha < \frac{a_1}{b} + \frac{a_2}{b^2} + \dots + \frac{a_n}{b^n} + \frac{1}{b^n}.$$

In this case we shall write⁷

$$\alpha =_b 0.a_1 a_2 \dots .$$

Recall that $\lfloor x \rfloor$ denotes the integer part of a number x and equals the largest integer that is less than or equal to x . Using the results in Section 1.5 about expansions of integers we may rewrite the equation (3.3.1) as follows:

$$\alpha =_b 0.a_1 a_2 \dots \Leftrightarrow (\forall n \geq 1)(\lfloor b^n \alpha \rfloor = a_1 b^{n-1} + \dots + a_n = (a_1 \dots a_n)_b).$$

For example,

$$\frac{1}{3} = 0.333 \dots \text{ because of } (\forall n \geq 1)(\lfloor 10^n / 3 \rfloor = \underbrace{333 \dots 3}_{n \text{ digits}}).$$

⁶For $b = 2$, $b = 10$ or $b = 16$ those fractions are called binary, decimal, or hexadecimal fractions, respectively.

⁷When $b = 10$, we will write $=$ instead of $=_{10}$.

The preceding definition applies also for finite representations. They occur in the case that $a_n \neq 0$ and $a_j = 0$ if $j > n$. But as we agreed we will always write $\alpha =_b 0.a_1 \dots a_n$ instead of $\alpha =_b 0.a_1 \dots a_n 00 \dots$. For example,

$$\frac{28}{125} = \frac{1}{5} + \frac{3}{125} =_5 0.103000 \dots = 0.103.$$

Geometrically condition (3.3.1) says the following. At each step n we divide the interval $[0, 1]$ into b^n intervals of length $1/b^n$. These intervals may be denoted by

$$(3.3.2) \quad I_{a_1, \dots, a_n} = \left[\sum_{j=1}^n \frac{a_j}{b^j}, \sum_{j=1}^n \frac{a_j}{b^j} + \frac{1}{b^n} \right)$$

where a_1, \dots, a_n are arbitrary integers in $\{0, \dots, b - 1\}$. With this notation, condition (3.3.1) can be rewritten as follows:

$$\alpha =_b 0.a_1 a_2 \dots \Leftrightarrow (\forall n \geq 1)(\alpha \in I_{a_1, \dots, a_n}).$$

Moreover, $\alpha \in [0, 1]$ possesses a finite representation if and only if α coincides with the left-hand endpoint of a suitable interval I_{a_1, \dots, a_n} .

Example 3.3.8. If, for instance, $b = 2$ and $n = 2$, then we get the 4 intervals

$$I_{0,0} = \left[0, \frac{1}{4} \right), \quad I_{0,1} = \left[\frac{1}{4}, \frac{1}{2} \right), \quad I_{1,0} = \left[\frac{1}{2}, \frac{3}{4} \right), \quad \text{and} \quad I_{1,1} = \left[\frac{3}{4}, 1 \right).$$

For example, if $\alpha = 1/3$, then $\alpha =_2 0.\overline{01}$, which is equivalent to

$$\frac{1}{3} \in I_{0,0}, \quad \frac{1}{3} \in I_{0,1}, \quad \frac{1}{3} \in I_{0,1,0}, \quad \frac{1}{3} \in I_{0,1,0,1}, \quad \frac{1}{3} \in I_{0,1,0,1,0}, \dots$$

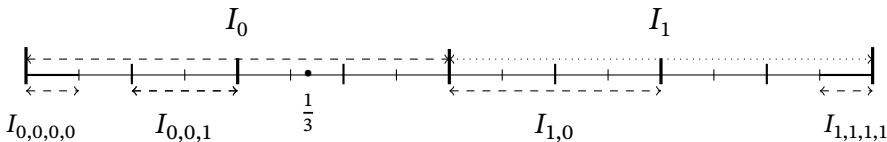


Figure 3.3.3. If $b = 2$ there are 2 intervals I_0 and I_1 of length $1/2$. There are 4 intervals of length $1/4$ and 8 of length $1/8$. In the next step one gets the intervals $I_{0,0,0,0}, \dots, I_{1,1,1,1}$, each of length $1/16$. All intervals are left closed and right open.

Remark 3.3.1. A basic question is which sequences a_1, a_2, \dots of integers may occur as expansion of a suitable rational number $\alpha \in [0, 1)$. We will investigate this question more thoroughly below. But already now we want to indicate that one type of sequence never occurs, namely the one with $a_j = b - 1$ for all $j > m$ for a certain $m \in \mathbb{N}$. In the decimal setting, this means that $a_j = 9$ for any $j > m$.

Let us assume the contrary. That is, there exists a rational number $\alpha \in [0, 1)$ such that for some $m \geq 1$ and $a_m < b - 1$ it follows that

$$\alpha =_b 0.a_1 \dots a_m(b-1)(b-1) \dots$$

By Definition 3.3.5 this implies that if $n > m$, then

$$\sum_{j=1}^m \frac{a_j}{b^j} + \sum_{j=m+1}^n \frac{b-1}{b^j} \leq \alpha < \sum_{j=1}^m \frac{a_j}{b^j} + \sum_{j=m+1}^n \frac{b-1}{b^j} + \frac{1}{b^n}.$$

Easy computations show that

$$\sum_{j=m+1}^n \frac{b-1}{b^j} = \frac{1}{b^m} - \frac{1}{b^n},$$

hence by using $a_m < b - 1$ the preceding estimates are equivalent to

$$(3.3.3) \quad \sum_{j=1}^{m-1} \frac{a_j}{b^j} + \frac{a_m + 1}{b^m} - \frac{1}{b^n} \leq \alpha < \sum_{j=1}^{m-1} \frac{a_j}{b^j} + \frac{a_m + 1}{b^m}.$$

In other words, we get for all $n > m$ that

$$0 \leq \alpha - 0.a_1 \dots a_{m-1}(a_m + 1) < \frac{1}{b^n}.$$

This is true for all $n > m$, so that b^{-n} can be made arbitrarily small, hence the only possibility is that

$$\alpha = 0.a_1 \dots a_{m-1}(a_m + 1),$$

which contradicts the right-hand estimate in (3.3.3).

A first important question is which rational numbers possess a representation as b -fractions and whether it is unique, provided it exists.

Proposition 3.3.2. *Fix a base $b \geq 2$. If $\alpha \in [0, 1)$ is a rational number, then there is a unique (finite or infinite) representation*

$$\alpha =_b 0.a_1 a_2 \dots$$

Proof: In a first step we prove the **existence** of the expansion. So take an arbitrary $\alpha \in [0, 1)$. Then $0 \leq b\alpha < b$ which implies that $a_1 = \lfloor b\alpha \rfloor \in \{0, \dots, b-1\}$. If the fractional part is

$$r_1 = b\alpha - a_1 = b\alpha - \lfloor b\alpha \rfloor,$$

then, on one hand $0 \leq r_1 < 1$, and on the other hand

$$\alpha = \frac{a_1}{b} + \frac{r_1}{b}.$$

In a next step set $a_2 = \lfloor br_1 \rfloor$ and $r_2 = br_1 - a_2$. Then

$$\alpha = \frac{a_1}{b} + \frac{a_2}{b^2} + \frac{r_2}{b^2}.$$

Suppose we have already constructed $a_1, \dots, a_{n-1} \in \{0, \dots, b-1\}$ as well as $r_1, \dots, r_{n-1} \in [0, 1)$ such that

$$\alpha = \sum_{j=1}^{n-1} \frac{a_j}{b^j} + \frac{r_{n-1}}{b^{n-1}}.$$

Setting $a_n = \lfloor br_{n-1} \rfloor$ and $r_n = br_{n-1} - a_n$, it follows $a_n \in \{0, \dots, b-1\}$, $0 \leq r_n < 1$ and

$$\alpha = \sum_{j=1}^n \frac{a_j}{b^j} + \frac{r_n}{b^n}.$$

Because of $0 \leq r_n < 1$ this implies

$$\frac{a_1}{b} + \frac{a_2}{b^2} + \dots + \frac{a_n}{b^n} \leq \frac{a_1}{b} + \frac{a_2}{b^2} + \dots + \frac{a_n}{b^n} + \frac{r_n}{b^n} < \frac{a_1}{b} + \frac{a_2}{b^2} + \dots + \frac{a_n}{b^n} + \frac{1}{b^n}.$$

Thus, by

$$\alpha = \frac{a_1}{b} + \frac{a_2}{b^2} + \cdots + \frac{a_n}{b^n} + \frac{r_n}{b^n}$$

for all $n \geq 1$ estimates (3.3.1) are satisfied, hence in view of (3.3.1) it follows that $\alpha =_b 0.a_1a_2\dots$ as asserted.

It remains to verify the **uniqueness** of the expansion. Thus, let us suppose that the rational number $\alpha \in (0, 1)$ possesses two representations

$$\alpha =_b 0.a_1a_2\dots =_b 0.a'_1a'_2\dots$$

with $0 \leq a_j, a'_j \leq b-1$. If we assume that these are different representations then there is a minimal n such that

$$a_1 = a'_1, \dots, a_{n-1} = a'_{n-1} \quad \text{but} \quad a_n \neq a'_n.$$

If $n = 1$ then this means that already $a_1 \neq a'_1$. We may assume $a_n < a'_n$, hence, because these are integers it follows that $a_n + 1 \leq a'_n$. This lets us conclude that

$$\alpha < \sum_{j=1}^n \frac{a_j}{b^j} + \frac{1}{b^n} = \sum_{j=1}^{n-1} \frac{a'_j}{b^j} + \frac{a_n}{b^n} + \frac{1}{b^n} \leq \sum_{j=1}^n \frac{a'_j}{b^j} \leq \alpha.$$

This contradiction proves the uniqueness. ■

Remark 3.3.2. The proof is constructive, that is, it gives an algorithm how to evaluate the representation as b -fraction. Moreover, it tells us that the representation is finite if and only if there is a step n for which the fractional part fulfills $r_n \neq 0$ but $r_{n+1} = 0$. In this case

$$\alpha =_b 0.a_1\dots a_n.$$

Example 3.3.9. $b = 2, \alpha = 0.3$

$$\begin{aligned} b\alpha &= 0.6, \quad \text{hence } a_1 = \lfloor 0.6 \rfloor = 0, \quad \text{thus } r_1 = 0.6. \\ br_1 &= 1.2, \quad \text{hence } a_2 = \lfloor 1.2 \rfloor = 1, \quad \text{thus } r_2 = 0.2. \\ br_2 &= 0.4, \quad \text{hence } a_3 = \lfloor 0.4 \rfloor = 0, \quad \text{thus } r_3 = 0.4. \\ br_3 &= 0.8, \quad \text{hence } a_4 = \lfloor 0.8 \rfloor = 0, \quad \text{thus } r_4 = 0.8. \\ br_4 &= 1.6, \quad \text{hence } a_5 = \lfloor 1.6 \rfloor = 1, \quad \text{thus } r_5 = 0.6. \end{aligned}$$

We observe that $r_5 = r_1$, consequently the calculations repeat, and we obtain that $a_6 = a_2 = 1, a_7 = a_3 = 0, a_8 = a_4 = 0$ and $a_9 = a_5 = 1$ and so on.

$$0.3 =_2 0.0100110011001\dots = 0.0\overline{1001}.$$

Example 3.3.10. Let $b = 2$ and $\alpha = 1/3$. Then

$$a_1 = \lfloor \alpha b \rfloor = \left\lfloor \frac{2}{3} \right\rfloor = 0.$$

Hence, $r_1 = 2/3 - 0 = 2/3$.

In the next step

$$a_2 = \lfloor r_1 b \rfloor = \left\lfloor \frac{4}{3} \right\rfloor = 1$$

and $r_2 = 4/3 - 1 = 1/3$. So we are back at the starting point and conclude that $a_3 = 0$, $a_4 = 1$, $a_5 = 0$ and so on. At the end we arrive at

$$\frac{1}{3} =_2 0.010101\dots = \overline{0.01}.$$

Example 3.3.11. Let $b = 16$ be the base of representation. Recall that, as usual in Computer Sciences, $A = 10$, $B = 11$ up to $F = 15$. Which representation does $\alpha = 0.12$ possess with respect to this base?

We start with

$$a_1 = [0.12 \cdot 16] = [1.92] = 1 \quad \text{hence} \quad r_1 = 0.92.$$

Next we obtain

$$a_2 = [0.92 \cdot 16] = [14.72] = 14 = E \quad \text{hence} \quad r_2 = 0.72.$$

Now we conclude

$$a_3 = [0.72 \cdot 16] = [11.52] = 11 = B \quad \text{hence} \quad r_3 = 0.52.$$

The next step gives

$$a_4 = [0.52 \cdot 16] = [8.32] = 8 \quad \text{hence} \quad r_4 = 0.32.$$

Proceeding further it follows that

$$a_5 = [0.32 \cdot 16] = [5.12] = 5 \quad \text{hence} \quad r_5 = 0.12.$$

Next

$$a_6 = [0.12 \cdot 16] = [1.92] = 1 \quad \text{hence} \quad r_6 = 0.92.$$

At this step we see that $r_6 = r_1$, hence $r_7 = r_2$ and so on. So we finally get

$$0.12 =_{16} 0.1EB851EB851\dots = \overline{0.1EB85}.$$

If we look at all these examples, we see that all presentations were periodic in the following sense.

Definition 3.3.6. The representation of $\alpha \in (0, 1)$ is said to be **periodic** provided that there are an $m \geq 0$ and a $k \geq 1$ such that

$$\alpha =_b 0.a_1 \dots a_m \overline{a_{m+1} \dots a_{m+k}}.$$

We hereby always assume that $m \geq 0$ and $k \geq 1$ are chosen minimal⁸. In this case we say α possesses a representation with a period of length $k \geq 1$ starting at $m + 1 \geq 1$.

$(\alpha \in [0, 1])$ has a periodic representation)

$$\Leftrightarrow (\exists m \geq 0, k \geq 1, 0 \leq a_j \leq b-1) (\alpha =_b 0.a_1 \dots a_m \overline{a_{m+1} \dots a_{m+k}}).$$

It is common to say that α has a **completely periodic representation** if $m = 0$. That is,

$$\alpha = 0.\overline{a_1 \dots a_k}.$$

Otherwise, i.e., if $m \geq 1$, then the representation is periodic with a **nonrepeating part**.

⁸A nonminimal expansion of α would for example be

$$\alpha =_b 0.a_1 \dots a_m a_{m+1} \dots a_{m+k} \overline{a_{m+k+1} \dots a_{m+2k}}$$

with $a_{m+k+1} = a_{m+1}, \dots, a_{m+2k} = a_{m+k}$.

Remark 3.3.3. When we talk about periodic representations we always mean true infinite expansions. Of course, one could also write

$$\alpha =_b a_1 \dots a_n = 0.a_1 \dots a_n 000 = 0.a_1 \dots a_n \bar{0},$$

but this is somehow artificial, and therefore we want to exclude this case here.

Example 3.3.12.

$$\begin{aligned} \frac{1}{10} &=_{10} 0.000\overline{1100} \Rightarrow m = 3 \text{ and } k = 4. \\ 0.12 &=_{16} 0.\overline{1EB85} \Rightarrow m = 0 \text{ and } k = 5. \end{aligned}$$

Thus, with respect to base $b = 2$, the number $1/10$ has a periodic representation with a nonrepeating part 000. On the other hand, 0.12 is completely periodic when expanded by base $b = 16$.

Proposition 3.3.3. Fix a base $b \geq 2$. The b -fraction of a rational number is either finite or periodic.

Proof: Let α be an arbitrary rational number. As we mentioned above by using $\alpha = [\alpha] + (\alpha - [\alpha])$ it follows that it suffices to investigate rational numbers in $[0, 1)$. So we may suppose that $\alpha = p/q$ for some $p, q \in \mathbb{N}$ with $p < q$. Furthermore, we may assume that α does not possess a finite representation as b -fraction. So we know that the fractional parts r_1, r_2, \dots occurring in the proof of Proposition 3.3.2 never vanish.

Now we follow the construction of the representation of α given in Proposition 3.3.2. In a first step set $r_0 = \alpha$. Then

$$qr_0 = p < q,$$

hence $qr_0 \in \{1, \dots, q-1\}$. Proceeding further as in the proof of Proposition 3.3.2 we have $r_1 = b\alpha - a_1$, hence

$$qr_1 = qb\alpha - qa_1 = pb - qa_1,$$

which implies that qr_1 is an integer. Moreover, since $0 < r_1 < 1$ (recall that we assume that α does not possess a finite representation) we get $qr_1 \in \{1, \dots, q-1\}$.

In the next step, by $qr_1 \in \mathbb{N}$ we also have (recall that $r_2 = br_1 - a_2$ and that b and a_2 are integers)

$$qr_2 = bqr_1 - qa_2 \in \mathbb{N},$$

and by the same arguments as before this implies that $qr_2 \in \{1, \dots, q-1\}$.

By induction, for all $n \geq 0$ it follows that

$$qr_n \in \{1, \dots, q-1\}.$$

Summing up, at step n there are $n+1$ (recall we start at zero) integers qr_0, \dots, qr_n which all belong to $\{1, \dots, q-1\}$, i.e., to a set of cardinality $q-1$. Consequently, the pigeonhole principle (cf. Proposition 1.1.3), lets us conclude the following: If $n+1 \geq q$, then at least two of the integers between qr_0 and qr_n must coincide.

In particular this is true for $n = q - 1$, thus we may choose a minimal $m \geq 0$ as well as a minimal $k \geq 1$ such that $m + k \leq q - 1$ and $qr_m = qr_{m+k}$. Summing up, there exist integers m and k with $0 \leq m < m + k \leq q - 1$ satisfying

$$qr_m = qr_{m+k}, \quad \text{hence} \quad r_m = r_{m+k}.$$

Recall that $a_{n+1} = \lfloor b r_n \rfloor$ and $r_{n+1} = b r_n - a_{n+1}$ for all $n \geq 0$. Thus, $r_m = r_{m+k}$ implies

$$a_{m+k+1} = a_{m+1} \quad \text{as well as} \quad r_{m+k+1} = r_{m+1}.$$

Proceeding further in the same way, we get

$$a_{m+k+2} = a_{m+2} \quad \text{as well as} \quad r_{m+k+2} = r_{m+2},$$

next

$$a_{m+2k+1} = a_{m+k+1} = a_{m+1} \quad \text{and} \quad r_{m+2k+1} = r_{m+k+1} = r_{m+1},$$

and so on. At the end we finally arrive at

$$\alpha =_b 0.a_1 \dots a_m \overline{a_{m+1} \dots a_{m+k}},$$

as asserted. Moreover, since $m \geq 0$ and $k \geq 1$ were chosen minimal, the period starts exactly at position $m + 1$ and has length k . ■

Corollary 3.3.4. *If $\alpha = p/q$ is a rational number in $(0, 1)$, then*

$$\alpha =_b 0.a_1 \dots a_m \overline{a_{m+1} \dots a_{m+k}},$$

with integers m and k satisfying $0 \leq m < m + k \leq q - 1$. In particular, the length of the period of p/q is at most $q - 1$.

Example 3.3.13. We have in decimal representation

$$\begin{aligned} \frac{1}{7} &= 0.\overline{142857}, & \frac{2}{7} &= 0.\overline{285714}, & \frac{3}{7} &= 0.\overline{428571} \\ \frac{4}{7} &= 0.\overline{571428}, & \frac{5}{7} &= 0.\overline{714285}, & \frac{6}{7} &= 0.\overline{857142}. \end{aligned}$$

Thus, in each of these cases the period starts at $m = 0$ and has maximal length $m + k = 6 = q - 1$. In particular, it shows that the estimate $0 \leq m < m + k \leq q - 1$ cannot be improved in general.

There is another interesting phenomenon about these numbers. All periods consist of the same digits and all are in the same cyclic order. This is not only so for $1/7$, but always whenever $1/q$ has a period of the maximal length $q - 1$. But for which natural numbers $q \geq 2$ does the reciprocal $1/q$ possess a representation as periodic b -fraction with maximal period length $q - 1$? As we saw above $q = 7$ is such a number with respect to base $b = 10$. Another example is $q = 5$ when taken to base $b = 2$. Note that $\frac{1}{5} =_2 0.\overline{0011}$.

So the precise question is as follows: Let $b \geq 2$ be a fixed base. For which natural numbers $q \geq 2$ does $1/q$ possess a period of maximal length $q - 1$? It is not difficult to prove that those numbers have to be prime. This follows, for example, by the fact that the length of the period of $1/q$ always divides the totient $\phi(q)$ (cf. Definition 2.7.1 for the definition and note that $\phi(q) = q - 1$ if and only if q is prime). On the other hand,

not every prime possesses this property. For example, with respect to base $b = 10$ the primes $p = 11$ and $q = 13$ satisfy

$$\frac{1}{p} = \frac{1}{11} = 0.\overline{09} \quad \text{and} \quad \frac{1}{q} = \frac{1}{13} = 0.\overline{076923}.$$

Hence the lengths of their periods are $2 < p - 1 = 10$ and $6 < q - 1 = 12$, respectively. Thus, the following definition makes sense.

Definition 3.3.7. A prime number $q \geq 2$ is called a **long prime**⁹ provided that its reciprocal $1/q$ has a decimal expansion with period of (maximal) length $q - 1$.

Examples of long primes are 23 and 29 because of

$$\begin{aligned}\frac{1}{23} &= 0.\overline{0434782608695652173913} \\ \frac{1}{29} &= 0.\overline{0344827586206896551724137931}.\end{aligned}$$

The first long primes in decimal system are known to be

$$\begin{aligned}7, 17, 19, 23, 29, 47, 59, 61, 97, 109, 113, 131, 149, 167, 179, \\181, 193, 223, 229, 233, 257, 263, 269, 313.\end{aligned}$$

Of course, one may ask also for long primes with respect to bases different from 10. If q is a long prime in base 10, it need not possess this property in bases $b \neq 10$. For instance, while 7 is a long prime with respect to base 10, the length of the period of $1/7$ in base $b = 2$ is $3 < 6$.

As an example, let us state the first long primes with respect to base $b = 2$. These are

$$\begin{aligned}3, 5, 11, 13, 19, 29, 37, 53, 59, 61, 67, 83, 101, 107, 131, 139, \\149, 163, 173, 179, 181, 197, 211, 227, 269.\end{aligned}$$

Remark 3.3.4. An open problem is whether there are infinitely many long primes. If yes, what is the asymptotic proportion of long primes among all primes? More precisely, if $c(n)$ is the number of long primes less than n divided by the number of all primes less than n , then one asks for the asymptotic behavior of $c(n)$ as n gets large. In the decimal¹⁰ case, the following conjecture of Emil Artin (1898–1962) is still open.

Conjecture 3.3.5. When n grows, the ratio $c(n)$ of all long primes among all primes tends to Artin's constant

$$C_{\text{Artin}} := \prod_{p \text{ prime}} \left(1 - \frac{1}{p(p-1)}\right) \approx 0.373955.$$

In other words, the conjecture asserts that in the long run about 37% of all primes are long primes with respect to base $b = 10$. Computational investigations (cf. A001913 - OEIS) endorse this conjecture.

⁹Some authors also call them **cyclic number** or **full-period prime** or **full reptend prime**.

¹⁰Of course, this question is also of interest (and open) for bases $b \neq 10$. But computational investigations suggest that in this case the constant C_{Artin} has to be changed suitably in dependence on the base b (see [6], p. 171).

We still have not answered the following question: For which a_1, \dots, a_m and a_{m+1}, \dots, a_{m+k} in $\{0, \dots, b-1\}$ is there a rational number $\alpha \in [0, 1)$ such that

$$\alpha =_b 0.a_1 \dots a_m \overline{a_{m+1} \dots a_{m+k}} ?$$

There are two special cases; one is trivial, the other one impossible.

(1) $k = 1$ and $a_{m+1} = 0$ which corresponds to finite expansions. Here we already know by Proposition 3.3.1 that for all a_1, \dots, a_m in $\{0, \dots, b-1\}$ there is a unique rational number α such that

$$\alpha =_b 0.a_1 \dots a_m =_b 0.a_1 \dots a_m \bar{0}.$$

(2) If $k = 1$ and $a_{m+1} = b-1$, we already saw in Remark 3.3.1 that there does not exist a rational number which may be represented as

$$\alpha =_b 0.a_1 \dots a_m \overline{b-1} =_b 0.a_1 \dots a_m(b-1)(b-1) \dots$$

For example, there are no rational numbers α and β which may be represented according to Definition 3.3.5 as

$$\alpha = 0.123999 \dots = 0.12\bar{3} \quad \text{or} \quad \beta =_2 0.100\bar{1}.$$

Definition 3.3.8. A finite sequence

$$a_1, \dots, a_m, a_{m+1}, \dots, a_{m+k}$$

of integers in $\{0, \dots, b-1\}$ is said to be **admissible** (for an expansion) if none of the above cases occurs. That is, we do not have $k = 1$ and $a_{m+1} = b-1$ nor $k = 1$ and $a_{m+1} = 0$.

To get a clue how to answer the question about the occurring sequences in the expansion of rational numbers, let us give some examples which easily may be checked by direct calculations.

Example 3.3.14. Using the notation introduced in Definition 1.5.2 one has

$$\frac{725}{999} = 0.\overline{725}, \quad \frac{7}{99} = 0.\overline{07}, \quad \frac{5}{7} = \frac{101_2}{7} =_2 0.\overline{101}, \quad \text{and} \quad \frac{23}{26} = \frac{212_3}{26} =_3 0.\overline{212}.$$

Observe that the denominators in the previous example are $10^3 - 1$, $10^2 - 1$, $2^3 - 1$, and $3^3 - 1$. This suggests a general result which we are going to prove now.

Proposition 3.3.6. Let $b \geq 2$ be a fixed base. Then for all a_1, \dots, a_k in $\{0, \dots, b-1\}$ with $a_j \neq b-1$ for at least one $1 \leq j \leq k$ it follows that

$$(3.3.4) \quad \frac{a_1 b^{k-1} + \dots + a_{k-1} b + a_k}{b^k - 1} = \frac{(a_1 \dots a_k)_b}{b^k - 1} =_b 0.\overline{a_1 \dots a_k}.$$

Proof: In a first step we prove that

$$(3.3.5) \quad 0 \leq \frac{(a_1 \dots a_k)_b}{b^k - 1} < 1.$$

The left-hand estimate is a consequence of $a_j \geq 0$ for all $j \leq k$.

The right-hand inequality easily follows from the fact that at least one of the integers $a_j < b - 1$. Hence, we get

$$(a_1 \dots a_k)_b = a_1 b^{k-1} + \dots + a_{k-1} b + a_k < (b-1) \sum_{j=0}^{k-1} b^j = (b-1) \cdot \frac{b^k - 1}{b-1} = b^k - 1.$$

Of course, this implies (3.3.5).

The crucial property we are going to prove now is

$$(3.3.6) \quad \frac{(a_1 \dots a_k)_b}{b^k - 1} = \frac{a_1}{b} + \frac{1}{b} \cdot \frac{(a_2 \dots a_k a_1)_b}{b^k - 1}.$$

For example, if $b = 10$ and $a = 2$, then it follows that

$$\frac{725}{999} = \frac{7}{10} + \frac{1}{10} \cdot \frac{257}{999} \quad \text{and} \quad \frac{6}{7} = \frac{110_2}{7} = \frac{1}{2} + \frac{1}{2} \cdot \frac{101_2}{7} = \frac{1}{2} + \frac{1}{2} \cdot \frac{5}{7}.$$

To see (3.3.6), write its right-hand side as

$$\frac{1}{b} \cdot \frac{[a_1(b^k - 1) + a_2 b^{k-1} + \dots + a_k b + a_1]}{b^k - 1} = \frac{(a_1 \dots a_k)_b}{b^k - 1}.$$

Combining this (3.3.6) with (3.3.5) (note that this estimate is valid for any choice of the a_j provided that not all of them coincide with $b - 1$) we obtain

$$\frac{a_1}{b} \leq \frac{(a_1 \dots a_k)_b}{b^k - 1} < \frac{a_1}{b} + \frac{1}{b}.$$

Consequently, the first digit in the b -expansion of $\frac{(a_1 \dots a_k)_b}{b^k - 1}$ is a_1 .

In the next step we apply (3.3.6) to the sequence a_2, \dots, a_k, a_1 and obtain

$$\frac{(a_1 \dots a_k)_b}{b^k - 1} = \frac{a_1}{b} + \frac{a_2}{b^2} + \frac{1}{b^2} \cdot \frac{(a_3 \dots a_k a_1 a_2)_b}{b^k - 1},$$

which as before implies

$$\frac{a_1}{b} + \frac{a_2}{b^2} \leq \frac{(a_1 \dots a_k)_b}{b^k - 1} < \frac{a_1}{b} + \frac{a_2}{b^2} + \frac{1}{b^2}.$$

Hence, the second digit in the expansion equals a_2 .

After k steps, we arrive at

$$\frac{(a_1 \dots a_k)_b}{b^k - 1} = \frac{a_1}{b} + \dots + \frac{a_k}{b^k} + \frac{1}{b^k} \cdot \frac{(a_1 \dots a_k)_b}{b^k - 1},$$

hence by (3.3.5) it follows that

$$\frac{a_1}{b} + \dots + \frac{a_k}{b^k} \leq \frac{(a_1 \dots a_k)_b}{b^k - 1} < \frac{a_1}{b} + \dots + \frac{a_k}{b^k} + \frac{1}{b^k}.$$

Now we see that the same game starts again, but this time with $\frac{1}{b^k} \cdot \frac{(a_1 \dots a_k)_b}{b^k - 1}$. Thus, the $(k+1)$ -th digit of the expansion is again a_1 the next a_2 , and so on. This proves

$$\frac{a_1 b^{k-1} + \dots + a_{k-1} b + a_k}{b^k - 1} = \frac{(a_1 \dots a_k)_b}{b^k - 1} =_b 0.\overline{a_1 \dots a_k}.$$

as asserted. ■

Remark 3.3.5. Maybe some readers will wonder why we did not prove Proposition 3.3.6 more directly in the following way: If $\alpha =_b 0.\overline{a_1 \dots a_k}$, then

$$b^k \alpha = b^k \cdot 0.\overline{a_1 \dots a_k} = a_1 b^{k-1} + \dots + a_k + 0.\overline{a_1 \dots a_k} = (a_1 \dots a_k)_b + \alpha,$$

hence

$$b^k \alpha - \alpha = (a_1 \dots a_k)_b,$$

which implies formula (3.3.4) asserting

$$\alpha = \frac{(a_1 \dots a_k)_b}{b^k - 1}.$$

There are two reasons why we cannot argue in this way. First, at this point we do not know whether there exists a rational number α with $\alpha =_b 0.\overline{a_1 \dots a_k}$. The existence of such an α will be proved in Proposition 4.7.2. But its proof rests upon the completeness of the real line which is not available in the current setting. Consequently, we cannot start the evaluation with a number α whose existence is not assured. We have to verify its existence directly by showing that there are integers p and q with $p/q =_b 0.\overline{a_1 \dots a_k}$.

Secondly, in the last step of the previous “proof” we added two infinite b -fractions. But it is not clear at all if

$$a_1 a_2 \dots a_1 \overline{a_1 \dots a_k} - 0.\overline{a_1 \dots a_k} = (a_1 a_2 \dots a_1)_b.$$

Therefore, we had to prove Proposition 3.3.6 directly within the framework of rational numbers which makes its proof a bit more involved.

An immediate consequence of Proposition 3.3.6 is as follows.

Theorem 3.3.7. *Let $a_1, \dots, a_m, a_{m+1}, \dots, a_{m+k}$ be an admissible sequence of integers in $\{0, \dots, b-1\}$. Then there is a unique rational $\alpha \in [0, 1)$ such that*

$$\alpha =_b 0.a_1 \dots a_m \overline{a_{m+1} \dots a_{m+k}}.$$

Moreover, it holds that

$$\begin{aligned} (3.3.7) \quad \alpha &= \frac{(a_1 \dots a_{m+k})_b - (a_1 \dots a_m)_b}{b^{m+k} - b^m} \\ &= \frac{[a_1 b^{m+k-1} + \dots + a_{m+k}] - [a_1 b^{m-1} + \dots + a_m]}{b^{m+k} - b^m}. \end{aligned}$$

Proof: In view of Proposition 3.3.6 applied to a_{m+1}, \dots, a_{m+k} there is a rational number α_0 such that

$$\alpha_0 = \frac{(a_{m+1} \dots a_{m+k})_b}{b^k - 1} =_b 0.\overline{a_{m+1} \dots a_{m+k}},$$

hence we get (compare Exercise 3.3.8)

$$\frac{\alpha_0}{b^m} = \frac{(a_{m+1} \dots a_{m+k})_b}{b^{m+k} - b^m} =_b 0, \underbrace{0 \dots 0}_{m \text{ times}} \overline{a_{m+1} \dots a_{m+k}}.$$

Define the b -rational number α_1 by

$$\alpha_1 =_b 0.a_1 \dots a_m = \frac{(a_1 \dots a_m)_b}{b^m}.$$

Then we get (compare again Exercise 3.3.8)

$$\alpha := \alpha_1 + \frac{\alpha_0}{b^m} = 0.a_1 \dots a_m \overline{a_{m+1} \dots a_{m+k}}.$$

This completes the proof of the existence of the rational number possessing the desired representation as b -fraction.

It remains to prove formula (3.3.7). But this follows by

$$\begin{aligned}\alpha &= \alpha_1 + \frac{\alpha_0}{b^m} = \frac{(a_1 \dots a_m)_b \cdot (b^k - 1) + (a_{m+1} \dots a_{m+k})_b}{b^{m+k} - b^m} \\ &= \frac{[a_1 b^{m+k-1} + \dots + a_m b^k] - [a_1 b^{m-1} + \dots + a_m] + [a_{m+1} b^{k-1} + \dots + a_{m+k}]}{b^{m+k} - b^m} \\ &= \frac{[a_1 b^{m+k-1} + \dots + a_m b^k + a_{m+1} b^{k-1} + \dots + a_{m+k}] - [a_1 b^{m-1} + \dots + a_m]}{b^{m+k} - b^m} \\ &= \frac{(a_1 \dots a_{m+k})_b - (a_1 \dots a_m)_b}{b^{m+k} - b^m}.\end{aligned}$$

■

Remark 3.3.6. In the case $m = 0$ formula (3.3.7) has to be understood as

$$\alpha = \frac{(a_1 \dots a_k)_b}{b^k - 1}.$$

This is nothing other than the assertion of Proposition 3.3.6.

Example 3.3.15. Suppose $b = 10$ and we want to evaluate $\alpha = 0.\overline{112}$. Then $m = 1$ and $k = 2$, hence $m + k = 3$ and we conclude that

$$\alpha = \frac{112 - 1}{10^3 - 10^1} = \frac{111}{990} = \frac{37}{330}.$$

Example 3.3.16. Let $b = 10$ and $\alpha = 0.3\overline{142}$. Then $m = 2$, $k = 2$, $a_1 = 3$, $a_2 = 1$, $a_3 = 4$ and $a_4 = 2$. So we get

$$\alpha = \frac{3142 - 31}{10^4 - 10^2} = \frac{3111}{10^4 - 10^2} = \frac{3111}{9900} = \frac{1037}{3300}.$$

Example 3.3.17. Let $b = 2$ and $\alpha = 0.0\overline{1011}$. Here $m = 2$ and $k = 4$, thus we get

$$\alpha = \frac{011011_2 - 01_2}{2^6 - 2^2} = \frac{26}{60} = \frac{13}{30}.$$

Example 3.3.18. Which rational number is the hexadecimal fraction $\alpha = {}_{16} 0.\overline{EB}$?

In this case $m = 0$ and $k = 2$. Consequently, it follows that

$$\alpha = \frac{EB_{16}}{16^2 - 1} = \frac{235}{16^2 - 1} = \frac{235}{255} = \frac{47}{51}.$$

Finally, let us briefly discuss the excluded case $\alpha = 0.a_1 \dots a_m \overline{b-1}$. Here $k = 1$ and $a_{m+1} = b - 1$, hence (3.3.7) leads in this case to

$$\begin{aligned}\alpha &= \frac{a_1 b^m + \dots + a_m b + a_{m+1} - [a_1 b^{m-1} + \dots + a_m]}{b^{m+1} - b^m} \\ &= \frac{a_1}{b} + \dots + \frac{a_m}{b^m} + \frac{a_{m+1}}{b^{m+1} - b^m} \\ &= \frac{a_1}{b} + \dots + \frac{a_m}{b^m} + \frac{b-1}{b^{m+1} - b^m} = \frac{a_1}{b} + \dots + \frac{a_m}{b^m} + \frac{1}{b^m}.\end{aligned}$$

This contradicts the right-hand estimate in (3.3.1) with $n = m$. But nevertheless, it gives the correct answer if we would allow expansions with period $b - 1$, namely we obtain $\alpha =_b 0.a_1 \dots a_{m-1}(a_m + 1)$.

Summary: Let $\alpha \in [0, 1)$ be an arbitrary rational number. Given a base $b \geq 2$, there exist unique integers a_1, a_2, \dots in $\{0, \dots, b - 1\}$ such that for all $n = 1, 2, \dots$ one has

$$(3.3.8) \quad \frac{a_1}{b} + \frac{a_2}{b^2} + \dots + \frac{a_n}{b^n} \leq \alpha < \frac{a_1}{b} + \frac{a_2}{b^2} + \dots + \frac{a_n}{b^n} + \frac{1}{b^n}.$$

In this case one writes $\alpha =_b 0.a_1 a_2 a_3 \dots$. The sequence a_1, a_2, \dots is either

- (1) finite, which happens if and only if α is b -fractional,
- (2) completely repeating, that is

$$\alpha =_b 0.\overline{a_1, \dots, a_k} \quad \text{for some } k \geq 1.$$

Then it follows that

$$\alpha = \frac{(a_1 \dots a_k)_b}{b^k - 1},$$

- (3) repeating with a nonrepeating part, meaning that

$$\alpha =_b 0.a_1 \dots a_m \overline{a_{m+1}, \dots, a_{m+k}} \quad \text{for some } k, m \geq 1.$$

Then

$$\alpha = \frac{(a_1 \dots a_{m+k})_b - (a_1 \dots a_m)_b}{b^{m+k} - b^m}.$$

Conversely, to any sequence a_1, a_2, \dots which is either finite, completely repeating, or repeating with a nonrepeating part and where the period is neither $\overline{0}$ nor $\overline{b-1}$ there exists a unique rational number $\alpha \in [0, 1)$ satisfying (3.3.8) for all $n \geq 1$.

Exercise 3.3.1. Write the following decimal fractions as p/q for suitable integers p and q :

$$0.3\bar{3}, 0.\bar{3}, 2.\bar{72}, 2.\overline{727}, 1.2\overline{343}, 1.\overline{234}.$$

Exercise 3.3.2. Prove that $1/7 = 0.\overline{142857}$.

Exercise 3.3.3. Find the decimal and dyadic fraction representation of

$$1/2, 1/4, 1/5, 1/6, 1/7, 1/8, 1/9.$$

Exercise 3.3.4. Write $\alpha = 7/10$ as b -fraction for $b \in \{2, 3, 5, 7, 12, 16\}$.

Exercise 3.3.5. Write α as a reduced fraction p/q with suitable integers p and q in each of the following cases:

$$\alpha =_{16} 0.\overline{BC}, \quad \alpha =_2 0.\overline{1001}, \quad \text{and} \quad \alpha =_3 0.\overline{2021}.$$

Exercise 3.3.6. Let $b \geq 3$ be some fixed base. Describe all rational numbers α which may be represented as

$$\alpha =_b 0.\overline{a},$$

for a suitable $a \in \{0, \dots, b-2\}$.

Exercise 3.3.7. Which rational numbers α can be written as

$$\alpha =_b 0.\overline{a_1 a_2},$$

with $a_1, a_2 \in \{0, \dots, b-1\}$ and $a_1 \neq a_2$.

Exercise 3.3.8. Let $b \geq 2$ be a fixed base. Suppose a number $\alpha \in \mathbb{Q}$ admits the representation $\alpha =_b 0.a_1 a_2 \dots$ according to Definition 3.3.5.

(1) Show that then for any $m \geq 1$ one has

$$\frac{\alpha}{b^m} = 0.\underbrace{0 \dots 0}_m a_1 a_2 \dots$$

(2) Let x be a rational number in $(0, 1)$ with finite representation $x =_b 0.x_1 \dots x_m$.

Verify that then for $\alpha =_b 0.a_1 a_2 \dots$ one gets

$$x + \frac{\alpha}{b^m} =_b 0.x_1 \dots x_m a_1 a_2 \dots$$

Exercise 3.3.9. Define a sequence $(a_j)_{j \geq 1}$ of integers in $\{0, \dots, 9\}$ as follows:

$$a_j = \begin{cases} 5 & : j = 2^k \text{ for some } k \geq 1 \\ 0 & : \text{otherwise} \end{cases}$$

So, the first elements in the sequence are

$$0,\underset{2}{\cancel{5}}, 0,\underset{4}{\cancel{5}}, 0, 0, 0,\underset{8}{\cancel{5}}, 0 \dots, 0,\underset{16}{\cancel{5}}, 0 \dots, 0,\underset{32}{\cancel{5}}, 0, \dots$$

Prove that there is **no** rational number α with

$$\alpha = 0.a_1 a_2 \dots$$

Exercise 3.3.10. Let $\alpha =_2 0.a_1 a_2 \dots$ and $\beta =_2 0.b_1 b_2 \dots$ be two rational numbers expanded as dyadic fractions. Hereby, all integers a_i and b_j are either 0 or 1. Show the following:

$$\alpha < \beta \Leftrightarrow (\exists m \geq 0)(a_1 = b_1, \dots, a_m = b_m \text{ and } a_{m+1} = 0, b_{m+1} = 1).$$

In the case $m = 0$ this means $a_1 = 0$ and $b_1 = 1$.

Exercise 3.3.11. Is it possible to change Definition 3.3.5 in the following way?

The rational number $\alpha \in [0, 1]$ admits an expansion as b -fraction $\alpha =_b 0.a_1 a_2 \dots$ if there exist integers $0 \leq a_j < b$ such that for all $n \geq 1$ it follows that

$$(3.3.9) \quad \frac{a_1}{b} + \frac{a_2}{b^2} + \dots + \frac{a_n}{b^n} < \alpha \leq \frac{a_1}{b} + \frac{a_2}{b^2} + \dots + \frac{a_n}{b^n} + \frac{1}{b^n}.$$

For which rational numbers α does condition (3.3.9) lead to the same expansion as (3.3.1) and for which is this not so?

3.4. Finite Continued Fractions

In this section¹¹, we discuss an alternative representation of rational numbers that sometimes is preferred to the usual representation of the form $\frac{m}{n} = m/n$ or to the decimal representation. Before giving the definition of continued fractions, we describe some simple examples.

Example 3.4.1. Let us start with $\frac{1}{2}$. We can write it as $0 + \frac{1}{2}$. Because we have the numerator 1 in the fraction $\frac{1}{2}$, we stop here and our representation is $0 + \frac{1}{2}$.

Let us try now the example $\frac{8}{3}$. By integer division of 8 by 3, we have that $\frac{8}{3} = 2 + \frac{2}{3}$. The numerator is not 1 so we rewrite the previous expression as

$$\frac{8}{3} = 2 + \frac{2}{3} = 2 + \frac{1}{\frac{3}{2}}.$$

We now focus on the denominator $3/2$ which we can write as $\frac{3}{2} = 1 + \frac{1}{2}$. Putting these things together, we get that

$$\frac{8}{3} = 2 + \frac{1}{1 + \frac{1}{2}}.$$

Now that we've gotten our feet wet, let us dive a bit deeper and start another example.

Example 3.4.2. Let us try the fraction $\frac{16}{41}$. It is easy to see that

$$\frac{16}{41} = 0 + \frac{1}{\frac{41}{16}} = 0 + \frac{1}{2 + \frac{9}{16}} = 0 + \frac{1}{2 + \frac{1}{1 + \frac{7}{9}}} = 0 + \frac{1}{2 + \frac{1}{1 + \frac{1}{2 + \frac{1}{1 + \frac{7}{9}}}}}.$$

Before our sequence of fractions becomes larger and larger, we encourage the reader to finish this process.

¹¹The origin of continued fractions is hard to determine. For many years any use of continued fractions was restricted to special examples. Rafael Bombelli (1526–1572) and Pietro Cataldi (1548–1626) were the first who used continued fractions to approximate $\sqrt{13}$ and $\sqrt{18}$, respectively. In 1696, John Wallis (1616–1703), in his book *Opera Mathematica*, laid some basic groundwork for continued fractions. He also introduced the notations “continued fraction” and “convergent”. The first practical application of continued fractions was made by Christiaan Huygens (1629–1695) when he used them to evaluate some gear ratios. Important contributions to the theory are due to Leonhard Euler (1707–1783), Johann Heinrich Lambert (1728–1777) and to Joseph Louis Lagrange (1716–1813). Euler used continued fractions to show that e and e^2 are irrational; Lambert extended Euler's approach to prove the irrationality of π . Since that time the theory of continued fractions has been further developed. A basic standard book about continued fractions was published in 1913 by Oskar Perron (1880–1975); *Die Lehre von den Kettenbrüchen* (cf. [27]). Nowadays continued fractions have made their appearance in many other fields of mathematics, physics, and computer science. Their importance to this day derives essentially from the fact that continued fractions provide the best approximation of irrational numbers by rational ones. Another advantage is that rational numbers may always be represented as **finite** continued fractions.

Definition 3.4.1. Let n be a nonnegative integer. Let a_0 be an integer and a_1, \dots, a_n be natural numbers. The number

$$a_0 + \cfrac{1}{a_1 + \cfrac{1}{\ddots + \cfrac{1}{a_{n-1} + \cfrac{1}{a_n}}}}$$

is called a **finite continued fraction**, denoted by $[a_0; a_1, \dots, a_n]$.

Our previous examples show that

$$1/2 = [0; 2], \quad 7/2 = [3; 2], \quad 8/3 = [2; 1, 2].$$

Example 3.4.3. A simple calculation gives us that

$$\begin{aligned} [1; 2, 3, 4] &= 1 + \cfrac{1}{2 + \cfrac{1}{3 + \cfrac{1}{4}}} \\ &= 1 + \cfrac{1}{2 + \cfrac{1}{\frac{13}{4}}} = 1 + \cfrac{1}{2 + 4/13} = 1 + \cfrac{1}{30/13} = 1 + \cfrac{13}{30} = 43/30. \end{aligned}$$

Example 3.4.4. We invite the reader to verify that

$$-12/5 = [-3; 1, 1, 2] \text{ and } 12/5 = [2; 2, 2].$$

The next result is not surprising.

Proposition 3.4.1. *Any finite continued fraction is a rational number.*

Proof: We use induction on n to show that any continued fraction $[a_0; a_1, \dots, a_n]$ is a rational number. The base case is when $n = 0$. In this case, our continued fraction is a_0 . Since a_0 is an integer, the base case is true. Let n be a natural number and assume that the statement is true for any continued fraction of the form $[b_1; b_2, \dots, b_n]$. Let $[a_0; a_1, \dots, a_n]$ be a continued fraction. By our induction hypothesis, the continued fraction $[a_1; a_2, \dots, a_n]$ is a rational number. Since $a_1, \dots, a_n > 1$, the number $[a_1; a_2, \dots, a_n]$ is also positive and denote it by k/ℓ , where k, ℓ are natural numbers. Note that

$$\begin{aligned} [a_0; a_1, \dots, a_n] &= a_0 + \cfrac{1}{a_1 + \cfrac{1}{\ddots + \cfrac{1}{a_{n-1} + \cfrac{1}{a_n}}}} \\ &= a_0 + \cfrac{1}{[a_1; a_2, \dots, a_n]}. \end{aligned}$$

Therefore, substituting $[a_1; a_2, \dots, a_n]$ by k/ℓ , we deduce that

$$[a_0; a_1, \dots, a_n] = a_0 + \frac{1}{\frac{k}{\ell}} = a_0 + \frac{\ell}{k} = \frac{ka_0 + \ell}{k}.$$

Since $ka_0 + \ell$ is an integer and k is a natural number, we deduce that $[a_0; a_1, \dots, a_n]$ is a rational number. This concludes our proof. \blacksquare

The converse of the previous proposition is also true.

Proposition 3.4.2. *Every rational number has a representation as a finite continued fraction.*

Proof: This result is a consequence of integer division. We show that any rational number of the form k/ℓ can be written as a continued fraction using strong induction on ℓ . The base case corresponds to $\ell = 1$. In this case, our rational number k/ℓ is an integer k which can be written as $[k]$ or as $[k - 1; 1]$. Either way, we get a finite continued fraction representation for k and this proves the base case. Let $\ell \geq 2$ be a natural number. Assume that any rational number of the form s/t with s integer and $t < \ell$ natural number, has a finite continued fraction representation. Consider the number k/ℓ . Using integer division, there exist integers q and r such that

$$k = q\ell + r, \quad 0 \leq r \leq \ell - 1.$$

Therefore,

$$\frac{k}{\ell} = \frac{q\ell + r}{\ell} = q + \frac{r}{\ell} = q + \frac{1}{\ell/r}.$$

The rational number ℓ/r has the denominator r that is strictly less than ℓ . By our induction hypothesis, ℓ/r has a finite continued representation $\ell/r = [a_0; a_1, \dots, a_n]$ for $n \geq 0$, integer a_0 and natural numbers a_1, \dots, a_n . Therefore,

$$\begin{aligned} \frac{k}{\ell} &= q + \frac{1}{\ell/r} = q + \frac{1}{[a_0; a_1, \dots, a_n]} = q + \cfrac{1}{a_0 + \cfrac{1}{a_1 + \cfrac{1}{\ddots + \cfrac{1}{a_{n-1} + \cfrac{1}{a_n}}}}} \\ &= [q; a_0, a_1, \dots, a_n]. \end{aligned}$$

This concludes our proof. \blacksquare

Before proving some general results regarding finite continued fractions, note that the representations above are not unique and that we can also write

$$\begin{aligned}\frac{1}{2} &= [0; 2] = 0 + \frac{1}{2} = 0 + \frac{1}{\frac{1}{1}} = [0; 1, 1], \quad \frac{7}{2} = [3; 2] = 3 + \frac{1}{2} = 3 + \frac{1}{\frac{1}{1}} = [3; 1, 1], \\ \frac{8}{3} &= [2; 1, 2] = 2 + \frac{1}{1 + \frac{1}{2}} = 2 + \frac{1}{1 + \frac{1}{1 + \frac{1}{1}}} = [2; 1, 1, 1], \\ \frac{43}{30} &= [1; 2, 3, 4] = 1 + \frac{1}{2 + \frac{1}{3 + \frac{1}{4}}} = 1 + \frac{1}{2 + \frac{1}{3 + \frac{1}{3 + \frac{1}{1}}}} = [1; 2, 3, 3, 1].\end{aligned}$$

We leave it to the reader to show the following general result.

Proposition 3.4.3. *Let $n \in \mathbb{N}$. For any nonnegative integer a_0 and natural numbers a_1, \dots, a_n with $a_n \geq 2$, the following equality holds:*

$$[a_0; a_1, \dots, a_{n-1}, a_n] = [a_0; a_1, \dots, a_{n-1}, a_n - 1, 1].$$

It turns out that these are only two possible ways of writing rational numbers as finite continued fractions.

Proposition 3.4.4. *If $[a_0; a_1, \dots, a_m] = [b_0; b_1, \dots, b_n]$, $a_m > 1$ and $b_n > 1$, then $m = n$ and $a_j = b_j$ for $0 \leq j \leq m$.*

Proof: Let $r = [a_0; a_1, \dots, a_m] = [b_0; b_1, \dots, b_n]$. For $0 \leq j \leq m$ and define r_j by $r_j = [a_j; a_{j+1}, \dots, a_m]$. Hence, $r_m = a_m > 1$, $r_{m-1} = [a_{m-1}; a_m] = a_{m-1} + \frac{1}{a_m} > a_{m-1}$ and

$$(3.4.1) \quad r_{m-\ell} = a_{m-\ell} + \frac{1}{r_{m-\ell+1}}, \quad 0 \leq \ell \leq m-1.$$

Using induction on ℓ , we can show that

$$a_{m-\ell} < r_{m-\ell} < a_{m-\ell} + 1, \forall \ell \in \{1, \dots, m-1\}.$$

From above, we see that the base case $\ell = 1$ is true and this is the crucial place where the inequality $a_m > 1$ is used. For the induction step, assume that $\ell \geq 2$ and that the inequality above is true for $\ell - 1$. Thus, $a_{m-\ell+1} < r_{m-\ell+1} < a_{m-\ell+1} + 1$ and therefore, $r_{m-\ell+1} \geq 2$ since $a_{m-\ell+1} \geq 1$. Using (3.4.1), we deduce that $a_{m-\ell} < r_{m-\ell} < a_{m-\ell} + 1/2$ which implies the desired result. Therefore, $a_j = \lfloor r_j \rfloor$ for any $0 \leq j \leq m$. Since $r_0 = r$, we get that $a_0 = \lfloor r \rfloor$. By a similar argument applied to $r = [b_0; b_1, \dots, b_n]$, we obtain that $b_0 = \lfloor r \rfloor$ and thus, $a_0 = b_0$. Because $r = a_0 + \frac{1}{[a_1, \dots, a_m]} = b_0 + \frac{1}{[b_1, \dots, b_n]}$, we

deduce that $[a_1; \dots, a_m] = [b_1; \dots, b_n] = r_1$. Repeating the argument above, we get that $a_1 = b_1 = |r_1|$. By continuing this process, we can obtain that $m = n$ and $a_j = b_j$ for any $0 \leq j \leq m$. ■

Summary: Any rational number can be expressed as a **finite** continued fraction. Conversely, any finite continued fraction represents a rational number. The correspondence between continued fractions and rational numbers is essentially unique (see Proposition 3.4.4).

Thus, continued fractions provide another way of representing rational numbers and possess many other interesting properties. Moreover, they can be used to approximate irrational numbers or to solve diophantine equations, for example.

Definition 3.4.2. Let n be a natural number and $[a_0; a_1, \dots, a_n]$ be a continued fraction. For $0 \leq m \leq n$, the **m -th convergent** of $[a_0; a_1, \dots, a_n]$ is defined as a_0 if $m = 0$ and as the continued fraction $[a_0; a_1, \dots, a_m]$ when $1 \leq m \leq n$.

This is a simple definition, but let us understand it better via an example.

Example 3.4.5. In Example 3.4.3, we found out that $43/30 = [1; 2, 3, 4]$. In this case, $n = 3$ and the convergents of $[1; 2, 3, 4]$ are

$$\begin{aligned} [1] &= 1, \\ [1; 2] &= 1 + \frac{1}{2} = 3/2, \\ [1; 2, 3] &= 1 + \frac{1}{2 + \frac{1}{3}} = 1 + 3/7 = 10/7, \\ [1; 2, 3, 4] &= 43/30. \end{aligned}$$

We list the convergents again below:

$$1/1, 3/2, 10/7, 43/30.$$

If we cross multiply the numerators and denominators of consecutive fractions, we get the following pairs:

$$(2, 3), (21, 20), (300, 301).$$

Earlier in this section, we also observed that

$$43/30 = [1; 2, 3, 3, 1].$$

In this case, $n = 4$ and the convergents of $[1; 2, 3, 3, 1]$ are

$$\begin{aligned}[1] &= 1 = 1/1, \\ [1; 2] &= 1 + \frac{1}{2} = 3/2, \\ [1; 2, 3] &= 1 + \frac{1}{2 + \frac{1}{3}} = 1 + \frac{1}{7/3} = 10/7, \\ [1; 2, 3, 3] &= 1 + \frac{1}{2 + \frac{1}{3 + \frac{1}{3}}} = 1 + \frac{1}{2 + \frac{1}{10}} = 1 + \frac{1}{23/10} = 33/23, \\ [1; 2, 3, 3, 1] &= 43/30.\end{aligned}$$

We list them horizontally as well:

$$1/1, 3/2, 10/7, 33/23, 43/30.$$

If we cross multiply the numerators and denominators of consecutive fractions, we get the following pairs:

$$(2, 3), (21, 20), (230, 231), (990, 989).$$

We explain the pattern in the following result.

Proposition 3.4.5. *Let $n \in \mathbb{N}$ and $x = [a_0; a_1, \dots, a_n]$ be a positive continued fraction. For $m \in \{0, \dots, n\}$, assume that the m -th convergent of x in lowest terms equals $\frac{b_m}{c_m}$, where b_m and c_m are natural numbers such that $\gcd(b_m, c_m) = 1$. The following statements are true.*

(1) *For any $2 \leq m \leq n$,*

$$\begin{aligned}b_m &= a_m b_{m-1} + b_{m-2}, \\ c_m &= a_m c_{m-1} + c_{m-2}.\end{aligned}$$

(2) *It holds $0 \leq b_0 \leq b_1 < b_2 < \dots < b_n$ and $1 = c_0 \leq c_1 < c_2 < \dots < c_n$.*

(3) *For any $1 \leq m \leq n$,*

$$b_m c_{m-1} - b_{m-1} c_m = (-1)^{m-1}.$$

(4) *For any $1 \leq m \leq n$, the convergents $\frac{b_m}{c_m}$ satisfy*

$$\frac{b_m}{c_m} - \frac{b_{m-1}}{c_{m-1}} = \frac{(-1)^{m-1}}{c_m c_{m-1}}.$$

Proof: First note that property (2) is a direct consequence of the first one. Recall that all c_m are natural numbers ($c_0 = 1$ because a_0 is an integer), hence positive. Moreover, we have $b_0 = a_0 \geq 0$ and $b_m, a_m > 0$ if $m \geq 1$.

We will prove the first and the third identity by induction on m . For the base case, we prove the first identity for $m = 2$. We know that

$$\begin{aligned}\frac{b_0}{c_0} &= [a_0] = a_0 = \frac{a_0}{1}, \\ \frac{b_1}{c_1} &= [a_0; a_1] = a_0 + \frac{1}{a_1} = \frac{a_0 a_1 + 1}{a_1}, \\ \frac{b_2}{c_2} &= [a_0; a_1, a_2] = a_0 + \frac{1}{a_1 + \frac{1}{a_2}} = a_0 + \frac{a_2}{a_1 a_2 + 1} = \frac{a_0(a_1 a_2 + 1) + a_2}{a_1 a_2 + 1}.\end{aligned}$$

Since $\gcd(b_0, c_0) = 1$, we must have that $b_0 = a_0$ and $c_0 = 1$. Also, $\gcd(a_0 a_1 + 1, a_1) = 1$ implies that $b_1 = a_0 a_1 + 1$ and $c_1 = a_1$. Note that $\gcd(a_0(a_1 a_2 + 1) + a_2, a_1 a_2 + 1) = 1$. To see this, if d is a natural divisor of both $a_0(a_1 a_2 + 1) + a_2$ and $a_1 a_2 + 1$, then d divides a_2 and therefore, d must divide $a_1 a_2 + 1 - a_1 a_2 = 1$. Hence, $d = 1$ which proves our assertion. Therefore, $b_2 = a_0(a_1 a_2 + 1) + a_2 = a_0 a_1 a_2 + a_0 + a_2$ and $c_2 = a_1 a_2 + 1$. We can now confirm that

$$\begin{aligned}a_2 b_1 + b_0 &= a_2(a_0 a_1 + 1) + a_0 = a_0 a_1 a_2 + a_2 + a_0 = b_2, \\ a_2 c_1 + c_0 &= a_2 a_1 + 1 = c_2,\end{aligned}$$

which finishes the proof of the base case for the first identity. The third identity is true also for $m \in \{1, 2\}$ and we leave it to the reader to check these calculations.

For the induction step, let m be a natural number between 2 and $n - 1$ and assume that the induction hypothesis is true for any $2 \leq j \leq m$. We want to prove the first identity for $m + 1$. From the definition of the convergents, we know that

$$\begin{aligned}\frac{b_m}{c_m} &= [a_0; a_1, \dots, a_{m-1}, a_m] \\ \frac{b_{m+1}}{c_{m+1}} &= [a_0; a_1, \dots, a_{m-1}, a_m, a_{m+1}] = [a_0; a_1, \dots, a_{m-2}, a_{m-1}, a_m + \frac{1}{a_{m+1}}].\end{aligned}$$

From our induction hypothesis, we know that

$$\begin{aligned}b_m &= a_m b_{m-1} + b_{m-2}, \\ c_m &= a_m c_{m-1} + c_{m-2}, \\ (-1)^{m-1} &= b_m c_{m-1} - b_{m-1} c_m.\end{aligned}$$

Note that the j -th convergents of $[a_0; a_1, \dots, a_{m-1}, a_m]$ and $[a_0; a_1, \dots, a_{m-1}, a_m + \frac{1}{a_{m+1}}]$ are the same for $j < m$. To calculate the m -th convergent of $[a_0; a_1, \dots, a_{m-1}, a_m + \frac{1}{a_{m+1}}]$, we use the induction hypothesis, and we can write it as

$$\begin{aligned}\frac{\left(a_m + \frac{1}{a_{m+1}}\right)b_{m-1} + b_{m-2}}{\left(a_m + \frac{1}{a_{m+1}}\right)c_{m-1} + c_{m-2}} &= \frac{(a_m a_{m+1} + 1)b_{m-1} + a_{m+1} b_{m-2}}{(a_m a_{m+1} + 1)c_{m-1} + a_{m+1} c_{m-2}} \\ &= \frac{a_{m+1}(a_m b_{m-1} + b_{m-2}) + b_{m-1}}{a_{m+1}(a_m c_{m-1} + c_{m-2}) + c_{m-1}} \\ &= \frac{a_{m+1}b_m + b_{m-1}}{a_{m+1}c_m + c_{m-1}}.\end{aligned}$$

However, the m -th convergent of $[a_0; a_1, \dots, a_{m-1}, a_m + \frac{1}{a_{m+1}}]$ is the $(m+1)$ -th convergent of $[a_0; a_1, \dots, a_n]$ and equals $\frac{b_{m+1}}{c_{m+1}}$. Note that

$$\begin{aligned} c_m(a_{m+1}b_m + b_{m-1}) - b_m(a_{m+1}c_m + c_{m-1}) &= b_{m-1}c_m - b_m c_{m-1} \\ &= (-1) \cdot (b_m c_{m-1} - b_{m-1} c_m) \\ &= (-1)^m. \end{aligned}$$

This implies that $\gcd(a_{m+1}b_m + b_{m-1}, a_{m+1}c_m + c_{m-1}) = 1$ and therefore,

$$b_{m+1} = a_{m+1}b_m + b_{m-1} \text{ and } c_{m+1} = a_{m+1}c_m + c_{m-1}.$$

This finishes our proof of properties (1) and (3). Finally, assertion (4) follows from (3) by dividing this equation by $c_m c_{m-1}$. \blacksquare

In Section 2.3, we showed how to solve diophantine equations of the form:

$$(3.4.2) \quad 43x + 30y = 1, \quad x, y \in \mathbb{Z}.$$

A key part of that work was coming up with one pair of integers (x_0, y_0) such that $43x_0 + 30y_0 = 1$. The previous proposition and our work before it can give us such a pair.

In Example 3.4.3, we saw that $43/30 = [1; 2, 3, 4]$. The second convergent of is $[1; 2, 3] = 10/7$ and we observed earlier that

$$43 \cdot 7 - 30 \cdot 10 = 1.$$

Hence, $(7, -10)$ is a particular solution of the equation (3.4.2). We give a quick recap of how one solves this equation. If (x, y) is an arbitrary solution, then

$$43x + 30y = 1 = 43 \cdot 7 + 30 \cdot (-10)$$

implies that

$$43(x - 7) = 30(-y - 10).$$

Since $\gcd(43, 30) = 1$, we get that 30 divides $x - 7$. Therefore, $x - 7 = 30k$ for some integer k . This implies that $30(-y - 10) = 43(x - 7) = 43 \cdot 30k$ and therefore, $y = -43k - 10$. It can be easily checked that any pair of the form $(30\ell + 7, -43\ell - 10)$ is a solution of the equation (3.4.2). Hence, the set of solutions of this equation is

$$\{(30k + 7, -43k - 10) : k \in \mathbb{Z}\}.$$

As mentioned earlier, we can also write $43/30$ as $[1; 2, 3, 3, 1]$. The third convergent is $[1; 2, 3, 3] = 33/23$. Proposition 3.4.5 and our earlier calculations guarantee that

$$43 \cdot 33 - 30 \cdot 23 = (-1)^3 = -1,$$

which implies that

$$43 \cdot (-33) + 30 \cdot 23 = 1.$$

Hence, $(-33, 23)$ is a solution of our equation (3.4.2). One can use this solution to solve the above equation. In the end, the set of solutions will be the same.

Proposition 3.4.6. Let n be a natural number and $x = [a_0; a_1, \dots, a_n]$ be a positive continued fraction. For $m \in \{0, \dots, n\}$, assume that the m -th convergent of x in lowest terms equals $\frac{b_m}{c_m}$, where b_m and c_m are natural numbers such that $\gcd(b_m, c_m) = 1$. The following statements are true:

- (1) The even convergents $\frac{b_{2m}}{c_{2m}}$ strictly increase with m for $0 \leq m \leq \lfloor n/2 \rfloor$.
- (2) The odd convergents $\frac{b_{2m+1}}{c_{2m+1}}$ strictly decrease with m for $0 \leq m \leq \lfloor n/2 \rfloor$.
- (3) Except for the n -th convergent, x is strictly greater than any even convergent and strictly smaller than any odd convergent.

Proof: Let k be an integer between 2 and n . From Proposition 3.4.5, we deduce that

$$\begin{aligned} \frac{b_k}{c_k} - \frac{b_{k-2}}{c_{k-2}} &= \left(\frac{b_k}{c_k} - \frac{b_{k-1}}{c_{k-1}} \right) + \left(\frac{b_{k-1}}{c_{k-1}} - \frac{b_{k-2}}{c_{k-2}} \right) \\ &= \frac{b_k c_{k-1} - b_{k-1} c_k}{c_k c_{k-1}} + \frac{b_{k-1} c_{k-2} - b_{k-2} c_{k-1}}{c_{k-1} c_{k-2}} \\ &= \frac{(-1)^{k-1}}{c_k c_{k-1}} + \frac{(-1)^{k-2}}{c_{k-1} c_{k-2}} \\ &= \frac{(-1)^k (c_k - c_{k-2})}{c_k c_{k-1} c_{k-2}}. \end{aligned}$$

Using part (1) of Proposition 3.4.5, we get that $c_k - c_{k-2} = a_k c_{k-1}$ and therefore,

$$\frac{b_k}{c_k} - \frac{b_{k-2}}{c_{k-2}} = \frac{(-1)^k a_k c_{k-1}}{c_k c_{k-1} c_{k-2}} = \frac{(-1)^k a_k}{c_k c_{k-2}}.$$

This equation implies the first two assertions.

If n is even, then the previous inequalities imply that

$$\frac{b_0}{c_0} < \frac{b_2}{c_2} < \dots < \frac{b_n}{c_n} = x.$$

Using part (2) of Proposition 3.4.5, we get that

$$\frac{b_{n-1}}{c_{n-1}} - x = \frac{b_{n-1}}{c_{n-1}} - \frac{b_n}{c_n} = \frac{b_{n-1} c_n - b_n c_{n-1}}{c_{n-1} c_n} = \frac{(-1)^{n-2}}{c_{n-1} c_n} > 0.$$

Combining this inequality with part (2) above, we deduce that x is strictly greater than any odd convergent. The proof for n odd is similar and is left as an exercise. ■

We end this section with some food for thought regarding the behavior of certain continued fractions.

Example 3.4.6. Let us consider the behavior of a sequence of finite fractions of the simplest form $[1; 1, \dots, 1]$. We have that

$$\begin{aligned}[1] &= 1 = \frac{1}{1}, \\ [1; 1] &= 1 + \frac{1}{1} = 2 = \frac{2}{1}, \\ [1; 1, 1] &= 1 + \frac{1}{[1; 1]} = \frac{3}{2}, \\ [1; 1, 1, 1] &= 1 + \frac{1}{[1; 1, 1]} = \frac{5}{3}, \\ [1; 1, 1, 1, 1] &= 1 + \frac{1}{[1; 1, 1, 1]} = \frac{8}{5}. \end{aligned}$$

Perhaps a pattern is apparent to the attentive reader who remembers the Fibonacci numbers from the previous chapters:

$$F_0 = 0, F_1 = 1, F_2 = 1, F_3 = 2, F_4 = 3, F_5 = 5, F_6 = 8, F_7 = 13,$$

and $F_{n+1} = F_n + F_{n-1}$ for any $n \geq 1$. We can prove using induction on n that

$$\underbrace{[1; 1, \dots, 1]}_{n \text{ terms}} = F_{n+2}/F_{n+1}.$$

This is true for $n \in \{0, 1, 2, 3, 4\}$ as the above calculations prove. For the induction step, assume that

$$\underbrace{[1; 1, \dots, 1]}_{n} = F_{n+2}/F_{n+1}.$$

It is clear that

$$\underbrace{[1; 1, \dots, 1]}_{n+1} = 1 + \frac{1}{\underbrace{[1; 1, \dots, 1]}_n} = 1 + F_{n+1}/F_{n+2} = F_{n+3}/F_{n+2}.$$

A natural question is to understand the behavior of the F_{n+2}/F_{n+1} sequence as n gets larger. As we saw earlier,

$$(3.4.3) \quad F_n = \frac{1}{\sqrt{5}} \left[\left(\frac{1+\sqrt{5}}{2} \right)^n - \left(\frac{1-\sqrt{5}}{2} \right)^n \right],$$

for every n . Given that the golden ratio $\varphi = \frac{1+\sqrt{5}}{2}$ is about 1.618 and $\frac{1-\sqrt{5}}{2}$ close to -0.618, we may convince ourselves that most of the *mass* of the Fibonacci number F_n is in the term $\left(\frac{1+\sqrt{5}}{2} \right)^n$ and that F_n is close to $\frac{1}{\sqrt{5}} \left(\frac{1+\sqrt{5}}{2} \right)^n$. With these approximations in mind, one educated guess is that the sequence F_{n+2}/F_{n+1} converges to the limit $\frac{1+\sqrt{5}}{2}$ as n gets large. This is actually the truth, and we will be able to prove this statement formally in Proposition 5.3.4 after learning about limits of sequences of real numbers (see also Exercise 3.4.10).

Let us take a different approach in our next example.

Example 3.4.7. Consider the irrational number $\alpha = \sqrt{2}$. We cannot write $\sqrt{2}$ as a finite continued fraction since $\sqrt{2}$ is not a rational number. However, let us putter around for a bit and see how close can we get to $\sqrt{2}$ using continued fractions. From $\alpha^2 = 2$ we can deduce that $(\alpha + 1)(\alpha - 1) = \alpha^2 - 1 = 1$ and therefore,

$$(3.4.4) \quad \alpha = 1 + (\alpha - 1) = 1 + \frac{1}{1 + \alpha}.$$

The right-hand side almost looks like a continued fraction except for the part involving α which is not rational. We can copy the formula from equation (3.4.4) and paste it in the right-hand side where α is. It is a bit like a dream in a dream if you wish.

$$\alpha = 1 + \cfrac{1}{2 + \cfrac{1}{1 + \alpha}}.$$

Of course, nothing (except the decreasing size of fonts perhaps) can stop us from repeating this process:

$$\alpha = 1 + \cfrac{1}{2 + \cfrac{1}{2 + \cfrac{1}{2 + \cfrac{1}{2 + \cfrac{1}{2 + \cfrac{1}{1 + \alpha}}}}}} = 1 + \cfrac{1}{2 + \cfrac{1}{2 + \cfrac{1}{2 + \cfrac{1}{2 + \cfrac{1}{2 + \cfrac{1}{2 + \cfrac{1}{1 + \alpha}}}}}}}.$$

A natural question is what can we say about

$$r_n = [1; \underbrace{2, \dots, 2}_n],$$

as n grows large.

If we write $r_n = b_n/c_n$, then we list below these numbers for small values of n :

| n | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|-------|---|---|---|----|----|----|-----|-----|
| b_n | 1 | 3 | 7 | 17 | 41 | 99 | 239 | 577 |
| c_n | 1 | 2 | 5 | 12 | 29 | 70 | 169 | 408 |

The reader is invited to double-check this work and to calculate other values of r_n . Note that already for $99/70$ we get a very good approximation of $\sqrt{2}$ as both numbers agree on the first four decimal digits

$$1.41428 < 99/70 < 1.41429 \text{ and } 1.41421 < \sqrt{2} < 1.41422.$$

The attentive reader might have noticed a similarity between these numbers and the ones in the equation (3.2.1). This leads us to the possibility that

$$b_m = b_{m-1} + 2c_{m-1} \text{ and } c_m = b_{m-1} + c_{m-1},$$

for any $m \geq 1$. We will prove the statements above by strong induction on m . These assertions can be checked easily for $m \in \{0, 1, 2\}$. Let $m \geq 3$ and assume that the

equations above are true for any $k \leq m - 1$. Our induction hypothesis implies that

$$\begin{aligned} b_{m-1} &= b_{m-2} + 2c_{m-2} \text{ and } c_{m-1} = b_{m-2} + c_{m-2}, \\ b_{m-2} &= b_{m-3} + 2c_{m-3} \text{ and } c_{m-2} = b_{m-3} + c_{m-3}. \end{aligned}$$

From Proposition 3.4.5, we know that

$$b_{m-1} = 2b_{m-2} + b_{m-3} \text{ and } c_{m-1} = 2c_{m-2} + c_{m-3}.$$

Therefore,

$$\begin{aligned} b_m &= 2b_{m-1} + b_{m-2} = 2(b_{m-2} + 2c_{m-2}) + b_{m-3} + 2c_{m-3} \\ &= (2b_{m-2} + b_{m-3}) + 2(2c_{m-2} + c_{m-3}) = b_{m-1} + 2c_{m-1}. \end{aligned}$$

The proof that $c_m = b_{m-1} + c_{m-1}$ is similar, and we leave it as an exercise. We deduce another useful relation:

$$b_m^2 - 2c_m^2 = (b_{m-1} + 2c_{m-1})^2 - 2(b_{m-1} + c_{m-1})^2 = -(b_{m-1}^2 - 2c_{m-1}^2).$$

This implies that $|b_m^2 - 2c_m^2| = 1$ for any $m \geq 0$. We can use this result to show that the fractions b_m/c_m give a very good approximation of $\sqrt{2}$:

$$\left| \sqrt{2} - \frac{b_m}{c_m} \right| = \frac{|b_m - c_m\sqrt{2}|}{c_m\sqrt{2}} = \frac{|b_m^2 - 2c_m^2|}{c_m\sqrt{2}(b_m + c_m\sqrt{2})} = \frac{1}{c_m\sqrt{2}(b_m + c_m\sqrt{2})}.$$

Note that $b_m > c_m > 2^m$ for any $m \geq 0$. This can also be proved by induction on m and combined with the previous inequality, gives that

$$\left| \sqrt{2} - \frac{b_m}{c_m} \right| < \frac{1}{c_m^2(2 + \sqrt{2})}.$$

Hence, the sequence of fractions b_m/c_m has increasing denominators c_m and is *getting closer* to $\sqrt{2}$ as m gets larger.

We will discuss infinite continued fractions in Section 5.7.

Exercise 3.4.1. Find the rational number corresponding to each of the following continued fractions:

- (1) $[1; 2, 3]$,
- (2) $[3; 7, 15]$,
- (3) $[3; 7, 15, 1]$,
- (4) $[3; 7, 15, 1, 292]$,
- (5) $[2; 1, 2, 1, 1, 4, 1, 1, 6]$.

Exercise 3.4.2. For each of the rational numbers above, calculate all their convergents.

Exercise 3.4.3. Find a representation as finite continued fractions of

$$\frac{16}{41}, \frac{22}{7}, \frac{333}{106}, \frac{335}{113}, \frac{23}{17}, \frac{79}{101}.$$

Exercise 3.4.4. Find a representation as finite continued fractions of

$$\frac{41}{16}, \frac{7}{22}, \frac{106}{333}, \frac{113}{335}, \frac{17}{23}, \frac{101}{79}.$$

Exercise 3.4.5. Find a representation as finite continued fractions of

$$\frac{-41}{16}, \frac{-7}{22}, \frac{-106}{333}, \frac{-113}{335}, \frac{-17}{23}, \frac{-101}{79}.$$

Exercise 3.4.6. Let $a_0 \in \mathbb{Z}$ and $a_1, a_2, k \in \mathbb{N}$. Prove that

$$[a_0; a_1] > [a_0; a_1 + k] \text{ and } [a_0; a_1, a_2] < [a_0; a_1, a_2 + k].$$

Exercise 3.4.7. Let $n \in \mathbb{N}$. If $a_0 \in \mathbb{Z}, a_1, \dots, a_n, k \in \mathbb{N}$, prove that

$$[a_0; a_1, \dots, a_n] > [a_0; a_1, \dots, a_n + k] \text{ if } n \text{ is odd,}$$

$$[a_0; a_1, \dots, a_n] < [a_0; a_1, \dots, a_n + k] \text{ if } n \text{ is even.}$$

Exercise 3.4.8. Find the rational numbers whose continued fraction representations are

- (1) $[2; 4, 4]$,
- (2) $[2; 4, 4, 4]$,
- (3) $[2; 4, 4, 4, 4]$,
- (4) $[2; 4, 4, 4, 4, 4]$.

What do you think is the behavior of $[2; \underbrace{4, \dots, 4}_{n \text{ terms}}]$ as n gets large?

Exercise 3.4.9. Find the rational numbers whose continued fraction representations are

- (1) $[1; 1, 2]$,
- (2) $[1; 1, 2, 1, 2]$,
- (3) $[1; 1, 2, 1, 2, 1, 2]$,
- (4) $[1; 1, 2, 1, 2, 1, 2, 1, 2]$.

What do you think is the behavior of $[1; \underbrace{1, 2, \dots, 1, 2}_{2n \text{ terms}}]$ as n gets large?

Exercise 3.4.10. Let $\varphi = \frac{1+\sqrt{5}}{2}$ denote the golden ratio. Using the formula (3.4.3), prove that

$$\left| \frac{F_{n+1}}{F_n} - \varphi \right| = \frac{1}{\sqrt{5}F_n^2} \left(1 - \frac{(-1)^n}{\varphi^{2n}} \right),$$

for any $n \in \mathbb{N}$. If n is even, show that

$$\left| \frac{F_{n+1}}{F_n} - \varphi \right| \leq \frac{1}{\sqrt{5}F_n^2}.$$

3.5. Farey Sequences and Pick's Formula

A rational number a/b is called reduced or in lowest terms or irreducible whenever $\gcd(a, b) = 1$ (see Definition 3.2.1). Throughout this section, the fractions we will consider are assumed to be nonnegative with positive denominators.

Definition 3.5.1. Let $n \in \mathbb{N}$. **The Farey sequence** \mathcal{F}_n of order n is the increasing sequence of reduced fractions between 0 and 1 whose denominator is at most n .

We describe these sequences in Figure 3.5.1 by drawing the Farey sequence $\mathcal{F}_1 = \{0/1, 1/1\}$ at the top and then proceeding to add the new fractions as n grows from 1 to 7. For example, $\mathcal{F}_5 = \{0/1, 1/5, 1/4, 1/3, 2/5, 1/2, 3/5, 2/3, 3/4, 4/5, 1/1\}$. A recursive way of constructing the Farey sequence \mathcal{F}_n is suggested by Figure 3.5.1. For $n \geq 2$, the Farey sequence \mathcal{F}_n can be obtained from the preceding Farey sequence \mathcal{F}_{n-1} by adding the fractions of the form $\frac{a}{n}$, where $1 \leq a \leq n$ and $\gcd(a, n) = 1$.

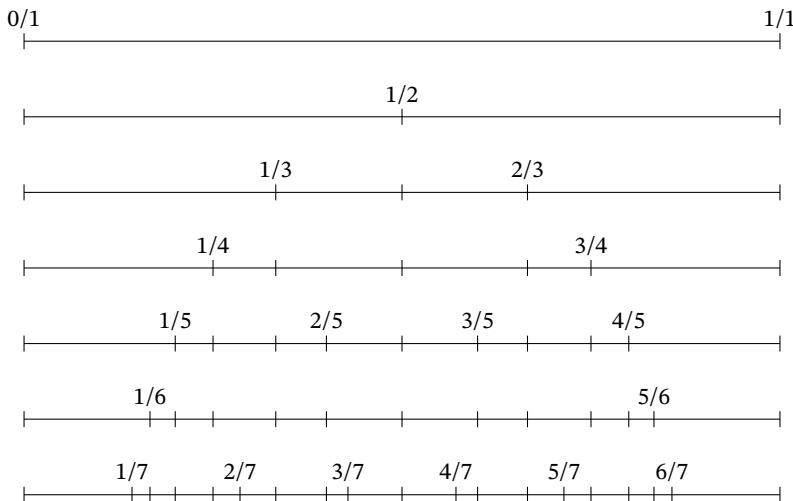


Figure 3.5.1. The Farey sequences \mathcal{F}_n for $1 \leq n \leq 7$.

In 1816, the English geologist John Farey (1766–1826) wrote a letter to the *Philosophical Magazine* in which he described some empirical observations about *vulgar fractions* (another name for reduced fractions) and asked some questions about these sequences. His questions, which we will discuss a bit later, were answered by the French mathematician Augustin-Louis Cauchy (1789–1857) around 1816. Cauchy assigned the name Farey to these sequences, but it turns out that, unknown to Cauchy, around 1802, another French mathematician, Charles Haros¹² already investigated these sequences. Charles Haros was a mathematician who worked in the French *Bureau de Cadastre* and was involved in creating conversion tables between the fraction and the decimal representations of rational numbers. Among other things, he created the sequence \mathcal{F}_{99} consisting of 3003 fractions.

To discuss these results, the following notion is important.

Definition 3.5.2. The **mediant** of two reduced fractions $\frac{a}{b}$ and $\frac{c}{d}$ with $b, d > 0$ is defined as $\frac{a+c}{b+d}$.

Example 3.5.1. The mediant of $\frac{1}{2}$ and $\frac{2}{3}$ is $\frac{1+2}{2+3} = \frac{3}{5}$ and the one of $\frac{1}{2}$ and $\frac{2}{5}$ is $\frac{1+2}{2+5} = \frac{3}{7}$. Lastly, the mediant of $\frac{2}{3}$ and $\frac{2}{5}$ is $\frac{2+2}{3+5} = \frac{4}{8} = \frac{1}{2}$.

¹²A detailed description of Haros's work can be found in the book *A Motif in Mathematics* by Scott Guthery [13].

The next result contains a key property of the mediant. We give a more general result that applies to all positive rational numbers, not just reduced ones.

Proposition 3.5.1. *If $\frac{a}{b} < \frac{c}{d}$ are two positive rational numbers, then*

$$(3.5.1) \quad \frac{a}{b} < \frac{a+c}{b+d} < \frac{c}{d}.$$

Proof: From $\frac{a}{b} < \frac{c}{d}$, we get that $bc > ad$. Thus, $a(b+d) = ab+ad < ab+bc = b(a+c)$, implying that $\frac{a}{b} < \frac{a+c}{b+d}$. Also, $c(b+d) = bc+cd > ad+cd = d(a+c)$, meaning that $\frac{a+c}{b+d} < \frac{c}{d}$. ■

The notion of mediant goes back to the ancient Greeks. For example, Proposition VII.12 of Euclid's Elements states that if $\frac{a}{b} = \frac{c}{d}$, then $\frac{a}{b} = \frac{a+c}{b+d} = \frac{c}{d}$ (assuming that all denominators are nonzero, of course). We give a geometric illustration of the mediant in Figure 3.5.2. The ratio $\frac{a}{b}$ is the slope of the line OM and the fraction $\frac{c}{d}$ is the slope of the line ON . The point P is the fourth corner of the parallelogram containing the points O, M , and N . The mediant $\frac{a+c}{b+d}$ is the slope of OP . The diagonal OP is sandwiched between the sides OM and ON and this is exactly the meaning of Proposition 3.5.1.

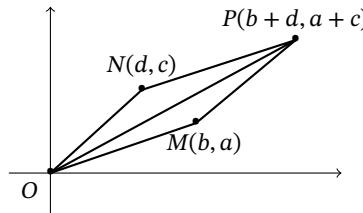


Figure 3.5.2. A geometric interpretation of Proposition 3.5.1.

Example 3.5.2. The reader will likely have noticed this fact already, but we will point it out as well: the mediant of two positive fractions is unrelated to their sum and is always smaller than the sum. For example, the mediant of $\frac{1}{2}$ and $\frac{1}{3}$ is $\frac{2}{5}$ while their sum is $\frac{1}{2} + \frac{1}{3} = \frac{5}{6} > \frac{2}{5}$.

Another French mathematician, Nicholas Chuquet (1445–1488) wrote about the mediant between 1480 and 1484, obtained Proposition 3.5.2, calling it the *régule des nombres moyens* (the rule of intermediate numbers) and used it to find rational approximations of certain irrational numbers.

Example 3.5.3. To illustrate this method for approximating $\sqrt{2}$, for example, one would start with some approximation of $\sqrt{2}$, say $1/1 < \sqrt{2} < 2/1$. By Proposition 3.5.1, the mediant $3/2$ is between $1/1$ and $2/1$. Next, one would compare $3/2$ with $\sqrt{2}$, note that $1/1 < \sqrt{2} < 3/2$ and then repeat. Compute the mediant $4/3$, check that $4/3 < \sqrt{2} < 3/2$ and so on. At each step, the interval containing $\sqrt{2}$ would become smaller and smaller. We leave to the motivated reader to verify that in ten steps, one would get $41/29 < \sqrt{2} < 99/70$.

In his letter, Farey describes his observation that in several sequences \mathcal{F}_n that he has checked (giving as examples \mathcal{F}_5 and a small part of \mathcal{F}_{99}), each fraction equals the mediant of its predecessor and of its successor, and asks whether anyone has observed or proved this result for any \mathcal{F}_n . Before we prove it in Proposition 3.5.4, let us deduce some simpler properties of the sequence \mathcal{F}_n .

Proposition 3.5.2. *Let $n \geq 2$. If $\frac{a}{b} < \frac{c}{d}$ are consecutive fractions in the Farey sequence \mathcal{F}_n , then $b+d > n$ and $b \neq d$.*

Proof: The inequality $b+d > n$ follows from the inequality (3.5.1). For the second part, we use proof by contradiction. Assume that $b = d$ and that $\frac{a}{b}$ and $\frac{c}{b}$ are consecutive fractions in \mathcal{F}_n . We must have that $b \leq n$ and $c = a + 1$. From the first part, we get that $b+b > n \geq 2$ implying that $b \geq 2$. Consider the fraction $\frac{a}{b-1}$ or the reduced form of it. Clearly, $\frac{a}{b} < \frac{a}{b-1}$ and a short calculation yields that $\frac{a}{b-1} < \frac{a+1}{b}$ which contradicts our assumption that $\frac{a}{b}$ and $\frac{c}{b}$ were consecutive. ■

Before we answer to Farey's assertion regarding the mediants, we prove the following important property of consecutive fractions in Farey sequences.

Proposition 3.5.3. *Let n be a natural number. If $\frac{a}{b} < \frac{c}{d}$ are consecutive fractions in the Farey sequence \mathcal{F}_n , then*

$$bc - ad = 1.$$

Proof: We will give a recipe for finding the successor of $\frac{a}{b}$, namely the fraction immediately following $\frac{a}{b}$ in \mathcal{F}_n . Consider the equation

$$(3.5.2) \quad bx - ay = 1,$$

in integer variables x and y . Because $\gcd(a, b) = 1$, we know that this equation has infinitely many solutions (see Proposition 2.3.2). Let (x_0, y_0) be an arbitrary solution of (3.5.2). For any integer k , the pair $(x_0 + ka, y_0 + kb)$ is a solution of (3.5.2) because

$$b(x_0 + ak) - a(y_0 + bk) = bx_0 - ay_0 = 1.$$

Actually, any solution of the equation (3.5.2) is of this form. Choose k such that $0 \leq n - b < y_0 + kb \leq n$. Denote $a' = x_0 + ka$ and $b' = y_0 + kb$. Because $ba' - ab' = 1$, we get that $\gcd(a', b') = 1$ and $a', b' > 0$. Hence, $\frac{a'}{b'}$ is a reduced fraction.

We claim that $\frac{a'}{b'}$ is the successor of $\frac{a}{b}$ in the Farey sequence \mathcal{F}_n . We will show that $\frac{a}{b} < \frac{a'}{b'}$ and that there is no other fraction $\frac{a''}{b''}$ in \mathcal{F}_n such that $\frac{a}{b} < \frac{a''}{b''} < \frac{a'}{b'}$.

Note that $n - b < b' \leq n$ from the definition of b' . Because $ba' - ab' = 1$, we deduce that

$$(3.5.3) \quad \frac{a'}{b'} - \frac{a}{b} = \frac{1}{bb'}.$$

Because $b, b' > 0$, we get that $\frac{a}{b} > 0$. Also, $b \leq n$ and $\frac{a}{b} < 1 - \frac{1}{n}$ (as $\frac{a}{b}$ is not the last term in \mathcal{F}_n). We deduce that

$$\begin{aligned}\frac{a'}{b'} &= \frac{a}{b} + \frac{1}{bb'} \leq 1 - \frac{1}{n} + \frac{1}{b(n-b+1)} \\ &= 1 - \frac{b(n-b+1)-n}{nb(n-b+1)} = 1 - \frac{(n-b)(b-1)}{nb(n-b+1)} \leq 1.\end{aligned}$$

Hence, $\frac{a'}{b'}$ is a member of \mathcal{F}_n .

All it remains to be proved is that there is no reduced fraction $\frac{a''}{b''}$ between $\frac{a}{b}$ and $\frac{a'}{b'}$. We prove this part by contradiction. Assume that there is $\frac{a''}{b''}$ in \mathcal{F}_n such that

$$\frac{a}{b} < \frac{a''}{b''} < \frac{a'}{b'}.$$

These inequalities imply that

$$a''b - ab'' \geq 1 \text{ and } a'b'' - a''b' \geq 1.$$

Therefore, we have that

$$\frac{a'}{b'} - \frac{a}{b} = \left(\frac{a'}{b'} - \frac{a''}{b''} \right) + \left(\frac{a''}{b''} - \frac{a}{b} \right) = \frac{a'b'' - a''b'}{b'b''} + \frac{a''b - ab''}{bb''} \geq \frac{1}{b'b''} + \frac{1}{bb''} = \frac{b+b'}{bb'b''}.$$

Because $n - b < b'$, the inequality above implies that $\frac{a}{b} - \frac{a'}{b'} > \frac{n}{bb'b''}$. On the other hand, equation (3.5.3) gives us that $\frac{a}{b} - \frac{a'}{b'} = \frac{1}{bb'}$. Hence, $\frac{1}{bb'} > \frac{n}{bb'b''}$ which means that $b'' > n$, contradiction with the definition of the Farey sequence \mathcal{F}_n . ■

Example 3.5.4. What is the successor of the fraction $\frac{7}{10}$ in \mathcal{F}_{26} ? As in the previous proof, consider the integer equation

$$10x - 7y = 1.$$

Knowing your multiples of 7 helps, and we can figure out the solution $(5, 7)$ without too much trouble. For any integer k , $(5 + 7k, 7 + 10k)$ is also a solution of the equation above. We are looking for a k such that $n - b = 26 - 10 = 16 < 7 + 10k \leq n = 26$. The only k that works is $k = 1$ and the corresponding solution is $(12, 17)$. Hence, the successor of $\frac{7}{10}$ in \mathcal{F}_{26} is $\frac{12}{17}$.

We now give the proof of a general result answering Farey's question in the affirmative.

Proposition 3.5.4. Let n be a natural number. If $\frac{a}{b} < \frac{e}{f} < \frac{c}{d}$ are three consecutive terms of the Farey sequence \mathcal{F}_n , then

$$\frac{e}{f} = \frac{a+c}{b+d}.$$

Proof: From the previous proposition, we get that

$$be - af = 1 \text{ and } cf - de = 1.$$

Using these equations, we can calculate

$$\begin{aligned} e(bc - ad) &= c(be) - a(de) = c(af + 1) - a(cf - 1) = acf + a - acf + c \\ &= a + c, \end{aligned}$$

and

$$\begin{aligned} f(bc - ad) &= b(fc) - d(af) = b(de + 1) - d(be - 1) = bde + b - bde + d \\ &= b + d. \end{aligned}$$

These last two equations imply that $\frac{e}{f} = \frac{a+c}{b+d}$. ■

Example 3.5.5. What is the predecessor of the fraction $\frac{7}{10}$ in \mathcal{F}_{26} ? By Proposition 3.5.3, if $\frac{a}{b}$ is the predecessor of $\frac{7}{10}$, then we must have that

$$7b - 10a = 1 \quad \text{and} \quad 1 \leq a, b \leq 26.$$

It is not hard to see $(2, 3)$ is a solution and therefore, for any integer k , $(2 + 7k, 3 + 10k)$ is a solution of the equation above. The condition $1 \leq a, b \leq 26$ limits our possible solutions to $(2, 3), (9, 13), (16, 23)$. Which one do we choose? We can check that $2/3 < 9/13 < 16/23$, and thus, $16/23$ is the predecessor of $7/10$ in \mathcal{F}_{26} . Using Example 3.5.4, the fractions $16/23 < 7/10 < 12/17$ are consecutive in \mathcal{F}_{26} .

We mention now Haros's construction of \mathcal{F}_{99} . He started with the sequence

$$1/99, 1/98, 1/97, \dots, 1/4, 1/3, 1/2, 2/3, 3/4, 5/6, \dots, 96/97, 97/98, 98/99,$$

and, for any two consecutive fractions, he inserted their mediant between them, as long as, in its reduced form, its denominator was at most 99.

Farey sequences can be used to approximate irrational numbers by rational numbers with *small* denominators. The following result is due to the German mathematician Peter Gustav Lejeune Dirichlet (1805–1859).

Theorem 3.5.5. *Let α be an irrational number. For any natural number $n \in \mathbb{N}$, there exists a rational number $\frac{u_n}{v_n}$ such that*

$$(3.5.4) \quad \left| \alpha - \frac{u_n}{v_n} \right| \leq \frac{1}{v_n(n+1)} \quad \text{and} \quad v_n \leq n.$$

Proof: Because $|\alpha| \in \mathbb{Z}$ and $0 \leq \alpha - |\alpha| < 1$, without loss of generality, we may assume that $\alpha \in (0, 1)$. Let $n \in \mathbb{N}$. There exist two consecutive fractions $\frac{a_n}{b_n}$ and $\frac{c_n}{d_n}$ in the Farey sequence \mathcal{F}_n such that $\frac{a_n}{b_n} < \alpha < \frac{c_n}{d_n}$. The mediant $\frac{e_n}{f_n} = \frac{a_n+c_n}{b_n+d_n}$ also lies between $\frac{a_n}{b_n}$ and $\frac{c_n}{d_n}$. Since $\frac{a_n}{b_n}$ and $\frac{c_n}{d_n}$ are consecutive in \mathcal{F}_n , we must have that $b_n + d_n \geq n + 1$. Because $\alpha \notin \mathbb{Q}$, $\alpha \neq \frac{a_n+c_n}{b_n+d_n}$ and therefore,

$$\frac{a_n}{b_n} < \alpha < \frac{e_n}{f_n} \quad \text{or} \quad \frac{e_n}{f_n} < \alpha < \frac{c_n}{d_n}.$$

In the first case,

$$\left| \alpha - \frac{a_n}{b_n} \right| \leq \frac{e_n}{f_n} - \frac{a_n}{b_n} = \frac{1}{f_n b_n} \leq \frac{1}{b_n(n+1)},$$

and we can take $u_n = a_n$ and $v_n = b_n$. In the second case,

$$\left| \alpha - \frac{c_n}{d_n} \right| \leq \frac{e_n}{f_n} - \frac{c_n}{d_n} = \frac{1}{f_n d_n} \leq \frac{1}{d_n(n+1)},$$

and we can take $u_n = c_n$ and $v_n = d_n$. ■

We use the previous result to show that irrational numbers can be approximated well by rational numbers.

Theorem 3.5.6. *Let α be an irrational number. There exist infinitely many rational numbers $\frac{a}{b}$ such that*

$$(3.5.5) \quad \left| \alpha - \frac{a}{b} \right| < \frac{1}{b^2}.$$

Proof: Proposition 3.5.5 shows that for any $n \in \mathbb{N}$, there are $u_n, v_n \in \mathbb{N}$ such that $0 < v_n \leq n$ and $\left| \alpha - \frac{u_n}{v_n} \right| < \frac{1}{v_n(n+1)} < \frac{1}{v_n^2}$. Thus, the rational numbers $\frac{u_n}{v_n}, n \geq 1$ satisfy (3.5.5). The only problem we may run into is if these numbers repeat and take finitely many values. We use proof by contradiction to show that this will not happen and that there are infinitely many values of $\frac{u_n}{v_n}$ for $n \geq 1$. If there were finitely many values of $\frac{u_n}{v_n}$, then there would be finitely many values of $\left| \alpha - \frac{u_n}{v_n} \right|$. One of these values, say $\left| \alpha - \frac{u_k}{v_k} \right|$ for some $k \geq 1$, must be the smallest and is a positive number as α is irrational and $\frac{u_k}{v_k}$ is rational. Because $\left| \alpha - \frac{u_k}{v_k} \right| > 0$, there exists a natural number n such that $\frac{1}{n+1} < \left| \alpha - \frac{u_k}{v_k} \right|$. Using (3.5.4), we deduce that

$$\frac{1}{n+1} < \left| \alpha - \frac{u_k}{v_k} \right| \leq \left| \alpha - \frac{u_n}{v_n} \right| \leq \frac{1}{v_n(n+1)} \leq \frac{1}{n+1},$$

where the last inequality follows from $v_n \geq 1$. This is a contradiction that proves our assertion and the theorem. ■

A famous theorem of the German mathematician Adolf Hurwitz (1859–1919) gives a sharper bound.

Theorem 3.5.7. *For any irrational number α , there are infinitely many rational numbers $\frac{a}{b}$ such that*

$$\left| \alpha - \frac{a}{b} \right| < \frac{1}{\sqrt{5} b^2},$$

The constant $\frac{1}{\sqrt{5}}$ may not be replaced by a smaller number¹³.

It is perhaps surprising, but it turns out that rational numbers cannot be approximated as well as irrational ones.

Proposition 3.5.8. *Let r be a positive rational number. There exist only finitely many rational numbers $\frac{a}{b}$ such that*

$$(3.5.6) \quad \left| r - \frac{a}{b} \right| < \frac{1}{b^2}.$$

¹³Let $\varphi = (1 + \sqrt{5})/2$ be the golden ratio. Then for any constant $A < 1/\sqrt{5}$ the number of rationals $\frac{a}{b}$ with $\left| \varphi - \frac{a}{b} \right| < \frac{A}{b^2}$ is finite. In this sense the golden ratio φ is the irrational number which can be approximated worst by rational numbers. See [25], Theorem 6.12, for a proof.

Proof: Let $r = \frac{c}{d}$ for some $c, d \in \mathbb{N}$. If $\frac{a}{b} \neq \frac{c}{d}$ is a positive rational number with $b \geq d$, then $|ad - bc| \geq 1$ and

$$\left|r - \frac{a}{b}\right| = \left|\frac{c}{d} - \frac{a}{b}\right| = \frac{|ad - bc|}{bd} \geq \frac{1}{b^2}.$$

Hence, if $\frac{a}{b}$ satisfies (3.5.6), then $b < d$ and there are only finitely many such fractions. ■

Example 3.5.6. Consider $r = 1/6$. The previous result states that if a rational number a/b satisfies $|1/6 - a/b| < 1/b^2$, then $b < 6$. When $b = 5$, $1/5$ is the only number that satisfies the previous inequality. We leave to the reader to verify whether there are other such numbers with $b \leq 4$.

We switch gears now and turn to a seemingly unrelated topic: determining areas of certain polygons.

Definition 3.5.3. A point in the Cartesian plane is called a **lattice point** if its coordinates are integers.

Consider the polygon in Figure 3.5.3 whose corners are lattice points. What is its area?

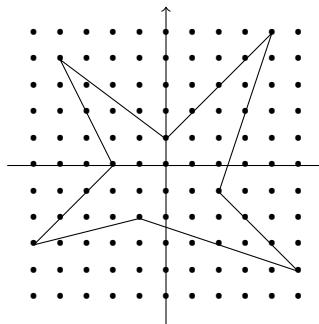


Figure 3.5.3. A polygon whose corners are lattice points.

The following beautiful theorem due to the Austrian mathematician Georg Alexander Pick (1859-1942) gives the answer by counting the number of certain lattice points.

Theorem 3.5.9 (Pick, 1899). *Let P be a polygon whose corners are lattice points. If i is the number of its interior points and b is the number of its border points, then the area of P equals $i + b/2 - 1$.*

We first consider a particular case of the theorem for which we need the following definition.

Definition 3.5.4. A triangle whose corners are lattice points is called **primitive** if it contains no lattice points in its interior and has exactly three lattice points on its boundary, namely its corners.

For a primitive triangle, we have that $i = 0$ and $b = 3$.

Theorem 3.5.10. *If T is a primitive triangle, then its area equals $1/2$.*

Before we give the proof of Theorem 3.5.10, we show how it implies Theorem 3.5.9.

Proof of Pick's Theorem: Triangulate the polygon into primitive triangles using the lattice points from its interior and from its boundary. If the number of triangles is t , then the sum of the angles in all these triangles is $t \cdot 180^\circ$. On the other hand, if we sum up the angles around each interior vertex, we get $i \cdot 360^\circ$ and summing the angles around the boundary points gets us $(b - 2) \cdot 180^\circ$. Hence,

$$t \cdot 180^\circ = i \cdot 360^\circ + (b - 2) \cdot 180^\circ,$$

and therefore, $t = 2i + b - 2$. Since the area of each primitive triangle is $1/2$, this gives the desired formula.

Another way to see that $t = 2i + b - 2$ is by using Euler's formula for planar graphs (Theorem 1.2.10). The triangulated polygon can be regarded as a planar graph. If v is the number of its vertices, e is the number of its edges and t the number of its triangles, Euler's formula implies that $t = e - v + 1$ (since we ignore the infinite face). The number of its vertices equals $i + b$. Each interior edge is contained in two triangles and each edge on the boundary is contained in one triangle. It follows that $2e = 3t + b$. Combining these equations, we get that

$$t = e - v + 1 = \frac{3t + b}{2} - (i + b) + 1 = \frac{3t}{2} - i - b/2 + 1,$$

which implies that $t = 2i + b - 2$. ■

The polygon in Figure 3.5.3 has $b = 18$ and $i = 21$. Therefore, its area is $i+b/2-1 = 21 + 18/2 - 1 = 29$.

In Figure 3.5.4, we illustrate some examples of primitive triangles. For some of them, finding the area is really easy. Each triangle in the second quadrant (negative x -coordinate and positive y -coordinate) has one side parallel to the y -axis and the opposite point at distance one from that side. The area of each such triangle can be computed via *the height times the base over two* formula. How about the triangles from the other quadrants? The two triangles in the first quadrant (both coordinates positive) are congruent and proving that the area of one is $1/2$ would be enough. The triangles in the third quadrant (both coordinates negative) and in the fourth quadrant (the x -coordinate positive and the y -coordinate negative) are all congruent to each other.

In order to prove Theorem 3.5.10, we will need the following result.

Proposition 3.5.11. *Let $M(a, b)$ and $N(c, d)$ be two points in the Cartesian plane whose origin is denoted by O . The area of the triangle OMN equals $|ad - bc|/2$.*

Proof: By translation, rotation, and reflection, we may assume that our triangle OMN is in one of the two situations sketched in Figure 3.5.5. We can calculate the area of the triangle OMN by *framing* this triangle inside a rectangle or inside a right-angled triangle.

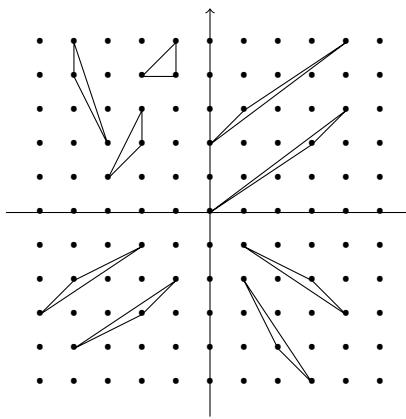


Figure 3.5.4. Some primitive triangles.

In the case on the left,

$$\begin{aligned} \text{area}(OMN) &= \text{area}(OABC) - \text{area}(OAN) - \text{area}(NBM) - \text{area}(MCO) \\ &= cb - \frac{cd}{2} - \frac{(c-a)(b-d)}{2} - \frac{ab}{2} = \frac{2bc - cd - bc - ad + ab + cd - ab}{2} \\ &= \frac{bc - ad}{2}. \end{aligned}$$

For the reader wondering why we need the absolute value in the formula above, note that if we switch the points $M(a, b)$ and $N(c, d)$, then the previous calculations would yield that $\text{area}(ONM) = \frac{ad-bc}{2}$.

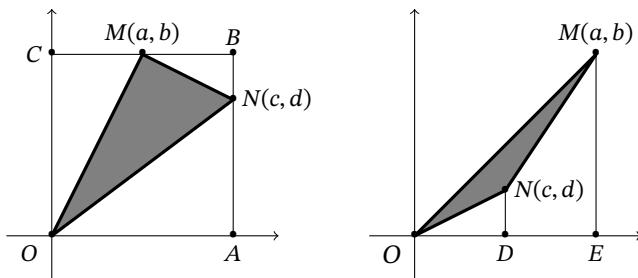


Figure 3.5.5. Two *framed* triangles.

In the case on the right,

$$\begin{aligned} \text{area}(OMN) &= \text{area}(OEM) - \text{area}(ODN) - \text{area}(DEM) \\ &= \frac{ab}{2} - \frac{cd}{2} - \frac{(a-c)(d+b)}{2} = \frac{ab - cd - ad - ab + cd + cb}{2} \\ &= \frac{bc - ad}{2}. \end{aligned}$$

Similarly to the earlier comment, switching the points M and N would give the formula $\text{area}(ONM) = \frac{ad-bc}{2}$. In all cases, $\text{area}(OMN) = |ad - bc|/2$. ■

Note that the previous result applies to any points M and N in the Cartesian plane, not just for lattice points. To complete our proof of Theorem 3.5.10, we will need the following result.

Proposition 3.5.12. *Consider the lattice points $A(m, 0)$, $B(m, n)$ and $C(0, n)$, where m and n are two natural numbers.*

- (1) *The number of lattice points in the interior of the rectangle $OABC$ equals $(m-1)(n-1)$.*
- (2) *The number of lattice points on the segment OB (including its endpoints O and B) equals $\gcd(m, n) + 1$.*
- (3) *The number of lattice points in the interior of the triangle OAB is*

$$\frac{(m-1)(n-1) - (\gcd(m, n) - 1)}{2}.$$

Proof: For the first part, a lattice point in the interior of the rectangle is of the form (x, y) with $1 \leq x \leq m-1$ and $1 \leq y \leq n-1$. Hence, there are $(m-1)(n-1)$ such points.

For the second part, each point (r, s) on the segment OB must satisfy the equation $rn = sm$ which implies that $r \cdot \frac{n}{\gcd(m, n)} = s \cdot \frac{m}{\gcd(m, n)}$. Since $\frac{m}{\gcd(m, n)}$ and $\frac{n}{\gcd(m, n)}$ are coprime, we deduce that $\frac{m}{\gcd(m, n)}$ divides r . Since $0 \leq r \leq m$, we get that $r = \frac{m}{\gcd(m, n)} \cdot t$ for an integer t with $1 \leq t \leq \gcd(m, n)$. This shows that there are $\gcd(m, n) + 1$ lattice points on this segment.

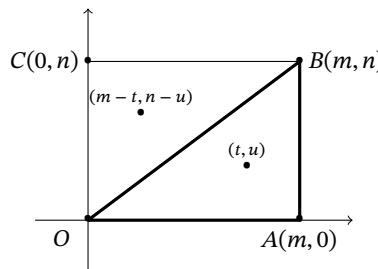


Figure 3.5.6. Counting the lattice points in the triangle OAB .

For the third part, for any lattice point (t, u) in the interior of the triangle OAB , the point $(m-t, n-u)$ is in the interior of the triangle OBC and vice versa. This proves that the number of lattice points in these two triangles must be equal. If we add them up to the number of lattice points in the interior of the segment OB , we get the number of lattice points in the interior of the rectangle $OABC$. Using parts (1) and (2), we deduce that the number of lattice points in the interior of OAB equals $\frac{(m-1)(n-1) - (\gcd(m, n) - 1)}{2}$. ■

When $\gcd(m, n) = 1$, the answer to part (3) is $\frac{(m-1)(n-1)}{2}$. Now we can proceed to giving the proof of Theorem 3.5.10. All we have to prove is that if $(0, 0), (a, b)$, and (c, d) are the corners of a primitive triangle, then $|ad - bc| = 1$. We definitely know that $|ad - bc| \geq 1$ because a, b, c , and d are integers and the three points of coordinates $(0, 0), (a, b)$, and (c, d) are not collinear.

We will use Figure 3.5.7 to show that $|ad - bc| = 1$. For a given polygon P , let $i(P)$ denote the number of lattice points in the interior of T .

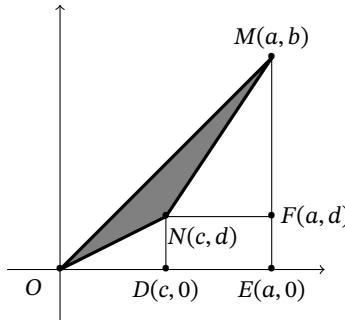


Figure 3.5.7. A primitive triangle.

By Proposition 3.5.12, we have that

$$\begin{aligned} i(OEM) &= \frac{(a-1)(b-1)}{2}, \\ i(ODN) &= \frac{(c-1)(d-1)}{2}, \\ i(NFM) &= \frac{(a-c-1)(b-d-1)}{2}, \\ i(DEFN) &= (a-c-1)(d-1). \end{aligned}$$

Here we used that $\gcd(a, b) = 1$, $\gcd(c, d) = 1$ and $\gcd(a-c, b-d) = 1$ for the triangle computations. Because ONM is a primitive triangle, there are no lattice points in its interior and therefore,

$$i(OEM) = i(OND) + i(NFM) + i(DEFN) + i(ND) + i(NF) + 1,$$

where the 1 on the right-hand side accounts for the point N . Since $i(ND) = d-1$ and $i(NF) = a-c-1$, we get that

$$\frac{(a-1)(b-1)}{2} = \frac{(c-1)(d-1)}{2} + \frac{(a-c-1)(b-d-1)}{2} + (a-c-1)(d-1) + d + a - c - 1.$$

After some algebraic simplifications, the right-hand side is

$$\frac{ab - a - b + ad - bc}{2}.$$

Since the left-hand side is $\frac{ab-a-b+1}{2}$, this means that $ad - bc = 1$ which finishes our proof that any primitive triangle has area $1/2$.

We give now some applications of Pick's formula.

Example 3.5.7. In Exercise 1.1.8, we asked to determine the number of ways of changing one dollar (100 cents) using quarters (25 cents), dimes (10 cents), and nickels (5 cents). This is equivalent to determining the number of nonnegative integer solutions (q, d, n) to the equation:

$$(3.5.7) \quad 25q + 10d + 5n = 100.$$

One way to solve this equation is to list methodically all the solutions. We can observe that the maximum value that q can take is 4 since $q \geq 5$ implies that $100 = 25q + 10d + 5n \geq 25 \cdot 5 = 125$, contradiction. The reader can proceed to list all the solutions in each of the five cases corresponding to $q \in \{0, 1, 2, 3, 4\}$ and count the total number of solutions. This should not be too hard, although a little tedious. One problem we face is how to generalize this result to larger amounts? What if we wanted to change 2 or 10 dollars? This may not be the most practical math problem, especially now that we can pay for parking by card almost everywhere, but it is an interesting problem nevertheless. One also may wonder what do these types of problems have to do with Pick's formula.

To see the connection with Pick's formula, we write the equation (3.5.7) in an equivalent, yet different form

$$25q + 10d \leq 100.$$

For each nonnegative integer pair (q, d) satisfying the equation above, we can get exactly one solution (q, d, n) of the equation (3.5.7) by setting $5n = 100 - 25q - 10d$ or $n = 25 - 5q - 2d$. Conversely, for each solution (q, d, n) of the equation (3.5.7), we get precisely one solution of the equation above. The number of nonnegative integer solutions (q, d) of the inequality above equals the sum of the number i of lattice points in the interior and the number b of lattice points on the boundary of the triangle whose corners have coordinates $(0, 0)$, $(4, 0)$ and $(0, 10)$. By Pick's formula, the area of this triangle equals $i + b/2 - 1$. On the other hand, this is a right-angled triangle, and we can calculate its area as *half of base times height* giving us $4 \cdot 10/2 = 20$. We calculate b as follows. There are 5 lattice points on the horizontal side and 11 lattice points on the vertical side with $(0, 0)$ being counted in both. We leave it to the reader to show that there is exactly one lattice point that is interior to the segment AB . Hence, for the

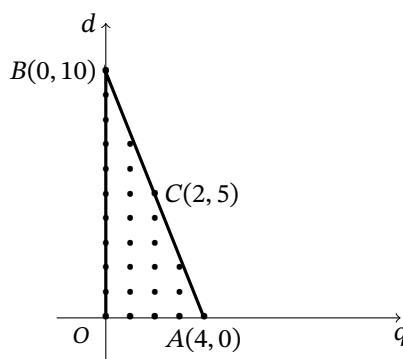


Figure 3.5.8. Changing one dollar with quarters, nickels, and dimes.

triangle OAB in Figure 3.5.8 has exactly $11 + 4 + 1 = 16$ lattice points on the boundary. Using Pick's formula $\text{area}(OAB) = i + b/2 - 1$, we deduce that $20 = i + 16/2 - 1$ which implies that $i = 13$. Hence, there are $i + b = 13 + 16 = 29$ lattice points on the boundary or in the interior of the triangle OAB which means that there are exactly 29 ways to make change for one dollar using quarters, dimes, and nickels.

Another example illustrating Pick's formula comes from baseball.

Example 3.5.8. In this sport, the batting average is defined as the ratio between the number of hits and the number of at-bats rounded up to three decimal points. For example, a player with 4 hits for 13 at-bats will have a ratio of

$$\frac{4}{13} = 0.\overline{307692},$$

which becomes the batting average of 0.308. Of course, this batting average can be attained in different ways such as 153/492 or 77/250 or 40/130. A natural question is in how many ways can one attain such a 0.308 batting average with at most 2000 at-bats?

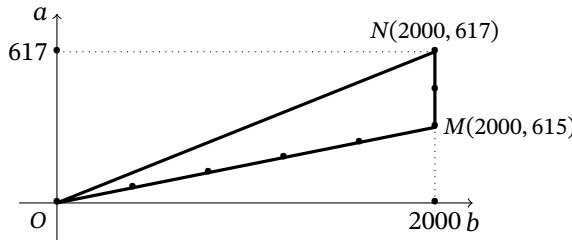


Figure 3.5.9. Attaining 0.308 batting average.

The answer is the number of pairs of nonnegative integers (a, b) (where a is the number of possible hits and b is the number of possible at-bats) such that

$$0.3075 \leq \frac{a}{b} < 0.3085, 1 \leq b \leq 2000.$$

Figure 3.5.9 shows that the number of pairs (a, b) equals the sum of the number of lattice points in the interior of the triangle OMN and the number of lattice points on its boundary (except those lattice points on the closed segment ON). Pick's formula means that $2000 \cdot 2/2 = 2000 = \text{area}(OMN) = i + b/2 - 1$. There are 3 lattice points on the closed segment MN . Because $\gcd(2000, 617) = 1$, there are 2 lattice points on the segment ON . Because $\gcd(615, 2000) = 5$, there are $5 + 1 = 6$ lattice points on the segment OM . These points are of the form

$$((615k)/5, (2000k)/5) = (123k, 400k), \quad 0 \leq k \leq 5.$$

Hence, $b = 8$. The above equation implies that $i = 2001 - b/2 = 1997$. Thus, the answer is $1997 + 5 + 1 = 2003$.

Another consequence of Pick's theorem is the property of consecutive fractions in a Farey sequence, namely Proposition 3.5.3.

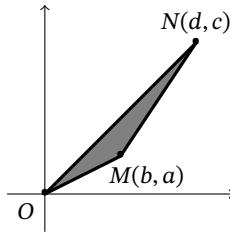


Figure 3.5.10. Pick implies Farey Proposition 3.5.3.

Assume that $\frac{a}{b} < \frac{c}{d}$ are two consecutive fractions in the Farey sequence \mathcal{F}_n for some $n \geq \max(b, d)$ (see Figure 3.5.10). Therefore, $\gcd(a, b) = \gcd(c, d) = 1$ and there is no lattice point in the interior or on the sides of the triangle OMN (except for the points O, M and N , of course) because $\frac{a}{b}$ and $\frac{c}{d}$ are consecutive in \mathcal{F}_n . Thus, the triangle OMN is primitive and by Pick's theorem, its area must be $1/2$. On the other hand, by Proposition 3.5.11, the area of the triangle OMN equals $(bc - ad)/2$. Hence, $bc - ad = 1$, proving Proposition 3.5.3.

We now show that Pick's theorem also implies Proposition 3.5.4¹⁴.

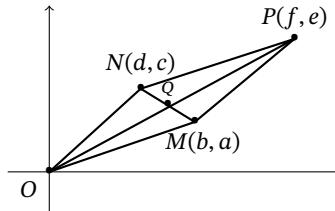


Figure 3.5.11. Pick implies Farey Proposition 3.5.4.

Assume that $\frac{a}{b} < \frac{e}{f} < \frac{c}{d}$ are three irreducible consecutive fractions in \mathcal{F}_n , for some $n \in \mathbb{N}$. Because $\frac{a}{b} < \frac{e}{f}$ are consecutive in \mathcal{F}_n , it means that $be - af = 1$ and the area of the triangle OMP is $1/2$. By a similar argument, the area of the triangle ONP is $1/2$. Because these triangles have the same area and share the side OP , we deduce that the distance from M to OP equals the distance from N to OP . We leave it to the reader to conclude from here that the intersection point Q of the segments OP and MN is the midpoint of MN and therefore its coordinates are $(\frac{b+d}{2}, \frac{a+c}{2})$. Since O, Q , and P are collinear, $\frac{e}{f} = \frac{a+c}{b+d}$, finishing our proof.

Exercise 3.5.1. Let $\frac{a}{b} < \frac{c}{d}$ be two positive rational numbers. If α and β are positive numbers, prove that

$$\frac{a}{b} < \frac{\alpha a + \beta c}{\alpha b + \beta d} < \frac{c}{d}.$$

¹⁴This argument is a geometric version of the proof of Proposition 3.5.4 and appears in the article *A visual approach to some elementary number theory*, The Mathematical Gazette, 1995, by Maxim Bruckheimer and Abraham Arcavi.

Exercise 3.5.2. Let $\frac{a}{b} < \frac{c}{d}$ be two positive rational numbers such that $bc - ad = 1$. Assume that $\frac{e}{f}$ is a rational number such that $\frac{a}{b} < \frac{e}{f} < \frac{c}{d}$. Prove that there exist nonnegative integers α and β such that

$$e = \alpha a + \beta c \text{ and } f = \alpha c + \beta d.$$

Exercise 3.5.3. Write down the Farey sequences \mathcal{F}_8 , \mathcal{F}_9 , and \mathcal{F}_{10} .

Exercise 3.5.4. How many terms are in \mathcal{F}_n for $1 \leq n \leq 10$? How many terms are in the Farey sequence \mathcal{F}_n for $n \in \mathbb{N}$?

Exercise 3.5.5. Determine the predecessor and the successor of $\frac{17}{82}$ in \mathcal{F}_{99} .

Exercise 3.5.6. If Q is a quadrilateral whose vertices are lattice points, prove that its area is at least 1.

Exercise 3.5.7. Determine (with proof) the number of ways to achieve batting average 0.413 with at most 2000 at-bats.

Exercise 3.5.8. If P is a pentagon whose vertices are lattice points, prove that it contains at least one lattice point in its interior. Prove that the area of P is at least $5/2$.

Exercise 3.5.9. What is the number of ways to make change for two dollars using quarters, nickels, and dimes? How about for D dollars where D is a natural number?

Exercise 3.5.10. Let S be a lattice square whose side has length 5. Prove that S has at most 36 and at least 28 lattice points. Give one example for each bound attaining equality.

3.6. Ford Circles and Stern–Brocot Trees

Lester Randolph Ford Sr.(1886–1967) was an American mathematician who introduced a graphical representation of rational number using circles¹⁵. Ford was the editor of the American Mathematical Monthly from 1942 to 1946, and president of the Mathematical Association of America from 1947 to 1948. One of the most prestigious honors of this association, *the Lester R. Ford award*¹⁶, is presented annually to authors of articles of expository excellence published in The American Mathematical Monthly.

Consider two positive reduced fractions $\frac{a}{b} < \frac{c}{d}$. Draw the circle of radius $1/2b^2$ that lies above the x -axis and is tangent to it at the point of coordinates $(\frac{a}{b}, 0)$. Let us call this circle $\mathcal{C}\left(\frac{a}{b}\right)$. Denote by M its center. Do the same for the rational number $\frac{c}{d}$ with N denoting the center of its corresponding circle $\mathcal{C}\left(\frac{c}{d}\right)$.

¹⁵Ford's work is contained in the article Fractions, *American Mathematical Monthly* **45** (1938), 586–601, which is the source for most of the material in this section.

¹⁶Since 2012, this award has been renamed *the Paul Halmos–Lester R. Ford award*.

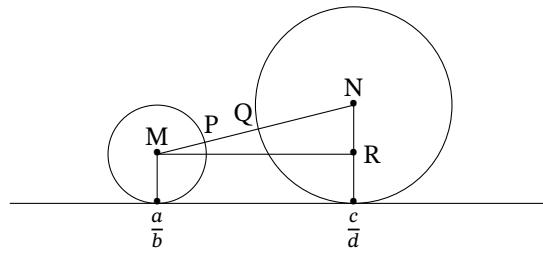


Figure 3.6.1. Two circles tangent to the x -axis.

Because $0 < \frac{a}{b} < \frac{c}{d}$, $bc - ad \geq 1$. The distance MN between the centers of these two circles can be obtained by Pythagoras's theorem:

$$\begin{aligned} MN^2 &= MR^2 + RN^2 = \left(\frac{c}{d} - \frac{a}{b}\right)^2 + \left(\frac{1}{2d^2} - \frac{1}{2b^2}\right)^2 \\ &= \frac{(bc - ad)^2}{b^2 d^2} + \frac{(b^2 - d^2)^2}{4b^2 d^2} \geq \frac{1}{b^2 d^2} + \frac{(b^2 - d^2)^2}{4b^2 d^2} \\ &= \left(\frac{1}{2b^2} + \frac{1}{2d^2}\right)^2 = (MP + NQ)^2. \end{aligned}$$

Hence, $MN \geq MP + NQ$ with equality if and only if $bc - ad = 1$. This means that the two circles above are disjoint when $bc - ad > 1$ and tangent when $bc - ad = 1$. Thus, we have proved the following result.

Proposition 3.6.1. *The Ford circles representing distinct reduced fractions are either tangent or disjoint.*

If $\frac{a}{b}$ and $\frac{c}{d}$ are consecutive fractions in a Farey sequence \mathcal{F}_n , then these circles are tangent. The previous proposition motivated Ford to introduce the following definition.

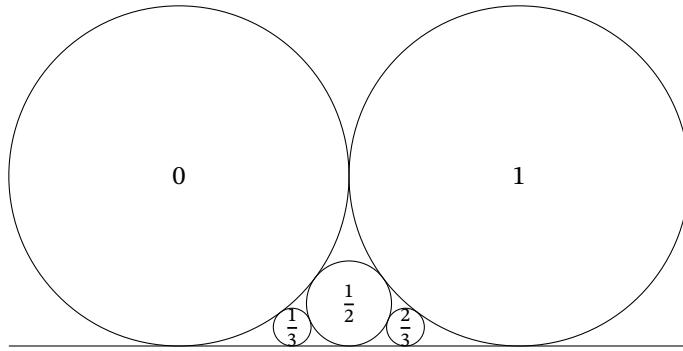


Figure 3.6.2. Ford circles for the Farey sequence \mathcal{F}_3 .

Definition 3.6.1. Two reduced fractions $\frac{a}{b}$ and $\frac{c}{d}$ are **adjacent** if their corresponding circles are tangent.

From the proof of the previous proposition, we can deduce the next result.

Corollary 3.6.2. *The fractions $\frac{a}{b}$ and $\frac{c}{d}$ are adjacent if and only if $|ad - bc| = 1$.*

Proposition 3.6.3. *Let $\frac{a}{b}$ be a reduced fraction.*

- (1) *The fraction $\frac{a}{b}$ has at least one adjacent fraction that is larger than $\frac{a}{b}$ and at least one adjacent fraction that is smaller than $\frac{a}{b}$.*
- (2) *If $\frac{c}{d}$ is a reduced fraction that is adjacent to $\frac{a}{b}$, then all reduced fractions that are adjacent to $\frac{a}{b}$ are of the form*

$$\frac{c_n}{d_n} = \frac{c + na}{d + nb}, n \in \mathbb{Z}.$$

- (3) *For any $n \geq 1$, the circles $C\left(\frac{c_n}{d_n}\right)$ and $C\left(\frac{c_{n+1}}{d_{n+1}}\right)$ are tangent.*

Proof: For the first part, to find an adjacent fraction that is larger than $\frac{a}{b}$, we search for fractions $\frac{c}{d}$ such that $\frac{a}{b} < \frac{c}{d}$ and $|ad - bc| = 1$. This means that $bc - ad = 1$. Since $\gcd(a, b) = 1$, we can use Bézout's lemma (see Proposition 2.3.2) and its proof using the Euclidean algorithm, to find such a fraction. Finding a smaller adjacent fraction can be done similarly and is left as an exercise.

For the second part, assume that $\frac{c}{d}$ is a reduced fraction that is adjacent to $\frac{a}{b}$. Therefore, $|ad - bc| = 1$. Let $n \in \mathbb{N}$. We can verify that

$$|a(d + nb) - b(c + na)| = |ad - bc + anb - anb| = |ad - bc| = 1,$$

which implies that $\frac{c+na}{d+nb}$ is adjacent to $\frac{a}{b}$.

We now show that any fraction that is adjacent to $\frac{a}{b}$ must be of the form $\frac{c+na}{b+nb}$ for some integer n . If $\frac{x}{y}$ is a fraction that is adjacent to $\frac{a}{b}$, we must have that $|ay - bx| = 1$. Since $|ad - bc| = 1$, we have that

$$ay - bx = ad - bc \text{ or } ay - bx = -ad + bc.$$

If $ay - bx = ad - bc$, we get that $a(y - d) = b(x - c)$. Since $\gcd(a, b) = 1$, we get that $a \mid x - c$ implying that $x - c = ak$ for some integer k . Plugging $x - c = ak$ back into $a(y - d) = b(x - c)$, we obtain that $a(y - d) = bak$. Thus, $y = d + bk$ and $\frac{x}{y} = \frac{c+ka}{d+kb}$.

If $ay - bx = -ad + bc$, we get that $a(y + d) = b(x + c)$. Since $\gcd(a, b) = 1$, we get that $a \mid x + c$ implying that $x + c = ak$ for some integer k . Plugging $x + c = ak$ back into $a(y + d) = b(x + c)$, we obtain that $a(y + d) = bak$. Thus, $y = -d + bk$ and $\frac{x}{y} = \frac{-c+ka}{-d+kb} = \frac{c+(-k)a}{d+(-k)b}$, finishing our proof.

For the third, we verify that

$$|c_n d_{n+1} - d_n c_{n+1}| = |(c + na)(d + (n + 1)b) - (c + (n + 1)a)(d + nb)| = |bc - ad| = 1.$$

This finishes our proof. ■

For any integer n ,

$$(3.6.1) \quad \frac{a}{b} - \frac{c + na}{d + nb} = \frac{ad + anb - bc - anb}{b(d + nb)} = \frac{ad - bc}{b(d + nb)}.$$

Since $|ad - bc| = 1$, this equation implies that when n gets large, $\frac{c+na}{d+nb}$ moves closer to $\frac{a}{b}$ from one side, and as n gets smaller, $\frac{c+na}{d+nb}$ approaches $\frac{a}{b}$ from the other side.

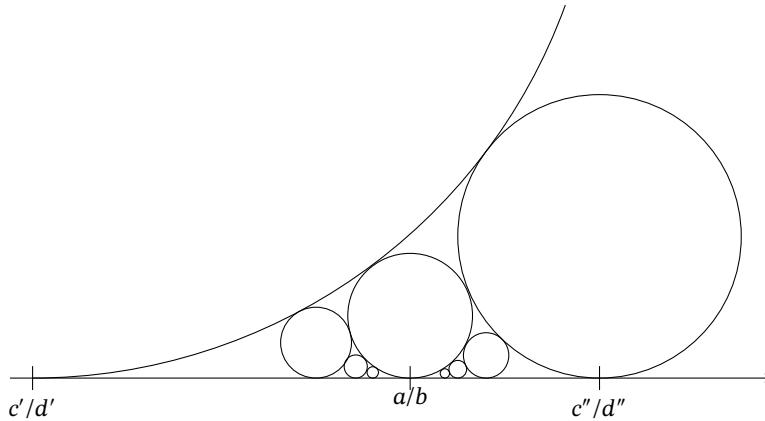


Figure 3.6.3. Illustration of Proposition 3.6.3 and Corollary 3.6.4.

Proposition 3.6.3 has another consequence. See Figure 3.6.3 for an illustration.

Corollary 3.6.4. Let $\frac{a}{b}$ be a reduced fraction with $|b| > 1$. Among all fractions adjacent to $\frac{a}{b}$, there are exactly two fractions whose denominators are smaller than b in absolute value.

Proof: Let $\frac{c}{d}$ be a fraction adjacent to $\frac{a}{b}$. By Proposition 3.6.3, every fraction that is adjacent to $\frac{a}{b}$ is of the form $\frac{c_n}{d_n} = \frac{c+na}{d+nb}$. We are interested in determining those values of n such that $|d + nb| < |b|$. Dividing both sides by $|b|$, we get that $|d/b + n| < 1$. There are exactly two integers n with this property, namely those two integers that lie closest to $-d/b$. They are $n' = \lfloor -d/b \rfloor$ and $n'' = n' + 1$. The corresponding fractions $\{c'/d', c''/d''\} = \{c_{n'}/d_{n'}, c_{n''}/d_{n''}\}$ are illustrated in Figure 3.6.3. From equation (3.6.1), we deduce that the signs of $a/b - c'/d'$ and $a/b - c''/d''$ are opposite and therefore, one of the fractions c'/d' and c''/d'' is smaller than a/b and the other one is larger. ■

Note that the diameter of the circle $C\left(\frac{a}{b}\right)$ is $\frac{1}{b^2}$. Let n be a natural number and β a number between $\frac{1}{(n+1)^2}$ and $\frac{1}{n^2}$. If we draw a horizontal line from $(0, \beta)$ to $(1, \beta)$, then this line will intersect all the circles corresponding to the fractions in the Farey sequence \mathcal{F}_n and no other circles.

Recall from Section 1.2 that a tree is a connected graph without any cycles.

Definition 3.6.2. A **rooted tree** is a pair (T, r) where T is a tree and r is a vertex of T called its **root**.

The distance between two vertices in a connected graph is defined as the shortest length of a path between them. Note that in a tree T , for any two distinct vertices, there is a unique path between them.

Definition 3.6.3. Let (T, r) be a rooted tree. The **level** $\ell(x)$ of a vertex x of (T, r) is defined as the distance between x and the root r .

In some situations, it is convenient to draw a rooted tree by placing the root at the top and the rest of the neighbors below the root, with their position depending on their level (closer vertices to the root being higher on the page). See Figure 3.6.4 for some examples. For the rooted tree on the left, $\ell(x) = 0$ as x is the root and $\ell(y) = \ell(z) = \ell(w) = 1$ for any other vertex. For the rooted tree on the right, $\ell(w) = 0$ as w is the root, $\ell(x) = 1$ and $\ell(y) = \ell(z) = 2$. See the following definition.

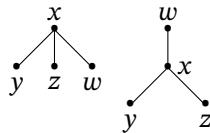


Figure 3.6.4. Same labeled tree rooted at different vertices.

Definition 3.6.4. Let (T, r) be a rooted tree. A vertex u is called an **ancestor** of a vertex v if u is contained in the path between the root r and v . If this happens, we also say that v is a **descendant** of u . If u is the neighbor of v on the path from v to the root r , we say that u is the **parent** of v and that v is the **child** of u .

Example 3.6.1. For the rooted tree in the right of Figure 3.6.4, y is a descendant of w , but is not a child of w . Also, y and z are the children of x .

Definition 3.6.5. A rooted tree is called **binary** if every vertex of it has two or zero descendants. If u is a vertex that has two descendants, the **left-child** of u is its descendant situated to the left of u and the **right-child** of u is the other descendant of u .

Example 3.6.2. The rooted tree in Figure 3.6.5 has a as a root and is binary as each vertex has zero or two descendants. The left-child of the vertex b is d and the right-child of b is e .

Definition 3.6.6. The **height** of a rooted tree (T, r) is the largest level of a vertex. A finite binary rooted tree is called **complete** if for any nonnegative integer k that is less than or equal to its height, the number of vertices at level k is 2^{k-1} .

Example 3.6.3. The tree in Figure 3.6.5 is a complete binary tree of height 2. It is not too hard to prove that a complete binary tree of height n must have exactly $2^n - 1$ vertices.

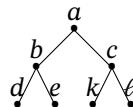


Figure 3.6.5. A binary tree with seven vertices.

Definition 3.6.7. A rooted tree (T, r) is an **infinite complete binary tree** if for any nonnegative integer k , the number of vertices at level k equals 2^{k-1} .

The Stern–Brocot tree is an infinite complete binary tree whose vertices are in bijective correspondence with the positive reduced fractions. It was discovered independently by the German mathematician Moritz Stern (1807–1894) in 1858 and by the French clockmaker Achille Brocot (1817–1878) in 1860. We now sketch the construction of this tree. The root of the tree is $\frac{1}{1}$. It has two ancestors: the left one is $\frac{0}{1}$ and $\frac{1}{0}$ (which we cannot calculate, but should be understood as a symbol for infinity ∞) which are not part of the Stern–Brocot tree, but are used in its construction. The vertex $\frac{0}{1}$ is the leftmost vertex and the vertex $\frac{1}{0}$ the rightmost vertex of the tree.

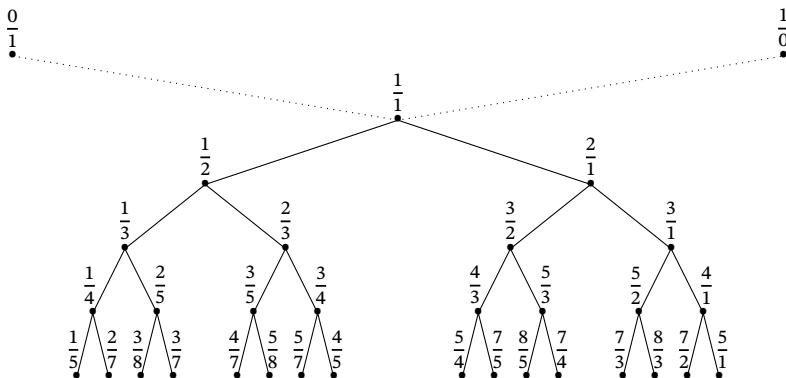


Figure 3.6.6. A few levels in the Stern–Brocot tree.

A left ancestor of a vertex u is an ancestor of u that is to the left of u . The nearest left ancestor of u is the left ancestor of u whose distance to u is the smallest. A right ancestor and the nearest right ancestor can be defined similarly.

Example 3.6.4. In Figure 3.6.5, d has no left ancestors, but has two right ancestors: b and a , while e has b as a left ancestor and a as a right ancestor.

Example 3.6.5. In Figure 3.6.6, the left ancestors of the vertex $\frac{3}{5}$ are $\frac{1}{2}$ and $\frac{0}{1}$. The right ancestors of $\frac{3}{5}$ are $\frac{2}{3}$, $\frac{1}{1}$, and $\frac{1}{0}$.

In the Stern–Brocot tree, the label of a vertex u equals the mediant of the label of its nearest left ancestor v and its nearest right ancestor w and we draw u to the left of v and to the right of w (see Figure 3.6.6 for an illustration).

Example 3.6.6. The left child of $\frac{1}{1}$ is labeled by the mediant $\frac{0+1}{1+1}$ of $\frac{0}{1}$ and $\frac{1}{1}$.

Note that in a rooted tree (T, r) , the set of ancestors of a given vertex u , consists of the vertices of the path from u to r . Thus, for any two ancestors v and w of a given vertex u , either v is an ancestor of w or w is an ancestor of v .

Proposition 3.6.5. *The following statements hold in the Stern–Brocot tree.*

- (1) *Any fraction is reduced.*
- (2) *Let $\frac{a}{b}$ be the label of a vertex in the tree. If $\frac{c}{d}$ is the nearest ancestor of $\frac{a}{b}$, then*

$$ad - bc = \begin{cases} +1, & \text{if } \frac{c}{d} \text{ is a left ancestor of } \frac{a}{b}, \\ -1, & \text{if } \frac{c}{d} \text{ is a right ancestor of } \frac{a}{b}. \end{cases}$$

- (3) *Let $\frac{a}{b}$ be a fraction in the Stern–Brocot tree. The left descendants of $\frac{a}{b}$ are smaller than $\frac{a}{b}$ and the right descendants of $\frac{a}{b}$ are greater than $\frac{a}{b}$.*

Proof: The root has level 0 and with an abuse of the definition, let us just call the vertices $\frac{0}{1}$ and $\frac{1}{0}$ (which are technically not part of the tree) as being at level -1 .

For an integer $k \geq -1$, we prove that the first two statements are true for all vertices of the tree at levels k or less. We use strong induction on k . For the base case, one visually inspect the levels -1 and 0 (the reader can check more if interested) in Figure 3.6.6. For the induction step, let k be a natural number and assume that our statement is true for all the vertices at levels $k - 1$ or less. If $\frac{e}{f}$ is a fraction at level k , then $\frac{e}{f}$ be obtained from the mediant of two fractions $\frac{a}{b}$ and $\frac{c}{d}$ at levels $k - 1$ or less. One of these fractions, say $\frac{a}{b}$, is the parent of $\frac{e}{f}$.

If $\frac{a}{b}$ is a left ancestor of $\frac{e}{f}$, then $\frac{c}{d}$ must be a right ancestor of $\frac{e}{f}$, and is a right ancestor of $\frac{a}{b}$ as well. By the induction hypothesis, $bc - ad = 1$, $\gcd(a, b) = 1$ and $\gcd(c, d) = 1$. We have that

$$\begin{aligned} (a+c)b - (b+d)a &= ab + bc - ab - ad = bc - ad = 1, \\ (a+c)d - (b+d)c &= ad + cd - bc - cd = ad - bc = -1. \end{aligned}$$

These equations imply that $\frac{a+c}{b+d}$ is reduced, confirming property (1). Thus, $e = a + c$ and $f = b + d$. The property (2) is also true for $\frac{a+c}{b+d}$ and its ancestors $\frac{a}{b}$ and $\frac{c}{d}$. Note that in this case, $\frac{a}{b} < \frac{a+c}{b+d} < \frac{c}{d}$.

If $\frac{a}{b}$ is a right ancestor of $\frac{e}{f}$, then $\frac{c}{d}$ must be a left ancestor of $\frac{e}{f}$, and is a left ancestor of $\frac{a}{b}$ as well. By the induction hypothesis, $ad - bc = 1$, $\gcd(a, b) = 1$ and $\gcd(c, d) = 1$. We have that

$$\begin{aligned} (a+c)b - (b+d)a &= ab + bc - ab - ad = bc - ad = -1, \\ (a+c)d - (b+d)c &= ad + cd - bc - cd = ad - bc = 1. \end{aligned}$$

These equations imply that $\frac{a+c}{b+d}$ is reduced, confirming property (1). Thus, $e = a + c$ and $f = b + d$. The property (2) is also true for $\frac{a+c}{b+d}$ and its ancestors $\frac{a}{b}$ and $\frac{c}{d}$. Note that in this case, $\frac{c}{d} < \frac{a+c}{b+d} < \frac{a}{b}$. This finishes our proof.

The proof of the third part is left an exercise. ■

Earlier in this section, we claimed that one can find every positive reduced fraction in this tree. We now prove this assertion.

Proposition 3.6.6. *Every positive reduced fraction appears once in the Stern–Brocot tree.*

Proof: Let $\frac{m}{n}$ be a positive and reduced fraction. To search for it in the Stern–Brocot tree, we start at the root $\frac{1}{1}$ of the tree. If $\frac{m}{n} = \frac{1}{1}$, then we are done. If $\frac{m}{n} < \frac{1}{1}$, then we move to the left child of $\frac{1}{1}$ and we can call this step *L* (left). If $\frac{m}{n} > \frac{1}{1}$, then we move to the right child of $\frac{1}{1}$ and we can call this step *R* (right). We repeat this procedure. Assume that at some point $\frac{a}{b} < \frac{m}{n} < \frac{c}{d}$, where $\frac{a}{b}$ and $\frac{c}{d}$ are the closest ancestors of a new position in the tree. We have that $mb - na \geq 1$, $nc - md \geq 1$ and $bc - ad = 1$. Multiplying the first inequality by d and the second inequality by b and adding up the result inequalities, we get that

$$d(mb - na) + b(nc - md) \geq b + d.$$

The left-hand side equals $n(bc - ad) \geq b + d$ which means that $n \geq b + d$. If $\frac{m}{n}$ equals the mediant $\frac{a+c}{b+d}$, then $m = a + c$, $n = b + d$ and we are done. If $\frac{a}{b} < \frac{m}{n} < \frac{a+c}{b+d}$, we replace $\frac{c}{d}$ by $\frac{a+c}{b+d}$ and repeat the argument. Note that the sum of the new denominators is $2b+d > b+d$. If $\frac{a+c}{b+d} < \frac{m}{n} < \frac{c}{d}$, we replace $\frac{a}{b}$ by $\frac{a+c}{b+d}$ and repeat the argument. Again, the sum of the new denominators is $b+2d > b+d$. Since the sum of the denominators strictly increases in the last two cases, the process cannot continue indefinitely as n is an upper bound for the sum of the denominators. Therefore, we are guaranteed to land in the first case above and every reduced fraction appears in the Stern–Brocot tree. ■

If we use the convention from the previous proof, we can represent each fraction $\frac{a}{b}$ in the Stern–Brocot tree by the sequence/word $w\left(\frac{a}{b}\right)$ of *Ls* and *Rs* that describes the path from the root to the vertex labeled $\frac{a}{b}$. Thus, w is a function whose domain is the set of positive rational numbers and whose codomain consists of the words over the alphabet *L* and *R*. This set of words includes *the empty word* which we denote by *I*. Hence, $w\left(\frac{1}{1}\right) = I$, $w\left(\frac{1}{2}\right) = L$, $w\left(\frac{2}{1}\right) = R$. Continuing this process, we get that

$$w\left(\frac{1}{3}\right) = LL, w\left(\frac{2}{3}\right) = LR, w\left(\frac{3}{2}\right) = RL, w\left(\frac{3}{1}\right) = RR.$$

We do not fully explore this connection in this book (except for some exercises at the end of the section), but the interested reader can read more about it in the wonderful book [11].

We finish this section by briefly describing a connection between the Stern–Brocot tree and the finite continued fractions. When writing the fractions of the first few levels of the Stern–Brocot tree in this form, we obtain Figure 3.6.7.

Recall that every rational number has two finite continued fractions representations: $[a_0; a_1, \dots, a_n] = [a_0; a_1, \dots, a_n - 1, 1]$ if $a_n > 1$ (see Proposition 3.4.3 and Proposition 3.4.4). The children of this node are obtained from the representations above by adding a one to the last entry:

$$[a_0; a_1, \dots, a_n + 1] \text{ and } [a_0; a_1, \dots, a_n - 1, 2].$$

Which child is left and which is right depends on the parity of n (see Exercise 3.4.6 and Exercise 3.4.7).

Exercise 3.6.1. Let $\frac{a}{b} < \frac{c}{d}$ be two positive fractions such that $bc - ad = 1$. Show that the circle $C\left(\frac{a+c}{b+d}\right)$ of the mediant $\frac{a+c}{b+d}$ is tangent to the circles $C\left(\frac{a}{b}\right)$ and $C\left(\frac{c}{d}\right)$.

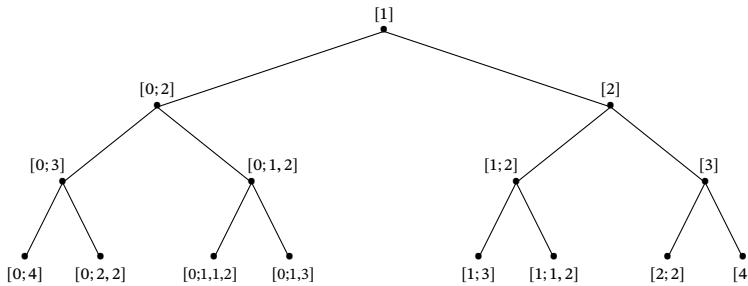


Figure 3.6.7. The first levels as finite continued fractions.

Exercise 3.6.2. What are the next two levels of the tree in Figure 3.6.6?

Exercise 3.6.3. Recall our earlier assumption that $\frac{0}{1}$ and $\frac{1}{0}$ are at the level -1 of the tree and $\frac{1}{1}$ is at level 0. For $k \in \mathbb{N}$, prove that any fraction $\frac{e}{f}$ at the level k of the Stern–Brocot tree must satisfy $e + f \geq k + 2$.

Exercise 3.6.4. For any $n \in \mathbb{N}$, determine the fractions that correspond to the words $\underbrace{L \dots L}_{n \text{ terms}}$ and $\underbrace{R \dots R}_{n \text{ terms}}$.

Exercise 3.6.5. For any $n \in \mathbb{N}$, determine the fractions that correspond to the words $\underbrace{LR \dots LR}_{2n \text{ letters}}$ and $\underbrace{RL \dots RL}_{2n \text{ letters}}$.

Exercise 3.6.6. Determine the LR words corresponding to $7/9, 11/10, 12/17$, and $14/9$ in the Stern–Brocot tree.

Exercise 3.6.7. Let $k \in \mathbb{N}$. Show that if a rational number $[a_0; \dots, a_n]$ is at level k in the Stern–Brocot tree, then $a_0 + \dots + a_n = k + 1$.

Exercise 3.6.8. Let A be the set of rational numbers in the interval $(0, 1)$ and let B denote the set of rational numbers in the interval $(1, +\infty)$. Construct a bijective function from A and B . Is your function related to the Stern–Brocot tree?

Exercise 3.6.9. What are the next two levels of the tree in Figure 3.6.7?

Exercise 3.6.10. Consider a rational number with finite continued fraction representations: $[a_0; a_1, \dots, a_n] = [a_0; a_1, \dots, a_n - 1, 1]$ if $a_n > 1$. Prove that the children of this node in the Stern–Brocot tree are:

$$[a_0; a_1, \dots, a_n + 1] \text{ and } [a_0; a_1, \dots, a_n - 1, 2].$$

3.7. Egyptian Fractions

We start with some puzzles. Assume that we have a supply of ropes and a lighter. Each rope burns in exactly 60 minutes. How do you measure 30 minutes? Since $\frac{30}{60} = \frac{1}{2}$, that is not so hard: light a rope at both ends. When the rope finishes burning, 30 minutes will have passed. How about measuring 45 minutes? Now $\frac{45}{60} = \frac{3}{4} = \frac{1}{2} + \frac{1}{4}$. At the same time, light the first rope at both ends and one end of the second rope. Once 30

minutes have passed, light the other end of the second rope. When the second rope finishes burning, 45 minutes will have gone by.

Writing $\frac{7}{8}$ as

$$\frac{7}{8} = \frac{1}{2} + \frac{1}{4} + \frac{1}{8}$$

has the useful interpretation as a recipe on how to divide 7 units or 7 pizzas for that matter, into 8 equal parts, by dividing each pizza into eights and giving each person a half of a pizza, a quarter of a pizza and an eighth of a pizza.

Definition 3.7.1. A **unit fraction** is a fraction of the form $\frac{1}{n}$, where $n \in \mathbb{N}$.

For some reasons¹⁷, the ancient Egyptians were interested in writing any rational number as sums of unit fractions.

Definition 3.7.2. An Egyptian fraction is a rational number that can be written as a sum of distinct unit fractions:

$$\frac{1}{a_1} + \cdots + \frac{1}{a_k},$$

for some natural number k and natural numbers $a_1 < \dots < a_k$.

Any rational number with numerator one is a unit fraction, so that is easy. Note however that even a unit fraction can be written as a sum of distinct unit fractions in many ways. For example, 1 can be written as

$$1 = \frac{1}{2} + \frac{1}{3} + \frac{1}{6},$$

but also as

$$\begin{aligned} 1 &= \frac{1}{2} + \frac{1}{3} + \frac{1}{12} + \frac{1}{18} + \frac{1}{36} \\ 1 &= \frac{1}{2} + \frac{1}{3} + \frac{1}{12} + \frac{1}{18} + \frac{1}{72} + \frac{1}{108} + \frac{1}{216} \end{aligned}$$

or

$$1 = \frac{1}{6} + \frac{1}{7} + \frac{1}{8} + \frac{1}{9} + \frac{1}{10} + \frac{1}{14} + \frac{1}{15} + \frac{1}{18} + \frac{1}{20} + \frac{1}{24} + \frac{1}{28} + \frac{1}{30}.$$

For a rational number of the form $\frac{2}{n}$, with $n \geq 2$, we can write $\frac{2}{n}$ as $\frac{1}{k}$ when $n = 2k$ is even.

Example 3.7.1. The Rhind Papyrus dating back to around 1650 BCE contains several formulas for fractions of the form $\frac{2}{n}$, when $n \leq 101$ is odd, written as sums of distinct unit fractions. We list below some small calculations:

$$\begin{aligned} \frac{2}{3} &= \frac{1}{2} + \frac{1}{6}, & \frac{2}{5} &= \frac{1}{3} + \frac{1}{15}, & \frac{2}{7} &= \frac{1}{4} + \frac{1}{28}, \\ \frac{2}{9} &= \frac{1}{6} + \frac{1}{18}, & \frac{2}{11} &= \frac{1}{6} + \frac{1}{66}, & \frac{2}{13} &= \frac{1}{7} + \frac{1}{91}. \end{aligned}$$

¹⁷They took a wrong turn according to the French mathematician André Weil (1906–1998).

How would one continue the previous sequence of unit fractions representations for fractions of the form $\frac{2}{n}$ with $n \geq 13$ odd? Is it possible to write any rational number $\frac{m}{n}$ with $m < n$ as a sum of distinct unit fractions? If that is possible, then what is the smallest number of unit fractions in the decomposition of $\frac{m}{n}$?

Theorem 3.7.1. *If $m < n$ are natural numbers, then $\frac{m}{n}$ is an Egyptian fraction.*

Proof: The proof is by strong induction on m and its main idea goes back to Fibonacci's book *Liber Abaci* from 1202. In the base case $m = 1$, there is nothing to prove as $\frac{1}{n}$ is a unit fraction for any n . For the induction step, let $m \geq 2$ be a natural number. Assume that n is a natural number such that $m < n$. If m divides n , then $\frac{m}{n}$ is a unit fraction, and we are done. Otherwise, define a_1 as the natural number such that

$$a_1 > \frac{n}{m} > a_1 - 1.$$

Equivalently, $a_1 - 1 = \lfloor \frac{n}{m} \rfloor$ or $a_1 - 1$ is the quotient of the integer division of n by m . Hence, there is a natural number r such that

$$n = m(a_1 - 1) + r \text{ and } 0 < r < m.$$

Therefore,

$$\frac{1}{a_1} < \frac{m}{n} < \frac{1}{a_1 - 1} \quad \text{and} \quad \frac{m}{n} - \frac{1}{a_1} = \frac{ma_1 - n}{na_1} = \frac{m - r}{na_1}.$$

Now $m - r$ is a natural number less than m and by our induction hypothesis, $\frac{m-r}{na_1}$ must be an Egyptian fraction. The only thing left to show is that $\frac{1}{a_1}$ does not appear in any unit fraction representation of $\frac{m-r}{na_1}$. This happens because

$$\frac{m-r}{na_1} < \frac{n}{na_1} = \frac{1}{a_1}.$$

Hence, combining $\frac{1}{a_1}$ with any distinct unit fraction decomposition of $\frac{ma_1-n}{na_1}$ will give us a distinct unit fraction decomposition of $\frac{m}{n}$. This finishes our proof. ■

We illustrate the mechanism of the previous proof on some concrete examples.

Example 3.7.2. Let us try $\frac{2}{13}$. The first denominator a_1 is defined as

$$a_1 - 1 < \frac{13}{2} < a_1,$$

and thus, $a_1 = 7$. It follows that

$$\frac{2}{13} - \frac{1}{7} = \frac{14 - 13}{91} = \frac{1}{91}.$$

It seems that we got lucky (see Exercise 3.7.5 about that) and we get an Egyptian fraction representation:

$$\frac{2}{13} = \frac{1}{7} + \frac{1}{91}.$$

The following example will require a bit more work.

Example 3.7.3. Consider the fraction $\frac{4}{17}$. For the first step, we define a_1 such that $a_1 - 1 < \frac{17}{4} < a_1$. Hence, $a_1 = 5$. To get the next fraction, we calculate

$$\frac{4}{17} - \frac{1}{5} = \frac{20 - 17}{85} = \frac{3}{85}.$$

Now a_2 is defined as the natural number such that

$$a_2 - 1 < \frac{85}{3} < a_2,$$

which leads to $a_2 = 29$. Therefore,

$$\frac{3}{85} - \frac{1}{29} = \frac{87 - 85}{85 \cdot 29} = \frac{2}{2465}.$$

Since our numerator is not equal to one yet, we must continue the procedure. We take a_3 as the natural number such that

$$a_3 - 1 < \frac{2465}{2} < a_3$$

meaning that $a_3 = 1233$. Next, we calculate

$$\frac{2}{2465} - \frac{1}{1233} = \frac{2466 - 2465}{2465 \cdot 1233} = \frac{1}{3039345}.$$

We finally have arrived at a unit fraction, and we write down the result of our work:

$$\frac{4}{17} = \frac{1}{5} + \frac{1}{29} + \frac{1}{1233} + \frac{1}{3039345}.$$

The readers are encouraged to try their own examples of this procedure. One can obtain another decomposition of m/n into Egyptian fractions by the following procedure. The method is also recursive, and its proof is by induction on m . We sketch the details and give a concrete illustration of it below.

If $m = 1$, then we are done. Otherwise, because $\gcd(m, n) = 1$, there exist integers a and b such that $ma = nb + 1$. This follows from Proposition 2.3.2. If (a_0, b_0) is an arbitrary solution of this equation, then the set of solutions consists of all pairs $(a_0 + nk, b_0 + mk)$ for $k \in \mathbb{Z}$. Therefore, there exists a solution (a, b) such that $0 < a < n$ and $0 < b < m$. Dividing both sides of the equation $ma = nb + 1$ by na , we deduce that

$$\frac{m}{n} = \frac{b}{a} + \frac{1}{na}.$$

The fraction $\frac{1}{na}$ will be part of our decomposition, and we can repeat our procedure above.

Example 3.7.4. Let us try the algorithm above on the fraction $\frac{4}{17}$. First, we need integers $0 < a < 17$ and $0 < b < 4$ such that $4a = 17b + 1$. This is the same as a being the inverse of 4 modulo 17. Such integers can be found using the Euclidean algorithm, and it is not hard to see that $a = 13$ and $b = 3$. From the equation $4 \cdot 13 = 17 \cdot 3 + 1$, dividing both sides by $13 \cdot 17$, we get that

$$\frac{4}{17} = \frac{3}{13} + \frac{1}{13 \cdot 17}.$$

We keep $\frac{1}{13 \cdot 17} = \frac{1}{221}$ in our decomposition and repeat the procedure for $\frac{3}{13}$. We need integers $0 < c < 13$ and $0 < d < 3$ such that $3c = 13d + 1$. Again, it is not hard to

figure out that $c = 9$ and $d = 2$. From the identity $3 \cdot 9 = 13 \cdot 2 + 1$, dividing both sides by $9 \cdot 13$, we get that

$$\frac{3}{13} = \frac{2}{9} + \frac{1}{9 \cdot 13}.$$

We keep $\frac{1}{9 \cdot 13} = \frac{1}{91}$ and we deal with $\frac{2}{9}$ next. It is not hard to see that $2 \cdot 5 = 9 \cdot 1 + 1$ and therefore,

$$\frac{2}{9} = \frac{1}{5} + \frac{1}{45}.$$

Hence, we obtain the following decomposition into Egyptian fractions:

$$\frac{4}{17} = \frac{1}{5} + \frac{1}{45} + \frac{1}{91} + \frac{1}{221}.$$

This is a different decomposition of $\frac{4}{17}$ than the one obtained earlier. It also has four terms, but the denominators seem smaller. Actually, it can be proved that any denominator in this procedure is at most $n(n - 1)$.

As hinted a bit earlier in this section, it is not too hard to show that each positive rational number less than one, has infinitely many representations as an Egyptian fraction (see Exercise 3.7.6). A natural question is to determine the smallest number of distinct unit fractions in such representations.

Despite their apparent simplicity, there are still many unsolved problems involving the Egyptian fractions. One of the most famous ones is following conjecture due to the Hungarian mathematician Paul Erdős (1913–1996) and the German-American mathematician Ernst Straus (1922–1983).

Conjecture 3.7.2 (Erdős–Straus 1948). *For any natural number $n \geq 2$, there exist natural numbers a, b and c such that*

$$\frac{4}{n} = \frac{1}{a} + \frac{1}{b} + \frac{1}{c}.$$

This conjecture has been verified for $n \leq 10^{17}$, but is still open for general n . A similar open conjecture is due to the Polish mathematicians Andrzej Schinzel (1937–2021) and Waclaw Sierpiński (1882–1969).

Conjecture 3.7.3 (Schinzel–Sierpiński 1956). *For any natural number $n \geq 5$, there exist natural numbers x, y and z such that*

$$\frac{5}{n} = \frac{1}{x} + \frac{1}{y} + \frac{1}{z}.$$

Exercise 3.7.1. Let n be a natural number. Prove that

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

Exercise 3.7.2. Find Egyptian fraction representations of

$$\frac{4}{5}, \frac{4}{7}, \frac{4}{9}, \frac{4}{11}, \frac{4}{13}, \frac{4}{15}.$$

Exercise 3.7.3. Find Egyptian fraction representations of

$$\frac{5}{6}, \frac{5}{7}, \frac{5}{8}, \frac{5}{9}, \frac{5}{11}, \frac{5}{12}, \frac{5}{13}.$$

Exercise 3.7.4. Find Egyptian fraction representations of the fractions in Exercise 3.7.2 and Exercise 3.7.3 with three terms.

Exercise 3.7.5. Let $n \geq 3$ be a natural number. Show that the algorithm in the proof of Theorem 3.7.1 produces an Egyptian fraction representation of $\frac{2}{n}$ with two terms.

Exercise 3.7.6. Let $m < n$ be two natural numbers. Show that $\frac{m}{n}$ has infinitely many representations as an Egyptian fraction.

Exercise 3.7.7. Let $2 \leq x < y < z$ be natural numbers such that

$$1 = \frac{1}{x} + \frac{1}{y} + \frac{1}{z}.$$

Find x, y , and z .

Exercise 3.7.8. Let $2 \leq x < y < z$ be natural numbers such that

$$\frac{2}{3} = \frac{1}{x} + \frac{1}{y} + \frac{1}{z}.$$

Find x, y , and z .

Exercise 3.7.9. Let $2 \leq m < n$ be two natural numbers. Show that the algorithm described in the proof of Theorem 3.7.1 produces an Egyptian fraction representation of $\frac{m}{n}$ with at most m distinct unit fractions.

Exercise 3.7.10. Let $2 \leq x < y < z$ be natural numbers such that

$$\frac{3}{4} = \frac{1}{x} + \frac{1}{y} + \frac{1}{z}.$$

Find x, y , and z .

3.8. More Exercises

Exercise 3.8.1. Determine with proofs which ones of the following numbers are rational:

$$\sqrt{\sqrt{5+2}}, \sqrt{3+2\sqrt{2}}, \text{ and } \sqrt{3+2\sqrt{2}} - \sqrt{3-2\sqrt{2}}.$$

Exercise 3.8.2. Prove that $\sqrt{2}, \sqrt{3}$ and $\sqrt{5}$ cannot be terms of the same arithmetic progression.

Exercise 3.8.3. Determine with proofs which ones of the following numbers are rational:

$$\log_5(9), \log_2(6) - \log_2(3), \log_3(5) + \log_3(25).$$

Exercise 3.8.4. Prove or disprove the following statement. If a, b, c, d, A, B, C, D are natural numbers such that

$$\frac{a}{b} < \frac{A}{B} \text{ and } \frac{c}{d} < \frac{C}{D},$$

then

$$\frac{a+c}{b+d} < \frac{A+C}{B+D}.$$

Exercise 3.8.5. Let a, b, c, d, m , and n be natural numbers such that $\frac{a}{b} < \frac{m}{n} < \frac{c}{d}$ and $bc - ad = 1$. Show that $m \geq a + c$ and $n \geq b + d$. Is this result true if $bc - ad > 1$?

Exercise 3.8.6. Let a and b be two natural numbers. Prove that there exists some natural number k that does not depend on a and b such that

$$\left| \sqrt{5} - \frac{a}{b} \right| \geq \frac{1}{kb^2}.$$

Exercise 3.8.7. Determine which of the numbers $\sin 0^\circ$, $\sin 10^\circ$, $\cos 20^\circ$, $\cos 40^\circ$ are rational.

Exercise 3.8.8. Let n be a natural number. What are the predecessor and the successor of $1/2$ in the Farey sequence \mathcal{F}_n ?

Exercise 3.8.9. Let n be a natural number. What are the first two terms and the last two terms of the Farey sequence \mathcal{F}_n ?

Exercise 3.8.10. Let n be a natural number. If $\frac{a}{b} < \frac{c}{d}$ are two fractions in the Farey sequence \mathcal{F}_n such that $bc - ad = 1$, does it mean that $\frac{a}{b}$ and $\frac{c}{d}$ are consecutive in \mathcal{F}_n ? Prove it or give a counterexample.

Exercise 3.8.11. Let $\frac{a}{b} < \frac{a'}{b'}$ be two consecutive terms in the Farey sequence \mathcal{F}_n for some $n \geq \max(b, b')$. Prove that

$$\frac{1}{n(n-1)} \leq \frac{a'}{b'} - \frac{a}{b} \leq \frac{1}{n}.$$

Exercise 3.8.12. Let n be a natural number. Assume that the elements of \mathcal{F}_n are:

$$\frac{a_1}{b_1} < \frac{a_2}{b_2} < \dots < \frac{a_k}{b_k}.$$

Prove that

$$\sum_{j=1}^{k-1} \frac{1}{b_j b_{j+1}} = 1.$$

Exercise 3.8.13. Let S be a lattice square whose side has length 13. What are the maximum and the minimum number of lattice points that S may contain? Give one example for each bound attaining equality.

Exercise 3.8.14. Pick five points in a square whose side has length one. Show that there are two points among them at distance at most $1/\sqrt{2}$.

Exercise 3.8.15. Pick five points in an equilateral triangle whose side has length one. Show that there are two points among them at distance at most $1/2$.

Exercise 3.8.16. Prove that there is no equilateral triangle whose vertices are lattice points. How about a square or regular pentagon or a regular hexagon?

Exercise 3.8.17 (Problem B1, Putnam contest 1993). Find the smallest positive integer n such that for every integer m , with $0 < m < 1993$, there exists an integer k such that

$$\frac{m}{1993} < \frac{k}{n} < \frac{m+1}{1994}.$$

Exercise 3.8.18. Let a and b be two coprime natural numbers. Show that

$$\left\lfloor \frac{a}{b} \right\rfloor + \left\lfloor \frac{2a}{b} \right\rfloor + \dots + \left\lfloor \frac{(b-1)a}{b} \right\rfloor = \frac{(a-1)(b-1)}{2}.$$

Exercise 3.8.19. Let p be a prime. If a and b are natural numbers such that $\frac{3}{p} = \frac{1}{a} + \frac{1}{b}$, show that $p \equiv 2 \pmod{3}$.

Exercise 3.8.20. Let n be a natural number such that $n \equiv 2 \pmod{3}$. Show that there exist natural numbers a and b such that $\frac{3}{n} = \frac{1}{a} + \frac{1}{b}$.

Exercise 3.8.21. Consider the rooted tree whose root is labeled $1/1$ and constructed as follows. If a node is labeled by the irreducible fraction a/b , its left child is labeled by $a/(a+b)$ and its right child is labeled by $(a+b)/b$.

- (1) Draw the first four levels of this tree.
- (2) Prove that every positive rational number appears exactly once in this tree¹⁸.

¹⁸This is called the Calkin-Wilf tree after the American mathematicians Neil Calkin (Clemson University, SC) and Herbert Wilf (1931–2012) (see [3]).

Real Numbers \mathbb{R}

When you have mastered numbers, you will in fact no longer be reading numbers, any more than you read words when reading books You will be reading meanings.

Harold Geneen

4.1. Basic Properties

The question *What is a real number?* occupied humans over centuries. Natural numbers, integers, and rational numbers already had occurred several hundred and thousands years ago, and we discussed their properties in our previous chapters. In the study of these numbers, it became clear that they are not sufficient to describe all phenomena occurring in nature. As shown in Figure 4.1.1, a square whose sides have length one, has the property that its diagonal d satisfies the equation $d^2 = 2$. Drawing a circle centered at the origin with radius d , we obtain a point on the number line and therefore, a number, that corresponds to d . As proved in Theorem 3.2.1, d is not a rational number. This was known to the Pythagoreans in the 6th century BCE (and likely, before that) and caused quite a shock in their ranks. Nowadays, we use $\sqrt{2}$ to denote d . The oldest known rational approximation of this number $\sqrt{2} \approx 1.41421296$, was found on a Babylonian clay table dating back between 1600 and 1800 BCE.

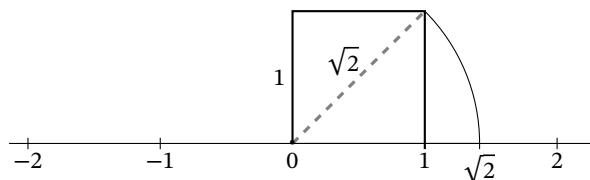


Figure 4.1.1. A square of length one creates $\sqrt{2}$.

In the previous chapter, we saw other examples of irrational numbers such as some trigonometric values and logarithms. For other numbers, figuring out if they are rational or not can be a difficult task. Take for example, the number π which is the ratio between the circumference or the length of a circle of radius one and its diameter¹. To get an intuition on why the ratio between the circumference and the radius of any circle is the same, we go back to the ancient Greeks who saw a circle of radius r as the *culmination* or *pinnacle* of regular polygons with n sides inscribed in a circle of the same radius r , when n grows. For given $n \geq 3$, the ratio between the perimeter of such polygon to the diameter $2r$ of the circle is the same for any r (see Exercise 4.1.1). The perimeter of the polygon can be made arbitrarily close to the circumference of the circle as n grows and therefore, the ratio between the circumference of a circle and the diameter $2r$ is the same for any $r > 0$. The name π (derived from the first Greek letter of the word *periphery*) was first used by the Welsh mathematician William Jones (1675–1749) in 1709 and became popular after its use by Leonhard Euler in his book *Mechanica* in 1736.

Archimedes (287–212 BCE) approximated π by using the perimeters of regular inscribed and circumscribed polygons. The idea is to start with an equilateral triangle inscribed in the circle of radius one, then subdivide each arc into two equal parts to obtain a regular hexagon, and then repeat the procedure. See Figure 4.1.2 for an illustration of the first three steps. The perimeter of each regular polygon will be a lower bound for the length of the circle that is a better approximation than the perimeter of the polygon used in the previous step.

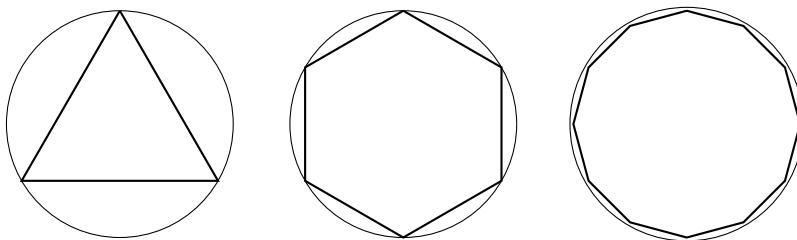


Figure 4.1.2. Approximating the length of a circle from below.

This procedure yields lower bounds for π as follows. If we use a regular polygon with $n \geq 3$ sides inscribed in a circle of radius one, then consider two consecutive nodes of the polygon A and B as in Figure 4.1.3.

The length of the arc from A to B is larger than the length of the segment AB which equals $2 \sin \theta$, where $\theta = 180^\circ/n$. The circumference of the circle is at least $2n \sin \theta$ leading to the lower bound

$$\pi > n \sin\left(\frac{180^\circ}{n}\right).$$

When $n = 3$, $\theta = 60^\circ$, $\sin \theta = \sqrt{3}/2$, and $\pi > \frac{3\sqrt{3}}{2}$, which perhaps is not too impressive. If $n = 6$, we have $\theta = 30^\circ$ which leads to $\pi > 3$. When $n = 12$, we

¹The book [1] is an exciting source for the history of the number π .

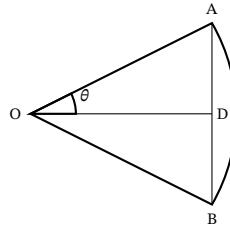


Figure 4.1.3. Approximating the length of the unit circle.

get that $\pi > 12 \sin(15^\circ)$. We can determine $\sin(15^\circ)$ from the trigonometric identity $\sqrt{3}/2 = \cos(30^\circ) = 1 - 2 \sin^2(15^\circ)$ which leads to

$$\pi > 6 \cdot \sqrt{2 - \sqrt{3}} \approx 3.10583,$$

another improvement over the previous estimate using hexagons. Using circumscribed circles and a similar doubling argument, one can get upper bounds for the length of the circle (see Exercise 4.1.6). Archimedes used 96-regular polygons and obtained the following estimates for the number π :

$$3.1408450 \approx 3 + \frac{10}{71} < \pi < 3 + \frac{10}{70} \approx 3.1428571.$$

Archimedes used the *pinnacle* idea mentioned earlier to show that the area of a circle of radius r equals πr^2 . This is called the method of exhaustion and is usually attributed to another Greek mathematician Eudoxus (408–355 BCE). This idea is a precursor of the notion of a limit which we will discuss formally in the next chapter. Informally, its use can be summed up by stating that the difference between the area of a circle of radius r and the area of a regular polygon with n sides inscribed in this circle, can be made arbitrarily small and positive by taking n large enough. From the other direction, the difference between the area of a circumscribed regular polygon with n sides and the area of the circle can also be made arbitrarily small and positive for n sufficiently large. These ideas led Archimedes to prove the following result.

Proposition 4.1.1. *Let r be a positive real number. The area of a circle of radius r equals the area of the right-angle triangle where the two small sides have length r and the perimeter of the circle, respectively. The area of the circle is πr^2 .*

Proof: Let C be the area of the circle of radius r . Denote by $P = 2\pi r$ the perimeter or circumference of the circle of radius r and let T be the area of the right-angle triangle with smaller sides of length r and P , respectively. We want to show that $C = T$. The proof is by contradiction.

Assume that $C > T$. Thus, $C - T > 0$. By the method of exhaustion, there is a regular polygon with n sides inscribed in the circle such that the difference between C and the area of the polygon (denoted by $A_n(r)$ in Exercise 4.1.4) is less than $C - T$. Hence, $C - A_n(r) < C - T$ and therefore, $A_n(r) > T$. Now $T = \frac{rP}{2}$ and $A_n(r) = \frac{h_n I_n(r)}{2}$ (see Exercise 4.1.1 and Exercise 4.1.4). Hence, $h_n I_n(r) > rP$. However, $h_n < r$

and $I_n(r) < P$ because the polygon is inscribed in the circle. Thus, $h_n I_n(r) < rP$, contradiction to the previous inequality.

Assume that $C < T$. Hence, $T - C > 0$. By the method of exhaustion, there is a regular polygon with m sides circumscribed to the circle such that the difference between its area (denoted $B_m(r)$ in Exercise 4.1.5) and C is less than $T - C$. Thus, $B_m(r) - C < T - C$ leading to $B_m(r) < T$. With the notation from Exercise 4.1.2, this implies that $\frac{rC_n(r)}{2} < \frac{rP}{2}$ and therefore, $C_n(r) < P$. However, the polygon is circumscribed to the circle and $C_n(r) > P$, in contradiction with the previous inequality. The only possibility is that $C = T$ and therefore, $C = \frac{\pi r^2}{2}$. \blacksquare

The first proof that π is irrational, was given by the Swiss mathematician Johann Heinrich Lambert (1728–1777) in 1760, and is a bit more involved than the proofs for square roots, logarithms or trig functions. See [23] or [24, Chapter 2] for an elementary calculus-based proof of this fact.

To get a first informal idea of how real numbers can be defined, we recall the main facts about the representation of rational numbers as presented in Section 3.3, where we showed that for each **rational number** α , there exist unique integers $a \in \mathbb{Z}$ and a_1, a_2, \dots in $\{0, \dots, 9\}$ such that for all $n \in \mathbb{N}$,

$$(4.1.1) \quad a + \frac{a_1}{10^1} + \cdots + \frac{a_n}{10^n} \leq \alpha < a + \frac{a_1}{10^1} + \cdots + \frac{a_n}{10^n} + \frac{1}{10^n}.$$

We wrote that $\alpha = a.a_1a_2\dots$ in this case. Moreover, as proved in Section 3.3 (see also the summary at page 191), the digits a_1, a_2, \dots in the expansion of $\alpha \in \mathbb{Q}$ occur periodically, meaning that there exist integers $m \geq 0$ and $k \geq 1$ such that

$$\alpha = a.a_1 \dots a_m \overline{a_{m+1} \dots a_{m+k}}.$$

Example 4.1.1. Let us consider the sequence a_1, a_2, \dots with

$$a_j = \begin{cases} 1 & : \text{ if } j = 2^k \text{ for some } k \geq 1 \\ 0 & : \text{ otherwise} \end{cases}$$

That is, the sequence equals $0, 1, 0, 1, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 1, 0, \dots$. This is definitely not a periodic sequence. Hence, by the previous observation, there is **no** rational number α such that $\alpha = 0.a_1a_2,\dots$ or, equivalently, there is **no** $\alpha \in \mathbb{Q}$ such that for all $n \geq 1$,

$$\frac{a_1}{10^1} + \cdots + \frac{a_n}{10^n} \leq \alpha < \frac{a_1}{10^1} + \cdots + \frac{a_n}{10^n} + \frac{1}{10^n}.$$

In particular, there is no rational number α such that

$$\alpha \in I_n = [0.a_1 \dots a_n, 0.a_1 \dots a_n 9] \text{ for any } n \geq 1.$$

The sequence of intervals $(I_n)_{n \geq 1}$ of rational numbers is nested, meaning that we have $I_n \supseteq I_{n+1}$ for every n . Moreover, the length of I_n becomes arbitrarily small as n grows. So one expects that they eventually *shrink* to a single point. However, by our construction, there is no $\alpha \in \mathbb{Q}$ belonging to all I_n . Therefore, the nested intervals *shrink* to the empty set or, equivalently, they *shrink* to a hole inside the set \mathbb{Q} . Imagine a funnel whose diameter becomes (in dependence of its length) arbitrarily small. In the end, there should be exactly one single point in the funnel. But here the funnel is empty.

Example 4.1.2. We described in the introduction the geometric interpretation of $\sqrt{2}$. An algebraic way to define a number whose square equals 2 would be to follow the approach indicated in Example 3.3.5. If we proceed as in that example, then we will find rational numbers $a_n < b_n$ such that $b_n - a_n = 10^{-n}$ and $a_n^2 < 2 < b_n^2$ for any $n \geq 1$. In Example 3.3.5 we got

$$a_1 = 1.4, \quad b_1 = 1.5, \quad a_2 = 1.41, \quad b_2 = 1.42 \quad \text{till} \quad a_5 = 1.41421 \quad \text{and} \quad b_5 = 1.41422.$$

As in the previous example, we have a sequence of closed intervals $I_n = [a_n, b_n]$ that are nested: $I_1 \supseteq I_2 \supseteq \dots$. Moreover, the length of I_n is 10^{-n} which becomes arbitrarily small as n grows. These intervals *get smaller and smaller* and *shrink* to a single point, which we define as $\sqrt{2}$. Note that if we do the same procedure within the set of rational numbers, then these intervals *shrink* to the empty set as there is no rational number which belongs to all the I_n .

Suppose now we want to complete \mathbb{Q} and to fill such holes in \mathbb{Q} . To this end we have to add objects (called irrational numbers) which satisfy (4.1.1) for all $n \geq 1$. We will do this later in Section 4.7. For now, for a first rough idea about real numbers, the reader should think of them as numbers which admit an infinite expansion $a.a_1a_2\dots$ with $a \in \mathbb{Z}$ and arbitrary, not necessarily periodic, integers $0 \leq a_j \leq 9$ for $j \geq 1$.

But we should not conceal that then the real unpleasant problems start. For example, if we say real numbers are those α which may be written as $\alpha = a.a_1a_2\dots$, how to define addition or multiplication of two of those objects, how do we order them? If we are able to solve these problems successfully, do these operations satisfy the rules we learned at school?

Many more of those technical problems occur. Nevertheless, for better understanding, always think of real numbers as objects which satisfy (4.1.1) for $a \in \mathbb{Z}$ and for certain, not necessarily periodic, integers $a.a_1, a_2, \dots$ with $0 \leq a_j \leq 9$.

$$\alpha \text{ real number} \Leftrightarrow \alpha = a.a_1a_2\dots \quad a \in \mathbb{Z}, \quad 0 \leq a_j \leq 9.$$

Already this short introduction about the definition of real numbers shows that its construction is by far not an easy task. Therefore, it is not too surprising that it took a very long time before mathematicians gave a solid description of real numbers. The essential progress occurred only in the second half of the 19th century. At this time three different theories of real numbers were developed by several German mathematicians: one by Richard Dedekind (1831–1916), another one by Georg Cantor (1845–1918) and Eduard Heine (1821–1881) and, finally, a third approach by Karl Weierstrass (1815–1897). In each of these cases, the authors assumed that the rational numbers are well-understood, and they used these numbers as a starting point for their theory².

Nowadays, there exist two main approaches to introduce the field \mathbb{R} of real numbers, a constructive and an axiomatic one. The **constructive** approach proceeds as follows. The starting point stems from the rational numbers which are completed by appropriate techniques. For instance, one possibility³ for the construction of \mathbb{R} , the set of real numbers, is to deal with so-called Dedekind cuts. Another related way to extend

²More about the history of this topic can be found in the interesting article [31]. We also refer to the classical book [7] for an overview about the different possibilities for the construction of the field of real numbers.

³Those who want to know more about this technique are referred to Subsection A.8.1.

the field of rational numbers is to add missing least upper bounds of certain subsets in \mathbb{Q} (cf. Section 4.4 below). Other techniques involve adding Cauchy sequences of rational numbers (cf. Remark 5.5.3) or defining real numbers as limits of monotone bounded sequences (cf. Section 5.3). As indicated earlier, real numbers may also be constructed by investigating nested closed intervals whose lengths tend to zero (cf. Proposition 5.4.7). Finally, one may interpret real numbers as infinite decimal or binary fractions (cf. Theorem 4.7.3 and/or Proposition 5.6.19). Every time one gets an extension (completion) of \mathbb{Q} and all these models are known to describe the same object up to an isomorphic transformation. That is, there exist bijections between these models which preserve addition, multiplication and order. Unfortunately, the rigorous approach to all these constructions is quite technical and cumbersome; this is by no means a topic suited for beginners in mathematics.

The second approach is the **axiomatic** one. One extracts the basic structural properties of the field of real numbers and puts them as axioms at the beginning of the theory. The great advantage of this approach is that it reveals in a very precise way which properties of the real numbers are essential and which are not, that is, which can be derived from the stated axioms. In other words, the axioms tell you everything you really need to know without worrying about the details of the model of the real numbers. The drawback of this approach is that if you do not construct a mathematical model satisfying the desired axioms, then you do not tell the full story. Theoretically one could work with a set of axioms which is not satisfied by any concrete object.

Our approach for the introduction of the real numbers will be the axiomatic one. Besides that, at several positions we will indicate how a concrete completion of \mathbb{Q} can be performed, mostly without giving all the technical details.

There are three types of axioms describing the set \mathbb{R} of real numbers:

- (I) Algebraic Properties (II) Order Properties (III) Completeness

We will describe these axioms in the following sections.

Exercise 4.1.1. Let $I_n(r)$ denote the perimeter of a regular polygon with $n \geq 3$ sides that is inscribed in a circle of radius r . Prove that $\frac{I_n(r)}{I_n(s)} = \frac{2r}{2s} = \frac{r}{s}$ for any natural number $n \geq 3$ and real numbers $r, s > 0$.

Exercise 4.1.2. Let $C_n(r)$ denote the perimeter of a regular polygon with $n \geq 3$ sides that is circumscribed to a circle of radius r . Prove that $\frac{C_n(r)}{C_n(s)} = \frac{2r}{2s} = \frac{r}{s}$ for any $n \geq 3$ and $r, s > 0$.

Exercise 4.1.3. With the notations from Exercise 4.1.1 and Exercise 4.1.2, prove that for any $n \geq 3$ and any $r > 0$,

$$I_{2n}(r) = \sqrt{I_n(r)C_{2n}(r)} \quad \text{and} \quad C_{2n}(r) = \frac{2I_n(r)C_n(r)}{I_n(r) + C_n(r)}.$$

Exercise 4.1.4. Let $A_n(r)$ denote the area of a regular polygon with $n \geq 3$ sides that is inscribed in a circle of radius r .

- (1) Show that $A_n(r) = \frac{h_n I_n(r)}{2}$, where h_n is the *height* of the polygon (the distance from the center of the circle to one of the sides of the polygon) and $I_n(r)$ is the perimeter of the polygon (as in Exercise 4.1.1).
- (2) Prove that $\frac{A_n(r)}{A_n(s)} = \frac{r^2}{s^2}$ for any $n \geq 3$ and $r, s > 0$.

Exercise 4.1.5. Let $B_n(r)$ denote the area of a regular polygon with $n \geq 3$ sides that is circumscribed to a circle of radius r .

- (1) Show that $B_n(r) = \frac{r C_n(r)}{2}$, where $C_n(r)$ is the perimeter of the polygon (as in Exercise 4.1.2).
- (2) Prove that $\frac{B_n(r)}{B_n(s)} = \frac{r^2}{s^2}$ for any $n \geq 3$ and $r, s > 0$.

Exercise 4.1.6. Consider a regular polygon with $n \geq 3$ sides that is circumscribed around a circle of radius one. Show that the length of the side of the polygon equals $2 \tan\left(\frac{180^\circ}{n}\right)$. How does this lead to the upper bound $\pi < n \tan\left(\frac{180^\circ}{n}\right)$?

Exercise 4.1.7. Start with inscribed and circumscribed squares to a circle of radius one. What estimates can you get for π ? What happens when you double the number of sides?

Exercise 4.1.8. Prove that $\frac{265}{153} < \sqrt{3} < \frac{1351}{780}$.

Exercise 4.1.9. Using the notation from Section 3.4, for each $n \in \mathbb{N}_0$, define

$$r_n = [1; \underbrace{2, \dots, 2}_n].$$

Prove that

- (1) $r_{2n} < r_{2n+2}$, for each $n \geq 0$,
- (2) $r_{2n+1} < r_{2n-1}$ for each $n \geq 1$,
- (3) $r_{2n+1} < \sqrt{2} < r_{2n}$ for each $n \geq 0$.

Exercise 4.1.10. Using the notation from Section 3.4, for each $n \in \mathbb{N}_0$, define

$$s_n = [2; \underbrace{4, \dots, 4}_n].$$

Prove that

- (1) $s_{2n} < s_{2n+2}$, for each $n \geq 0$,
- (2) $s_{2n+1} < s_{2n-1}$ for each $n \geq 1$,
- (3) $s_{2n+1} < \sqrt{5} < s_{2n}$ for each $n \geq 0$.

Exercise 4.1.11. Let x and a be positive real numbers. Show that

$$x + \frac{a}{2\sqrt{x^2 + a}} < \sqrt{x^2 + a} < x + \frac{a}{2x}.$$

Use this formula to approximate $\sqrt{10}$.

4.2. The Real Numbers Form a Field

A basic property of the real numbers is the existence of two algebraic operations, called addition and multiplication. These two operations denoted by “+” and “ \times ” (or by \cdot) are both mappings from $\mathbb{R} \times \mathbb{R}$ to \mathbb{R} and they possess exactly the same properties as they do on \mathbb{Z}_p or \mathbb{Q} , respectively, and as we were taught them at school.

Thus, our **first axiom** about \mathbb{R} is as follows: The collection of real numbers \mathbb{R} is a set endowed with two binary operations denoted by + and \times or \cdot , such that the following properties (I a), (I b), and (I c) are satisfied.

(I a) The addition of two real numbers is associative and commutative and there exists a neutral element denoted by 0, i.e.,

$$(\forall x \in \mathbb{R})(x + 0 = 0 + x = x).$$

Moreover, each $x \in \mathbb{R}$ has an (additive) inverse element, denoted by $-x$. That is

$$(\forall x \in \mathbb{R})(x + (-x) = (-x) + x = 0).$$

(I b) The multiplication of two real numbers is associative and commutative, and there exists a neutral element denoted by 1, i.e.,

$$(4.2.1) \quad (\forall x \in \mathbb{R})(x \cdot 1 = 1 \cdot x = x).$$

Moreover, each nonzero real number possesses a (unique) inverse element, denoted by x^{-1} or $1/x$, such that

$$(\forall x \in \mathbb{R}, x \neq 0)(x \cdot x^{-1} = x^{-1} \cdot x = 1).$$

(I c) Addition and multiplication are linked by the distributive law asserting that

$$(\forall x, y, z \in \mathbb{R})(x \cdot (y + z) = x \cdot y + x \cdot z).$$

Another way to express the validity of (I a), (I b), and (I c) is as follows:

The real numbers form a **field**.

Conventions and Notation:

(1) Since the operations + and \cdot are assumed to be associative, brackets which show the order of the application of these operations are not needed. For all x, y , and z in \mathbb{R} ,

$$\begin{aligned} x + (y + z) &= (x + y) + z = x + y + z \quad \text{and} \\ x \cdot (y \cdot z) &= (x \cdot y) \cdot z = x \cdot y \cdot z. \end{aligned}$$

Furthermore, if there is no risk of confusion, let us write xy instead of $x \cdot y$.

(2) Subtraction is defined as follows:

$$x - y = x + (-y).$$

In other words, $x - y$ is the unique number in \mathbb{R} satisfying $(x - y) + y = x$.

If $y \neq 0$, the division of x by y may be introduced as $x/y = x y^{-1}$. Hence, x/y is completely described by $(x/y) \cdot y = x$.

$$\begin{aligned} (\forall x, y \in \mathbb{R})(x - y = x + (-y)) \quad &\text{and} \\ (\forall x, y \in \mathbb{R}, y \neq 0)(x/y = x \cdot y^{-1}). \end{aligned}$$

(3) Given $x \in \mathbb{R}$ and $n \geq 1$, the n th power of x and of x^{-1} are defined by

$$x^n = \underbrace{x \cdots x}_{n \text{ times}} \quad \text{and by} \quad x^{-n} = \underbrace{x^{-1} \cdots x^{-1}}_{n \text{ times}} = \underbrace{(1/x) \cdots (1/x)}_{n \text{ times}}.$$

Summary: The real numbers form an algebraic object called a field, where one may add, subtract, multiply, and divide (by nonzero element) in the same manner as one learned in school. We were taught these operations as if it were properties of the numbers. But these are not their properties; these are the rules for combining real numbers, an important difference.

Let us state and prove some well-known properties of addition and multiplication in \mathbb{R} . Most of them are easy to verify, but one has to be very careful to use only facts stated in properties (I a), (I b), and (I c) above, but nothing else.

Proposition 4.2.1. *Addition and multiplication in \mathbb{R} have the following properties.*

- (a₁) $(\forall x \in \mathbb{R})(x \cdot 0 = 0 \cdot x = 0)$ (a₂) $(x \cdot y = 0) \Rightarrow (x = 0 \text{ or } y = 0)$
- (b₁) $(\forall x, y \in \mathbb{R})((-x)y = x(-y) = -xy)$
- (b₂) $(\forall x, y \in \mathbb{R})((-x)(-y) = xy)$ (c) $(\forall x \in \mathbb{R} \setminus \{0\})\left(\frac{1}{-x} = -\frac{1}{x}\right)$
- (d) $(\forall x \in \mathbb{R} \setminus \{0\})\left(x^{-n} = \frac{1}{x^n}\right)$.

Proof: (a) Since $x \cdot 0 = x \cdot (0 + 0) = x \cdot 0 + x \cdot 0$, one gets $x \cdot 0 = 0$ as asserted. Suppose that $x \cdot y = 0$ and assume $x \neq 0$. Then x^{-1} exists, and therefore,

$$y = x^{-1} x y = x^{-1} \cdot 0 = 0.$$

The same argument applies assuming $y \neq 0$. Hence, at least one of x or y has to be zero.

(b) From

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

we get that $x \cdot (-y) = -xy$. Similarly, it follows that $(-x) \cdot y = -xy$.

A direct application is $(-1)x = -1x = -x$. Furthermore, a two times application of the previous formula implies

$$(-x)(-y) = -x(-y) = -(-xy) = xy.$$

(c) We have to show that

$$(4.2.2) \quad x^{-1} + (-x)^{-1} = 0.$$

Note that

$$(-x)(x^{-1} + (-x)^{-1}) = (-x)x^{-1} + 1 = -x x^{-1} + 1 = -1 + 1 = 0.$$

Because $-x \neq 0$, by property (a) we derive (4.2.2) as asserted.

(d) Using distributivity and associativity of multiplication, we have that

$$x^n \cdot x^{-n} = \underbrace{x \cdots x}_n \cdot \underbrace{x^{-1} \cdots x^{-1}}_n = \underbrace{(x x^{-1}) \cdots (x x^{-1})}_n = 1.$$

This finishes our proof. ■

Exercise 4.2.1. Prove that the field properties imply that for all $x, y \in \mathbb{R}$ and all $n, m \in \mathbb{Z}$

$$x^n \cdot x^m = x^{n+m}, \quad (x^n)^m = x^{nm} \quad \text{and} \quad (x \cdot y)^n = x^n \cdot y^n.$$

Exercise 4.2.2. Let x and y be two real numbers such that $x + y = 10$ and $xy = 20$. Calculate $x^2 + y^2$, $x^3 + y^3$ and $x^4 + y^4$.

Exercise 4.2.3. Show that properties (I a), (I b), and (I c) imply the validity of the binomial formula

$$(x + y)^n = \sum_{k=0}^n \binom{n}{k} x^k y^{n-k}, \quad \forall x, y \in \mathbb{R}, \quad \forall n \in \mathbb{N}.$$

Exercise 4.2.4. Deduce from properties (I a), (I b), and (I c) the following identity:

$$x^n - y^n = (x - y)(x^{n-1} + x^{n-2}y + \cdots + xy^{n-2} + y^{n-1}), \quad \forall x, y \in \mathbb{R}, \quad \forall n \in \mathbb{N}.$$

Exercise 4.2.5. Let $n \in \mathbb{N}$ and $x, y \in \mathbb{R}$. Prove that

$$x^{2n+1} + y^{2n+1} = (x + y)(x^{2n} - x^{2n-1}y + \cdots - xy^{2n-1} + y^{2n}).$$

Exercise 4.2.6. Prove the summation formula for the sum of geometric progressions:

$$1 + x + \cdots + x^n = \frac{x^{n+1} - 1}{x - 1}, \quad x \neq 1, n \in \mathbb{N}_0,$$

from the field properties of \mathbb{R} . Here 1 denotes the neutral element satisfying (4.2.1).

Exercise 4.2.7. Show that

$$(a^2 + b^2)(x^2 + y^2) = (ax + by)^2 + (ay - bx)^2,$$

for any real numbers a, b, x, y .

Exercise 4.2.8. Let n be a natural number. If a_1, \dots, a_n and b_1, \dots, b_n are real numbers, prove that

$$\left(\sum_{j=1}^n a_j^2 \right) \left(\sum_{k=1}^n b_k^2 \right) - \left(\sum_{\ell=1}^n a_\ell b_\ell \right)^2 = \sum_{1 \leq j < k \leq n} (a_j b_k - a_k b_j)^2.$$

Exercise 4.2.9. Let a_1, a_2, a_3, a_4 and b_1, b_2, b_3, b_4 be real numbers. Show that

$$\begin{aligned} & (a_1^2 + a_2^2 + a_3^2 + a_4^2)(b_1^2 + b_2^2 + b_3^2 + b_4^2) \\ &= (a_1 b_1 - a_2 b_2 - a_3 b_3 - a_4 b_4)^2 + (a_1 b_2 + a_2 b_1 + a_3 b_4 - a_4 b_3)^2 \\ & \quad + (a_1 b_3 - a_2 b_4 + a_3 b_1 + a_4 b_2)^2 + (a_1 b_4 + a_2 b_3 - a_3 b_2 + a_4 b_1)^2. \end{aligned}$$

Exercise 4.2.10. Let x_1, x_2, x_3 , and x_4 be real numbers. Show that

$$(x_1 - x_2)(x_3 - x_4) + (x_1 - x_3)(x_2 - x_4) + (x_1 - x_4)(x_3 - x_2) = 0.$$

4.3. Order and Absolute Value

Another important characteristic of \mathbb{R} is the fact that we can order real numbers. Moreover, this order is compatible with the existing algebraic operations. For example, we know that if x is less than y , then this is also so for $2x$ and $2y$ or for $x + 2$ and $y + 2$. Such an order also exists on \mathbb{Q} . On the other hand, it is impossible to order the field \mathbb{Z}_p in a way such that $a < b$ always implies $a + c < b + c$. The abstract notation for fields with a compatible order is that of an ordered field. In this context, \mathbb{Q} and \mathbb{R} are ordered fields, but \mathbb{Z}_p is not.

Which way is the best to describe the basic properties of the (natural) order on the set \mathbb{R} ? Already every child knows that 5 sweets are more (better) than 2 of them. We may express this by saying that the difference $5 - 2$ is positive⁴. So we see that the order on \mathbb{R} is completely determined by the set \mathbb{R}_+ of positive elements.

Thus, our **second axiom** about \mathbb{R} is as follows:

(II) There is a subset $\mathbb{R}_+ \subset \mathbb{R}$ (called the set or the cone⁵ of positive elements) satisfying the following:

- (1) $(\forall x, y \in \mathbb{R}_+)(x + y \in \mathbb{R}_+)$
- (4.3.1) (2) $(\forall x, y \in \mathbb{R}_+)(x \cdot y \in \mathbb{R}_+)$
- (3) $(\forall x \in \mathbb{R})(\text{either } x \in \mathbb{R}_+ \text{ or } -x \in \mathbb{R}_+ \text{ or } x = 0).$

Remark 4.3.1. Letting $\mathbb{R}_- := \{x \in \mathbb{R} : -x \in \mathbb{R}_+\}$, property (3) above may be reformulated as follows: \mathbb{R} is the disjoint union of the following three sets:

$$\mathbb{R} = \mathbb{R}_+ \cup \mathbb{R}_- \cup \{0\}.$$

The properties of \mathbb{R}_+ are quite natural and agree completely with what we learned in school about positive numbers. The sum and product of positive numbers are again positive. If elements in \mathbb{R}_- are called negative, then a number is either positive, negative or zero. Also, a number cannot be positive and negative at the same time, and 0 is neither positive nor negative.

How does \mathbb{R}_+ lead to an order on \mathbb{R} ? Set

$$x < y \quad \text{if} \quad y - x \in \mathbb{R}_+.$$

Furthermore,

$$x \leq y \quad \text{if either} \quad x < y \quad \text{or} \quad x = y.$$

Instead of $x < y$ we also write $y > x$ and $x \leq y$ is equivalent to $y \geq x$.

$$x < y \Leftrightarrow y - x \in \mathbb{R}_+ \quad \text{and} \quad x \leq y \Leftrightarrow (x < y \text{ or } x = y).$$

In this notation,

$$\mathbb{R}_+ = \{x \in \mathbb{R} : x > 0\} \quad \text{and} \quad \mathbb{R}_- = \{x \in \mathbb{R} : x < 0\}.$$

⁴When one has 5 sweets, a positive number of sweets remains after eating two of them.

⁵There is a more general notion of a cone and in higher dimensions, the geometric representation of it resembles a cone.

Proposition 4.3.1. *The relation “ \leq ” on \mathbb{R} is a (total) order. That is, it satisfies the following properties:*

(1) *It is reflexive, i.e.,*

$$(\forall x \in \mathbb{R})(x \leq x).$$

(2) *The relation is anti-symmetric,*

$$(\forall x, y \in \mathbb{R})([x \leq y \text{ and } y \leq x] \Rightarrow x = y).$$

(3) *It is transitive,*

$$(\forall x, y, z \in \mathbb{R})([x \leq y \text{ and } y \leq z] \Rightarrow x \leq z).$$

(4) *The relation \leq is a total (complete) order,*

$$(\forall x, y \in \mathbb{R})(x \leq y \text{ or } y \leq x).$$

Proof: We have that $x \leq x$ by the definition of \leq .

Note that it is impossible that for two elements $x, y \in \mathbb{R}$ at the same time $x < y$ and $y < x$. Indeed, this would imply $y - x \in \mathbb{R}_+$ and $-(y - x) = x - y \in \mathbb{R}_+$. But this is excluded by property (3) of \mathbb{R}_+ . Thus, the only possibility for $x \leq y$ and $y \leq x$ is when $x = y$.

Suppose $x \leq y$ and $y \leq z$ for some $x, y, z \in \mathbb{R}$. If either $x = y$ or $y = z$, then $x \leq z$. Assume that $x < y$ and $y < z$. Therefore, $z - y \in \mathbb{R}_+$ and $y - x \in \mathbb{R}_+$. By property (1) of the subset \mathbb{R}_+ , we get that

$$z - x = (z - y) + (y - x) \in \mathbb{R}_+.$$

Hence, $x < z$ as asserted. Thus, we have shown that \leq is a partial order on \mathbb{R} .

It remains to show that any two elements in $x, y \in \mathbb{R}$ are comparable. By property (3) of \mathbb{R}_+ we either have $y - x \in \mathbb{R}_+$ or $x - y = -(y - x) \in \mathbb{R}_+$ or $y - x = 0$. That is, either $y > x$ or $y < x$ or $x = y$. This completes the proof. ■

Let us summarize the main properties of the order relation on \mathbb{R} .

Proposition 4.3.2. *Let the order on \mathbb{R} be defined by the cone \mathbb{R}_+ . Then for all x, y, z and u, v in \mathbb{R} the following are valid:*

- | | |
|---|--|
| (a) $x \leq y \Rightarrow x + z \leq y + z.$ (c) $(x \leq y \text{ and } u \leq v) \Rightarrow x + u \leq y + v.$ (d) $(0 \leq x \leq y \text{ and } 0 \leq u \leq v) \Rightarrow xu \leq yv.$ (e) $x \leq y \Leftrightarrow -y \leq -x.$ (g) $(\forall x \in \mathbb{R}, x \neq 0)(x^2 > 0).$ (i) $0 < x \Leftrightarrow 0 < x^{-1}.$ | (b) $(z \geq 0 \text{ and } x \leq y) \Rightarrow xz \leq yz.$ (f) $(x \leq y \text{ and } z \leq 0) \Rightarrow yz \leq xz.$ (h) $0 < 1.$ (j) $0 < x \leq y \Rightarrow 0 < y^{-1} \leq x^{-1}.$ |
|---|--|

Proof: Properties (a) and (b) follow directly from the properties of \mathbb{R}_+ and by $x < y$ if $y - x \in \mathbb{R}_+$.

Apply (a) or (b) twice to prove (c) and (d), respectively.

Property (e) is a direct consequence of $-x - (-y) = y - x$.

To prove (f) note that $-z \geq 0$. From (b) we conclude that $(-z)x \leq (-z)y$. But $(-z)x = -xz$ and $(-z)y = -yz$. Hence, (f) is a consequence of (e).

If $x > 0$, then by the properties of \mathbb{R}_+ also $x^2 = x \cdot x \in R_+$, i.e., $x^2 > 0$. If $x < 0$, by the previous observation $(-x)(-x) > 0$. Yet, $(-x)(-x) = x^2$, which completes the proof of (g).

Since $1 = 1^2$, property (h) follows from (g).

To prove (i) note that $x^{-1} < 0$ and $x > 0$ by (f) would imply $1 = xx^{-1} < 0$, contradicting (h). Therefore, x^{-1} cannot be negative, proving $x^{-1} > 0$. To prove the other implication, note that by the first step $x^{-1} > 0$ implies $x = (x^{-1})^{-1} > 0$. This completes the proof of (i).

Finally, using property (b), then $0 < x \leq y$ and $x^{-1} > 0$ imply $0 < x^{-1}x \leq x^{-1}y$. Another application of (b), this time with y^{-1} leads to $0 < y^{-1} \leq x^{-1}$ as asserted. ■

Let us state a useful corollary.

Corollary 4.3.3. *Let x and y be two positive real numbers and $n \in \mathbb{N}$. Then $x^n < y^n$ if and only if $x < y$.*

$$(\forall x, y > 0)(x < y \Leftrightarrow x^n < y^n).$$

Proof: We first show that $x < y$ implies $x^n < y^n$. This is done by induction over n . Of course, this is true by trivial reason for $n = 1$.

Now suppose we already know that $x^n < y^n$ for a certain $n \geq 1$. An application of property (d) in Exercise 4.3.1 with $u = x^n$ and $v = y^n$ implies that

$$x^{n+1} = x x^n < y y^n = y^{n+1}.$$

Consequently, the assertion is valid for $n + 1$, hence by induction for all $n \geq 1$.

Assume now that we have $x^n < y^n$ for a certain $n \geq 1$. Our goal is to show that then $x < y$. We verify this by contraposition. That is we show, that if not $x < y$, then we cannot have $x^n < y^n$. But if we do not have $x < y$, then $y \leq x$. Thus, $y = x$ or $y < x$. Hence, either $x^n = y^n$, or by applying the first step with x and y interchanged, we get that $y^n < x^n$. In conclusion, we arrive at $y^n \leq x^n$ which is the contrary of $x^n < y^n$. This completes the proof. ■

As we saw, the above rules are easy to prove. Nevertheless, their incorrect application is one of the most common errors occurring in calculations, whether it is in school, at the university or in mathematical research. Therefore, let us emphasize once more the main rules.

- (1) If $x, y > 0$, then $\frac{x}{y}$ becomes **larger** when x is increased or $y > 0$ is decreased.
- (2) The fraction $\frac{x}{y}$ becomes **smaller** by decreasing $x > 0$ or increasing $y > 0$.
- (3) If $x \leq y$, then it follows that $ax \leq ay$ if $a > 0$ while one gets $ay \leq ax$ if $a < 0$, no matter if x and/or y are positive or negative.
- (4) For all $x, y > 0$ one has $x \leq y$ if and only if $\frac{1}{y} \leq \frac{1}{x}$.

Before proceeding further let us still introduce the notion of intervals in \mathbb{R} . There exist finite and infinite intervals, closed, open and also semi-closed ones.

Definition 4.3.1. Let $a < b$ be two real numbers. Then the following finite (closed, open and half-open) intervals are defined by

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

Similarly, closed infinite intervals are given by

$$\begin{aligned}[a, \infty) &= \{x \in \mathbb{R} : x \geq a\} \quad \text{and} \\ (-\infty, b] &= \{x \in \mathbb{R} : x \leq b\}.\end{aligned}$$

The notion of **absolute value** of an element $x \in \mathbb{R}$ depends on the introduced order.

Definition 4.3.2. Given $x \in \mathbb{R}$, define the **absolute value** $|x|$ of x as

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

Remark 4.3.2. Although easy to define, the absolute value plays an important role in calculus. This is mainly due to the fact that $|x - y|$ is nothing else as the distance between the two numbers x and y , no matter if $x < y$ or $y < x$.

Let us summarize some easy properties of the absolute value.

Proposition 4.3.4. *The absolute value satisfies the following:*

- (1) $(\forall x \in \mathbb{R})(|x| \geq 0 \text{ and } |x| = |-x|)$.
- (2) $|x| = 0 \Leftrightarrow x = 0$.
- (3) $(\forall x, y \in \mathbb{R})(|x \cdot y| = |x| \cdot |y|)$.
- (4) $(\forall x \in \mathbb{R}, \forall n \in \mathbb{N})(|x^n| = |x|^n)$.
- (5) $(\forall x, a \in \mathbb{R}, a > 0)[(|x| < a) \Leftrightarrow (-a < x < a)]$.
- (6) $(\forall x, a \in \mathbb{R}, a > 0)[(|x| \leq a) \Leftrightarrow (-a \leq x \leq a)]$.

Proof: The first property follows directly from the definition of the absolute value. Recall that $x < 0$ if and only if $-x > 0$.

Property (2) is a direct consequence of the definition of the absolute value. Note that $x \neq 0$ if and only if $-x \neq 0$.

To prove (3) let us distinguish between different cases.

Case 1: If $x, y > 0$, then $xy > 0$. Therefore, $|xy| = xy = |x||y|$ and we are done.

Case 2: One element is positive, the other negative. Say $x > 0$ and $y < 0$. Then $xy < 0$, and therefore,

$$|xy| = -(xy) = x(-y) = |x||y|,$$

as asserted.

Case 3: $x, y < 0$ Then $x \cdot y > 0$, hence the result follows by

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

Property (4) follows from (3) by induction over n via

$$|x^n| = |\underbrace{x \cdots \cdots x}_n| = \underbrace{|x| \cdots \cdots |x|}_n = |x|^n.$$

Property (5) may be proved as follows: If $x > 0$, then $|x| < a$ if and only if $x < a$, and this happens if and only if $-a < x < a$. Recall that we investigate the case $x > 0$ and that $a > 0$, hence $-a < x$ is always satisfied.

For $x < 0$ the proof is as follows: Here one has $|x| < a$ if and only if $-x < a$ or, equivalently, if $x > -a$. But the latter happens (recall $x < 0$) if and only if $-a < x < a$. Thus, (5) is proved.

Property (6) follows by exactly the same methods as (5). ■

Example 4.3.1. Assume we want to characterize those $x \in \mathbb{R}$ which satisfy

$$||x| - 2| < 1.$$

Using Property (5) of Proposition 4.3.4 we obtain

$$||x| - 2| < 1 \Leftrightarrow -1 < |x| - 2 < 1 \Leftrightarrow 1 < |x| < 3 \Leftrightarrow -3 < x < -1 \text{ or } 1 < x < 3.$$

Hence, the set of solutions of the inequality above is $(-3, -1) \cup (1, 3)$.

Let us still give another similar example.

Example 4.3.2. Which real numbers satisfy

$$|x - 1||x + 1| \leq 3 ?$$

Note that the left-hand side coincides with $|x^2 - 1|$. Consequently, we obtain

$$\begin{aligned} |x - 1||x + 1| \leq 3 &\Leftrightarrow |x^2 - 1| \leq 3 \Leftrightarrow -3 \leq x^2 - 1 \leq 3 \\ &\Leftrightarrow x^2 \leq 4 \Leftrightarrow -2 \leq x \leq 2 \Leftrightarrow |x| \leq 2. \end{aligned}$$

The set of solutions of our inequality is $[-2, 2]$.

The next result is an important inequality valid for the absolute value.

Proposition 4.3.5 (Triangle Inequality). *For any $x, y \in \mathbb{R}$,*

$$(4.3.2) \quad |x + y| \leq |x| + |y|.$$

Equality happens if $x, y \geq 0$ or $x, y < 0$.

Proof: First note that for $x \in \mathbb{R}$, $|x| = \max(x, -x)$. To see this, consider the cases $x > 0$ and $x < 0$ separately. Consequently, $|x| \geq x$ and $|x| \geq -x$. Therefore,

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

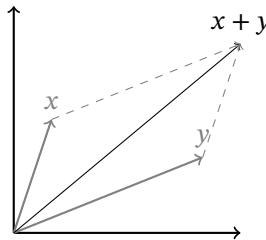
Combining both estimates leads to

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

The proof is finished by using property (6) in Proposition 4.3.4 with $a = |x| + |y|$.

If the equality $|x+y| = |x|+|y|$ happens, either $x+y = |x|+|y|$ or $x+y = -|x|-|y|$. The identity $x+y = |x|+|y|$ holds if and only if $x = |x|$ and $y = |y|$ which occurs exactly when $x, y \geq 0$. Similarly, $x+y = -|x|-|y|$ if and only if $x, y \leq 0$. ■

Remark 4.3.3. One may wonder why estimate (4.3.2) is called the triangle inequality. This stems from its higher-dimensional analogue where $|x|$ denotes the length of a vector x . In this case it says that the sum of the length of two sides of a triangle is always greater than or equal the remaining third one. The triangle inequality (6.2.1) for complex numbers can be interpreted in this way (see Theorem 6.2.3).



Corollary 4.3.6 (Reverse Triangle Inequality). *We have*

$$(\forall x, y \in \mathbb{R}) (||x| - |y|| \leq |x - y|).$$

Proof: An application of the triangle inequality leads to

$$|x| = |x - y + y| \leq |x - y| + |y|, \quad \text{hence to} \quad |x| - |y| \leq |x - y|.$$

In the same way,

$$|y| = |x - y - x| \leq |x - y| + |-x| = |x - y| + |x|, \quad \text{implying} \quad |y| - |x| \leq |x - y|.$$

Combining both estimates gives

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

An application of property (6) in Proposition 4.3.4 with $a = |x - y|$ completes the proof. ■

Example 4.3.3. Do there exist real numbers x satisfying

$$|x - 2| + |x + 2| < 4?$$

The answer is negative. A plausibility consideration is as follows: $|x - 2|$ and $|x + 2|$ measure the distance of x to 2 and -2 , respectively. So their sum is the distance from

-2 to 2 when choosing the way via the point $x \in \mathbb{R}$. But since the distance between -2 and 2 is 4 , going via x cannot shorten the way from -2 to 2 . The precise mathematical argument is as follows. By the triangle inequality we get

$$4 = (2 - x) + (2 + x) \leq |(2 - x) + (2 + x)| \leq |2 - x| + |x + 2| = |x - 2| + |x + 2|,$$

This proves our assertion.

Exercise 4.3.1. Prove the following modification of Proposition 4.3.2.

- (a) $x < y \Rightarrow x + z < y + z$.
- (b) $(z > 0 \text{ and } x < y) \Rightarrow xz < yz$.
- (c) $(x < y \text{ and } u \leq v) \Rightarrow x + u < y + v$.
- (d) $(0 < x < y \text{ and } 0 < u \leq v) \Rightarrow xu < yv$.
- (e) $x < y \Leftrightarrow -y < -x$.
- (f) $(x < y \text{ and } z < 0) \Rightarrow yz < xz$.
- (g) $0 < x < y \Leftrightarrow 0 < y^{-1} < x^{-1}$.

Exercise 4.3.2. Let x and y be two nonzero real numbers. In which cases is $x \leq y$ equivalent to $\frac{1}{y} \leq \frac{1}{x}$? For example, is this true if both numbers are negative or if only one of these numbers is so?

Exercise 4.3.3. (1) For any $0 < x < 1$, prove that

$$1 > x > x^2 > x^3 > \dots > x^n > x^{n+1} > \dots > 0.$$

(2) For any $1 < x < \infty$, show that

$$1 < x < x^2 < x^3 < \dots < x^n < x^{n+1} < \dots$$

Exercise 4.3.4. Show that for any $n \in \mathbb{N}$ and any $x_1, \dots, x_n \in \mathbb{R}$,

$$|x_1 + \dots + x_n| \leq |x_1| + \dots + |x_n| \quad \text{and} \quad |x_1 \cdot \dots \cdot x_n| = |x_1| \cdot \dots \cdot |x_n|.$$

Exercise 4.3.5. Characterize the equality case in the previous left-hand inequality.

Exercise 4.3.6. Verify the following assertion:

$$(\forall x \in \mathbb{R}, x \neq 0 \text{ and } \forall n \in \mathbb{N})(|x^{-n}| = |x|^{-n}).$$

Exercise 4.3.7. Determine all real numbers x such that $|x - 3| > 5$.

Exercise 4.3.8. Determine all real numbers x such that $|x + 1| + |3 - 2x| < 4$.

Exercise 4.3.9. Determine all real numbers x such that $|x - 1| \cdot |x + 1| < 1$.

Exercise 4.3.10. Find all real numbers x such that $2 < |x| + 2 < 3$.

Exercise 4.3.11. Given n real numbers x_1, \dots, x_n , then

$$\max\{x_1, \dots, x_n\} \quad \text{and} \quad \min\{x_1, \dots, x_n\}$$

denote the largest and the smallest number of x_1, \dots, x_n , respectively. So, for example $\max\{3, 1\} = 3$ while $\min\{3, 1\} = 1$.

(1) Prove the following formulas for the maximum and the minimum:

$$\max\{x, y\} = \frac{x + y + |x - y|}{2} \quad \text{and} \quad \min\{x, y\} = \frac{x + y - |x - y|}{2}, \quad x, y \in \mathbb{R}.$$

(2) Show that for all $x, y, z \in \mathbb{R}$ it follows that

$$\begin{aligned}\max\{x, \max\{y, z\}\} &= \max\{\max\{x, y\}, z\} = \max\{x, y, z\} \\ \min\{x, \min\{y, z\}\} &= \min\{\min\{x, y\}, z\} = \min\{x, y, z\}.\end{aligned}$$

(3) Why do the two latter properties imply that the two binary operations

$$(x, y) \mapsto \max\{x, y\} \quad \text{and} \quad (x, y) \mapsto \min\{x, y\}$$

are both associative?

4.4. Completeness

So far, we have seen that the rational numbers and the real numbers share the same properties. It is natural to ask if there are differences between these two classes of numbers. To answer this question, let us start with an example which shows again that there are *holes* in \mathbb{Q} . Moreover, at the same time this example gives a clue on how these holes could be *filled*.

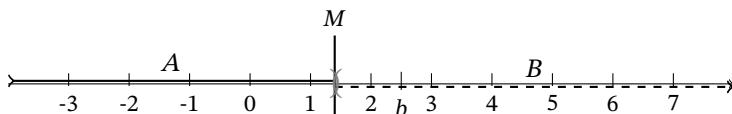
Example 4.4.1. Consider the set

$$(4.4.1) \quad A := \{a \in \mathbb{Q} : a^2 < 2\} \cup \{a \in \mathbb{Q} : a < 0\}.$$

First we note that A is bounded above, i.e., there exists at least one $b \in \mathbb{Q}$ so that all elements in A are less than or equal to this number b . This easily follows by the fact that

$$(4.4.2) \quad (b \in \mathbb{Q}_+, b^2 \geq 2) \Rightarrow (\forall a \in A)(a \leq b).$$

We will call a number b satisfying (4.4.2) an upper bound of A . Let B be the set of all these upper bounds. The crucial question is now the following. Does there exist a smallest element in the set B of upper bounds? In other words, does the set A in (4.4.1) possess a least upper bound?



It is very plausible, and we shall give a rigorous proof soon (cf. Theorem 4.6.3 below), that such a least upper bound M of A should satisfy $M^2 = 2$. But as we saw before, there is no such number $M \in \mathbb{Q}$. Consequently, it follows that in \mathbb{Q} there is no least upper bound of the set A defined by (4.4.1). Thus, one way to add $\sqrt{2}$ to \mathbb{Q} would be to define it as the missing least upper bound of A .

In order to formulate the third axiom about \mathbb{R} we need the following notation.

Definition 4.4.1. A nonempty subset $A \subseteq \mathbb{R}$ is **bounded above** if there exists $M \in \mathbb{R}$ such that

$$(4.4.3) \quad x \leq M \quad \text{for all } x \in A.$$

Similarly, the set A is **bounded below** if there exists $m \in \mathbb{R}$ such that

$$(4.4.4) \quad m \leq x \quad \text{for all } x \in A.$$

Finally, A is **bounded**, if it is bounded above and below.

$$A \text{ is bounded above} \Leftrightarrow [(\exists M \in \mathbb{R})(\forall x \in A)(x \leq M)].$$

$$A \text{ is bounded below} \Leftrightarrow [(\exists m \in \mathbb{R})(\forall x \in A)(m \leq x)].$$

$$A \text{ is bounded} \Leftrightarrow [(\exists m, M \in \mathbb{R})(\forall x \in A)(m \leq x \leq M)].$$

Remark 4.4.1. Note that

$$A \text{ is bounded} \Leftrightarrow (\exists M \in \mathbb{R})(\forall x \in A)(|x| \leq M).$$

We leave the proof of this equivalence as an exercise (see Exercise 4.4.4).

Definition 4.4.2. If $A \subseteq \mathbb{R}$ is bounded above, then each $M \in \mathbb{R}$ satisfying (4.4.3) is said to be an **upper bound** of A .

If $A \subseteq \mathbb{R}$ is bounded below, a number $m \in \mathbb{R}$ with (4.4.4) is called a **lower bound** of A .

$$M \text{ is an upper bound of } A \Leftrightarrow (\forall x \in A)(x \leq M).$$

$$m \text{ is a lower bound of } A \Leftrightarrow (\forall x \in A)(m \leq x).$$

Remark 4.4.2. Suppose $A \subseteq \mathbb{R}$ is bounded above. Of course, if $M \in \mathbb{R}$ is an upper bound of A , then any $M' \geq M$ is an upper bound as well. Thus, the set of all possible upper bounds is a certain interval located on the right-hand side of A on the number axis. While this interval is unbounded on its right border, it is far from clear what happens at its left border.

Now we are in position to state the **third** (and last) **axiom** about \mathbb{R} :

(III) **Order Completeness:** Each nonempty subset A of real numbers, bounded above, possesses a least upper bound.

$$(\forall A \subseteq \mathbb{R}, A \neq \emptyset, A \text{ bounded above}) \Rightarrow (\exists \text{ least upper bound of } A).$$

In other words, whenever A is a nonempty subset of \mathbb{R} which is bounded above, then there is $M \in \mathbb{R}$ such that

- For all $x \in A$, $x \leq M$.
- If $M' \in \mathbb{R}$ satisfies $x \leq M'$ for all $x \in A$, then $M \leq M'$.

Remark 4.4.3. Another way to express the existence of a least upper bound of a set A is as follows: If B denotes the set of upper bounds of A , i.e.,

$$B := \{b \in \mathbb{R} : x \leq b \text{ for all } x \in A\},$$

then its minimum $\min(B)$ exists. In particular this implies that for any nonempty set, bounded above, its least upper bound is a uniquely determined real number.

Remark 4.4.4. There is one case where the least upper bound can be easily determined. Assume that $A \subseteq \mathbb{R}$ possesses a maximum element $a_0 = \max(A)$. That is, $a_0 \in A$ satisfies

$$x \leq a_0, \quad \forall x \in A.$$

In this case a_0 is also an upper bound of A and, of course, no upper bound can be strictly smaller than a_0 . Recall that $a_0 \in A$. Hence, in this situation we see that

$$a_0 = \max(A) \text{ is the least upper bound of } A.$$

Example 4.4.2.

(1) If

$$A = [0, 1] = \{x \in \mathbb{R} : 0 \leq x \leq 1\},$$

then $1 = \max(A)$, hence 1 is also the smallest upper bound of A .

(2) On the contrary, if

$$A = (0, 1) = \{x \in \mathbb{R} : 0 < x < 1\},$$

then A does not possess a maximum element. But since a number $M \in \mathbb{R}$ is an upper bound of A if and only if $M \geq 1$, also in this case the least upper bound is 1 .

(3) Suppose the set A consists of the following fractions:

$$A = \left\{ \frac{n^2 - 1}{n^2} : n \in \mathbb{N} \right\}.$$

Any number $M \geq 1$ is an upper bound for A because $M \geq 1 > \frac{n^2 - 1}{n^2}$ for any $n \in \mathbb{N}$. It seems plausible that 1 is the smallest one among all upper bounds. But perhaps there is an upper bound $M' < 1$, that is, 1 would not be the smallest one. How we can show that such an $M' < 1$ cannot exist? We have to show that any $M' < 1$ is no longer an upper bound or, equivalently, there is at least one $a \in A$ for which $a \leq M'$ is violated, i.e., where $M' < a$. In our case this means, we have to show the following:

$$(\forall M' < 1)(\exists n \geq 1) \left(M' < \frac{n^2 - 1}{n^2} \right).$$

We encourage the reader to think more about this problem, and we postpone the proof of it to the next section.

Summary: The set of real numbers, denoted by \mathbb{R} , is characterized by the existence of two binary operations $+$ and \cdot and by a subset \mathbb{R}_+ of positive elements satisfying properties (I a), (I b), (I c), (II), and (III). That's all!

Exercise 4.4.1. Find all upper and lower bounds of the sets A and B defined by

$$A = \left\{ \frac{1}{\sqrt{n+1}} : n \geq 1 \right\} \quad \text{and} \quad B = \left\{ \frac{3n+2}{n+1} : n \geq 1 \right\}.$$

Exercise 4.4.2. Find all upper and lower bounds for the sets C and D defined by

$$C = (-\infty, 5] \text{ and } D = \{x : x^2 \geq 9\}.$$

Exercise 4.4.3. Find all upper and lower bounds for the sets E and F defined by

$$E = \{1.01^n : n \in \mathbb{N}\} \text{ and } F = \{0.99^n : n \in \mathbb{N}\}.$$

Exercise 4.4.4. Prove that a nonempty set $A \subseteq \mathbb{R}$ is bounded if and only if there exists $M \geq 0$ such that $|x| \leq M$, for any $x \in A$.

Exercise 4.4.5. Let $A \subseteq \mathbb{R}$ be a nonempty set. Show that A is not bounded above if and only if for any $M \in \mathbb{R}$, there exists $x \in A$ (that depends on M) such that $x > M$.

Exercise 4.4.6. State and prove a similar statement to the one above characterizing when a nonempty set of real numbers is not bounded below.

Exercise 4.4.7. Let $A \subseteq B$ be two subsets of \mathbb{R} .

- (1) If A is bounded above, then is B bounded above? Prove or give a counterexample.
- (2) If A is bounded below, then is B bounded below? Prove or give a counterexample.
- (3) How are the sets of upper/lower bounds of A and B related?

Exercise 4.4.8. Let $A \subseteq \mathbb{R}$ be a nonempty set. Determine, with proof or counterexample, the truth value of each of the following statements.

- (1) If A is bounded above, then $\mathbb{R} \setminus A$ is not bounded above.
- (2) If A is not bounded above, then $\mathbb{R} \setminus A$ is not bounded above.
- (3) If A is not bounded above, then $\mathbb{R} \setminus A$ is bounded above.
- (4) If A is bounded above, then $\mathbb{R} \setminus A$ is bounded below.

Exercise 4.4.9. Let $A, B \subseteq \mathbb{R}$ be two nonempty sets. Determine, with proof or counterexample, the truth value of each of the following statements.

- (1) If A and B are bounded above, then $A \cup B$ is bounded above.
- (2) If A is bounded above and B is bounded below, then $A \cup B$ is bounded.
- (3) If A is bounded above and B is bounded below, then $A \cap B$ is bounded.
- (4) If A and B are not bounded below, then $A \cap B$ is not bounded below.

Exercise 4.4.10. Suppose we introduce Dedekind cuts as in Definition A.8.1 but now on \mathbb{R} instead on \mathbb{Q} . Why can every such (real) cut A be written as

$$A = \{x \in \mathbb{R} : x < M\}$$

for a suitable $M \in \mathbb{R}$? Note that this is in complete contrast to the rational case where there are cuts (so-called irrational cuts) which do not allow such a representation for some $M \in \mathbb{Q}$.

4.5. Supremum and Infimum of a Set

Definition 4.5.1. Let $A \subseteq \mathbb{R}$ be a set bounded above. Its least upper bound is called the **supremum** of A and denoted by $\sup(A)$. In other words, if $A \subseteq \mathbb{R}$ is bounded above, then

$$M = \sup(A) \Leftrightarrow [(\forall x \in A)(x \leq M) \text{ and } (\forall M' \in \mathbb{R}) \text{ upper bound of } A(M \leq M')].$$

In this setting, the order completeness as stated in property (III) of \mathbb{R} may also be formulated as follows⁶.

(III) Every nonempty subset of \mathbb{R} bounded above has a supremum.

As already mentioned, there is one special case where it is easy to determine the supremum of a set A and where its existence does not depend on the validity of property (III) of the real line:

Proposition 4.5.1. *If $A \subseteq \mathbb{R}$ possesses a maximum element $\max(A)$, then*

$$\sup(A) = \max(A).$$

Proof: Let x_0 be the maximum element in A , i.e., $x_0 \in A$ and $x \leq x_0$ for all $x \in A$. Then it is obvious that a number $M \in \mathbb{R}$ is an upper bound of A if and only if $M \geq x_0$. That is, the set of upper bounds coincides with the interval $[x_0, \infty)$. Clearly, then x_0 is the least upper bound, hence by the definition of the supremum it follows that $x_0 = \sup(A)$ which completes the proof. ■

Remark 4.5.1. If there exists a maximum element in A , then $\sup(A) = \max(A)$ is an element of A . In this case one says that the supremum of A is **attained**.

$$\sup(A) \text{ is attained} \Leftrightarrow \sup(A) \in A \Leftrightarrow \max(A) \text{ exists.}$$

Example 4.5.1. If $A_1 = \left\{ \frac{n}{n+1} : n \geq 1 \right\}$ and $A_2 = \left\{ \frac{1}{n} : n \geq 1 \right\}$, then it follows that $\sup(A_1) = \sup(A_2) = 1$. Here the supremum is attained for A_2 , but not for A_1 .

In order to determine the supremum of a given set, it is useful to characterize the supremum in a slightly different way.

Proposition 4.5.2. *Let $A \subseteq \mathbb{R}$ be a subset bounded above. The following statements are equivalent.*

- (1) $M = \sup(A)$.
- (2) M is an upper bound of A , and for any $M' < M$, there exists $x \in A$ (depending on M') such that $M' < x \leq M$.
- (3) M is an upper bound of A , and for each $\varepsilon > 0$, there is $x \in A$ (depending on ε) such that $M - \varepsilon < x \leq M$.

$$\begin{aligned} M = \sup(A) &\Leftrightarrow [M \text{ upper bound and } (\forall M' \in \mathbb{R}, M' < M)(\exists x \in A)(x > M')] \\ &\Leftrightarrow [M \text{ upper bound of } A \text{ and } (\forall \varepsilon > 0)(\exists x \in A)(x > M - \varepsilon)]. \end{aligned}$$

Proof: The equivalence of (2) and (3) is obvious. Indeed, going from (2) to (3) choose $M' = M - \varepsilon < M$ while in the other direction one sets $\varepsilon = M - M' > 0$.

(1) \Rightarrow (2) : If $M = \sup(A)$ and $M' < M$, then M' cannot be an upper bound. Consequently, since M' is not an upper bound, we cannot have $x \leq M'$ for all $x \in A$.

⁶Verbally, this can be expressed as follows: For any nonempty set bounded above there exists a (unique) least number (called supremum) being bigger than all elements of the given set. If the set is the pan, then the supremum is the lid on the top of the pan. Hereby, the lid may or may not belong to the pan.

But this implies that there is at least one $x \in A$ (depending on M') where $x \not\leq M'$, i.e., $x > M'$. The estimate $x \leq M$ is satisfied since M is assumed to be an upper bound. This proves (2).

(2) \Rightarrow (1) : Let $M \in \mathbb{R}$ satisfy property (2). Then M is an upper bound. But why it is also the least one? This easily follows by property (2). There is no smaller upper bound than M because by (2) any $M' < M$ is no longer an upper bound. This completes the proof. ■

Let us give a first application of the completeness of \mathbb{R} . It asserts an important property of the set of natural numbers that we have seen at the beginning of our book. We look at it now through the lenses of real numbers.

Theorem 4.5.3 (Eudoxos⁷). *The set of natural numbers is **not** bounded above.*

Proof: Let us assume the contrary. Then $x = \sup(\mathbb{N})$ would exist. An application of property (3) in Proposition 4.5.2 with $\varepsilon = 1$ implies the following. There is $n \in \mathbb{N}$ with $x - 1 < n$. Equivalently, $x < n + 1$, but since $n + 1 \in \mathbb{N}$, x cannot be an upper bound of \mathbb{N} . This contradiction shows that our assumption \mathbb{N} is *bounded above* was wrong, which proves the proposition. ■

Remark 4.5.2. Which properties of \mathbb{N} did we use to prove Theorem 4.5.3? Firstly we needed that $\mathbb{N} \neq \emptyset$. Otherwise, $\sup(\mathbb{N})$ would not make sense. And secondly, we had to know that $n \in \mathbb{N}$ implies $n + 1 \in \mathbb{N}$. Nothing else was used during the proof of $\sup(\mathbb{N}) = \infty$. Thus, every nonempty subset of \mathbb{R} which contains with some $x \in \mathbb{R}$ also $x + 1$ cannot be bounded above.

Another important consequence of Theorem 4.5.3 is a result which is usually called **Archimedean property**⁸ of the real numbers.

Theorem 4.5.4 (Archimedean property of \mathbb{R}). *For each real number $x > 0$ there is a natural number $n \geq 1$ such that $0 < 1/n < x$.*

$$(\forall x \in \mathbb{R}, x > 0)(\exists n \in \mathbb{N})(0 < 1/n < x).$$

Proof: Let us assume that the Archimedean property is not valid. Then there exists a real number $x > 0$ such that $1/n \geq x$ for any $n \in \mathbb{N}$. But this is equivalent to $n \leq 1/x$ for all $n \geq 1$. Consequently, $1/x$ is an upper bound of \mathbb{N} , thus \mathbb{N} is bounded above. But this contradicts Theorem 4.5.3 and completes the proof. ■

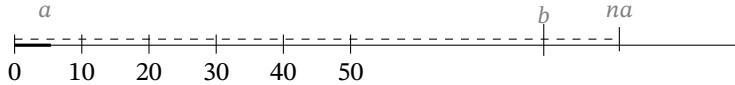
Corollary 4.5.5. *Let $a, b \in \mathbb{R}$ be two positive numbers. Then there is $n \in \mathbb{N}$ such that $b < n \cdot a$.*

Proof: Set $x = a/b$ and choose $n \in \mathbb{N}$ such that $1/n < x = a/b$. ■

⁷Eudoxos of Cnidus, c. 400 BCE – 347 BCE.

⁸In honor of Archimedes of Syracuse (c. 287 BCE – 212 BCE), but Archimedes credited it to Eudoxos. Therefore, sometimes it is also called the theorem of Eudoxos. In the formulation of Corollary 4.5.5 it appeared in Euclid's book *Elements*, volume V, around 300 BCE

Remark 4.5.3. The previous corollary says the following: Suppose we are given two straight lines, one of length $a > 0$ and another one of length $b > 0$. Putting together sufficiently many copies of lines of length $a > 0$, one exceeds the line of length $b > 0$. This is always possible, even if $a > 0$ is very small and $b > 0$ is quite big⁹.



Corollary 4.5.6. Suppose a real number $x \geq 0$ satisfies $x < 1/n$ for all $n \in \mathbb{N}$. Then this implies $x = 0$.

Proof: The proof is straightforward. Assume to the contrary that $x \neq 0$, hence $x > 0$. By the Archimedean property there is at least one $n_0 \in \mathbb{N}$ for which $1/n_0 < x$. But this contradicts $x < 1/n$ for all $n \in \mathbb{N}$. ■

Remark 4.5.4. The proof of the Archimedean property of \mathbb{R} rests upon the theorem of Eudoxos (cf. Theorem 4.5.3). And to prove this result the order completeness of \mathbb{R} is needed, i.e., we used the existence of lowest upper bounds of sets bounded above. The Archimedean property cannot be proved without this completeness property. There exist ordered fields (sets satisfying axioms (I a), (I b), (I c), and (II)) containing a copy of \mathbb{N} and where the Archimedean property is not valid. That is, such an ordered field contains positive elements which are less than every $1/n$, $n \geq 1$, (difficult to imagine but not too complicated to construct). Thus, it is a little bit surprising that, as stated in Exercise 3.1.9, the ordered field \mathbb{Q} fulfills the Archimedean property, although \mathbb{Q} does not satisfy the completeness axiom (III).

Example 4.5.2. Let

$$A = \left\{ \frac{n}{n+1} : n \geq 1 \right\} = \left\{ \frac{1}{2}, \frac{2}{3}, \frac{3}{4}, \dots \right\}.$$

We claim that $\sup(A) = 1$.

Of course, for all $n \geq 1$ it follows that $\frac{n}{n+1} < 1$, hence 1 is an upper bound of A .

We still have to show that 1 is the least upper bound. To this end we verify property (3) of Proposition 4.5.2. That is, given $\varepsilon > 0$ we have to find at least one $x = x(\varepsilon) \in A$ for which $1 - \varepsilon < x$. Since A consists of numbers $\frac{n}{n+1}$, we have to look for an integer $n = n(\varepsilon) \geq 1$ with

$$1 - \varepsilon < \frac{n}{n+1}.$$

Equivalently, we have to find at least one $n \geq 1$ satisfying

$$\frac{1}{n+1} = 1 - \frac{n}{n+1} < \varepsilon.$$

Let us choose an integer n with $n > 1/\varepsilon - 1$. Theorem 4.5.3 assures that such a natural number n exists. Then $1 - \varepsilon < \frac{n}{n+1}$. This proves that $\sup(A) = 1$.

Let us discuss another example mentioned earlier.

⁹Even from small grains of sand, you can build mountains of any size. All you have to do is have enough grains of sand.

Example 4.5.3. Let

$$A := \left\{ \frac{n^2 - 1}{n^2} : n \geq 1 \right\}.$$

We claim that $\sup(A) = 1$.

Of course, any $M \geq 1$ is an upper bound of A . In order to show that $M = 1$ is the least upper bound we have to find for each $\varepsilon > 0$ an element $x \in A$ for which $1 - \varepsilon < x$. In other words, we have to show that

$$(4.5.1) \quad (\forall \varepsilon > 0)(\exists n \in \mathbb{N})\left(\frac{n^2 - 1}{n^2} > 1 - \varepsilon\right).$$

But since

$$\frac{n^2 - 1}{n^2} > 1 - \varepsilon \Leftrightarrow n^2 - 1 > n^2 - n^2\varepsilon \Leftrightarrow \frac{1}{n^2} < \varepsilon,$$

any natural number n with $1/n < \varepsilon$ satisfies (4.5.1). Indeed, if $1/n < \varepsilon$, then also $1/n^2 < 1/n < \varepsilon$ and, as we saw above, this implies the desired estimate

$$\frac{n^2 - 1}{n^2} > 1 - \varepsilon.$$

The existence of natural numbers n with $1/n < \varepsilon$ is ensured by $\varepsilon > 0$ and an application of the Archimedean property.

We turn now to the investigation of lower bounds of a given set $A \subseteq \mathbb{R}$.

Definition 4.5.2. Let $A \subseteq \mathbb{R}$ bounded below. A number $m \in \mathbb{R}$ is said to be the **infimum** of A (write $m = \inf(A)$) if it is the greatest lower bound. In other words

$$m = \inf(A) \Leftrightarrow [(\forall x \in A)(m \leq x) \text{ and } (\forall m' \in \mathbb{R})(m' \leq x, x \in A)(m' \leq m)]$$

There exists a characterization of the infimum of a set similar as in Proposition 4.5.2 for the supremum. Since its proof is almost identical with that of Proposition 4.5.2, we leave it as an exercise.

Proposition 4.5.7. Let $A \subseteq \mathbb{R}$ be a subset bounded below. The following statements are equivalent.

- (1) $m = \inf(A)$.
- (2) m is a lower bound of A and for any $m' > m$, there exists $x \in A$ (depending on m') such that $m \leq x < m'$.
- (3) m is a lower bound of A and for each $\varepsilon > 0$, there is $x \in A$ (depending on ε) such that $m \leq x < m + \varepsilon$.

$$\begin{aligned} m = \inf(A) &\Leftrightarrow [m \text{ lower bound and } (\forall m' \in \mathbb{R}, m < m')(\exists x \in A)(x < m')] \\ &\Leftrightarrow [m \text{ lower bound of } A \text{ and } (\forall \varepsilon > 0)(\exists x \in A)(x < m + \varepsilon)] \end{aligned}$$

Until now, we do not know whether every set bounded below possesses an infimum. The next proposition answers this question.

Proposition 4.5.8. Let $A \subseteq \mathbb{R}$ be a nonempty set. If A is bounded below, then $\inf(A)$ exists.

Proof: The idea to prove the result is to reflect A around the origin so that lower bounds become upper bounds and vice versa. More precisely, define

$$-A := \{-x : x \in A\}.$$

We claim that $-A$ is bounded above and that

$$(4.5.2) \quad \inf(A) = -\sup(-A).$$

In particular, $\inf(A)$ exists.

To prove this claim, note first that a number M is an upper bound of $-A$ if and only if $-x \leq M$, or, equivalently if $-M \leq x$ for all $x \in A$. Hence, M is an upper bound of $-A$ if and only if $-M$ is a lower bound of A . In particular, $-A$ is bounded above.

Now let $M = \sup(-A)$ and set $m := -M$. We claim that $m = \inf(A)$. Since M is an upper bound of $-A$, by the previous observation $m = -M$ is a lower bound of A . It remains to show that m is the greatest one. To this end use Proposition 4.5.7. Take any $m' > m$. Then $-m' < -m = M$, and since $M = \sup(-A)$ by Proposition 4.5.2 there is $x \in A$ such that $-m' < -x$. Of course, then $m' > x$. This being true for each $m' > m$, by Proposition 4.5.7 it follows that m is the infimum of A . Recalling that $m = -M = -\sup(-A)$, this proves (4.5.2) and completes the proof. ■

Corollary 4.5.9. Let $A \subseteq \mathbb{R}$ be a nonempty subset. Then A is bounded above if and only if $-A$ is bounded below and, moreover,

$$\inf(-A) = -\sup(A).$$

Proof: Apply (4.5.2) with A replaced by $-A$. ■

To complete the picture we extend the definition of supremum and infimum to unbounded sets.

Definition 4.5.3. Let $A \subseteq \mathbb{R}$ be nonempty. If A is **not** bounded above, let us write $\sup(A) = \infty$. Similarly, if A is **not** bounded below, then express this by $\inf(A) = -\infty$.

$$\sup(A) = \infty \Leftrightarrow A \text{ not bounded above}$$

$$\inf(A) = -\infty \Leftrightarrow A \text{ not bounded below}$$

Remark 4.5.5. Using this notation it follows that a nonempty $A \subseteq \mathbb{R}$ is bounded above if and only if $\sup(A) < \infty$. Similarly, A is bounded below if and only if $\inf(A) > -\infty$.

$$\sup(A) < \infty \Leftrightarrow A \text{ bounded above},$$

$$\inf(A) > -\infty \Leftrightarrow A \text{ bounded below}$$

The following characterization of unbounded sets is quite useful.

Proposition 4.5.10. Let $A \subseteq \mathbb{R}$ be a nonempty subset. The following statements are equivalent.

$$(1) \sup(A) = \infty.$$

$$(2) \text{For each real number } \Lambda \in \mathbb{R} \text{ there is } x \in A, x = x(\Lambda), \text{ such that } x > \Lambda.$$

(3) For each natural number $N \in \mathbb{N}$ there is $x \in A$, $x = x(N)$, with $x > N$.

$$\sup(A) = \infty \Leftrightarrow (\forall \Lambda \in \mathbb{R})(\exists x \in A)(x > \Lambda) \Leftrightarrow (\forall N \in \mathbb{N})(\exists x \in A)(x > N).$$

Proof: (1) \Leftrightarrow (2) : The set A is unbounded above if and only if there is no upper bound of A . That is, given $\Lambda \in \mathbb{R}$, then it cannot be an upper bound or, equivalently, there is at least one $x \in A$ depending on Λ with $x > \Lambda$. Hence, (2) is nothing else as a reformulation of (1).

(2) \Rightarrow (3) : Since (2) is true for all $\Lambda \in \mathbb{R}$, it is also satisfied for all $N \in \mathbb{N}$.

(3) \Rightarrow (2) : Let us prove this by contraposition. If (2) is not true, then there is at least one $\Lambda \in \mathbb{R}$ for which $x \leq \Lambda$, $x \in A$. Now Proposition 4.5.3 implies, that there is an $N \in \mathbb{N}$ with $\Lambda < N$, hence $x \leq N$ for all $x \in A$. This shows that (3) is not satisfied, which completes the proof. ■

Remark 4.5.6. Of course, it suffices if property (2) is satisfied for *large* $\Lambda \in \mathbb{R}$. That is, suppose that there is some (big) $\Lambda_0 \in \mathbb{R}$ such that (2) is valid for all $\Lambda \geq \Lambda_0$, then $\sup(A) = \infty$. Similarly, it suffices whenever (3) holds for *large* integers $N \in \mathbb{N}$.

Example 4.5.4. Let $A = \left\{ \frac{x^3+1}{x^2+1} : x > 0 \right\}$. We claim that $\sup(A) = \infty$. Thus, given an $N \in \mathbb{N}$, we have to find an $x > 0$ for which $\frac{x^3+1}{x^2+1} > N$. To this end we observe that for $x \geq 1$

$$\frac{x^3+1}{x^2+1} \geq \frac{x^3}{x^2+x^2} = \frac{x}{2}.$$

Hence, if $\frac{x}{2} > N$, then also $\frac{x^3+1}{x^2+1} > N$. Thus, choosing a fixed $x > 2N$ and setting $a = \frac{x^3+1}{x^2+1}$, then $a \in A$ and $a > N$. Since $N \in \mathbb{N}$ was arbitrarily chosen, we conclude $\sup(A) = \infty$ as claimed above.

Exercise 4.5.1. Show that for all (bounded or unbounded) nonempty subsets A of \mathbb{R}

$$\inf(A) \leq \sup(A).$$

Moreover, equality holds if and only if $|A| = 1$.

Exercise 4.5.2. Let A and B be subsets of \mathbb{R} such that $x < y$ for any $x \in A$ and any $y \in B$. Prove that

$$\sup(A) \leq \inf(B).$$

Give an example of two sets with the property above such that $\sup(A) = \inf(B)$.

Exercise 4.5.3. Let $A, B \subseteq \mathbb{R}$ be two nonempty sets in \mathbb{R} , both bounded above. Let

$$A + B = \{x + y : x \in A, y \in B\}.$$

Prove that then $A + B$ is also bounded above and, moreover,

$$\sup(A + B) = \sup(A) + \sup(B).$$

Exercise 4.5.4. Let $A \subseteq \mathbb{R}$ and $\alpha \in \mathbb{R}$. Denote $\alpha A = \{\alpha \cdot x : x \in A\}$. Show that

$$\sup(\alpha A) = \alpha \cdot \sup(A) \quad \text{if } \alpha > 0 \quad \text{and} \quad \sup(\alpha A) = \alpha \cdot \inf(A) \quad \text{if } \alpha < 0.$$

Exercise 4.5.5. Let A and B be two sets of positive real numbers such that both of them are bounded above. Define

$$A \cdot B = \{x \cdot y : x \in A, y \in B\}.$$

Prove that $A \cdot B$ is bounded above and that

$$\sup(A \cdot B) = \sup(A) \cdot \sup(B).$$

Exercise 4.5.6. Let $A \subseteq B$ be two nonempty subsets of \mathbb{R} . Prove that

$$\sup(A) \leq \sup(B) \quad \text{and} \quad \inf(B) \leq \inf(A),$$

no matter if the infima and/or suprema are finite or infinite.

Exercise 4.5.7. (\star) Let $A \subseteq \mathbb{R}$ be a nonempty set which is bounded above. Suppose that its supremum $M = \sup(A)$ is **not** attained. Show the following:

There exist elements $x_1 < x_2 < \dots$ in A such that

$$M - \frac{1}{n} < x_n < M, \quad n = 1, 2, \dots$$

Exercise 4.5.8. Evaluate the supremum and the infimum of the set

$$A = \left\{ \frac{n^2 + 1}{n^2 + 2} : n \in \mathbb{N} \right\}.$$

Exercise 4.5.9. Let

$$B = \left\{ (-1)^n \frac{n^2 - 1}{n + 4} : n \in \mathbb{N} \right\}.$$

Show that $\sup(B) = \infty$ and $\inf(B) = -\infty$.

Exercise 4.5.10. Let

$$C = \left\{ (-1)^n \frac{n^2 - 1}{n^2 + 4} : n \in \mathbb{N} \right\}.$$

Determine $\sup(C)$ and $\inf(C)$.

4.6. Roots and Powers

The aim of this section is to define arbitrary powers x^α for any $x > 0$ and any $\alpha \in \mathbb{R}$. Until now, we only know how x^n is defined for $n \in \mathbb{Z}$. Recall that, given $x > 0$ and $n \in \mathbb{N}$, then

$$x^n = \underbrace{x \cdot x \cdots \cdots x}_{n \text{ times}}, \quad x^{-n} = (x^{-1})^n \quad \text{and} \quad x^0 = 1.$$

The basic properties of this operation are

$$(\forall x > 0, n, m \in \mathbb{Z})(x^n \cdot x^m = x^{n+m}) \quad \text{and} \quad (\forall x, y > 0, n \in \mathbb{Z})(x^n \cdot y^n = (x \cdot y)^n)$$

In particular, this implies

$$(\forall x > 0, n, m \in \mathbb{Z})((x^n)^m = x^{n \cdot m})$$

Our next objective is to verify the existence of roots of a given positive number. As we saw in Theorem 3.2.1 roots do in general not exist within the framework of rational numbers. This suggests that we have to use the completeness property of the real numbers in order to construct roots of any order.

For the investigation of this question we need the following lemma.

Lemma 4.6.1. *If $a, b \in \mathbb{R}$ and $n \in \mathbb{N}$, then*

$$|a^n - b^n| \leq |a - b| \cdot n \cdot \max\{|a|, |b|\}^{n-1}.$$

Proof: In order to simplify the notation let us assume $|b| \leq |a|$ so that we have $\max\{|a|, |b|\} = |a|$. We start with formula (1.3.2) asserting that

$$a^n - b^n = (a - b)(a^{n-1} + a^{n-2}b + \cdots + ab^{n-2} + b^{n-1}).$$

Using the triangle inequality for the absolute value this implies (recall that we assumed $|b| \leq |a|$)

$$\begin{aligned} |a^n - b^n| &= |a - b| \cdot |a^{n-1} + a^{n-2}b + \cdots + ab^{n-2} + b^{n-1}| \\ &\leq |a - b| \cdot (|a|^{n-1} + |a|^{n-2}|b| + \cdots + |a||b|^{n-2} + |b|^{n-1}) \\ &\leq |a - b| \cdot (\underbrace{|a|^{n-1} + \cdots + |a|^{n-1}}_{n \text{ times}}) = |a - b| \cdot n \cdot |a|^{n-1}. \end{aligned}$$

This completes the proof. ■

Corollary 4.6.2. *For all $s > 0$ the following inequalities are true:*

$$(4.6.1) \quad (1) \quad (s + \varepsilon)^n - s^n \leq n \cdot \varepsilon \cdot (s + 1)^{n-1}, \quad 0 < \varepsilon < 1,$$

$$(4.6.2) \quad (2) \quad s^n - (s - \varepsilon)^n \leq n \cdot \varepsilon \cdot s^{n-1}, \quad 0 < \varepsilon < s.$$

Proof: To verify estimate (1) we use Lemma 4.6.1 with $a = s + \varepsilon$ and $b = s$. Then by $0 < \varepsilon < 1$ we obtain

$$|a^n - b^n| = (s + \varepsilon)^n - s^n, \quad |a - b| = \varepsilon, \quad \text{and} \quad \max\{|a|, |b|\}^{n-1} = (s + \varepsilon)^{n-1} \leq (s + 1)^{n-1} \leq (s + 1)^{n-1}.$$

This immediately proves the first estimate.

The proof of inequality (2) follows by the same scheme. Here we choose $a = s$ and $b = s - \varepsilon$. Then

$$|a^n - b^n| = s^n - (s - \varepsilon)^n, \quad |a - b| = \varepsilon, \quad \text{and} \quad \max\{|a|, |b|\}^{n-1} = s^{n-1}.$$
■

We state and prove the existence of roots of any positive real number.

Theorem 4.6.3. *Let n a natural number. For any positive real number x , there exists a unique positive real number w such that $w^n = x$.*

$$(\forall n \in \mathbb{N}, \forall x > 0)(\exists! w > 0 \text{ s.t. } w^n = x).$$

Proof: Define

$$A := \{a > 0 : a^n < x\}.$$

We will prove the following claims:

- (1) The set A is nonempty and bounded above. Consequently, $w := \sup(A)$ exists and is positive.
- (2) Any number $b > 0$ with $b^n \geq x$ is an upper bound of A .

- (3) If $b^n > x$, then there is an upper bound b' of A with $b' < b$.
(4) Any number $b > 0$ with $b^n < x$ is not an upper bound of A .

Why do these facts imply $w^n = x$? First note that we cannot have $w^n > x$. Indeed, by (3) this would imply the existence of an upper bound of A strictly smaller than w . But w is the supremum of A , hence its least upper bound. So we conclude that $w^n > x$ is impossible and, thus, we have to have $w^n \leq x$. But also $w^n < x$ is not possible. Indeed, the number w is the supremum of the set A , in particular it is an upper bound of A . But if $w^n < x$ then property (4) tells us that w cannot be an upper bound. Consequently, the only possibility is $w^n = x$. Summing up, to prove the existence of an n th root of $x > 0$ it suffices to verify properties (1) - (4).

Proof of claim (1): Set

$$a_0 := \frac{x}{x+1}.$$

Then $0 < a_0 < 1$ as well as $a_0 < x$. Hence, we derive (use the first estimate in Exercise 4.3.3) that

$$a_0^n < a_0 < x \Rightarrow a_0 \in A \Rightarrow A \neq \emptyset.$$

Furthermore, setting $b = \max\{1, x\}$ we get $b \geq 1$, hence $b^n \geq b$, consequently by $b \geq x$ we conclude that

$$a \in A \Rightarrow a^n < x \Rightarrow a^n < b \leq b^n \Rightarrow a < b \Rightarrow b \text{ is upper bound of } A.$$

Thus, A is bounded above which proves assertion (1).

Proof of claim (2): This easily follows by

$$b^n \geq x \quad \text{and} \quad a \in A \Rightarrow a^n < x \leq b^n \Rightarrow a < b.$$

That is, any $b > 0$ with $b^n \geq x$ is an upper bound of A .

Proof of claim (3): Suppose now that a positive number $b \in \mathbb{R}$ satisfies $b^n > x$. Our aim is to show that there is an $\varepsilon > 0$ such that also $(b - \varepsilon)^n > x$. Then by (2) the number $b' := b - \varepsilon$ is an upper bound of A as well and (3) would be true. To this end set $\delta := b^n - x$ which, by assumption, is a positive number. Choose an arbitrary $\varepsilon < b$ such that

$$0 < \varepsilon < \frac{\delta}{n b^{n-1}}.$$

An application of inequality (4.6.2) with $s = b$ then leads to

$$(b - \varepsilon)^n \geq b^n - n\varepsilon b^{n-1} > b^n - n \cdot \frac{\delta}{n b^{n-1}} \cdot b^{n-1} = b^n - \delta = x.$$

Hence, by property (2) the number $b - \varepsilon$ is also an upper bound of A , as we claimed.

Proof of claim (4): Here we have to show the following: If $b^n < x$, then there is $a \in A$ such that $b < a$, i.e., b cannot be an upper bound of A . Equivalently, we have to verify the existence of $\varepsilon > 0$ for which $(b + \varepsilon)^n < x$. Indeed, then $a := b + \varepsilon$ belongs to A and, moreover, by $\varepsilon > 0$ it follows that $b < a$.

Set $\delta := x - b^n > 0$ and choose $\varepsilon \leq 1$ for which

$$0 < \varepsilon < \frac{\delta}{n(b+1)^{n-1}}.$$

Now let us apply inequality (4.6.1) with $s = b$. Doing so we obtain

$$(b + \varepsilon)^n \leq b^n + n \cdot \varepsilon(b + 1)^{n-1} < b^n + n \cdot \frac{\delta}{n(b + 1)^{n-1}} \cdot (b + 1)^{n-1} = b^n + \delta = x.$$

By the definition of the set A it follows that $b + \varepsilon \in A$, and this completes the proof of property (4).

As mentioned above, these four properties imply that $w = \sup(A)$ exists and satisfies $w^n = x$.

It remains to prove the uniqueness of the number $w > 0$ for which $w^n = x$. Thus, let us assume that there exists still another $z > 0$ for which $z^n = x$. If $z \neq w$, then either $z < w$ or $w < z$. In the first case an application of Corollary 4.3.3 implies that

$$x = z^n < w^n = x,$$

which, of course, cannot happen. Similarly, $w < z$ leads to a contradiction as well, hence it follows that $w = z$, which proves the uniqueness of the number $w > 0$ with $w^n = x$. This completes the proof of the theorem. ■

Definition 4.6.1. Given a positive number x and an $n \in \mathbb{N}$, the unique number $w > 0$ satisfying $w^n = x$ is said to be the *n*th **root** of x . It is denoted by $w = \sqrt[n]{x}$ or by $w = x^{1/n}$.

$$(\forall x > 0, \forall n \in \mathbb{N}) \left[(w = \sqrt[n]{x}) \Leftrightarrow (w = x^{1/n}) \Leftrightarrow (w^n = x) \right]$$

Let us state some easy properties of roots.

Proposition 4.6.4.

(1) *For any two positive numbers x and y it follows that*

$$x < y \Leftrightarrow x^{1/n} < y^{1/n}.$$

In particular, since $1^{1/n} = 1$ this implies

$$x > 1 \Leftrightarrow x^{1/n} > 1 \quad \text{and} \quad 0 < x < 1 \Leftrightarrow 0 < x^{1/n} < 1.$$

(2) *For all $x > 1$ it follows that*

$$(4.6.3) \quad 0 < x^{1/n} - 1 \leq \frac{x - 1}{n}.$$

Proof: The first assertion is a direct consequence of Corollary 4.3.3.

To prove estimate (4.6.3) we apply the Bernoulli inequality (cf. Exercise 1.2.5) to $x^{1/n} - 1$. Then we get

$$x = (1 + (x^{1/n} - 1))^n \geq 1 + n(x^{1/n} - 1),$$

which easily implies (4.6.3). ■

Definition 4.6.2. Let $x > 0$ be an arbitrary real number. If $n \in \mathbb{N}$, then

$$x^{-1/n} := (x^{-1})^{1/n} = \frac{1}{\sqrt[n]{x}}.$$

More general, for all $n, m \in \mathbb{Z}$ with $n \neq 0$ we set

$$x^{m/n} := (x^m)^{1/n}.$$

In other words, $x^{m/n}$ is the unique number $y > 0$ such that $y^n = x^m$.

Let us state some easy properties of this construction.

Proposition 4.6.5. (1) For all $x > 0$ and integers n, m with $n \neq 0$ we have

$$x^{m/n} = (x^{1/n})^m.$$

(2) Whenever $\frac{m_1}{n_1} = \frac{m_2}{n_2}$, then this implies

$$x^{m_1/n_1} = x^{m_2/n_2}.$$

Proof: To prove the first assertion we have to verify that

$$(x^m)^{1/n} = (x^{1/n})^m \Leftrightarrow x^m = ((x^{1/n})^m)^n.$$

But this easily follows by

$$((x^{1/n})^m)^n = (x^{1/n})^{mn} = ((x^{1/n})^n)^m = x^m.$$

In order to prove the second property we must show that

$$x^{m_1/n_1} = x^{m_2/n_2} \Leftrightarrow x^{m_1} = (x^{m_2/n_2})^{n_1} = (x^{1/n_2})^{m_2 n_1}.$$

Since by assumption $m_2 n_1 = m_1 n_2$ this is a consequence of

$$(x^{1/n_2})^{m_2 n_1} = (x^{1/n_2})^{m_1 n_2} = ((x^{1/n_2})^{n_2})^{m_1} = x^{m_1}.$$

■

In view of the second property the following definition makes sense.

Definition 4.6.3. Given $x > 0$ and a rational number $q = m/n$ with integers m and $n \neq 0$. Then we set

$$x^q := x^{m/n}.$$

Example 4.6.1. In this setting we have

$$2^{3/2} = (\sqrt{2})^3 = \sqrt{2^3} = 2^{9/6} = (\sqrt[6]{2})^9.$$

Let us state some properties of powers with respect to rational numbers Since the proofs follow by standard arguments, we leave them as an exercise.

Proposition 4.6.6. Let x be a positive number. Then for all rational numbers p and q the following are valid:

(1) If $0 \leq p < q$, then

$$\begin{aligned} x^q &< x^p & \text{if } & 0 < x < 1 \\ x^p &< x^q & \text{if } & 1 < x < \infty \end{aligned}$$

(2) For all $p, q \in \mathbb{Q}$ we have

$$x^p \cdot x^q = x^{p+q} \quad \text{and} \quad (x^p)^q = x^{pq}.$$

(3) For all $x, y > 0$ it follows that

$$x^q \cdot y^q = (xy)^q.$$

It remains to answer the question of how x^α is defined for irrational numbers α . For example, what are $2^{\sqrt{2}}$ or 3^π ? There are different possibilities to treat this question. For example, when logarithm and exponential functions are already available, one may set

$$x^\alpha := \exp(\alpha \log x).$$

Another possible approach is by using sequences. One approximates the irrational number α by a sequence $(q_n)_{n \geq 1}$ of rational numbers and defines x^α as limit of x^{q_n} . The crucial point here is to show that the limit does not depend on the chosen approximating sequence $(q_n)_{n \geq 1}$. Finally, we may define arbitrary powers also as follows:

If $1 < x < \infty$, set

$$A_\alpha := \{x^q : q < \alpha, q \in \mathbb{Q}\}.$$

Then A_α is a nonempty set which is bounded above. Then we set

$$x^\alpha = \sup(A_\alpha).$$

In the case $0 < x < 1$ one has to modify the set A_α slightly as follows.

$$A_\alpha := \{x^q : \alpha < q, q \in \mathbb{Q}\}.$$

The problem of this construction is to verify that this general power function possesses exactly the same properties as stated in Proposition 4.6.6. Some proofs of these properties are quite technical.

Example 4.6.2. We have

$$2^{\sqrt{2}} = \sup\{2^{m/n} : m, n \in \mathbb{N}, m/n < \sqrt{2}\} = \sup\{2^{m/n} : m, n \in \mathbb{Z}, m^2 < 2n^2\}.$$

For example, if $n = 100$ and $m = 141$, then

$$m^2 = 19,881 < 20,000 = 2n^2,$$

hence

$$2^{\sqrt{2}} > 2^{m/n} = 2^{1.41}.$$

Exercise 4.6.1. Given natural numbers m and n , verify that for any positive real number x

$$x^{1/n} \cdot x^{1/m} = x^{1/n+1/m} \quad \text{and} \quad x^{1/n} \cdot x^{-1/m} = x^{1/n-1/m}.$$

Exercise 4.6.2. Which of the following equations are correct for all $n, m \in \mathbb{N}$? Which ones are false? Justify your answers.

- | | |
|--|---|
| (a) $\sqrt{2^n} = 2^n$ or $\sqrt{2^n} = 2^{n/2}$ | (b) $\sqrt[n]{2^n} = \sqrt[n]{2}$ |
| (c) $2^n + 2^{n-1} = 3 \cdot 2^{n-1}$ | (d) $2^n + 2^{n-2} = 5 \cdot 2^{n-2}$ |
| (e) $2^n - 2^{n-1} = 2^{n-1}$ | (f) $\sqrt[n]{\sqrt[n]{2}} = \sqrt[n^2]{2}$ or $\sqrt[n]{\sqrt[n]{2}} = \sqrt[2n]{2}$ |
| (g) $2^n + 3^n = 3^n \left(1 + \left(\frac{2}{3}\right)^n\right)$ | (h) $(2^n)^{-1} = \left(\frac{1}{2^{n/2}}\right)^2$ |
| (i) $\frac{2^n}{2^{n-3}} = 8$ or $\frac{2^n}{2^{n-3}} = \frac{1}{8}$ | (j) $(2^n)^n = 2^{n^2}$ or $(2^n)^n = 2^{2n}$ |
| (k) $(\sqrt{2})^3 + \sqrt{8} = \sqrt{32}$ | (l) $(2^n)^m \cdot (2^m)^n = 4^{mn}$ |
| (m) $2^{n+3} - 2^{n+1} = 6 \cdot 2^n$ | (n) $2^{-n} - 2^{-n-1} = 2^{-n-1}$ |
| (o) $\frac{2^n}{2^{n-1}} = 2$ or $\frac{2^n}{2^{n-1}} = \frac{1}{2}$ | (p) $(2^5)^{1/4} = (\sqrt[8]{2})^{10} = (\sqrt{2})^{5/2}$ |
| (q) $\underbrace{2^n \cdot \dots \cdot 2^n}_{n \text{ times}} = 2^{n^2}$ | (r) $\underbrace{3^{-n} \cdot \dots \cdot 3^{-n}}_{m \text{ times}} = 3^{-mn}$ |

Exercise 4.6.3. For each $\alpha \in \mathbb{R}$ the function $f_\alpha : \mathbb{R}_+ \rightarrow \mathbb{R}_+$ is defined by

$$f_\alpha(x) = x^\alpha, \quad x > 0.$$

Then f_α is said to be a **power function** with exponent $\alpha \in \mathbb{R}$. Prove the following (see Section A.3 for necessary notations and definitions).

(1) If x and y are two positive real numbers, then

$$f_\alpha(x \cdot y) = f_\alpha(x) \cdot f_\alpha(y).$$

(2) For all $\alpha, \beta \in \mathbb{R}$ it follows that¹⁰

$$f_\alpha \cdot f_\beta = f_{\alpha+\beta}.$$

(3) For all $\alpha, \beta \in \mathbb{R}$ one has

$$f_\alpha \circ f_\beta = f_\beta \circ f_\alpha = f_{\alpha \cdot \beta}.$$

(4) For all $\alpha \in \mathbb{R}$ follows that

$$f_\alpha \circ f_1 = f_1 \circ f_\alpha = f_\alpha.$$

(5) If $\alpha \neq 0$, then f_α is a bijection from \mathbb{R}_+ to \mathbb{R}_+ with inverse function

$$f_\alpha^{-1} = f_{\alpha^{-1}}.$$

Exercise 4.6.4. Given $a > 0$, the function E_a from \mathbb{R} to \mathbb{R}_+ with

$$(4.6.4) \quad E_a(x) = a^x, \quad x \in \mathbb{R},$$

is called an **exponential function**¹¹ with base $a > 0$.

¹⁰As usual the product of two real-valued functions f and g is defined by $f \cdot g : x \mapsto f(x) \cdot g(x)$.

¹¹Some authors define only the function $E_e(x) = e^x$ with Euler number e as exponential one.

Prove the following properties of E_a .

- (1) The function E_a is increasing if $a > 1$ and decreasing in the case $0 < a < 1$.
- (2) For all $x, y \in \mathbb{R}$ it follows that

$$E_a(x + y) = E_a(x) \cdot E_a(y),$$

- (3) For any $\beta \in \mathbb{R}$ and each $x > 0$ one has

$$E_a(\beta x) = E_a(x)^\beta.$$

Exercise 4.6.5. Let a be a positive real number that is not equal to 1. Consider the exponential function $E_a : \mathbb{R} \rightarrow \mathbb{R}_+$ defined by (4.6.4). Prove that E_a is a bijection and, therefore, it possesses an inverse function $E_a^{-1} : \mathbb{R}_+ \rightarrow \mathbb{R}$. Show that this inverse function satisfies

$$E_a^{-1}(1) = 0 \quad \text{and} \quad E_a^{-1}(a) = 1.$$

$$E_a^{-1}(x \cdot y) = E_a^{-1}(x) + E_a^{-1}(y), \quad x, y > 0.$$

$$E_a^{-1}(x^\beta) = \beta \cdot E_a^{-1}(x), \quad x > 0, \beta \in \mathbb{R}.$$

The inverse function $E_a^{-1} : \mathbb{R}_+ \rightarrow \mathbb{R}$ is usually called **the logarithm in base a** and is denoted by \log_a . In the case $a = e$ it is common to write \log instead of \log_e . Some authors also denote it by \ln .

Exercise 4.6.6. Solve the following equations:

- (1) $2^x + 2^{x+2} = 5 \cdot 2^{x+1}$.
- (2) $5^{x-2} - 5^{\frac{2x-1}{3}} + 5^{\frac{x+1}{3}} = 1$.
- (3) $\log_3(x+1) + \log_3(x-1) = 2 \log_9(8)$.
- (4) $\log_2(x) + \log_4(x) = \frac{3}{2}$.

Exercise 4.6.7. Prove the following inequalities:

- (1) $\log_{10}(2) > 0.3, \log_{10}(3) < 0.48$.
- (2) $\frac{20}{13} < \log_6 17 < \frac{9}{5}$.
- (3) $\log_3(4) < 3/2 < \log_2(3)$.
- (4) $\log_{10}^2(19) > 2 \log_{10} 6$.

Exercise 4.6.8. Let a, b, c , and x be some real numbers in the interval $(1, +\infty)$. Prove that $\log_a x, \log_b x$, and $\log_c x$ are consecutive terms of a geometric progression if and only if $\log_a b = \log_b c$.

Exercise 4.6.9. If a and b are real numbers in the interval $(1, +\infty)$, prove the following inequalities:

- (1) $\log_{a+1}(b) < \log_a(b) < \log_a(b+1)$.
- (2) $\log_a(b) + \log_b(a) \geq 2$.

Exercise 4.6.10. If a, x , and y are positive real numbers, prove that

- (1) $a^x + a^y \geq 2a^{\frac{x+y}{2}}$.
- (2) $x^x \cdot y^y > x^y \cdot y^x$.

4.7. Expansion of Real Numbers

We investigated in Chapter 3.3 representations of rational numbers as b -fractions. The aim of this section is to extend these results to the larger class of real numbers. Given $x \in \mathbb{R}$, we may write x as the sum of its integer and fractional part, i.e.,

$$x = [x] + x' \quad \text{where} \quad x' = x - [x].$$

Consequently, it suffices to ask for the fractional representation of real numbers $0 \leq x < 1$.

Definition 4.7.1. Fix a base $b \geq 2$. A real number $x \in [0, 1)$ admits an expansion as b -fraction provided that there are integers $x_j \in \{0, \dots, b-1\}$ such that for all $n \geq 1$ it follows that

$$\sum_{j=1}^n \frac{x_j}{b^j} \leq x < \sum_{j=1}^n \frac{x_j}{b^j} + \frac{1}{b^n}.$$

In this case we shall write

$$x =_b 0.x_1 x_2 \dots .$$

Looking carefully into the proof of Proposition 3.3.2 one observes that we never used the fact that the number under consideration is rational. The only reason for restricting ourselves to rational numbers was that real numbers were not available yet. So, in fact we proved the following more general result than Proposition 3.3.2

Proposition 4.7.1. Let $b \geq 2$ be a fixed base. Given a real number $x \in [0, 1)$. Then x admits a unique expansion as b -fraction, which may be either finite or infinite. The finite case occurs if and only if x is a b -rational number.

Remark 4.7.1. The rules for evaluating the integers x_1, x_2, \dots are exactly as presented in the proof of Proposition 3.3.2:

$$\begin{aligned} x_1 &= [bx], & r_1 &= bx - x_1 \\ x_2 &= [br_1], & r_2 &= br_1 - x_2 \\ x_3 &= [br_2], & r_3 &= br_2 - x_3 \end{aligned}$$

and so on.

A basic question remains unanswered: Which sequences x_1, x_2, \dots of integers may occur in the representation of real numbers as b -fractions? First note that we can only expect admissible sequences as introduced in Definition 3.3.8. In the slightly more general setting of infinite sequences this has to be understood as follows:

Definition 4.7.2. An infinite sequence x_1, x_2, \dots of integers in $\{0, \dots, b-1\}$ is said to be **admissible** if neither from a certain point all $x_j = 0$ nor do we have from a certain point always $x_j = b-1$. In other words, nonadmissible sequences are those which for some $m \geq 0$ can be written either as

$$x_1, \dots, x_m, 0, 0, 0, \dots \quad \text{or as} \quad x_1, \dots, x_m, b-1, b-1, \dots .$$

Proposition 4.7.2. Suppose we are given an infinite admissible sequence x_1, x_2, \dots of integers in $\{0, \dots, b - 1\}$. Then there is a (unique) real number $x \in [0, 1)$ with $x =_b 0.x_1x_2\dots$

Proof: Set

$$S_n := \sum_{j=1}^n \frac{x_j}{b^j}, \quad n = 1, 2, \dots$$

Of course, the S_n 's are nondecreasing, i.e.,

$$0 \leq S_1 \leq S_2 \leq \dots,$$

and, moreover, if $n < m$ by $x_j \in \{0, \dots, b - 1\}$ we also get

$$(4.7.1) \quad S_m - S_n = \sum_{j=n+1}^m \frac{x_j}{b^j} \leq \sum_{j=n+1}^m \frac{b-1}{b^j} = (b-1) \sum_{j=n+1}^m b^{-j} = \frac{1}{b^n} - \frac{1}{b^m}.$$

Define the subset $A \subset \mathbb{R}$ by

$$A := \{S_n : n \geq 1\}.$$

The set A is nonempty and because of

$$S_n \leq \sum_{j=1}^n \frac{b-1}{b^j} = 1 - \frac{1}{b^n} < 1$$

bounded above by 1. Consequently,

$$x := \sup(A)$$

is a well-defined real number in $[0, 1]$. We claim now that this supremum is the desired real number with expansion $x =_b 0.x_1x_2\dots$. To verify this we have to show the following

$$(4.7.2) \quad (\forall n \geq 1) \left(S_n \leq x < S_n + \frac{1}{b^n} \right).$$

The left-hand estimate $S_n \leq x$ is true by trivial reason. Indeed, x is an upper bound of A , hence by the definition of the set A bigger than all S_n 's.

The right-hand estimate in (4.7.2) is less obvious and its proof needs some preparation. Fix $n \geq 1$. Next we use that the sequence of the x_j 's is admissible. This implies the existence of an $k > n$ (depending on n) such that $x_k \leq b - 2$. Arguing as in (4.7.1), for any $m \geq k$ we obtain

$$\begin{aligned} S_m - S_n &= \sum_{j=n+1}^m \frac{x_j}{b^j} \leq \sum_{j=n+1: j \neq k}^m \frac{b-1}{b^j} + \frac{b-2}{b^k} \\ &= \sum_{j=n+1}^m \frac{b-1}{b^j} - \frac{1}{b^k} = \frac{1}{b^n} - \frac{1}{b^m} - \frac{1}{b^k} < \frac{1}{b^n} - \frac{1}{b^k}. \end{aligned}$$

In other words, if m is sufficiently large, i.e., if $m \geq k$ for some k depending on n , it follows that

$$(4.7.3) \quad S_m + \frac{1}{b^k} < S_n + \frac{1}{b^n}.$$

Next we use that $x = \sup(A)$. Thus, making x a bit smaller, there exists an element in A which exceeds this lesser supremum x . Let $k > n$ be as before and consider $x - \varepsilon$ where

$$\varepsilon = \frac{1}{2b^k}.$$

Then there exists m with $S_m > x - \varepsilon$. Since the S_m 's are nondecreasing we may assume that m is large, bigger than the existing $k > n$. Combining this with (4.7.3) leads to

$$x < S_m + \varepsilon < S_n + \frac{1}{b^n} + \varepsilon - \frac{1}{b^k} = S_n + \frac{1}{b^n} + \frac{1}{2b^k} - \frac{1}{b^k} = S_n + \frac{1}{b^n} - \frac{1}{2b^k} < S_n + \frac{1}{b^n}.$$

This proves the right-hand estimate of (4.7.2) and completes the proof. \blacksquare

Remark 4.7.2. The proof of Proposition 4.7.2 could be considerably simplified by using properties of infinite sums. In this setting the number $x =_b x_1 x_2 \dots$ is nothing other than the infinite sum $\sum_{k=1}^{\infty} \frac{x_k}{b^k}$. But at this point those sums as well as their properties are not available yet. To overcome this difficulty we were forced to define $x \in \mathbb{R}$ as supremum of the set A .

Combining the previous results with those of Section 3.3 we obtain the following general statement.

Theorem 4.7.3. Fix a base $b \geq 2$. Each real number in $x \in [0, 1)$ admits a unique expansion as b -fraction.

- (1) The representation is finite if and only if x is b -rational.
- (2) The representation is infinite and periodic if and only if x is rational but not b -rational.
- (3) The representation is infinite and nonperiodic if and only if x is irrational.

Conversely, given any admissible sequence $(x_j)_{j \geq 1}$ of integers in $\{0, \dots, b-1\}$, then there exists a unique real number $x \in [0, 1)$ with

$$x =_b 0.x_1 x_2 \dots$$

Because of its importance let us state the binary case $b = 2$ separately.

Theorem 4.7.4. Any real number in $[0, 1)$ may be represented by an admissible sequence of 0's and 1's which is either finite or infinite periodic or infinite nonperiodic. This relation between numbers and sequences is one-to-one.

Example 4.7.1. The first 50 digits in the expansion of the irrational number $\sqrt{2}$ are

$$\sqrt{2} = 1.4142135623730950488016887242096980785696718753769 \dots$$

The first 50 digits in the expansion of the Euler number e are

$$e = 2.7182818284590452353602874713526624977572470937000 \dots$$

Both expansions are infinite and nonperiodic.

Example 4.7.2. There exists a real number $x \in [0, 1]$ with dyadic expansion

$$x =_2 0.10110111011110111110 \dots .$$

The number x is irrational because this expansion is nonperiodic. Its approximate value as a decimal fraction is $0.716733932495117\dots$. Note that this decimal expansion is nonperiodic as well.

Finally, let us give a geometric interpretation of the expansion into b -fractions. This is a direct consequence of the way how we defined those fractions.

Proposition 4.7.5. Fix a base $b \geq 2$. Then the following is valid.

- (a) For each $n \geq 1$ the interval $[0, 1)$ is divided into the following b^n disjoint half-open intervals, each of length $1/b^n$.

$$(4.7.4) \quad I_{a_1, \dots, a_n} = \left[\sum_{j=1}^n \frac{a_j}{b^j}, \sum_{j=1}^n \frac{a_j}{b^j} + \frac{1}{b^n} \right), \quad 0 \leq a_j < b.$$

The left as well as the right endpoints of these half-open intervals are either b -rational numbers in $[0, 1)$ or it is 1, which is the right endpoint of the most right interval¹².

This happens if $a_1 = a_2 = \dots = a_n = b - 1$.

- (b) A real number $x \in [0, 1)$ belongs to I_{a_1, \dots, a_n} if and only if its first n digits in the expansion as b -fraction are a_1, \dots, a_n .

$$x \in I_{a_1, \dots, a_n} \Leftrightarrow x =_b 0.a_1 \dots a_n x_{n+1} x_{n+2} \dots,$$

where the x_j s denote any integers in $\{0, \dots, b - 1\}$, with the exclusion that all of them equal $b - 1$. The case that all x_j are zero corresponds to the left endpoint of I_{a_1, \dots, a_n} .

- (c) The b^{n+1} disjoint intervals of length $1/b^{n+1}$ are generated by dividing each interval of level n into b disjoint intervals of length $1/b^{n+1}$.

$$I_{a_1, \dots, a_n} = \bigcup_{j=0}^{b-1} I_{a_1, \dots, a_n, j}$$

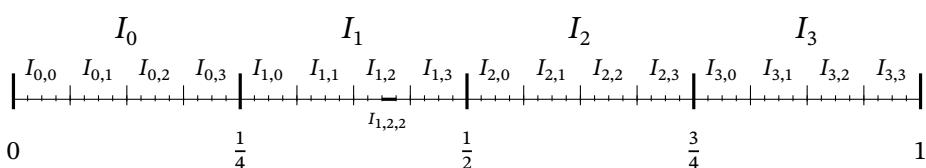


Figure 4.7.1. The first three divisions of $[0, 1)$ in the case of $b = 4$. At level 3 there are 64 intervals each of length $1/64$ denoted by $I_{0,0,0}$ to $I_{3,3,3}$.

We conclude this part by an important example resting upon the fractional expansion of real numbers. We are going to introduce the Cantor set, a subset of $[0, 1]$, possessing many surprising and interesting properties.

¹²If $a_n < b - 1$, then the right endpoint may also be expanded as $0.a_1 \dots a_{n-1}a_n + 1$

Example 4.7.3 (Cantor Set). One possible way to construct the Cantor set is as follows: According to Theorem 4.7.3 one may expand each number in $[0, 1)$ as fraction with respect to base 3 (called ternary expansion). That is, each $x \in [0, 1)$ admits either a finite representation (which happens if x is 3-rational) as

$$x =_3 0.x_1 \dots x_n \quad \text{with} \quad x_j \in \{0, 1, 2\} \text{ and } x_n \neq 0,$$

or an infinite representation

$$x =_3 0.x_1 x_2 x_3 \dots \quad \text{with} \quad x_j \in \{0, 1, 2\}.$$

with an admissible sequence $(x_j)_{j \geq 1}$. That is, we do not have $x_j = 2$ for some integer $m \geq 1$ and $j > m$. But in contrast to our former agreements we will add zeroes in the case of finite expansions. This turns out to be very useful during our next considerations. Thus, we will write $1/3 =_3 0.1000\dots$ but, of course, remember that in fact $1/3 =_3 0.1$.

The basic idea for the definition of the Cantor set is to sort out numbers whose ternary expansions contain the digit 1. This is easy to do if the investigated number admits an infinite ternary expansion. But it causes some technical difficulties for numbers in $[0, 1]$ with finite ternary expansion. For example, $1/3 =_3 0.100\dots$, thus its expansion contains an 1 although it is very desirable that $1/3$ belongs to the Cantor set. A careful look at the expansion of $1/3$ tells us that the digit 1 occurs at the very end of the ordinary expansion. And this is the crucial property we will use in the description of the Cantor set.

Definition 4.7.3. The Cantor set \mathcal{C} consists of numbers $x \in [0, 1]$ possessing one of the two following properties: Let x be expanded as

$$x =_3 0.x_1 x_2 \dots \quad \text{where} \quad x_j \in \{0, 1, 2\}.$$

Either all digits of x are different of 1, i.e., $x_j \neq 1$ for all $j \geq 1$, or there is exactly one of its digits equal to 1. But in this case it has to be the very last nonzero one. That is, for some $n \geq 1$ and certain $x_1, \dots, x_{n-1} \in \{0, 2\}$ we have

$$x =_3 0.x_1 \dots x_{n-1} 100000\dots$$

Of course, the latter case is only possible if x is 3-rational. Moreover, we add $\{1\}$ to \mathcal{C} because the number 1 is not covered by the definition. Written as formula this reads as follows:

$$\begin{aligned} \mathcal{C} &= \{x \in [0, 1] : x =_3 0.x_1 x_2 \dots \text{ s.t. } \forall j \geq 1, x_j \neq 1\} \\ &\cup \{x \in [0, 1] : \exists n \geq 1, \exists x_j \neq 1, x =_3 0.x_1 \dots x_{n-1} \underbrace{1}_{n} 000\dots\} \cup \{1\}. \end{aligned}$$

Let us state some typical examples.

$$\begin{aligned} \frac{1}{4} &\in \mathcal{C} \quad \text{because of} \quad \frac{1}{4} =_3 0.0202\dots =_3 0.\overline{02}. \\ \frac{1}{26} &\notin \mathcal{C} \quad \text{because of} \quad \frac{1}{26} =_3 0.001001\dots =_3 0.\overline{001}. \\ \frac{8}{27} &\in \mathcal{C} \quad \text{because of} \quad \frac{8}{27} =_3 0.022 =_3 0.022000\dots. \\ \frac{7}{9} &\in \mathcal{C} \quad \text{because of} \quad \frac{7}{9} =_3 0.21 =_3 0.210000\dots. \end{aligned}$$

Other examples of numbers in \mathcal{C} are

$$0.1 =_3 \frac{1}{3}, \quad 0.01 =_3 \frac{1}{9}, \quad 0.\overline{202} =_3 \frac{10}{13}, \quad 0.0201 =_3 \frac{19}{81}.$$

How can \mathcal{C} be visualized geometrically? One possible way is to describe the set of those $x \in [0, 1]$ which do **not** belong to \mathcal{C} .

Proposition 4.7.6. *A number $x \in [0, 1)$ does **not** belong to \mathcal{C} if and only if there is an $n \geq 1$ and there are $a_1, \dots, a_{n-1} \in \{0, 2\}$ such that*

$$(4.7.5) \quad \sum_{j=1}^{n-1} \frac{a_j}{3^j} + \frac{1}{3^n} < x < \sum_{j=1}^{n-1} \frac{a_j}{3^j} + \frac{2}{3^n}.$$

In the case $n = 1$ the above estimate has to be understood as

$$\frac{1}{3} < x < \frac{2}{3}.$$

Proof: Suppose first $x \notin \mathcal{C}$. Then, if $x =_3 0.x_1x_2\dots$, there exists at least one $n \geq 1$ with $x_n = 1$ and $x_j \neq 1$ if $1 \leq j < n$. In other words, the first occurrence of the digit 1 happens at position n . An application of property (b) in Proposition 4.7.5 with $a_1 = x_1, a_{n-1} = x_{n-1}$ and $a_n = 1$ implies

$$(4.7.6) \quad \sum_{j=1}^{n-1} \frac{x_j}{3^j} + \frac{1}{3^n} \leq x < \sum_{j=1}^{n-1} \frac{x_j}{3^j} + \frac{2}{3^n}.$$

Moreover, all $x_j, 1 \leq j < n$, are different of 1 because the digit 1 occurs at position n for the first time. Thus, choosing $a_1 = x_1, \dots, a_{n-1} = x_{n-1}$, we almost got the validity of (4.7.5). It remains to show that x does not coincide with the left-hand value in (4.7.6). Assume for a moment this would be so. That is, we suppose that

$$x = \sum_{j=1}^{n-1} \frac{x_j}{3^j} + \frac{1}{3^n} =_3 0.x_1\dots x_{n-1}1000\dots$$

By definition, this implies $x \in \mathcal{C}$ which contradicts the choice of x . Consequently, x cannot coincide with the left-hand expression in (4.7.6), hence (4.7.5) is satisfied with $a_j = x_j, 1 \leq j < n$.

Suppose now conversely that $x \in [0, 1]$ satisfies (4.7.5) with certain integers $a_j \in \{0, 2\}$. Another application of property (b) in Proposition 4.7.5 implies

$$x =_3 0.a_1\dots a_{n-1}\underbrace{1}_n x_{n+1}\dots$$

with certain x_j s in $\{0, 1, 2\}$, $j > n$. Moreover, not all the x_j s can be zero, because otherwise

$$x = \sum_{j=1}^{n-1} \frac{a_j}{3^j} + \frac{1}{3^n},$$

contradicting (4.7.5). Thus, the n th digit in the expansion of x equals 1 and, moreover, there is at least one $j > n$ with $x_j \neq 0$. This tells us that $x \notin \mathcal{C}$ and completes the proof. \blacksquare

Let us analyze the previous characterization of the complementary set of \mathcal{C} . To this end set

$$D_1 = \left(\frac{1}{3}, \frac{2}{3} \right) = \{x \in [0, 1] : x =_3 0.1 \times \times \times \dots\},$$

$$D_2 = \left(\frac{1}{9}, \frac{2}{9} \right) \cup \left(\frac{7}{9}, \frac{8}{9} \right) = \{x \in [0, 1] : x =_3 0.a_1 1 \times \times \times \dots, a_1 = 0 \text{ or } a_1 = 2\}$$

where “ \times ” means that at these positions may occur any integers in $\{0, 1, 2\}$, but not all of them equal 0 nor all equal 2 from a certain position.

For $n \geq 1$, we define D_n by

$$D_n = \bigcup_{a_1, \dots, a_{n-1} \in \{0, 2\}} \{x \in [0, 1] : x =_3 0.a_1 \dots a_{n-1} 1 \times \times \times \dots\},$$

with the same restrictions on the integers at the positions “ \times ” as before in the case of D_1 and D_2 , respectively.

Which properties do these sets possess?

- (1) By the construction the sets D_1, D_2, \dots are disjoint. That is, for all $n \neq m$ it follows that $D_n \cap D_m = \emptyset$.
- (2) The set D_n is the union of 2^{n-1} open intervals. These intervals have left and right endpoints (hereby a_1, \dots, a_n run through all choices of 0 and 2)

$$\sum_{j=1}^{n-1} \frac{a_j}{3^j} + \frac{1}{3^n} =_3 0.a_1 a_2 \dots a_{n-1} 1 \quad \text{and} \quad \sum_{j=1}^{n-1} \frac{a_j}{3^j} + \frac{2}{3^n} =_3 0.a_1 a_2 \dots a_{n-1} 2.$$

Moreover, these 2^n endpoints (left and right ones) belong to the Cantor set. For example, if $n = 3$, the left and right endpoints are

$$\frac{1}{27}, \quad \frac{2}{27}, \quad \frac{7}{27}, \quad \frac{8}{27}, \quad \frac{19}{27}, \quad \frac{20}{27}, \quad \frac{25}{27}, \quad \frac{26}{27}.$$

- (3) The length of each of the 2^{n-1} intervals generating D_n equals $1/3^n$. So the overall size of D_n (the sum of the lengths over all intervals generating D_n) equals

$$|D_n| = \sum_{a_1, \dots, a_{n-1} \in \{0, 2\}} \left| \left(\sum_{j=1}^{n-1} \frac{a_j}{3^j} + \frac{1}{3^n}, \sum_{j=1}^{n-1} \frac{a_j}{3^j} + \frac{2}{3^n} \right) \right| = \frac{2^{n-1}}{3^n}.$$

Thus, if n increases, the overall size of D_n becomes smaller and smaller.

To formulate the main property of the D_n let us introduce the sets

$$C_n := [0, 1] \setminus \left(\bigcup_{j=1}^n D_j \right).$$

So, C_n denotes the set of those numbers in $[0, 1]$ which remained after we took off D_1 to D_n . For example,

$$C_1 = \left[0, \frac{1}{3}\right] \cup \left[\frac{2}{3}, 1\right],$$

$$C_2 = \left[0, \frac{1}{9}\right] \cup \left[\frac{2}{9}, \frac{1}{3}\right] \cup \left[\frac{2}{3}, \frac{7}{9}\right] \cup \left[\frac{8}{9}, 1\right],$$

$$C_3 = \left[0, \frac{1}{27}\right] \cup \left[\frac{2}{27}, \frac{1}{9}\right] \cup \left[\frac{2}{9}, \frac{7}{27}\right] \cup \left[\frac{8}{27}, \frac{1}{3}\right] \cup \left[\frac{2}{3}, \frac{19}{27}\right] \cup \left[\frac{20}{27}, \frac{7}{9}\right] \cup \left[\frac{8}{9}, \frac{25}{27}\right] \cup \left[\frac{26}{27}, 1\right].$$

The following theorem is a direct consequence of Proposition 4.7.6.

Theorem 4.7.7. *The Cantor set \mathcal{C} can be written as follows:*

$$\mathcal{C} = \bigcap_{n=1}^{\infty} C_n = \bigcap_{n=1}^{\infty} \left[[0, 1] \setminus \left(\bigcup_{j=1}^n D_j \right) \right].$$

Remark 4.7.3. The Cantor set \mathcal{C} contains on one hand all left and right endpoints of withdrawn intervals (intervals generating the D_n for each $n \geq 1$). Besides, there are **infinite** points in \mathcal{C} . These are numbers admitting an infinite ternary expansion consisting only of 0 and 2. For example, such a number is the irrational one

$$x =_3 0.202002000200002\ldots.$$

Another examples are the rational numbers

$$y =_3 0.22\overline{020} =_3 \frac{107}{117} \quad \text{and} \quad \frac{1}{4} =_3 0.\overline{02}.$$

On the other hand, endpoints different of 0 and 1 are those which have for some $n \geq 1$ and certain integers $a_1, \dots, a_{n-1} \in \{0, 2\}$ either a ternary expansion as

$$0.a_1 \dots a_{n-1} 1 =_3 \sum_{j=1}^{n-1} \frac{a_j}{3^j} + \frac{1}{3^n} \quad \text{or one as} \quad 0.a_1 \dots a_{n-1} 2 =_3 \sum_{j=1}^{n-1} \frac{a_j}{3^j} + \frac{2}{3^n}.$$

Examples of this type are

$$x =_3 0.02020221 \quad \text{or} \quad y =_3 0.02222002202.$$

Remark 4.7.4. The geometric construction of \mathcal{C} can be done as follows:

- *Step 1:* Divide $[0, 1]$ into three intervals of equal length. Cut out the open middle interval D_1 . The remaining set is C_1 .
- *Step 2:* The set C_1 left after the first step consists of two intervals of length $1/3$. Divide each of these two intervals into three intervals of length $1/9$. In each of the two intervals take off the open interval in the middle (i.e., after D_1 now, we take off D_2).
- *Step n:* Assume we already cut out the sets D_1 to D_n . The remaining set $C_n = [0, 1] \setminus (D_1 \cup \dots \cup D_n)$ consists of 2^n closed intervals of length $1/3^n$. Divide each of these 2^n closed intervals into 3 intervals of equal length, then take off in each of the 2^n intervals the (open) one in the middle. The resulting set is C_{n+1} .

- *Step ∞ :* In view of Theorem 4.7.7 we finally end up in

$$C_1 \supseteq C_2 \supseteq C_3 \supseteq \dots \supseteq \mathcal{C} \quad \text{and} \quad \mathcal{C} = \bigcap_{n=1}^{\infty} C_n.$$

Note that at step n we sort out those numbers in $[0, 1]$ which have their first 1 in the ternary expansion at position n . For example, the number $x =_3 0.00212$ belongs to D_4 , thus it is eliminated at step 4. On the other hand, $y =_3 0.2002$ is not in any D_n , hence it is an element of \mathcal{C} .

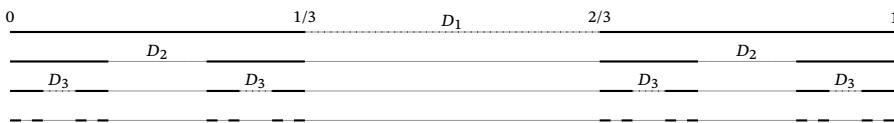


Figure 4.7.2. The first four steps in the construction of the Cantor set \mathcal{C} . The black lines describe the remaining sets C_1 to C_4 , after cutting out D_1 to D_4 , respectively.

As already mentioned, the Cantor set has many remarkable properties. We will state here a few of these; more will follow later on.

The set C_n is the union of 2^n disjoint intervals, each of length $1/3^n$. Hence, the overall length of C_n equals $(2/3)^n$. Now, if n increases to ∞ , then $(2/3)^n$ becomes arbitrarily small. Thus, one can say that the Cantor set has length 0. Does this say that \mathcal{C} is empty? Of course, that is not the case. Besides the endpoints of the withdrawn intervals there are many more elements in \mathcal{C} . Later on, in Proposition A.4.13 we will prove that \mathcal{C} contains uncountable many elements. Is this not a surprising result?

Another interesting property is that \mathcal{C} is a closed set in the sense as stated in Exercise 5.8.19. Moreover, \mathcal{C} is a nowhere dense set (also called a rare set) which says that it does not contain any nonempty open interval¹³.

Proposition 4.7.8. *Let $x \in \mathcal{C}$ be an arbitrary element. Then for each $\varepsilon > 0$ there exist infinitely many numbers $y_k \notin \mathcal{C}$ with*

$$|x - y_k| < \varepsilon, \quad k = 1, 2, \dots$$

In particular, \mathcal{C} does not contain any nonempty open interval.

Proof: We start the proof with an important consequence of property (b) in Proposition 4.7.5: Whenever the first n digits in the ternary expansion of two elements $x, y \in [0, 1]$ coincide, that is,

$$x =_3 0.x_1x_2\dots \quad \text{and} \quad y =_3 0,x_1x_2\dots x_ny_{n+1}\dots,$$

then this implies $|x - y| < \frac{1}{3^n}$. Recall that under this assumption both numbers belong to the same set I_{x_1, \dots, x_n} defined in (4.7.4).

Case 1: There are an $n \geq 1$ as well as $x_1, \dots, x_{n-1} \in \{0, 2\}$ such that

$$x =_3 0.x_1 \dots x_{n-1} \underbrace{1}_{n} 000 \dots$$

¹³Other closed sets with this property are for example \mathbb{N} or $\{1, 1/2, 1/3, \dots\} \cup \{0\}$.

Given $\varepsilon > 0$ choose an $N > n$ satisfying $1/3^N < \varepsilon$. Set

$$y_k =_3 0.x_1 \dots x_{n-1} \underbrace{1}_n 0 \dots 0 \underbrace{1}_{N+k} 00 \dots$$

By the construction, none of the y_k s belongs to \mathcal{C} and, furthermore, the first $N+k-1$ digits of x and y_k coincide. Hence, as we observed before,

$$|x - y_k| \leq \frac{1}{3^{N+k-1}} \leq \frac{1}{3^N} < \varepsilon.$$

Of course, whenever $k \neq \ell$, then $y_k \neq y_\ell$. So we found infinitely many elements in an ε -neighborhood of x , all belonging to $[0, 1] \setminus \mathcal{C}$.

Case 2: Assume that $x \in \mathcal{C}$ admits the ternary expansion

$$x =_3 0, x_1 x_2 \dots \text{ with all } x_j \neq 1.$$

Given $\varepsilon > 0$, we choose an integer $N \geq 1$ such that $1/3^N < \varepsilon$. Then, if $k \geq 1$ set

$$y_k =_3 0, x_1 x_2 \dots x_{N+k-1} \underbrace{1}_{N+k} \underbrace{1}_{N+k+1} x_{N+k+2} \dots$$

Again, all y_k do not belong to \mathcal{C} , and by exactly the same arguments as before it follows that

$$|x - y_k| \leq \frac{1}{3^{N+k-1}} \leq \frac{1}{3^N} < \varepsilon,$$

This completes the proof in this case.

It remains the case $x = 1$. For example, here we may choose $y_k \notin \mathcal{C}$ as follows:

$$y_k =_3 0.222 \dots 2 \underbrace{1}_{N+k} \underbrace{1}_{N+k+1} 000 \dots$$

We leave it as an exercise about infinite series (see Exercise 5.6.10) that also in this case

$$|1 - y_k| < \varepsilon, \quad k = 1, 2, \dots,$$

provided that $N \geq 1$ is chosen sufficiently large. ■

Remark 4.7.5. The previous result suggests that the elements in the Cantor set are somehow isolated points. This is not so. Indeed, it is not difficult to prove that any ε -neighborhood of an $x \in \mathcal{C}$ contains infinitely many different points belonging to \mathcal{C} as well. One may even choose these elements as endpoints of withdrawn intervals; compare Exercise 5.4.10 below.

Example 4.7.4. There exists an interesting two-dimensional analogue of the Cantor set, the so-called **Sierpiński carpet**. It bears the name of the Polish mathematician Waclaw Sierpiński (1882–1969) who constructed this special fractal type set in 1916.

This time the construction starts with the unit square $[0, 1] \times [0, 1]$, not with the unit interval as in the case of the Cantor set. In a first step one divides this unit square into 9 squares of side length $1/3$, then takes off the open square in the middle. We are left with 8 closed squares, each with side length $1/3$.

Next, by the same procedure, we divide each of the 8 remaining squares into 9 squares of equal size, then remove the open center square of side length $1/9$. Now we are left with 64 squares of side length $1/9$.

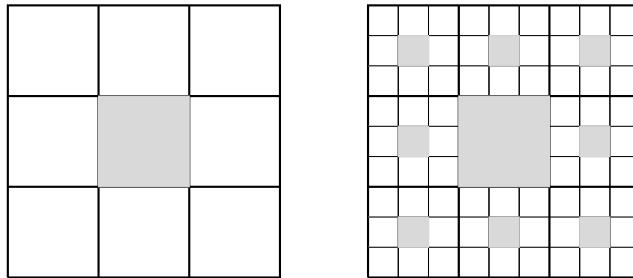


Figure 4.7.3. Step 1 and step 2.

Let us continue in this way as follows: Say in step $n - 1$ we have 8^{n-1} remaining squares of side length $1/3^{n-1}$. Each of these squares is divided into 9 squares of side length $1/3^n$. Take off the middle square. So in step n we eliminate 8^{n-1} squares of side length $1/3^n$. Thus, 8^n squares of side length $1/3^n$ remain. Finally, the Sierpiński carpet is the (infinite) intersection over all points in the unit square which remain after all cuts. Compare Figure 4.7.4 for the first four steps of the construction of the carpet¹⁴.

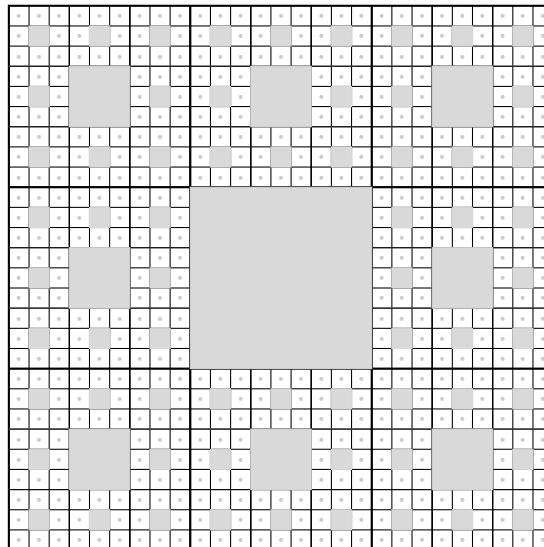


Figure 4.7.4. Construction of the Sierpiński carpet. In the first step, we delete one square of side length $1/3$. In step two, we cut out 8 squares of side length $1/9$. Next, in step three, we remove 64 squares of side length $1/27$. In the fourth step, 256 squares of side length $1/81$ are eliminated.

Finally, after we introduced and thoroughly investigated the field \mathbb{R} of real numbers, let us look back now from a different point of view to the way we went from \mathbb{N}

¹⁴In a commonly known web-dictionary one finds the following information about the usefulness of the Sierpiński carpet: “Mobile phone and Wi-Fi fractal antennas have been produced in the form of few iterations of the Sierpiński carpet. Due to their self-similarity and scale invariance, they easily accommodate multiple frequencies. They are also easy to fabricate and smaller than conventional antennas of similar performance, thus being optimal for pocket-sized mobile phones.”

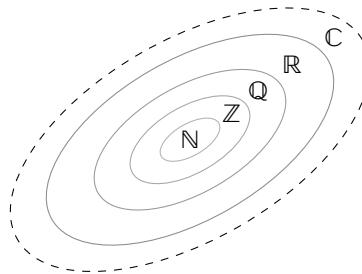


Figure 4.7.5. The field C is that of complex numbers treated in Section 6.

to \mathbb{R} . We started our construction of the real numbers with introducing \mathbb{N} , the set of natural numbers. In a next step we extended it to \mathbb{Z} , the set of integers, by adding (additive) inverse elements to \mathbb{N} . Then we arrived at \mathbb{Q} , the set of rational numbers, by introducing pairs of integers (m, n) with $n \neq 0$, usually written as m/n . And in an (almost final) step we completed \mathbb{Q} and ended up with the set \mathbb{R} of real numbers. So we have (cf. Figure 4.7.5)

$$\mathbb{N} \subset \mathbb{Z} \subset \mathbb{Q} \subset \mathbb{R}.$$

But this is not the whole truth. Indeed, we introduced \mathbb{R} as a set satisfying three types of axioms. This we could have done without knowing anything about natural numbers, about integers or about rational numbers. Thus, we have to answer the question in which way we can rediscover \mathbb{N} , \mathbb{Z} and \mathbb{Q} in the abstract model of real numbers.

Let us start with characterizing the natural numbers as subset of \mathbb{R} . To do so we need the following definition.

Definition 4.7.4. A subset $A \subseteq \mathbb{R}$ is said to be **inductive** provided that it possesses the two following properties

- (1) $1 \in A$
- (2) $(\forall a \in A)(a + 1 \in A)$.

Of course, the set \mathbb{N} is inductive, but also \mathbb{R}_+ or $\left\{1, \frac{3}{2}, 2, \frac{5}{2}, 3, \frac{7}{2}, \dots\right\}$ are so.

Proposition 4.7.9. *The set \mathbb{N} of natural numbers is the smallest inductive subset of \mathbb{R} . That is, whenever $A \subseteq \mathbb{R}$ is inductive, then this implies $\mathbb{N} \subseteq A$.*

Proof: This is an immediate consequence of the induction principle. Suppose that there is an inductive set A for which $\mathbb{N} \not\subseteq A$. Then $S := A \cap \mathbb{N}$ would be an inductive set strictly contained in \mathbb{N} . So we have $1 \in S$ and also $n + 1 \in S$ whenever $n \in S$, hence by the induction principle, $S = \mathbb{N}$ which contradicts the construction of $S = A \cap \mathbb{N}$. ■

Next we want to characterize \mathbb{Z} and \mathbb{Q} as subsets in \mathbb{R} . We will see that these sets are chosen as minimal extensions of \mathbb{N} and \mathbb{Z} , respectively.

Proposition 4.7.10. *The set $\mathbb{Z} \subseteq \mathbb{R}$ is the smallest set among all nonempty subsets $A \subseteq \mathbb{R}$ possessing the following two properties:*

- (1) $\mathbb{N} \subseteq A$.
- (2) $(\forall x, y \in A)(x - y \in A)$.

Proof: Of course, by the construction of \mathbb{Z} it possesses properties (1) and (2). Thus, it remains to prove that $\mathbb{Z} \subseteq A$ for any set A with (1) and (2). First note that property (2) implies $0 \in A$. To see this, choose any $x \in A$ (recall that A is nonempty) and use $0 = x - x \in A$. Next take any nonzero $x \in \mathbb{Z}$. Then either $x \in \mathbb{N}$, hence by (1) also $x \in A$, or $-x \in \mathbb{N}$. In the latter case we get $-x \in A$, hence by (2) also $0 - (-x) = x \in A$. Thus, $\mathbb{Z} \subseteq A$ for any set A with properties (1) and (2). This completes the proof. ■

A similar characterization is also true for the field of rational numbers. Here we have the following.

Proposition 4.7.11. *The set $\mathbb{Q} \subseteq \mathbb{R}$ is the smallest set among all nonempty subsets $A \subseteq \mathbb{R}$ possessing the following two properties:*

- (1) $\mathbb{Z} \subseteq A$.
- (2) $(\forall x, y \in A, y \neq 0)(xy^{-1} \in A)$.

Proof: Since \mathbb{Q} possesses properties (1) and (2) it remains to verify the following. Whenever a subset $A \subseteq \mathbb{R}$ has properties (1) and (2), then this implies $\mathbb{Q} \subseteq A$.

To this end choose any $x = p/q$ in \mathbb{Q} . Here $p, q \in \mathbb{Z}$ with $q \neq 0$. In view of property (1) we also have $p, q \in A$, hence (2) implies $x = pq^{-1} \in A$. Consequently, it follows that $\mathbb{Q} \subset A$ for any A satisfying (1) and (2). ■

Remark 4.7.6. One may ask for a similar characterization of \mathbb{R} as minimal extension of \mathbb{Q} . This can be done in several ways, as for example as smallest set containing \mathbb{Q} and least upper bounds of bounded above subsets in \mathbb{Q} . The better way is to characterize \mathbb{R} as smallest set containing limits of all converging sequences of rational numbers. We postpone this question until we investigated sequences and their limits (cf. Proposition 5.6.20 below).

A basic question is how rational and irrational numbers are distributed in \mathbb{R} . Are there gaps free of rational and/or irrational numbers? Or are both types of numbers strongly mixed. First note an easy property of irrational numbers.

Proposition 4.7.12. *If $x \in \mathbb{R}$ is irrational, then for any $c \in \mathbb{Q}$ also $x + c \notin \mathbb{Q}$ and if $c \neq 0$, this is also so for cx .*

Proof: We prove the first assertion by contradiction. Assume there exists a rational number c such that $x + c$ is rational. Because the sum of rational numbers is rational, we deduce that $x = (x + c) + (-c) \in \mathbb{Q}$, in contradiction with x being irrational. This proves the first part. The second part follows similarly using that the product of two rational numbers is rational. We leave the details as an exercise. ■

Example 4.7.5. The number $q + c\sqrt{2}$ is irrational for any $q \in \mathbb{Q}$ and for each nonzero $c \in \mathbb{Q}$.

The next result shows that the rational numbers as well as the irrational ones are dense in \mathbb{R} .

Proposition 4.7.13. *Let $x < y$ be two real numbers.*

- (1) *There exist infinitely many rational numbers q with $x < q < y$.*
- (2) *There are infinitely many irrational numbers w such that $x < w < y$.*

Proof: The Archimedean property yields the existence of a natural number n such that $n(y - x) > 1$ or, equivalently, with $ny > nx + 1$. Define the integer m by $m = \lfloor nx \rfloor + 1$. Then $nx < m \leq nx + 1 < ny$. This implies

$$x < \frac{m}{n} < y,$$

that is, the rational number $q_1 = \frac{m}{n}$ satisfies $x < q_1 < y$.

In the next step we apply the same construction to x and q_1 . Since $x < q_1$, by the first step we find a rational number q_2 satisfying $x < q_2 < q_1$, hence $x < q_2 < y$. Proceeding further, now with x and q_2 we get a rational number q_3 satisfying $x < q_3 < q_2$. At the end we find a decreasing sequence $(q_j)_{j \geq 1}$ of rational numbers such that

$$x < \dots < q_4 < q_3 < q_2 < q_1 < y.$$

This completes the proof of the first part.

To prove the second part we first note that it suffices to prove the result for rational numbers x and y . Indeed, by the first assertion we find for arbitrary $x < y$ rational \tilde{x} and \tilde{y} with $x < \tilde{x} < \tilde{y} < y$. Hence, if there are infinitely many irrational numbers between \tilde{x} and \tilde{y} , then, of course, this is also true for $x < y$.

The Archimedean Property implies the existence of an $n \in \mathbb{N}$ with $1/n < \frac{y-x}{\sqrt{2}}$ or, equivalently

$$x < x + \frac{\sqrt{2}}{n} < y.$$

The number $w = x + \frac{\sqrt{2}}{n}$ is irrational, since we assumed $x \in \mathbb{Q}$ (compare Example 4.7.5 above). Thus, we have shown the following: For any reals $x < y$ there is at least one irrational number $w \in \mathbb{R}$ with $x < w < y$. To get infinitely many irrational numbers between x and y we use exactly the same argument as in the proof of the first assertion, i.e., choose an irrational w_1 with $x < w_1 < y$, find an irrational number w_2 with $x < w_2 < w_1 < y$ and so on. ■

Definition 4.7.5. A real number x is said to be **algebraic** provided there exist **integers** a_0, \dots, a_n such that

$$a_0 + a_1 x + \dots + a_n x^n = 0.$$

Otherwise, $x \in \mathbb{R}$ is called **transcendental**.

For example, each rational number $q = m/n$ is algebraic because of $nq - m = 0$. But also $x = \sqrt{2}$ is algebraic by $x^2 - 2 = 0$. More general, for any $m \in \mathbb{N}$ and any $n > 1$ the n th root $x = \sqrt[n]{m}$ is algebraic because of $x^n - m = 0$. The first nonalgebraic, i.e., transcendental, number was constructed by Joseph Liouville (1809–1882) in 1844 as $\sum_{k=1}^{\infty} 10^{-k!}$. Later on, in 1873 Charles Hermite (1822–1901) proved that the Euler number e (cf. Definition 5.3.1) is transcendental, and in 1882 Ferdinand von Lindemann (1852–1939) showed that also π is not algebraic, thus transcendental.

In 1900, at the World Congress of mathematicians in Paris the German mathematician David Hilbert (1862–1943) gave a talk where he formulated 23 problems which were in his opinion the most important open ones. In Problem 7 he asked in which cases α^β is transcendental. For example, at this time it was even unknown whether $2^{\sqrt{2}}$ is transcendental or algebraic. The problem was finally solved in 1934/35 by the Russian mathematician Alexander Gelfond (1906–1968) and, independently, at the same time by the German mathematician Theodor Schneider (1911–1988).

Theorem 4.7.14 (Gelfond–Schneider Theorem). *If α and β are algebraic numbers with $\alpha \neq 0$, $\alpha \neq 1$ and $\beta \notin \mathbb{Q}$, then α^β is transcendental.*

In particular, the theorem implies that $2^{\sqrt{2}}$ and $\sqrt[3]{2^{\sqrt{2}}}$ are transcendental. On the other hand, there are important mathematical constants as the Euler–Mascheroni constant¹⁵ γ or

$$e \cdot \pi, \quad e + \pi, \quad \pi - e, \quad \frac{\pi}{e}, \quad \pi^\pi, \quad e^e, \quad \pi^e, \quad \pi^{\sqrt{2}},$$

where it is unknown whether they are rational or irrational, algebraic or transcendental.

Exercise 4.7.1. Suppose $x \in [0, 1]$ possesses the expansion $x =_b 0.x_1x_2\dots$ with respect to some base $b \geq 2$. Given $k \in \mathbb{N}$, which representation as b -fraction does x/b^k possess? Justify your answer.

Exercise 4.7.2. Let $x, y \in [0, 1]$ be expanded as $x =_b 0.x_1x_2\dots$ and $y =_b 0.y_1y_2\dots$ for certain $x_j, y_j \in \{0, \dots, b-1\}$. Show that $x < y$ if and only if there is an $m \in \mathbb{N}$ such that $x_1 = y_1, \dots, x_{m-1} = y_{m-1}$ and $x_m < y_m$. If $m = 1$, this has to be understood as $x_1 < y_1$. Estimate in this case the distance between x and y by a negative power of the base $b \geq 2$.

Exercise 4.7.3. Let as before x and y be real numbers in $[0, 1]$ with $x =_b 0.x_1x_2\dots$ and with $y =_b 0.y_1y_2\dots$ for certain $x_j, y_j \in \{0, \dots, b-1\}$. Suppose furthermore that $x_j \leq y_j$ for all $j \geq 1$. Does this imply

$$y - x =_b 0.z_1z_2\dots \quad \text{where } z_j = y_j - x_j, \quad j \geq 1 ?$$

Exercise 4.7.4. Use Proposition 4.7.2 to simplify the proof of Theorem 3.3.7 about the evaluation of $\alpha =_b 0.a_1 \dots a_m \overline{a_{m+1} \dots a_{m+k}}$. Note that due to Proposition 4.7.2 the existence of the real number α is now known which simplifies the proof of Theorem 3.3.7 considerably.

15

$$\gamma = \lim_{n \rightarrow \infty} \left(\sum_{k=1}^n \frac{1}{k} - \log n \right).$$

Exercise 4.7.5. Define the sets D_n as in Example 4.7.3. How does D_4 look like? Describe the set

$$C_4 = \bigcap_{j=1}^4 ([0, 1] \setminus D_j).$$

Exercise 4.7.6. Which of the following numbers belong to the Cantor set \mathcal{C} , and which do not?

$$\frac{19}{27}, \quad \frac{3}{5}, \quad \frac{4}{13}, \quad \frac{5}{12}, \quad \pi - 3.$$

Exercise 4.7.7. Let T_ℓ and T_r be the functions from $[0, 1]$ to $[0, 1]$ defined by

$$T_\ell(x) = \frac{x}{3} \quad \text{and} \quad T_r(x) = \frac{2+x}{3}, \quad 0 \leq x \leq 1.$$

Here “ ℓ ” stands for “left” and “ r ” for right.

Prove the following properties.

- (1) Both functions leave the Cantor set \mathcal{C} invariant. In other words, it follows that

$$T_\ell(\mathcal{C}) \subseteq \mathcal{C} \quad \text{and} \quad T_r(\mathcal{C}) \subseteq \mathcal{C}.$$

- (2) It holds

$$T_\ell(\mathcal{C}) \cap T_r(\mathcal{C}) = \emptyset \quad \text{and} \quad T_\ell(\mathcal{C}) \cup T_r(\mathcal{C}) = \mathcal{C}.$$

- (3) The sets $T_\ell(\mathcal{C})$ and \mathcal{C} as well as $T_r(\mathcal{C})$ and \mathcal{C} are isomorphic. That is, there exist bijections from $T_\ell(\mathcal{C})$ to \mathcal{C} and also from $T_r(\mathcal{C})$ to \mathcal{C} .

Exercise 4.7.8. The set $\mathcal{C} \times \mathcal{C}$ is called **Cantor dust**. Describe a method for the construction of this set. Which parts of $[0, 1] \times [0, 1]$ have to cut out in each step and which remain? Evaluate the area left over after n steps.

Hint: In a first step one takes out the gray area in Figure 4.7.6.

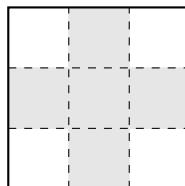


Figure 4.7.6. First step in the construction of the Cantor dust.

Exercise 4.7.9. Show that $\frac{8^n - 1}{7}$ is the number of squares eliminated after n steps in the construction of the Sierpiński carpet. Prove that in the construction of the carpet the remaining area after n steps equals $(8/9)^n$. What does this say about the area of the carpet?

Exercise 4.7.10. Let $A \subseteq \mathbb{R}$ and $B \subseteq \mathbb{R}$ be defined by

$$A = \{q > 0 : q \in \mathbb{Q}, q^2 < 2\} \quad \text{and} \quad B = \{r > 0 : r \notin \mathbb{Q}, r^2 < 2\}.$$

Determine (with proof) $\sup(A)$ and $\sup(B)$.

Exercise 4.7.11. Prove that $\frac{\sqrt{5}+1}{2}$ and $\frac{\sqrt{5}-1}{2}$ are algebraic numbers.

Exercise 4.7.12. Prove that $\sin(15^\circ)$, $\cos(40^\circ)$ and $\sin(70^\circ)$ are algebraic numbers.

Exercise 4.7.13. A real number x is called **an algebraic integer** if there exists a natural number n and integers a_0, \dots, a_{n-1} such that

$$a_0 + a_1x + \dots + a_{n-1}x^{n-1} + x^n = 0.$$

Give an example of an algebraic number that is not an algebraic integer.

Exercise 4.7.14. Let $m, n \in \mathbb{Z}$ such that $n \neq 0$. If m/n is an algebraic integer, then n must divide m .

Exercise 4.7.15. Prove that a real number x is algebraic if and only if there are **rational** coefficients a_0, \dots, a_n such that

$$a_0 + a_1x + \dots + a_nx^n = 0.$$

Why is x^{-1} algebraic if $x \neq 0$ is so?

Remark: It is also true that the sum and product of algebraic numbers are algebraic, but this is more difficult to prove.

4.8. More Exercises

Exercise 4.8.1. Find all real numbers $x \in \mathbb{R}$ for which

- | | |
|--|--|
| (a) $4 - x < 3 - 2x.$ | (b) $5 - x^2 < 8.$ |
| (c) $5 - x^2 < -2.$ | (d) $(x - 1)(x - 3) > 0.$ |
| (e) $x^2 - 2x + 2 > 0.$ | (f) $x^2 + x + 1 > 2.$ |
| (g) $(x - 1)(x + 5)(x - 3) > 0.$ | (h) $(x - \sqrt[3]{2})(x - \sqrt{2}) > 0.$ |
| (i) $\frac{1}{x} + \frac{1}{1-x} > 0.$ | (j) $\frac{x-1}{x+1} > 0.$ |

Exercise 4.8.2. Find all real numbers x which satisfy the following:

- | | |
|------------------------------|------------------------------|
| (a) $ x - 3 = 8.$ | (b) $ x - 3 < 8.$ |
| (c) $ x + 4 < 2.$ | (d) $ x - 1 + x - 2 > 1.$ |
| (e) $ x - 1 + x - 2 > 2.$ | (f) $ x - 1 + x + 1 < 1.$ |
| (g) $ x - 1 x + 1 = 0.$ | (h) $ x - 1 x + 2 = 3.$ |

Exercise 4.8.3. Find all real numbers x such that

$$|x + 1| + |x - 2| - |x + 3| < 5.$$

Exercise 4.8.4. Let x and y be arbitrary real numbers. Express each of the following without absolute signs. Treat different cases if necessary.

- | | |
|---------------------|---------------------|
| (1) $ x + y - y $ | (2) $ x - 1 $ |
| (3) $ x - x^2 $ | (4) $x - x - x $ |

Exercise 4.8.5. Let x and y be two real numbers satisfying $x^n = y^n$, for some $n \geq 1$.

- (1) If n is odd, prove that $x = y$.
- (2) If n is even, show that $x = y$ or $x = -y$.

Exercise 4.8.6. Prove that $0.69 < \log_{10}(5) < 0.7$.

Exercise 4.8.7. Prove that $\sqrt[3]{4} - \sqrt[3]{2} = 1 + \sqrt[3]{9\sqrt[3]{2} - 9}$.

Exercise 4.8.8. Let x be a nonzero real number. If $a = x + \frac{1}{x}$, then write down $x^2 + \frac{1}{x^2}$, $x^3 + \frac{1}{x^3}$ and $x^4 + \frac{1}{x^4}$ in terms of a .

Exercise 4.8.9. Solve the following equations for x .

- (1) $36^x - 2 \cdot 6^{x+1} - 28 = 0$.
- (2) $3^{2x} + \frac{1}{3^{2x}} - 4\left(3^x + \frac{1}{3^x}\right) + 5 = 0$.
- (3) $\log_3(x+1) + \log_3(x-2) = 81$.
- (4) $\log_3(x) = \log_2(x)$.

Exercise 4.8.10. Let a_1, a_2, b_1, b_2 be real numbers.

- (1) If $a_1 \leq a_2$ and $b_1 \leq b_2$, prove that $\frac{a_1 b_1 + a_2 b_2}{2} \geq \frac{a_1 + a_2}{2} \cdot \frac{b_1 + b_2}{2}$.
- (2) If $a_1 \leq a_2$ and $b_1 \geq b_2$, show that $\frac{a_1 b_1 + a_2 b_2}{2} \leq \frac{a_1 + a_2}{2} \cdot \frac{b_1 + b_2}{2}$.
- (3) Characterize the equality case in each of the previous inequalities.

Exercise 4.8.11. Let $n \in \mathbb{N}$ and $a_1, \dots, a_n, b_1, \dots, b_n \in \mathbb{R}$.

- (1) If $a_1 \leq \dots \leq a_n$ and $b_1 \leq \dots \leq b_n$, then prove that

$$\frac{a_1 b_1 + \dots + a_n b_n}{n} \geq \frac{a_1 + \dots + a_n}{n} \cdot \frac{b_1 + \dots + b_n}{n}.$$

- (2) If $a_1 \leq \dots \leq a_n$ and $b_1 \geq \dots \geq b_n$, then prove that

$$\frac{a_1 b_1 + \dots + a_n b_n}{n} \leq \frac{a_1 + \dots + a_n}{n} \cdot \frac{b_1 + \dots + b_n}{n}.$$

- (3) Characterize the equality case in each of the previous inequalities.

These results are known as the Chebyshev sum inequalities after the Russian mathematician Pafnuty Chebyshev (1821–1894).

Exercise 4.8.12. Let $n \in \mathbb{N}$ and $a_1, \dots, a_n \in \mathbb{R}$. Show that

$$\sqrt{\frac{a_1^2 + \dots + a_n^2}{n}} \geq \frac{a_1 + \dots + a_n}{n}.$$

Characterize the equality case.

Exercise 4.8.13. Let $k, \ell, n \in \mathbb{N}$ and $a_1, \dots, a_n \in \mathbb{R}$ be nonnegative numbers. Show that

$$\frac{a_1^{k+\ell} + \dots + a_n^{k+\ell}}{n} \geq \frac{a_1^k + \dots + a_n^k}{n} \cdot \frac{a_1^\ell + \dots + a_n^\ell}{n}.$$

Exercise 4.8.14. For $j \in \{1, 2\}$, let P_j be the point of coordinates (x_j, y_j) as in Figure 4.8.1. Denote by θ the angle between OP_1 and OP_2 . Prove that

$$\cos \theta = \frac{x_1 x_2 + y_1 y_2}{\sqrt{x_1^2 + y_1^2} \cdot \sqrt{x_2^2 + y_2^2}}.$$

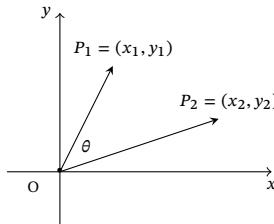


Figure 4.8.1. The angle between two lines.

Exercise 4.8.15. Prove that for any pairs (x_1, y_1) and (x_2, y_2) of real numbers

$$|x_1x_2 + y_1y_2| \leq \sqrt{x_1^2 + y_1^2} \cdot \sqrt{x_2^2 + y_2^2}.$$

When does equality happen? How is this exercise related to Exercise 4.8.14?

Exercise 4.8.16. Let n be a natural number. If $x_1, \dots, x_n, y_1, \dots, y_n$ are real numbers, prove that

$$|x_1y_1 + \dots + x_ny_n| \leq \sqrt{(x_1^2 + \dots + x_n^2)(y_1^2 + \dots + y_n^2)}.$$

Characterize the equality case. This result is called Cauchy-Schwarz inequality or also Cauchy-Bunyakovsky-Schwarz inequality.¹⁶

Exercise 4.8.17. Let $x_1, x_2, x_3, y_1, y_2, y_3$ be real numbers with $0 < x_1 \leq x_2 \leq x_3$ and $0 < y_1 \leq y_2 \leq y_3$. Prove that

$$x_1y_1 + x_2y_2 + x_3y_3 \leq x_1y_2 + x_2y_3 + x_3y_1 \leq x_1y_3 + x_2y_2 + x_3y_1.$$

Exercise 4.8.18. Let $n \in \mathbb{N}$ and $x_1, \dots, x_n, y_1, \dots, y_n \in \mathbb{R}$ be real numbers such that $0 < x_1 \leq \dots \leq x_n$ and $0 < y_1 \leq \dots \leq y_n$. If σ is a permutation in S_n , show that

$$x_1y_1 + \dots + x_ny_n \leq x_1y_{\sigma(1)} + \dots + x_ny_{\sigma(n)} \leq x_1y_n + \dots + x_ny_1.$$

Exercise 4.8.19. Let $n \in \mathbb{N}$ and $x_1, \dots, x_n, y_1, \dots, y_n$ be positive real numbers such that $x_1 < \dots < x_n$ and $y_1 < \dots < y_n$. If σ is a permutation such that

$$x_1y_1 + \dots + x_ny_n = x_1y_{\sigma(1)} + \dots + x_ny_{\sigma(n)},$$

then show that σ is the identity permutation.

Exercise 4.8.20. Let $n \in \mathbb{N}$ and $x_1, \dots, x_n, y_1, \dots, y_n$ be positive real numbers such that $x_1 < \dots < x_n$ and $y_1 < \dots < y_n$. If σ is a permutation such that

$$x_1y_1 + \dots + x_ny_n = x_1y_{\sigma(n)} + \dots + x_ny_{\sigma(1)},$$

then show that $\sigma(j) = n - j + 1$ for any $1 \leq j \leq n$.

¹⁶Named after the French mathematician Augustin-Louis Cauchy (1789–1857), the Russian mathematician Viktor Bunyakovsky (1804–1889) and the German mathematician Karl Hermann Amandus Schwarz (1843–1921).

Sequences of Real Numbers

Nothing comes easy, it takes much practice.

Nas

5.1. Basic Properties

What is a sequence x_1, x_2, \dots of objects? First, one must distinguish between finite and infinite sequences, written as x_1, \dots, x_n or as x_1, x_2, \dots , respectively. Thereby the entries x_j can be chosen quite general. For example, they can be letters from the alphabet, or they may be circles, dashes, and crosses, but most of the time in this book, these objects will be numbers. Our interest in this section is infinite sequences of real numbers which are, in addition to x_1, x_2, \dots , also denoted by $(x_n)_{n \geq 1}$. But how is such a sequence defined in a rigorous way? Does it suffice to know the first three or four entries in order to describe the sequence completely? For example, if one knows that a sequence starts with $1, 4, 9, 16, 25, \dots$, then most people will expect that the next elements in the sequence are $36, 49, 64, 81, \dots$. But this needs not to be true (see Example 1.7.6). We could proceed with any other numbers, for example with $1, 1, 1, \dots$. The key point is that one has to have a rule which defines **every** element in the sequence in a precise way, not only the first three or four ones. Mathematically this means the following.

Definition 5.1.1. An (infinite) **sequence** of real numbers is a function x from \mathbb{N} to \mathbb{R} . As usual, one writes x_n instead of $x(n)$, i.e.,

$$x_n = x(n), \quad n = 1, 2, \dots$$

$$((x_n)_{n \geq 1} \text{ real sequence}) \Leftrightarrow (\exists x : \mathbb{N} \rightarrow \mathbb{R})(\forall n \in \mathbb{N})(x(n) = x_n)$$

In this sense, x_1, x_2, \dots are the values of x in the natural order given on \mathbb{N} .

It is important to distinguish between the sequence x_1, x_2, \dots and the set $\{x_1, x_2, \dots\}$. For example, if a sequence is constant 1, i.e., we consider the sequence $1, 1, 1, \dots$, the set of values coincides with $\{1\}$.

Sometimes the enumeration of a sequence does not start at index 1, instead at zero or at some other integer $N \in \mathbb{Z}$. Then the defining function x is understood to map $\{N, N+1, \dots\}$ into \mathbb{R} . For example, we may have sequences enumerated as $(x_n)_{n \geq 0}$ or $(x_n)_{n \geq 5}$.

Example 5.1.1. In order that the sequence $\{1, 4, 9, 16, 25, \dots\}$ proceeds by $36, 49, \dots$ the function $x : \mathbb{N} \rightarrow \mathbb{R}$ has to be defined by $x(n) = n^2$.

Example 5.1.2. If $x : \mathbb{N} \rightarrow \mathbb{R}$ is given by $x(n) = 1/n$, then the sequence is $1, \frac{1}{2}, \frac{1}{3}, \frac{1}{4}, \dots$.

There exist interesting sequences that are defined inductively by certain rules. That is, we know the first element (sometimes also the first two or three) of the sequence while the remaining entries are defined inductively. In fact, also in this case the sequence is nothing other than as a function from \mathbb{N} or \mathbb{N}_0 to \mathbb{R} , but this function is either too difficult to describe or it is even impossible to give a simple formula for it. A typical example of such a sequence is that of Fibonacci numbers $(F_n)_{n \geq 0}$ which we introduced in (1.4.3). Recall that this sequence is defined by $F_0 = 0, F_1 = 1$ and

$$F_{n+1} = F_n + F_{n-1}, \quad n = 1, 2, \dots$$

Another interesting example of this type is as follows.

Example 5.1.3. The sequence $(x_n)_{n \geq 1}$ of real numbers is defined by $x_1 = 2$ and

$$x_{n+1} = \frac{x_n}{2} + \frac{1}{x_n}, \quad n \geq 2.$$

The first six entries of this sequence are

$$\begin{aligned} x_1 &= 2, & x_2 &= \frac{3}{2}, & x_3 &= \frac{17}{12}, & x_4 &= \frac{577}{408}, & x_5 &= \frac{665857}{470832}, \\ x_6 &= \frac{886731088897}{627013566048}, \end{aligned}$$

with numerical values

$$\begin{aligned} x_2 &= 1.5, & x_3 &= 1.4167, & x_4 &= 1.4142157, & x_5 &= 1.41421356 \\ x_6 &= 1.41421356237309504880168. \end{aligned}$$

Since infinite sequences of real numbers are functions from \mathbb{N} to \mathbb{R} , they allow several algebraic operations. They can be added, subtracted, multiplied and divided. For example, given two real sequences $(x_n)_{n \geq 1}$ and $(y_n)_{n \geq 1}$ and an $\alpha \in \mathbb{R}$, then

$$\begin{aligned} (x_n)_{n \geq 1} + (y_n)_{n \geq 1} &= (x_n + y_n)_{n \geq 1}, & (x_n)_{n \geq 1} - (y_n)_{n \geq 1} &= (x_n - y_n)_{n \geq 1}, \\ \alpha \cdot (x_n)_{n \geq 1} &= (\alpha x_n)_{n \geq 1}, & (x_n)_{n \geq 1} \cdot (y_n)_{n \geq 1} &= (x_n \cdot y_n)_{n \geq 1}, \\ \frac{(x_n)_{n \geq 1}}{(y_n)_{n \geq 1}} &= \left(\frac{x_n}{y_n} \right)_{n \geq 1} \text{ if } y_n \neq 0, n \geq 1. \end{aligned}$$

An important property of sequences is whether they are monotone in some sense. For example, if x_n denotes the price of an article in a store on day n , it may be important whether this sequence is decreasing or increasing.

Definition 5.1.2. A sequence $(x_n)_{n \geq 1}$ is said to be **increasing** (resp. **nondecreasing**) if

$$x_1 < x_2 < x_3 < \dots \quad \text{resp.} \quad x_1 \leq x_2 \leq x_3 \leq \dots .$$

It is said to be **decreasing** (resp. **nonincreasing**) if

$$x_1 > x_2 > x_3 > \dots \quad \text{resp.} \quad x_1 \geq x_2 \geq x_3 \geq \dots .$$

A sequence $(x_n)_{n \geq 1}$ of real numbers is

- | | | |
|---------------|-------------------|--|
| increasing | \Leftrightarrow | $(\forall n \geq 1)(x_n < x_{n+1})$. |
| nondecreasing | \Leftrightarrow | $(\forall n \geq 1)(x_n \leq x_{n+1})$. |
| decreasing | \Leftrightarrow | $(\forall n \geq 1)(x_n > x_{n+1})$. |
| nonincreasing | \Leftrightarrow | $(\forall n \geq 1)(x_n \geq x_{n+1})$. |
| monotone | \Leftrightarrow | it is either nonincreasing or nondecreasing. |

Note that nonincreasing is not the same as not increasing.

Example 5.1.4. The sequence $(x_n)_{n \geq 1}$ with $x_n = \frac{n}{n+1}$ is increasing while the sequence $(y_n)_{n \geq 1}$ with $y_n = 1/n^2$ is decreasing. On the other hand, if $z_n = \frac{(-1)^n}{n}$, then $(z_n)_{n \geq 1}$ is neither increasing nor decreasing. Sequences of this type with changing signs are called **alternating**.

A crucial property of a sequence is whether it is bounded or maybe unbounded. For example, if x_n denotes the water level gauge of a river at day n , a question of vital importance is if this sequence is bounded above by some upper bound.

Definition 5.1.3. A real sequence $(x_n)_{n \geq 1}$ is **bounded above** if there exists an $M \in \mathbb{R}$ with $x_n \leq M$ for all $n \geq 1$. Otherwise, it is said to be **unbounded above**.

The sequence $(x_n)_{n \geq 1}$ is **bounded below** if there exists an $m \in \mathbb{R}$ with $x_n \geq m$ for all $n \geq 1$.

Finally, the sequence is **bounded** if it is at the same time bounded above and below. Note that this happens if and only if there exists an $M \geq 0$ with $|x_n| \leq M$ for all $n \geq 1$.

Remark 5.1.1. If one defines the set $B \subseteq \mathbb{R}$ by

$$B := \{x_1, x_2, \dots\},$$

then the sequence $(x_n)_{n \geq 1}$ is bounded above or below or bounded if and only if this is so for B as introduced in Definition 4.4.1.

Example 5.1.5. The sequence $(x_n)_{n \geq 1}$ with $x_n = \frac{n}{n+1}$ satisfies $\frac{1}{2} \leq x_n \leq 1$. Hence, it is bounded above and below, thus bounded.

On the other hand, if $x_n = \frac{n^3 - 2}{n^2 + 1}$, then it is not difficult to show (compare Example 5.1.7 below) that this sequence is unbounded above.

A basic role in the investigation of sequences plays their supremum and their infimum. Fortunately, we can use the results presented in Section 4.5.

Definition 5.1.4. Let $(x_n)_{n \geq 1}$ be a sequence of real numbers bounded above. Define the subset $B \subseteq \mathbb{R}$ by

$$B := \{x_1, x_2, \dots\}.$$

Note that B is nothing else as the range of the function $x : \mathbb{N} \rightarrow \mathbb{R}$ with $x(n) = x_n$. Then we set

$$\sup_{n \geq 1} x_n := \sup(B).$$

In the same way we introduce the infimum of a sequence bounded below by

$$\inf_{n \geq 1} x_n := \inf(B).$$

We write

$$\sup_{n \geq 1} x_n = \infty \quad \text{or} \quad \inf_{n \geq 1} x_n = -\infty$$

provided that the sequence $(x_n)_{n \geq 1}$ is **not** bounded above or below, respectively.

If we transform the main assertions about supremum and infimum of sets in \mathbb{R} into the special case of sets generated by sequences, then we obtain the following basic characterization of the supremum and infimum of sequences.

Theorem 5.1.1.

- (i) $(\sup_{n \geq 1} x_n < \infty) \Leftrightarrow [(\exists M \in \mathbb{R})(\forall n \geq 1)(x_n \leq M)]$.
- (ii) $(\inf_{n \geq 1} x_n > -\infty) \Leftrightarrow [(\exists m \in \mathbb{R})(\forall n \geq 1)(x_n \geq m)]$.
- (iii) $(\sup_{n \geq 1} |x_n| < \infty) \Leftrightarrow [(\exists m, M \in \mathbb{R})(\forall n \geq 1)(m \leq x_n \leq M)]$.
- (iv) $(\sup_{n \geq 1} x_n = M) \Leftrightarrow [(\forall n \geq 1)(x_n \leq M) \text{ and } (\forall \varepsilon > 0)(\exists n \geq 1)(x_n > M - \varepsilon)]$.
- (v) $(\inf_{n \geq 1} x_n = m) \Leftrightarrow [(\forall n \geq 1)(x_n \geq m) \text{ and } (\forall \varepsilon > 0)(\exists n \geq 1)(x_n < m + \varepsilon)]$.
- (vi) $(\sup_{n \geq 1} x_n = \infty) \Leftrightarrow [(\forall K \in \mathbb{N})(\exists n \geq 1)(x_n > K)]$.
- (vii) $(\inf_{n \geq 1} x_n = -\infty) \Leftrightarrow [(\forall K \in \mathbb{N})(\exists n \geq 1)(x_n < -K)]$.

Example 5.1.6. Let $(x_n)_{n \geq 1}$ be defined by $x_n = (-1)^n \frac{n^2 - n}{n^2 + 3}$. We claim that $\sup_{n \geq 1} x_n = 1$. Because always $n^2 - n \leq n^2 + 3$, it follows that

$$x_n = \frac{n^2 - n}{n^2 + 3} \leq 1 \quad \text{for all even } n \geq 1.$$

If n is odd, then $x_n < 0$, hence in this case also $x_n \leq 1$. This shows that 1 is an upper bound of $(x_n)_{n \geq 1}$.

To prove that 1 is the smallest upper bound we have to verify the following: For each $\varepsilon > 0$ there is at least one $n \geq 1$ such that

$$(5.1.1) \quad x_n = (-1)^n \frac{n^2 - n}{n^2 + 3} > 1 - \varepsilon \quad \text{or, equivalently,} \quad 1 - x_n < \varepsilon.$$

Take any even $n > 3$. Then

$$1 - x_n = 1 - \frac{n^2 - n}{n^2 + 3} = \frac{n + 3}{n^2 + 3} \leq \frac{2n}{n^2} = \frac{2}{n}.$$

Hence, if $\frac{2}{n} < \varepsilon$, or, equivalently, $n > \frac{2}{\varepsilon}$, then $1 - x_n < \varepsilon$. Thus, if we choose any even $n > 3$ with $n > \frac{2}{\varepsilon}$, then (5.1.1) is satisfied. This completes the proof of the claim.

Let us present another example of interest.

Example 5.1.7. Suppose that

$$x_n = \frac{n^3 - 2}{n^2 + 1}, \quad n = 1, 2, \dots$$

We claim that $\sup_{n \geq 1} x_n = \infty$. To verify this we have to show the following:

$$(5.1.2) \quad (\forall K \in \mathbb{N})(\exists n \geq 1 \text{ s.t. } x_n > K).$$

The basic idea to prove (5.1.2) is as follows: Estimate the x_n 's from below by some y_n 's which are easier to handle. Indeed, if we find an $n \geq 1$ with $y_n > K$, then by $x_n \geq y_n$ this also implies $x_n > K$.

How do we get the suitable numbers y_n ? First note that for $n \geq 2$ it follows that $n^3/2 > 2$, hence we get

$$n^3 - 2 > n^3 - \frac{n^3}{2} = \frac{n^3}{2}.$$

To lessen the x_n 's we have to enlarge the denominator. Here we use $n^2 \geq 1$ and obtain

$$n^2 + 1 \leq n^2 + n^2 = 2n^2.$$

Combining the estimates for the nominator and the denominator leads to

$$x_n = \frac{n^3 - 2}{n^2 + 1} > \frac{n^3/2}{2n^2} = \frac{n}{4}.$$

We claim now that $y_n = n/4$ does the job. Indeed, if we choose any $n \geq 1$ satisfying $n > 4K$, then $y_n > K$ and therefore also $x_n > K$. This shows that (5.1.2) is true, hence $\sup_{n \geq 1} x_n = \infty$ as claimed.

Remark 5.1.2. In order to show that a sequence fulfills $\sup_{n \geq 1} x_n = \infty$, one has to find for each (large) $K \in \mathbb{N}$ at least one index n for which $x_n > K$. There is no need to find the smallest n for which this is true. Moreover, the choice of n in Example 5.1.7 was one possibility, but by far not the only one. Other estimates could be used, other sequences $(y_n)_{n \geq 1}$ with $y_n \leq x_n$ could have been chosen. We only have to verify that we have $x_n > K$ for a suitable $n \geq 1$.

The following properties of the supremum and infimum of sequences follow directly from the corresponding results proved in Section 4.5. Thus, we omit the proof and state it as an exercise.

Proposition 5.1.2. Let $(x_n)_{n \geq 1}$ and $(y_n)_{n \geq 1}$ be two sequences of real numbers. Then

- (1) $\sup_{n \geq 1} (x_n + y_n) \leq \sup_{n \geq 1} x_n + \sup_{n \geq 1} y_n$.
- (2) $\inf_{n \geq 1} (x_n + y_n) \geq \inf_{n \geq 1} x_n + \inf_{n \geq 1} y_n$.
- (3) $(\forall \alpha \geq 0)(\sup_{n \geq 1} (\alpha x_n) = \alpha \sup_{n \geq 1} x_n)$ and $(\forall \alpha < 0)(\sup_{n \geq 1} (\alpha x_n) = \alpha \inf_{n \geq 1} x_n)$.
- (4) $[(\forall n \geq 1)(x_n > 0)] \Rightarrow \left[\sup_{n \geq 1} \left(\frac{1}{x_n} \right) = \frac{1}{\inf_{n \geq 1} x_n} \text{ and } \inf_{n \geq 1} \left(\frac{1}{x_n} \right) = \frac{1}{\sup_{n \geq 1} x_n} \right]$.
- (5) $(\forall n \geq 1)(x_n \leq y_n) \Rightarrow \left[\sup_{n \geq 1} x_n \leq \sup_{n \geq 1} y_n \text{ and } \inf_{n \geq 1} x_n \leq \inf_{n \geq 1} y_n \right]$.

Exercise 5.1.1.

- (1) Suppose a sequence is given by $x_1 = 1$ and

$$x_n = n \cdot x_{n-1}, \quad n \geq 2.$$

Describe the function $x : \mathbb{N} \rightarrow \mathbb{R}$ corresponding to the sequence $(x_n)_{n \geq 1}$. Verify your answer by mathematical induction.

- (2) Answer the same question as in the first problem for the sequence $(x_n)_{n \geq 1}$ with $x_1 = 1$ and

$$x_n = 2x_{n-1} + 3, \quad n \geq 2.$$

Exercise 5.1.2. Which sequence $(x_n)_{n \geq 1}$ of real numbers satisfies $x_1 = 0$, $x_2 = 1$, and

$$x_n = \frac{x_{n-1} + x_{n+1}}{2}, \quad n \geq 2?$$

Exercise 5.1.3. Let

$$x_n = \frac{n-1}{n} \quad \text{and} \quad y_n = \frac{n+1}{n}, \quad n = 1, 2, \dots$$

Describe the sequences

$$(x_n + 2y_n)_{n \geq 1}, (x_n - y_n)_{n \geq 1}, (x_n \cdot y_n)_{n \geq 1}, \left(\frac{x_n}{y_n} \right)_{n \geq 1}, \text{ and } \left(\frac{x_n}{y_n^2} \right)_{n \geq 1}.$$

Exercise 5.1.4. Which of the following sequences are monotone:

$$\left(\frac{n^2 - 1}{n^2 + 1} \right)_{n \geq 1}, \left(\frac{n}{n^2 + 1} \right)_{n \geq 1}, (n 2^{-n})_{n \geq 1}, \left(1 + \frac{(-1)^n}{n} \right)_{n \geq 1}?$$

Exercise 5.1.5. Which of the following sequences are bounded above, bounded below, or bounded:

$$x_n = \frac{n^2 + 1}{n + 1}, \quad y_n = \frac{1}{n^4}, \quad z_n = (-1)^n \frac{n}{\sqrt{n+1}}, \quad n = 1, 2, \dots ?$$

Exercise 5.1.6. Let $(x_n)_{n \geq 1}$ and $(y_n)_{n \geq 1}$ be two sequences of real numbers. Prove that

$$\sup_{n \geq 1} (x_n + y_n) \leq \sup_{n \geq 1} x_n + \sup_{n \geq 1} y_n \quad \text{and} \quad \inf_{n \geq 1} (x_n + y_n) \geq \inf_{n \geq 1} x_n + \inf_{n \geq 1} y_n.$$

Give examples where the inequalities above are strict.

Exercise 5.1.7. Prove properties (3) and (4) in Proposition 5.1.2. Does property (4) remain valid in the cases $\inf_{n \geq 1} x_n = 0$ and $\sup_{n \geq 1} x_n = \infty$ by letting $1/0 = \infty$ and $1/\infty = 0$?

Exercise 5.1.8. Determine

$$\sup_{n \geq 1} \frac{(n+1)^3 - n^3}{n^2} \quad \text{and} \quad \inf_{n \geq 1} \frac{(n+1)^3 - n^3}{n^2}.$$

Exercise 5.1.9. Show that

$$\sup_{n \geq 1} \frac{(n-4)^2}{n+3} = \infty.$$

Exercise 5.1.10. Let $a > 1$ some real number. Prove that

$$\sup_{n \geq 1} a^n = \infty.$$

Hint: For example, use Bernoulli's inequality (cf. Exercise 1.2.5). Why does this imply that in the case $0 < a < 1$ it follows that

$$\inf_{n \geq 1} a^n = 0 ?$$

5.2. Convergent and Divergent Sequences

What does it mean that a sequence $(x_n)_{n \geq 1}$ converges to some number L in \mathbb{R} ? Or when does a sequence approach the number L ? Similarly, when does a sequence tend to some L ? All these questions describe the same problem. Heuristically, the answer is as follows¹: The larger the index n becomes the closer are the x_n 's to L . And although the x_n 's get closer and closer to its limit L , in general they will never reach the value L in finite time.

Let us explain the situation at some examples.

(1) Consider the sequence $1, -\frac{1}{2}, \frac{1}{3}, -\frac{1}{4}, \frac{1}{5}, \dots$. This sequence approaches the value 0, but it never reaches it.

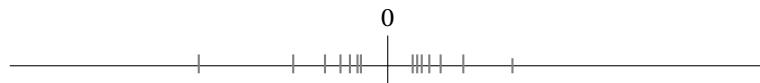


Figure 5.2.1. The first entries of the sequence $\left(\frac{(-1)^{n+1}}{n}\right)_{n \geq 1}$.

(2) The number $2/9$ has the decimal representation $0.2222\dots$. We can take as many often the digit 2 as we wish, every time we come closer and closer to $2/9$, but we will never reach the precise value $2/9$. To see this let

$$x_n = 0.\underbrace{222 \dots 2}_{n \text{ digits}} = \frac{2}{10} + \frac{2}{10^2} + \dots + \frac{2}{10^n}.$$

¹This heuristic approach to limits had been used for a long time. The first mathematical precise definition of the limit occurred in 1813 in a publication by Carl Friedrich Gauss. Later on in 1816 it was specified by B. Bolzano, and finally, in 1870 K. Weierstrass defined limits in the way we use them nowadays.

In this case we can even give the precise distance between x_n and its limit $2/9$. It is

$$\left| \frac{2}{9} - x_n \right| = \frac{2}{9} - 0.\underbrace{222\cdots 2}_{n \text{ digits}} = \frac{2}{9} \cdot \frac{1}{10^n}.$$

So we see that the x_n 's approach $2/9$ very fast, but even for extremely large values of n there remains a small gap between x_n and $2/9$. Nevertheless, in practical applications it suffices completely to work with a certain x_n .

(3) Define a sequence $(x_n)_{n \geq 1}$ by $1, \frac{1}{2}, 1, \frac{1}{4}, 1, \frac{1}{6}, 1, \frac{1}{8}, 1, \dots$. That is, for odd $n \in \mathbb{N}$ we have $x_n = 1$ while for even n 's it follows that $x_n = 1/n$. Does this sequence approach zero? The answer is negative. Why? Even if we look for x_n 's with very large $n \geq 1$, we will always find those with great distance to zero. Part of the sequence approaches zero, but another (infinite) part does not so.

The probably best way to introduce the convergence of a sequence is by using the notion of a neighborhood of a point in \mathbb{R} .

Definition 5.2.1. Let $L \in \mathbb{R}$ be some number. Given $\varepsilon > 0$ (we always think of ε as a very small positive number, say $\varepsilon = 10^{-100}$ or even much, much smaller). Then the ε -neighborhood $U_\varepsilon(L)$ of L is defined by

$$U_\varepsilon(L) = \{x \in \mathbb{R} : |x - L| < \varepsilon\} = (L - \varepsilon, L + \varepsilon).$$

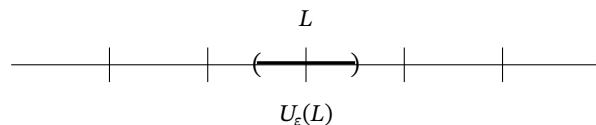


Figure 5.2.2. $x \in U_\varepsilon(L)$ if and only if $|x - L| < \varepsilon$.

Definition 5.2.2. A sequence $(x_n)_{n \geq 1}$ of real numbers **converges** to some real number L provided the following is satisfied: For any ε -neighborhood $U_\varepsilon(L)$ there exists an integer N such that $x_n \in U_\varepsilon(L)$ for all $n \geq N$. That is,

$$(5.2.1) \quad (\forall \varepsilon > 0)(\exists N \in \mathbb{N})(n \geq N \Rightarrow |x_n - L| < \varepsilon).$$

In this case we write

$$\lim_{n \rightarrow \infty} x_n = L \quad \text{or} \quad x_n \xrightarrow{n \rightarrow \infty} L.$$

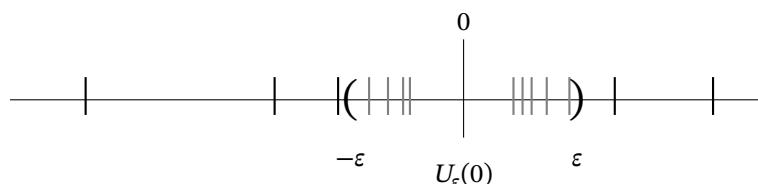


Figure 5.2.3. Convergence of $(-1)^{n+1}/n$ to 0.

Another way to formulate (5.2.1) is as follows:

$$L - \varepsilon < x_n < L + \varepsilon \Leftrightarrow -\varepsilon < x_n - L < \varepsilon \text{ provided that } n \geq N.$$

Consequently we get the following:

Proposition 5.2.1. *A sequence $(x_n)_{n \geq 1}$ converges to some $L \in \mathbb{R}$ if and only if $(L - x_n)_{n \geq 1}$ tends to zero. That is,*

$$x_n \xrightarrow{n \rightarrow \infty} L \Leftrightarrow L - x_n \xrightarrow{n \rightarrow \infty} 0$$

It is important to mention that, in general, the number $N = N(\varepsilon)$ depends heavily on the chosen $\varepsilon > 0$. Taking a smaller $\varepsilon' > 0$, it is very likely that the new $N = N(\varepsilon')$, this one for which (5.2.1) holds with ε' , has to be chosen bigger than $N(\varepsilon)$. Moreover, note that by no reason $N = N(\varepsilon)$ is unique. Indeed, if (5.2.1) is satisfied for some $N \in \mathbb{N}$, then it is also valid for any $N' \geq N$. In order to prove the convergence of a sequence, there is no need to find the *best* (smallest) $N \in \mathbb{N}$ for which (5.2.1) is satisfied. It completely suffices to find **one** $N = N(\varepsilon)$ for which (5.2.1) is valid.

Another equivalent formulation of the convergence of a sequence is as follows:

Proposition 5.2.2. *We have*

$$\lim_{n \rightarrow \infty} x_n = L \Leftrightarrow (\forall \varepsilon > 0)(|\{n \in \mathbb{N} : x_n \notin U_\varepsilon(L)\}| < \infty).$$

Proof: If $\lim_{n \rightarrow \infty} x_n = L$, then for each $\varepsilon > 0$ there exists an $N \geq 1$ such that $x_n \in U_\varepsilon(L)$ whenever $n \geq N$. Hence,

$$\{n \in \mathbb{N} : x_n \notin U_\varepsilon(L)\} \subseteq \{1, 2, \dots, N-1\}.$$

This shows that there exist at most finitely many of the x_n which are outside $U_\varepsilon(L)$.

Conversely, if the number of x_n 's outside $U_\varepsilon(L)$ is always finite, then there exists an $N \geq 1$ such that all $n \geq 1$ with $x_n \notin U_\varepsilon(L)$ are smaller than N . That is, on the right-hand side of N there are no exceptional elements or, equivalently, for all $n \geq N$ it follows that $x_n \in U_\varepsilon(L)$. This proves $\lim_{n \rightarrow \infty} x_n = L$. ■

Remark 5.2.1. It is important to mention that it does not suffice to say that for each $\varepsilon > 0$ there are infinitely many of the x_n 's inside $U_\varepsilon(L)$. Take for example the sequence $1, -1, 1, -1, 1, \dots$. Then for each $\varepsilon > 0$ there are infinitely many of the x_n 's inside $U_\varepsilon(1)$, but this sequence does not converge to 1. Thus, we really have to know that all but finitely many are inside the neighborhood.

Definition 5.2.3. A sequence $(x_n)_{n \geq 1}$ is said to be **convergent** provided that there is $L \in \mathbb{R}$ such that $\lim_{n \rightarrow \infty} x_n = L$. Otherwise, $(x_n)_{n \geq 1}$ is called **divergent**.

$$((x_n)_{n \geq 1} \text{ is convergent}) \Leftrightarrow (\exists L \in \mathbb{R})(\lim_{n \rightarrow \infty} x_n = L).$$

$$((x_n)_{n \geq 1} \text{ is divergent}) \Leftrightarrow ((x_n)_{n \geq 1} \text{ is not convergent}).$$

Before proceeding further in a first step we have to clarify whether a sequence may converge to more than one limit. The next proposition answers this question.

Proposition 5.2.3. *The limit of a sequence $(x_n)_{n \geq 1}$ is, whenever it exists, unique.*

$$(\lim_{n \rightarrow \infty} x_n = L_1 \text{ and } \lim_{n \rightarrow \infty} x_n = L_2) \Rightarrow (L_1 = L_2).$$

Proof: Suppose $\lim_{n \rightarrow \infty} x_n = L_1$ and $\lim_{n \rightarrow \infty} x_n = L_2$ for some $L_1 \neq L_2$. Because of $x_n \rightarrow L_1$, given $\varepsilon > 0$, there is an $N_1 = N_1(\varepsilon)$ such that

$$(5.2.2) \quad |x_n - L_1| < \varepsilon \quad \text{if } n \geq N_1.$$

By similar reasoning, there is an $N_2 = N_2(\varepsilon)$ such that

$$(5.2.3) \quad |x_n - L_2| < \varepsilon \quad \text{if } n \geq N_2.$$

Let $N = \max\{N_1, N_2\}$. Thus, if $n \geq N$, then (5.2.2) and (5.2.3) are both satisfied.

Next let $\varepsilon_0 = |L_1 - L_2| > 0$ (recall that we assume $L_1 \neq L_2$) and choose $\varepsilon < \varepsilon_0/2$. Consequently, if $n \geq N$, then (5.2.2) and (5.2.3) imply

$$\varepsilon_0 = |L_1 - L_2| = |L_1 - x_n + x_n - L_2| \leq |x_n - L_1| + |x_n - L_2| < \varepsilon + \varepsilon = 2\varepsilon < \varepsilon_0.$$

This contradiction shows that there cannot exist two different limits of a given sequence. ■

Let us give a first example of a convergent sequence. To make it as understandable as possible we will include all details.

Example 5.2.1. We claim that

$$\lim_{n \rightarrow \infty} \frac{n^2 - 1}{n^2 + 2} = 1.$$

To verify this we have to show the following: Given any $\varepsilon > 0$ there is an integer $N = N(\varepsilon)$ such that

$$(5.2.4) \quad \left| \frac{n^2 - 1}{n^2 + 2} - 1 \right| < \varepsilon \quad \text{if } n \geq N.$$

Since $\frac{n^2 - 1}{n^2 + 2} < 1$, condition (5.2.4) is equivalent to the following: For each $n \geq N$ one has

$$(5.2.5) \quad 1 - \frac{n^2 - 1}{n^2 + 2} < \varepsilon \quad \Leftrightarrow \quad \frac{n^2 - 1}{n^2 + 2} > 1 - \varepsilon \quad \Leftrightarrow \quad n^2 \varepsilon > 3 - 2\varepsilon \quad \Leftrightarrow \quad n^2 > \frac{3}{\varepsilon} - 2.$$

Now take any natural number N which satisfies

$$(5.2.6) \quad N > \sqrt{\frac{3}{\varepsilon} - 2}.$$

Since we are only interested in small $\varepsilon > 0$, we may always suppose that $\frac{3}{\varepsilon} > 2$. Indeed, if we found a suitable $N(\varepsilon)$ for some $\varepsilon > 0$, then the same $N(\varepsilon)$ works also for any bigger $\varepsilon' > \varepsilon$.

We claim now, that for N chosen as in (5.2.6) it follows that $n \geq N$ implies (5.2.4). Indeed, if $n \geq N$, then by the choice of N we get

$$n^2 \geq N^2 > \frac{3}{\varepsilon} - 2,$$

which implies by (5.2.5) that

$$1 - \frac{n^2 - 1}{n^2 + 2} < \varepsilon, \quad \text{hence} \quad \left| \frac{n^2 - 1}{n^2 + 2} - 1 \right| < \varepsilon.$$

The number $\varepsilon > 0$ was chosen arbitrarily, consequently we can conclude that, as asserted

$$\lim_{n \rightarrow \infty} \frac{n^2 - 1}{n^2 + 2} = 1.$$

For better understanding² let us give a second example.

Example 5.2.2. We want to show that

$$\lim_{n \rightarrow \infty} \frac{n}{n+1} = 1.$$

Thus, given $\varepsilon > 0$ we have to find an $N \geq 1$ such that

$$\left| \frac{n}{n+1} - 1 \right| < \varepsilon \quad \text{provided that} \quad n \geq N.$$

But

$$\left| \frac{n}{n+1} - 1 \right| = \frac{1}{n+1},$$

hence we have to have $\frac{1}{n+1} < \varepsilon$ or, equivalently, $1/\varepsilon < n+1$, whenever $n \geq N$. If we choose any fixed $N \in \mathbb{N}$ satisfying $N > \frac{1}{\varepsilon} - 1$, then $n \geq N$ implies $n+1 > 1/\varepsilon$, hence

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

This proves $\frac{n}{n+1} \xrightarrow[n \rightarrow \infty]{} 1$ as claimed above.

Remark 5.2.2. Observe that the convergence or divergence of a sequence $(x_n)_{n \geq 1}$ is completely independent of the first finitely many values of the sequence. More precisely,

$$(\exists N_0 \in \mathbb{N})(\forall n \geq N_0)(x_n = y_n) \Rightarrow [(\lim_{n \rightarrow \infty} x_n = L) \Leftrightarrow (\lim_{n \rightarrow \infty} y_n = L)].$$

Let us formulate what it means that a sequence does **not** converge to some $L \in \mathbb{R}$. In order that $x_n \xrightarrow[n \rightarrow \infty]{} L$, the convergence condition (5.2.1) has to be satisfied for **all** $\varepsilon > 0$. Hence, if the x_n 's do not converge to L , then condition (5.2.1) has to be violated for at least one $\varepsilon_0 > 0$. That is, it is not true that from a certain point all x_n 's are in the ε_0 -neighborhood of L . A way to formulate this mathematically is as follows:

$$(x_n \xrightarrow[n \rightarrow \infty]{} L) \Leftrightarrow (\exists \varepsilon_0 > 0)(\forall N \in \mathbb{N})(\exists n \geq N)(|x_n - L| \geq \varepsilon_0).$$

But when do we have that for each $N \geq 1$ there is an $n \geq N$ with $|x_n - L| \geq \varepsilon_0$? This happens if and only if there are infinitely many of the x_n 's outside this ε_0 -neighborhood of L . That is,

$$(x_n \xrightarrow[n \rightarrow \infty]{} L) \Leftrightarrow (\exists \varepsilon_0 > 0)(|\{n \geq 1 : |x_n - L| \geq \varepsilon_0\}| = \infty).$$

²Limits are one of the basic concepts in mathematics. For example, such important tools as the derivative of a function or its integral, or infinite sums of sequences, and so on, all these items are defined as suitable limits.

Thus, if a sequence $(x_n)_{n \geq 1}$ diverges, then this says that for all $L \in \mathbb{R}$ there is at least one ε_0 -neighborhood of L ($\varepsilon_0 > 0$ may depend on L) such that infinitely many of the x_n 's are outside this neighborhood. One (but by far not the only one) possible example for this happening is when the x_n 's become bigger and bigger in dependence of n . For instance, if $x_n = n$, then this sequence is divergent, but its type of divergence is completely different from the type of divergence in the case of the sequence $1, -1, 1, -1, 1, \dots$. To precise this we introduce the next definition.

Definition 5.2.4. A sequence $(x_n)_{n \geq 1}$ is said to be **properly divergent** if either

$$\lim_{n \rightarrow \infty} x_n = \infty \quad \text{or} \quad \lim_{n \rightarrow \infty} x_n = -\infty.$$

Remark 5.2.3. What does it mean that $\lim_{n \rightarrow \infty} x_n = \infty$? This says the following: For each $\Lambda \in \mathbb{R}$ (Λ big) there is $N = N(\Lambda) \geq 1$ such that $x_n > \Lambda$ whenever $n \geq N$. Or, equivalently, given any (big) natural number K , there is an integer $N = N(K)$ such that $x_n > K$ whenever $n \geq N$. In similar way $\lim_{n \rightarrow \infty} x_n = -\infty$ is defined.

$$\begin{aligned} (\lim_{n \rightarrow \infty} x_n = \infty) &\Leftrightarrow (\forall K \in \mathbb{N})(\exists N \in \mathbb{N})(\forall n \geq N)(x_n > K) \\ (\lim_{n \rightarrow \infty} x_n = -\infty) &\Leftrightarrow (\forall K \in \mathbb{N})(\exists N \in \mathbb{N})(\forall n \geq N)(x_n < -K). \end{aligned}$$

Remark 5.2.4. Another way to express that $\lim_{n \rightarrow \infty} x_n = \infty$ is as follows: For each (big) $K \in \mathbb{N}$ there are only finitely many of the x_n 's smaller than or equal K .

Proposition 5.2.4. If a sequence $(x_n)_{n \geq 1}$ converges to ∞ , then $\sup_{n \geq x_n} = \infty$. The converse is not valid.

$$(\lim_{n \rightarrow \infty} x_n = \infty) \Rightarrow (\sup_{n \geq 1} x_n = \infty), \text{ but } (\sup_{n \geq 1} x_n = \infty) \not\Rightarrow (\lim_{n \rightarrow \infty} x_n = \infty).$$

Proof: Suppose $\lim_{n \rightarrow \infty} x_n = \infty$. Given $K \in \mathbb{N}$, there is an $N \in \mathbb{N}$ with $x_n > K$ whenever $n \geq N$. Consequently, there is at least one $n \in \mathbb{N}$ (in fact there are even infinitely many such n 's) for which $x_n > K$. Thus, the sequence is not bounded above, that is $\sup_{n \geq 1} x_n = \infty$.

To prove that the converse is false, take the sequence $1, 1, 2, 1, 3, 1, 4, 1, \dots$. This sequence is not bounded above, but it also does not converge to ∞ . There is **no** $N \in \mathbb{N}$ such that $x_n \geq 2$ for all $n \geq N$. ■

The important difference between $\lim_{n \rightarrow \infty} x_n = \infty$ and $\sup_{n \geq 1} x_n = \infty$ is the following: Given a (big) integer K , in the case of the limit one has to have $x_n > K$ for **all** $n \geq N$ while the supremum is infinite whenever there is at least **one** $n \in \mathbb{N}$ with $x_n > K$. In fact, then there are even infinitely many of the x_n bigger than K (why?), but not necessarily all from a certain index. Hence, $\lim_{n \rightarrow \infty} x_n = \infty$ implies $\sup_{n \geq 1} x_n = \infty$. The converse implication is false.

Example 5.2.3. Let $x_n = \frac{n^3 - n^2}{n^2 + 2}$. We claim that $\lim_{n \rightarrow \infty} x_n = \infty$.

Thus, choose any $K \geq 1$. We have to show that $x_n > K$ whenever $n \geq N$ for a suitable $N \in \mathbb{N}$. Writing x_n as

$$x_n = n^2 \frac{n-1}{n^2+2}$$

and using $n-1 \geq n/2$ and $n^2+2 < 2n^2$ if $n \geq 2$, it follows that $x_n > n/4$ if $n \geq 2$. Hence, let us choose N as some fixed number satisfying $N > K/4$. Then $n \geq N$ implies $4n > K$, hence $x_n > n/4 > K$, which shows that $\lim_{n \rightarrow \infty} x_n = \infty$ as claimed.

Finally, we want to characterize the proper divergence of a sequence by the convergence of its reciprocal values.

Proposition 5.2.5. *Let $(x_n)_{n \geq 1}$ be a sequence of **positive** real numbers. Then*

$$\lim_{n \rightarrow \infty} x_n = \infty \Leftrightarrow \lim_{n \rightarrow \infty} \frac{1}{x_n} = 0.$$

Proof: Suppose first that $\lim_{n \rightarrow \infty} x_n = \infty$. Our goal is to show that $(1/x_n)_{n \geq 1}$ tends to zero. That is, given $\varepsilon > 0$, we have to find an $N \in \mathbb{N}$ such that

$$(5.2.7) \quad \left| \frac{1}{x_n} - 0 \right| < \varepsilon \Leftrightarrow 0 < \frac{1}{x_n} < \varepsilon \text{ if } n \geq N.$$

Since $x_n \xrightarrow[n \rightarrow \infty]{} \infty$, there exists an $N \in \mathbb{N}$ such that

$$x_n > \frac{1}{\varepsilon} \text{ provided that } n \geq N.$$

Of course, this implies $0 < 1/x_n < \varepsilon$. Hence, we found a suitable integer N for which (5.2.7) is satisfied. This proves that

$$\lim_{n \rightarrow \infty} \frac{1}{x_n} = 0$$

as claimed. The converse implication follows in the same way, and therefore we leave its verification as an exercise. ■

A first result tells us that convergent sequences are bounded. But recall that there are many bounded divergent sequences.

Proposition 5.2.6. *Each converging sequence is bounded.*

Proof: Suppose $\lim_{n \rightarrow \infty} x_n = L$. Taking $\varepsilon = 1$, there is an $N \in \mathbb{N}$ such that

$$|x_n - L| < 1 \text{ whenever } n \geq N.$$

From this we derive

$$|x_n| = |x_n - L + L| \leq |x_n - L| + |L| < |L| + 1 < \infty$$

whenever $n \geq N$. If we set

$$c := \max\{|x_1|, |x_2|, \dots, |x_{N-1}|, |L| + 1\},$$

then it follows that $|x_n| \leq c$ for all $n \geq 1$. Hence, $(x_n)_{n \geq 1}$ is bounded. ■

Proposition 5.2.7 (Sandwich Theorem³). Let $(x_n)_{n \geq 1}$, $(y_n)_{n \geq 1}$ and $(z_n)_{n \geq 1}$ be three sequences with

$$x_n \leq y_n \leq z_n.$$

If $x_n \xrightarrow[n \rightarrow \infty]{} L$ and $z_n \xrightarrow[n \rightarrow \infty]{} L$, then $y_n \xrightarrow[n \rightarrow \infty]{} L$.

$$[(\forall n \geq 1)(x_n \leq y_n \leq z_n) \text{ and } (x_n \xrightarrow[n \rightarrow \infty]{} L, z_n \xrightarrow[n \rightarrow \infty]{} L)] \Rightarrow (y_n \xrightarrow[n \rightarrow \infty]{} L).$$

Proof: By assumption, given $\varepsilon > 0$, there are N_1 and N_2 such that

$$L - \varepsilon < x_n < L + \varepsilon \quad \text{and} \quad L - \varepsilon < z_n < L + \varepsilon,$$

whenever $n \geq N_1$ or $n \geq N_2$, respectively. Consequently, if $n \geq N = \max\{N_1, N_2\}$, then

$$L - \varepsilon < x_n \leq y_n \leq z_n < L + \varepsilon,$$

i.e., $|y_n - L| < \varepsilon$ provided that $n \geq N$. This completes the proof. ■

Example 5.2.4. We claim that for any $k \geq 1$ it follows that $1/n^k \xrightarrow[n \rightarrow \infty]{} 0$. Indeed, if we let $x_n = 0$, $y_n = 1/n^k$ and $z_n = 1/n$, then $x_n \leq y_n \leq z_n$, and because of $x_n \xrightarrow[n \rightarrow \infty]{} 0$ and $z_n \xrightarrow[n \rightarrow \infty]{} 0$ it also follows $y_n \xrightarrow[n \rightarrow \infty]{} 0$.

The next result is a very useful tool for the evaluation of limits.

Proposition 5.2.8. Let $(x_n)_{n \geq 1}$ and $(y_n)_{n \geq 1}$ be two convergent sequences of real numbers. Then for each $\alpha, \beta \in \mathbb{R}$ the sequence $(\alpha x_n + \beta y_n)_{n \geq 1}$ is convergent as well. Furthermore, also $(x_n \cdot y_n)_{n \geq 1}$ is a convergent sequence. And if $\lim_{n \rightarrow \infty} y_n \neq 0$, then $\lim_{n \rightarrow \infty} (x_n / y_n)$ exists.

The corresponding limits may be evaluated by the following equations:

$$(1) \quad \lim_{n \rightarrow \infty} [\alpha x_n + \beta y_n] = \alpha \lim_{n \rightarrow \infty} x_n + \beta \lim_{n \rightarrow \infty} y_n.$$

$$(2) \quad \lim_{n \rightarrow \infty} [x_n y_n] = \left[\lim_{n \rightarrow \infty} x_n \right] \cdot \left[\lim_{n \rightarrow \infty} y_n \right].$$

$$(3) \quad \lim_{n \rightarrow \infty} y_n \neq 0 \Rightarrow \lim_{n \rightarrow \infty} \frac{x_n}{y_n} = \frac{\lim_{n \rightarrow \infty} x_n}{\lim_{n \rightarrow \infty} y_n}.$$

Proof: Denote the limit of the x_n 's by L and that of the y_n 's by M .

(1) Suppose that α and β are nonzero. Otherwise, the result is true by trivial reasons. Given $\varepsilon > 0$, choose N_1 and N_2 such that for $n \geq N_1$ or $n \geq N_2$

$$|x_n - L| < \frac{\varepsilon}{2|\alpha|} \quad \text{and} \quad |y_n - M| < \frac{\varepsilon}{2|\beta|},$$

respectively. Let $N = \max\{N_1, N_2\}$. Then, if $n \geq N$, both previous estimates hold and lead to

$$|(\alpha L + \beta M) - (\alpha x_n + \beta y_n)| \leq |\alpha| |L - x_n| + |\beta| |M - y_n| < |\alpha| \frac{\varepsilon}{2|\alpha|} + |\beta| \frac{\varepsilon}{2|\beta|} = \varepsilon.$$

³Consider the x_n 's and z_n 's as slices of bread and the y_n 's as the cheese between them. Look what happens if $n \rightarrow \infty$.

Since $\varepsilon > 0$ was arbitrary, this proves

$$\lim_{n \rightarrow \infty} (\alpha x_n + \beta y) = \alpha L + \beta M,$$

as asserted.

(2) First note that the convergent sequence $(x_n)_{n \geq 1}$ is bounded by Proposition 5.2.6. Hence, there is a $c > 0$ with $|x_n| < c$ for all $n \geq 1$.

Choose $\varepsilon > 0$. By assumption, there are integers N_1 and N_2 such that

$$|L - x_n| < \frac{\varepsilon}{2(|M| + 1)} \quad \text{and} \quad |M - y_n| < \frac{\varepsilon}{2c},$$

whenever $n \geq N_1$ or $n \geq N_2$, respectively. Hence, if $n \geq N = \max\{N_1, N_2\}$, then it follows that

$$\begin{aligned} |L \cdot M - x_n \cdot y_n| &= |x_n(M - y_n) + M(L - x_n)| \leq |x_n| |M - y_n| + |M| |L - x_n| \\ &\leq c |M - y_n| + |M| |L - x_n| < \frac{c\varepsilon}{2c} + \frac{|M|\varepsilon}{2(|M| + 1)} < \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon. \end{aligned}$$

Since $\varepsilon > 0$ was arbitrary, this completes the proof of property (2).

(3) In view of property (2) we only have to prove the following: Whenever $\lim_{n \rightarrow \infty} y_n = M$ with $M \neq 0$, then

$$(5.2.8) \quad \frac{1}{y_n} \xrightarrow{n \rightarrow \infty} \frac{1}{M}.$$

Indeed, then by (2)

$$\frac{x_n}{y_n} = x_n \cdot \frac{1}{y_n} \xrightarrow{n \rightarrow \infty} L \cdot \frac{1}{M} = \frac{L}{M}.$$

To verify (5.2.8) we need the following preliminary observation. Since $M \neq 0$, there is an integer N_1 such that

$$|M - y_n| < \frac{|M|}{2} \quad \text{whenever } n \geq N_1.$$

From this we derive

$$|M| = |M - y_n + y_n| \leq |y_n| + |M - y_n| < |y_n| + \frac{|M|}{2},$$

hence, $|y_n| > |M|/2$ whenever $n \geq N_1$.

Given $\varepsilon > 0$, choose $N \geq N_1$ such that (recall $|M| > 0$)

$$|M - y_n| < \frac{\varepsilon |M|^2}{2} \quad \text{whenever } n \geq N.$$

Then $n \geq N$ implies $|y_n| > |M|/2$, and by the previous estimate we conclude that

$$\left| \frac{1}{M} - \frac{1}{y_n} \right| = \frac{|M - y_n|}{|M||y_n|} < \frac{|M - y_n|}{|M|^2/2} < \frac{2}{|M|^2} \frac{\varepsilon |M|^2}{2} = \varepsilon.$$

Again, since $\varepsilon > 0$ was arbitrary, this proves (5.2.8) and completes the proof of the proposition. ■

Example 5.2.5. We want to apply the previous properties to show that

$$\lim_{n \rightarrow \infty} \frac{3n^3 - n^2 + 1}{2n^3 + n} = \frac{3}{2}.$$

To do so rewrite the sequence as follows:

$$\frac{3n^3 - n^2 + 1}{2n^3 + n} = \frac{3 - 1/n + 1/n^3}{2 + 1/n^2}.$$

Now by the linearity of the limit the nominator tends to 3 as $n \rightarrow \infty$ while the denominator converges to 2. Hence, by property (3) the limit is 3/2 as asserted.

Although Proposition 5.2.8 is very helpful for the evaluation of certain limits, some important cases are not covered by it. These are so-called **indeterminate forms**. Let us explain this by an example.

Example 5.2.6. Find

$$\lim_{n \rightarrow \infty} [\sqrt{n+1} - \sqrt{n}] .$$

Since

$$\lim_{n \rightarrow \infty} \sqrt{n+1} = \lim_{n \rightarrow \infty} \sqrt{n} = \infty ,$$

we get an indeterminate expression of the form $\infty - \infty$ which is not covered by Proposition 5.2.8. So a special treatment is necessary. In our case we multiply the expression by $\sqrt{n+1} + \sqrt{n}$ and obtain

$$\lim_{n \rightarrow \infty} [\sqrt{n+1} - \sqrt{n}] = \lim_{n \rightarrow \infty} \frac{[\sqrt{n+1} - \sqrt{n}][\sqrt{n+1} + \sqrt{n}]}{\sqrt{n+1} + \sqrt{n}} = \lim_{n \rightarrow \infty} \frac{1}{\sqrt{n+1} + \sqrt{n}} = 0.$$

Here, in the last step we used $\lim_{n \rightarrow \infty} \sqrt{n} = \infty$ as well as Proposition 5.2.5.

The previous example is one of seven special indeterminate forms. Besides $\infty - \infty$ these are

$$\frac{\infty}{\infty}, \quad \frac{0}{0}, \quad 0 \cdot \infty, \quad 0^0, \quad \infty^0 \quad \text{and} \quad 1^\infty.$$

What is meant by these forms? For example, $\frac{\infty}{\infty}$ means that there are two sequences $(x_n)_{n \geq 1}$ and $(y_n)_{n \geq 1}$, both converging to ∞ , and we want to evaluate

$$\lim_{n \rightarrow \infty} \frac{x_n}{y_n}.$$

In the same way a form 0^0 occurs if $\lim_{n \rightarrow \infty} x_n = \lim_{n \rightarrow \infty} y_n = 0$ and we want to know

$$\lim_{n \rightarrow \infty} (x_n)^{y_n}.$$

Summary: There are seven indeterminate forms, i.e., expressions where Proposition 5.2.8 does not apply. In general the limit of these expressions cannot be evaluated by only knowing the behavior of the underlying sequences $(x_n)_{n \geq 1}$ and $(y_n)_{n \geq 1}$. Mostly, these forms need a special treatment. For example, if both sequences converge to ∞ , then x_n/y_n may tend to infinity or to some real number or the quotient may even diverge. Everything is possible here.

Let us investigate now the behavior of some important indeterminate forms.

Proposition 5.2.9. *The following are valid:*

- (1) $\lim_{n \rightarrow \infty} \sqrt[n]{a} = 1 \text{ for any } a > 0.$
- (2) $\lim_{n \rightarrow \infty} \sqrt[n]{n} = 1.$
- (3) $\lim_{n \rightarrow \infty} \sqrt[n]{n!} = \infty.$
- (4) $\lim_{n \rightarrow \infty} \frac{n^a}{b^n} = 0 \text{ for each } b > 1 \text{ and each } a > 0.$
- (5) $\lim_{n \rightarrow \infty} \frac{a^n}{n!} = 0 \text{ for all } a > 1.$
- (6) $\lim_{n \rightarrow \infty} \frac{n!}{n^n} = 0.$

Proof: Assertions (1) and (2) are left as exercise (cf. Exercise 5.2.15 below).

In order to prove (3) we have to show that we find for any $K > 0$ an $N \in \mathbb{N}$ such that

$$\sqrt[n]{n!} > K \quad \text{provided that } n > N,$$

or, equivalently,

$$(5.2.9) \quad \frac{K^n}{n!} < 1 \quad \text{if } n > N.$$

To verify (5.2.9) we use property (5) proved below. An application of this with $K = a$ tells us that

$$\lim_{n \rightarrow \infty} \frac{K^n}{n!} = 0.$$

Thus, there is an integer N depending on K such that (5.2.9) is valid whenever $n > N$. This completes the proof of (3).

To prove (4) we may suppose that $a \in \mathbb{N}$. Note that $a \mapsto n^a$ is increasing. Define $x > 0$ by $b = 1 + x$. Recall that $b > 1$. Thus, we have to show that

$$\lim_{n \rightarrow \infty} \frac{n^a}{(1+x)^n} = 0$$

for any $a \in \mathbb{N}$ and each $x > 0$. The binomial formula as well as $x > 0$ imply

$$(1+x)^n = \sum_{j=0}^n \binom{n}{j} x^j \geq \binom{n}{k} x^k = n \cdots (n-k+1) \cdot \frac{x^k}{k!}$$

for any $0 \leq k \leq n$. Consequently, whenever $k \leq n$, then we get

$$\frac{n^a}{(1+x)^n} \leq \frac{n^a}{n \cdots (n-k+1)} \cdot \frac{k!}{x^k}.$$

Choose now some $n > a+1$. Applying the previous estimate with $k = a+1$ leads to

$$(5.2.10) \quad \frac{n^a}{(1+x)^n} \leq \underbrace{\frac{n}{n} \cdots \frac{n}{n-a+1}}_{a \text{ fractions}} \cdot \frac{1}{n-a} \cdot \frac{(a+1)!}{x^{a+1}}.$$

Now, if $n \rightarrow \infty$, the first a fractions all tend to 1 while $1/(n-a)$ tends to zero. So the whole right expression of (5.2.10) tends to zero which completes the proof of (4).

To verify (5) it suffices to investigate the behavior of $a^n/n!$ for large $n \in \mathbb{N}$. So we may assume $n > a+1$. Next we choose an arbitrary, but fixed integer $k < n$ with $k > a$. With this integer k we argue as follows:

$$0 \leq \frac{a^n}{n!} = \frac{a^k}{k!} \cdot \underbrace{\frac{a}{k+1} \cdots \frac{a}{n}}_{n-k \text{ fractions}} \leq \frac{a^{k+1}}{k!} \cdot \frac{1}{n} \xrightarrow{n \rightarrow \infty} 0.$$

Here we used that by the choice of k we always have

$$\frac{a}{k+1} \leq 1, \dots, \frac{a}{n-1} \leq 1.$$

This completes the proof of (5) by virtue of the sandwich theorem, Proposition 5.2.7.

Assertion (6) follows from

$$0 \leq \frac{n!}{n^n} = \frac{1}{n} \cdot \frac{2}{n} \cdots \frac{n}{n} \leq \frac{1}{n} \xrightarrow{n \rightarrow \infty} 0$$

via Proposition 5.2.7. ■

Let us shortly discuss the previous results:

- (i) Properties (1) and (2) become clearer in the logarithmic level. In this setting they say that

$$\lim_{n \rightarrow \infty} \frac{\log a}{n} = 0 \quad \text{and} \quad \lim_{n \rightarrow \infty} \frac{\log n}{n} = 0, \quad \text{respectively.}$$

In other words, as $n \rightarrow \infty$, the sequence $\log n$ tends to ∞ much slower than n .

- (ii) Property (4) says that an exponential increase is always stronger than a polynomial one. No matter, how big $a > 0$ and how small $b - 1 > 0$, in the long run b^n always beats n^a . So, for example, one gets

$$\lim_{n \rightarrow \infty} \frac{n^{1000}}{1.01^n} = 0,$$

although the first values of this sequence suggest that it converges to infinity.

- (iii) Properties (3), (5) and (6) are tightly related to Stirling's formula⁴. It asserts that

$$n! \sim \sqrt{2\pi n} \left(\frac{n}{e}\right)^n, \quad n \rightarrow \infty.$$

Here " \sim " means that the quotient of both sequences tends to 1 as $n \rightarrow \infty$ and e denotes the Euler number (cf. Definition 5.3.1 below).

It is not difficult to verify properties (3), (5) and (6) by virtue of Stirling's formula. Since we did not prove Stirling's formula, we preferred to give direct proofs of these assertions.

Remark 5.2.5. There is another very important indeterminate form not covered by Proposition 5.2.9. This is the sequence

$$\left(1 + \frac{x}{n}\right)^n, \quad x \in \mathbb{R}.$$

⁴Named after the Scottish mathematician James Stirling (1692–1770).

Note that this is an indeterminate form of type 1^∞ . We will thoroughly investigate the behavior of this sequence for $x = 1$ in the subsequent Theorem 5.3.7. The general case $x \in \mathbb{R}$ will be shortly discussed in Proposition 5.6.14 below.

Exercise 5.2.1. Let

$$x_n = \frac{n^2}{(n-1)^2 + 1}, \quad n = 1, 2, \dots$$

For any given $\epsilon > 0$, find an integer $N = N(\epsilon) \in \mathbb{N}$ such that

$$|x_n - 1| < \epsilon,$$

whenever $n \geq N$.

Exercise 5.2.2. Let $(x_n)_{n \geq 1}$ and $(y_n)_{n \geq 1}$ be two sequences of real numbers such that

$$\lim_{n \rightarrow \infty} |x_n - y_n| = 0.$$

Give a proof of the following fact: The sequence $(x_n)_{n \geq 1}$ converges to some $L \in \mathbb{R}$ if and only if $(y_n)_{n \geq 1}$ does so.

Exercise 5.2.3. Let $a < b$ be two real numbers. Suppose the sequence $(x_n)_{n \geq 1}$ converges to some $L \in \mathbb{R}$. Show that $a \leq x_n \leq b$ for all $n \in \mathbb{N}$ implies $a \leq L \leq b$. Another way to express this is to say that the interval $[a, b]$ is closed under taking limits of sequences.

Exercise 5.2.4. Suppose

$$\lim_{n \rightarrow \infty} x_n = L > 0.$$

Show that there is an $N \in \mathbb{N}$ such that $x_n > L/2$ if $n \geq N$. In particular, this implies $x_n > 0$ for $n \geq N$.

Exercise 5.2.5. Prove the following: If for some $L \in \mathbb{R}$ we have

$$\lim_{n \rightarrow \infty} x_n = L,$$

then for any given $b > L$ there exists an $N \in \mathbb{N}$ (depending on b) such that

$$x_n < b \quad \text{provided that } n \geq N.$$

Similarly, for any $a < L$ there is some (maybe different) $N \in \mathbb{N}$ such that

$$x_n > a \quad \text{provided that } n \geq N.$$

Why is Exercise 5.2.4 a direct consequence of the latter property?

Exercise 5.2.6. Let $(x_n)_{n \geq 1}$ be a sequence of real numbers. Fix some $m \in \mathbb{N}$. Show that $(x_n)_{n \geq 1}$ is convergent if and only if the shifted sequence $(x_{n+m})_{n \geq 1}$ is so, and, moreover,

$$\lim_{n \rightarrow \infty} x_{n+m} = \lim_{n \rightarrow \infty} x_n.$$

Exercise 5.2.7. Suppose we are given a sequence of **integers**. Give a necessary and sufficient condition for its convergence.

Exercise 5.2.8. Show the reverse implication in Proposition 5.2.5. That is, prove that for positive x_n 's with $1/x_n \xrightarrow[n \rightarrow \infty]{} 0$ always follows that $x_n \xrightarrow[n \rightarrow \infty]{} \infty$,

Exercise 5.2.9. Which of the implications in Proposition 5.2.5 remain valid in the case of arbitrary (not necessarily positive) $x_n \neq 0$, which one become false?

Exercise 5.2.10. Verify that

$$\lim_{n \rightarrow \infty} \frac{n-1}{\sqrt{n+1}} = \infty.$$

Exercise 5.2.11. Let $(x_n)_{n \geq 1}$ and $(y_n)_{n \geq 1}$ be two sequences of real numbers such that

$$x_n \leq y_n, \quad n = 1, 2, \dots$$

Show the following:

$$\lim_{n \rightarrow \infty} x_n = \infty \Rightarrow \lim_{n \rightarrow \infty} y_n = \infty.$$

Exercise 5.2.12. Let $(x_n)_{n \geq 1}$ and $(y_n)_{n \geq 1}$ be two sequences of positive real numbers with $\inf_{n \geq 1} y_n > 0$. Show that

$$\lim_{n \rightarrow \infty} x_n = \infty$$

implies

$$\lim_{n \rightarrow \infty} x_n y_n = \infty.$$

Exercise 5.2.13. Give an example of a sequence $(x_n)_{n \geq 1}$ such that

$$\inf_{n \geq 1} x_n = -\infty \text{ but } \lim_{n \rightarrow \infty} x_n \neq -\infty.$$

Exercise 5.2.14. Evaluate the following limits:

$$\begin{aligned} a) \quad & \lim_{n \rightarrow \infty} \frac{4 - 9n^2}{5n^2}, \quad b) \quad \lim_{n \rightarrow \infty} \frac{n+1}{n^2 - 1}, \quad c) \quad \lim_{n \rightarrow \infty} \frac{8n^3 - 4n + 5}{2n^2 - n + 4}, \\ d) \quad & \lim_{n \rightarrow \infty} \frac{2n + (-1)^n}{n}, \quad e) \quad \lim_{n \rightarrow \infty} \frac{1 - 0.1^n}{0.9}, \quad f) \quad \lim_{n \rightarrow \infty} \frac{(n+1)^6 - n^6}{n^5}. \end{aligned}$$

Exercise 5.2.15. Show that

$$(5.2.11) \quad \lim_{n \rightarrow \infty} \sqrt[n]{n} = 1.$$

Hint: Set $x_n = \sqrt[n]{n} - 1$, hence

$$n = (1 + x_n)^n,$$

and apply the binomial formula to conclude that $x_n \xrightarrow{n \rightarrow \infty} 0$.

Why does (5.2.11) imply

$$\lim_{n \rightarrow \infty} \sqrt[n]{a} = 1,$$

for any given $0 < a < \infty$?

5.3. The Monotone Convergence Theorem and Its Applications

We turn now to one of the most important facts about real sequences. Its validity essentially rests on the completeness axiom (III) of the real line. In fact one may prove that this result is even equivalent to the existence of lowest upper bounds of nonvoid sets which are bounded above.

Theorem 5.3.1 (Monotone Convergence Theorem). *A monotone sequence of real numbers either converges or diverges properly. The former happens if and only if the sequence is bounded.*

$$\begin{aligned} ((x_n)_{n \geq 1} \text{ monotone}) &\Rightarrow [(\lim_{n \rightarrow \infty} x_n \text{ exists}) \Leftrightarrow (\sup_{n \geq 1} |x_n| < \infty)]. \\ ((x_n)_{n \geq 1} \text{ nondecreasing}) &\Rightarrow [(\lim_{n \rightarrow \infty} x_n = \infty) \Leftrightarrow (\sup_{n \geq 1} x_n = \infty)]. \\ ((x_n)_{n \geq 1} \text{ nonincreasing}) &\Rightarrow [(\lim_{n \rightarrow \infty} x_n = -\infty) \Leftrightarrow (\inf_{n \geq 1} x_n = -\infty)]. \end{aligned}$$

Proof: We already know by Proposition 5.2.4 that each converging sequence is bounded. Let us conversely suppose that $(x_n)_{n \geq 1}$ is a monotone and bounded sequence. Without loss of generality assume that the sequence is nondecreasing. If not, consider $(-x_n)_{n \geq 1}$. Since $(x_n)_{n \geq 1}$ is bounded, its supremum $L = \sup_{n \geq 1} x_n$ is a well-defined real number. We claim now that $\lim_{n \rightarrow \infty} x_n = L$. Thus, take an arbitrary $\varepsilon > 0$. Since L is the supremum of the x_n 's, there exists an $N \geq 1$ such that

$$L - \varepsilon < x_N \leq L.$$

Because $(x_n)_{n \geq 1}$ is nondecreasing, it follows

$$x_N \leq x_{N+1} \leq \dots \leq L.$$

Recall that L is an upper bound of the x_n 's. Hence, if $n \geq N$,

$$L - \varepsilon < x_N \leq x_n \leq L < L + \varepsilon,$$

i.e., $|L - x_n| < \varepsilon$ whenever $n \geq N$. This proves $x_n \xrightarrow[n \rightarrow \infty]{} x$ as asserted.

It remains to prove that $x_n \xrightarrow[n \rightarrow \infty]{} \infty$ provided that $\sup_{n \geq 1} x_n = \infty$. The latter implies the following: Given $K \geq 1$ there is an N such that $x_N > K$. But since $(x_n)_{n \geq 1}$ is nondecreasing, it follows

$$K < x_N \leq x_n \quad \text{for all } n \geq N.$$

This proves $x_n \xrightarrow[n \rightarrow \infty]{} \infty$ as claimed. ■

Corollary 5.3.2. *If a sequence $(x_n)_{n \geq 1}$ is nondecreasing or nonincreasing, then*

$$\lim_{n \rightarrow \infty} x_n = \sup_{n \geq 1} x_n \quad \text{or} \quad \lim_{n \rightarrow \infty} x_n = \inf_{n \geq 1} x_n,$$

respectively.

Let us state a first result which shows how Theorem 5.3.1 applies.

Proposition 5.3.3. *Let $a \in \mathbb{R}$ be a real number with $|a| < 1$. Then this implies*

$$(5.3.1) \quad \lim_{n \rightarrow \infty} a^n = 0.$$

Proof: First note that it suffices to investigate positive numbers $0 < a < 1$. This follows directly by

$$\lim_{n \rightarrow \infty} a^n = 0 \quad \Leftrightarrow \quad \lim_{n \rightarrow \infty} |a|^n = 0.$$

Since

$$a > a^2 > a^3 > \dots > 0,$$

the sequence a, a^2, a^3, \dots is decreasing and bounded below. By Theorem 5.3.1 the limit

$$L = \lim_{n \rightarrow \infty} a^n = \inf_{n \geq 1} a^n$$

exists. We claim now that $L = 0$. To this end investigate $a \cdot L$. Using Exercise 5.2.6 it follows that

$$a \cdot L = a \cdot \lim_{n \rightarrow \infty} a^n = \lim_{n \rightarrow \infty} a^{n+1} = L.$$

Since $a \neq 1$, this proves $L = 0$ as asserted. ■

Remark 5.3.1. If $1 < a < \infty$, an application of Proposition 5.3.1 with $1/a$ implies

$$\lim_{n \rightarrow \infty} \frac{1}{a^n} = 0.$$

Thus, by Proposition 5.2.5 it follows

$$(5.3.2) \quad \lim_{n \rightarrow \infty} a^n = \infty.$$

As indicated in Exercise 5.1.10, another way to verify (5.3.2) is via Bernoulli's inequality as stated in Exercise 1.2.5.

Our next aim is to investigate the limit behavior of the sequence $(F_n)_{n \geq 0}$ of Fibonacci numbers as defined in (1.4.3). Recall that these numbers satisfy

$$F_0 = 0, F_1 = 1 \quad \text{and} \quad F_{n+1} = F_n + F_{n-1}, \quad n \geq 1.$$

As proved in Proposition 1.4.7, an explicit representation of these numbers is given by

$$(5.3.3) \quad F_n = \frac{\varphi^n - \psi^n}{\sqrt{5}}$$

with golden ratio φ and ψ defined by

$$\varphi := \frac{1 + \sqrt{5}}{2} \quad \text{and} \quad \psi := \frac{1 - \sqrt{5}}{2}.$$

Since $|\psi| < \varphi$, one easily gets

$$\lim_{n \rightarrow \infty} F_n = \infty.$$

Thus, the general behavior of the Fibonacci sequence is clarified. But if we look at the ratio between F_{n+1} and F_n for certain $n \geq 1$, we see that

$$\frac{F_2}{F_1} = 1, \quad \frac{F_3}{F_2} = 2, \quad \dots, \quad \frac{F_{11}}{F_{10}} = \frac{89}{55} = 1.61\overline{8}, \quad \dots, \quad \frac{F_{41}}{F_{40}} = \frac{165580141}{102334155} \approx 1.61803401.$$

Already these few ratios suggest⁵ that the ratio F_{n+1}/F_n should converge to

$$\varphi = \frac{1 + \sqrt{5}}{2} \approx 1.6180339887.$$

Let us verify this now.

⁵This was already observed by Johannes Kepler (1571–1630).

Proposition 5.3.4. *Let $(F_n)_{n \geq 0}$ be the sequence of Fibonacci numbers. Then it follows that*

$$\lim_{n \rightarrow \infty} \frac{F_{n+1}}{F_n} = \frac{1 + \sqrt{5}}{2} = \varphi.$$

Proof: Our first proof is related to the finite continued fraction expansion of $F_{n+1}/F_n = [1; \underbrace{1, \dots, 1}_{n-1 \text{ terms}}]$ that we discussed in Example 3.4.6. The sequence $(x_n)_{n \geq 1}$ with $x_n = F_{n+1}/F_n$ satisfies $x_1 = 1$ and

$$x_{n+1} = 1 + \frac{1}{x_n}, \quad n \geq 1.$$

If one is able to show the convergence of this sequence to some limit $L \in \mathbb{R}$, then necessarily L is nonnegative and

$$L = 1 + \frac{1}{L} \Rightarrow L = \frac{1 + \sqrt{5}}{2}.$$

The convergence of the sequence $(x_n)_{n \geq 0}$ follows easily by showing that the subsequence $(x_{2n+1})_{n \geq 0}$ with odd indices is increasing and the subsequence $(x_{2n})_{n \geq 0}$ with even indices is decreasing. This can easily be shown, but follows also by Proposition 3.4.6.

We prefer to give a direct proof which rests on the explicit representation (5.3.3) of the Fibonacci numbers. Using the same notation as in (5.3.3), it follows that

$$\frac{F_{n+1}}{F_n} = \frac{\varphi^{n+1} - \psi^{n+1}}{\varphi^n - \psi^n} = \varphi \cdot \frac{1 - (\psi/\varphi)^{n+1}}{1 - (\psi/\varphi)^n}.$$

Since $|\psi| = \frac{\sqrt{5}-1}{2} < \sqrt{5} + 1 = \varphi$, we get $|\psi/\varphi| < 1$. Hence, Proposition 5.3.1 applies and leads to

$$\lim_{n \rightarrow \infty} \left(\frac{\psi}{\varphi} \right)^n = 0.$$

Consequently, together with property (3) in Proposition 5.2.8 we get

$$\lim_{n \rightarrow \infty} \frac{F_{n+1}}{F_n} = \varphi \cdot \lim_{n \rightarrow \infty} \frac{1 - (\psi/\varphi)^{n+1}}{1 - (\psi/\varphi)^n} = \varphi = \frac{1 + \sqrt{5}}{2}$$

as asserted. ■

Our next example for the application of Theorem 5.3.1 is a sequence already introduced in Example 5.1.3. Sequences defined in this way are called Heron⁶ sequences. For the general case we refer to Exercise 5.3.1. We should also mention that one gets the same iteration formula by applying Newton's Method to the function $f(x) = x^2 - 2$. But Newton's method relies on basic facts from Calculus, hence it is not a topic of this textbook.

⁶Named after Heron of Alexandria (10–70 CE)

Proposition 5.3.5. *If the sequence $(x_n)_{n \geq 1}$ of real numbers is defined inductively by $x_1 = 2$ and*

$$x_{n+1} = \frac{x_n}{2} + \frac{1}{x_n}, \quad n \geq 2,$$

then it follows that

$$(5.3.4) \quad \lim_{n \rightarrow \infty} x_n = \sqrt{2}.$$

Proof: In a first step we show that the sequence is bounded from below. More precisely, we claim that

$$(5.3.5) \quad x_n \geq \sqrt{2}, \quad n \geq 1.$$

Since $x_1 = 2$ this is valid for $n = 1$.

If $n \geq 2$, then we apply the inequality in Proposition 1.1.4 asserting

$$(a + b)^2 \geq 4ab, \quad a, b \in \mathbb{R}.$$

Letting $a = x_{n-1}/2$ and $b = 1/x_{n-1}$ leads to

$$x_n^2 = \left(\frac{x_{n-1}}{2} + \frac{1}{x_{n-1}} \right)^2 \geq 4 \left(\frac{x_{n-1}}{2} \cdot \frac{1}{x_{n-1}} \right) = 2,$$

thus, (note that the x_n s are positive) we obtain $x_n \geq \sqrt{2}$ as claimed in (5.3.5).

Next we show that the x_n s are nonincreasing. Using (5.3.5) this follows because of

$$x_{n+1} = \frac{x_n}{2} + \frac{1}{x_n} = x_n \left(\frac{1}{2} + \frac{1}{x_n^2} \right) \leq x_n \left(\frac{1}{2} + \frac{1}{(\sqrt{2})^2} \right) = x_n.$$

So, our sequence $(x_n)_{n \geq 1}$ is nonincreasing and bounded below. Due to Theorem 5.3.1 it converges to some $L \geq \sqrt{2}$ (apply Exercise 5.2.3 together with (5.3.5)). Using the assertion of Exercise 5.2.6 we get

$$L = \lim_{n \rightarrow \infty} x_{n+1} = \lim_{n \rightarrow \infty} \left(\frac{x_n}{2} + \frac{1}{x_n} \right) = \frac{L}{2} + \frac{1}{L},$$

hence $L^2 = \frac{L^2}{2} + 1$ implying $L^2 = 2$. Since $L > 0$ it follows that $L = \sqrt{2}$ as asserted in (5.3.4). ■

Remark 5.3.2. The sequence $(x_n)_{n \geq 1}$ investigated in the previous example approximates $\sqrt{2}$ very fast. To see this compare x_6 with the first digits of the decimal expansion of $\sqrt{2}$. Here we have

$$\begin{aligned} x_6 &= 1,41421356237309504880168, \quad \text{while} \\ \sqrt{2} &= 1,41421356237309504880166. \end{aligned}$$

Our next objective is the investigation of a sequence tightly related to the continued fraction presented in Example 3.4.7. It answers the question about the convergence of the sequence of finite fractions.

Proposition 5.3.6. Suppose the sequence $(x_n)_{n \geq 1}$ of real numbers is defined by $x_1 = 1$ and

$$x_{n+1} = 1 + \frac{1}{1+x_n}, \quad n = 1, 2, \dots$$

Then this implies

$$\lim_{n \rightarrow \infty} x_n = \sqrt{2}.$$

Proof: Define $f : [0, \infty) \rightarrow \mathbb{R}$ by

$$f(x) := 1 + \frac{1}{1+x}, \quad x \geq 0.$$

This function possesses the following properties:

- (1) The function is decreasing.
- (2) $f : [0, \infty) \rightarrow [1, 2]$.
- (3) $(\forall n \geq 1)(x_{n+1} = f(x_n))$.

If

$$f^2(x) = (f \circ f)(x) = f(f(x)),$$

then f^2 is increasing and, moreover,

$$f^2(x_n) = x_{n+2}, \quad n = 1, 2, \dots$$

The first 4 values of the sequence are

$$x_1 = 1, \quad x_2 = 1.5, \quad x_3 = 1.4, \quad \text{and} \quad x_4 = 1.41667.$$

So, we observe that $x_1 < x_3$ and $x_2 > x_4$, hence an inductive application of f^2 leads to

$$x_1 < x_3 < x_5 < x_7 < \dots \quad \text{and} \quad x_2 > x_4 > x_6 > \dots$$

Let us mention that the monotonicity of these sequences would also follow by an application of Proposition 3.4.6. But we believe that in this special case it is easier and more natural to give a direct proof.

So we got that both sequences $(x_{2n})_{n \geq 1}$ and $(x_{2n+1})_{n \geq 1}$ are monotone and bounded. Recall that f attains values in $[1, 2)$. Now Theorem 5.3.1 applies and the limits $L, L' > 0$ defined by

$$L := \lim_{n \rightarrow \infty} x_{2n} \quad \text{as well as} \quad L' := \lim_{n \rightarrow \infty} x_{2n+1}$$

exist. By the definition of L as limit we get⁷

$$(5.3.6) \quad L = \lim_{n \rightarrow \infty} x_{2n+2} = \lim_{n \rightarrow \infty} f^2(x_{2n}) = f^2(L).$$

Elementary calculations give

$$f^2(x) = \frac{2 + f(x)}{1 + f(x)} = \frac{4 + 3x}{3 + 2x}, \quad x \geq 0.$$

Thus, by (5.3.6) it follows that

$$L = \frac{4 + 3L}{3 + 2L} \quad \Rightarrow \quad L^2 = 2 \quad \Rightarrow \quad L = \sqrt{2}.$$

⁷We use here a continuity argument which in the case of the concrete function f easily follows by Proposition 5.2.8.

Exactly by the same arguments for L' we get $L' = \sqrt{2}$, hence

$$\lim_{n \rightarrow \infty} x_n = \sqrt{2},$$

as asserted. ■

The next application of Theorem 5.3.1 is maybe the most important one. It proves the existence of the Euler number e , which plays an outstanding role not only in mathematics but also in many other sciences.

Theorem 5.3.7. *The sequence $(x_n)_{n \geq 1}$ defined by*

$$x_n = \left(1 + \frac{1}{n}\right)^n$$

is convergent.

Remark 5.3.3. The limit is an indeterminate form 1^∞ . Note that $1 + \frac{1}{n} \xrightarrow[n \rightarrow \infty]{} 1$.

Proof: Our goal is to show that $(x_n)_{n \geq 1}$ is increasing and bounded above. Then the assertion is a consequence of Theorem 5.3.1.

(1) In a first step we prove the monotonicity of the sequence. So choose $n \geq 2$ and consider

$$\frac{x_n}{x_{n-1}} = \left(\frac{n+1}{n}\right)^n \left(\frac{n-1}{n}\right)^{n-1} = \frac{n+1}{n} \left(1 - \frac{1}{n^2}\right)^{n-1}.$$

If we are able to verify that this quotient is strictly greater than 1, then we are done. Recall Bernoulli's inequality (cf. Exercise 1.2.5). It asserts that

$$(5.3.7) \quad (1+x)^n \geq 1+nx \quad \text{if } n \geq 1 \text{ and } x \geq -1.$$

Let us apply this estimate with $x = -1/n^2$ and with $n-1$ instead of n . Doing so we arrive at

$$\frac{x_n}{x_{n-1}} \geq \frac{n+1}{n} \left[1 - \frac{n-1}{n^2}\right] = 1 + \frac{1}{n^3} > 1$$

as asserted. Consequently, the sequence $(x_n)_{n \geq 1}$ is increasing.

(2) Our next aim is to show that $(x_n)_{n \geq 1}$ is bounded above. To this end we replace $(x_n)_{n \geq 1}$ by a slightly changed sequence $(y_n)_{n \geq 1}$ defined by

$$y_n = \left(1 + \frac{1}{n}\right)^{n+1}, \quad n \geq 1.$$

Of course $x_n < y_n$, and if we are able to show that the y_n 's are decreasing, then the boundedness of the sequence $(x_n)_{n \geq 1}$ follows by

$$x_n < y_n \leq y_1 = 4.$$

So it remains to prove that $(y_n)_{n \geq 1}$ is decreasing. Another application of Bernoulli's inequality (5.3.7), this time with $x = 1/(n^2 - 1)$, leads for $n \geq 2$ to the following estimate:

$$\begin{aligned} \frac{y_{n-1}}{y_n} &= \left(\frac{n}{n-1}\right)^n \left(\frac{n}{n+1}\right)^{n+1} = \frac{n}{n+1} \left(1 + \frac{1}{n^2-1}\right)^n \\ &\geq \frac{n}{n+1} \left[1 + \frac{n}{n^2-1}\right] > \frac{n}{n+1} \left[1 + \frac{1}{n}\right] = 1. \end{aligned}$$

Here, in the last step we used

$$\frac{n}{n^2 - 1} = \frac{1}{n} \cdot \frac{n^2}{n^2 - 1} > \frac{1}{n}.$$

Thus, $(y_n)_{n \geq 1}$ is decreasing, and as already mentioned, this implies $x_n < 4$ for all $n \geq 1$. Hence, $(x_n)_{n \geq 1}$ is increasing and bounded, thus convergent. This completes the proof. ■

Definition 5.3.1. The limit

$$e = \lim_{n \rightarrow \infty} \left(1 + \frac{1}{n}\right)^n$$

is called the **Euler Number**. It is approximately

$$e \approx 2.71828182845904523536028747135266249775724709369995 \dots .$$

In Corollary 5.6.13 we will show that e is irrational. Consequently, the decimal expansion of the Euler number e is nonperiodic.

Let us state an important property of the number e . As shown in the proof of Theorem 5.3.7 we have the following estimates:

Corollary 5.3.8.

$$\left(1 + \frac{1}{n}\right)^n < e < \left(1 + \frac{1}{n}\right)^{n+1}, \quad n \geq 1.$$

Moreover, the left sequence is increasing while the right one decreases, and both sequences converge to e .

Exercise 5.3.1. Let $b > 0$ be some fixed real number. Define a sequence $(x_n)_{n \geq 1}$ by $x_1 = b$ and

$$x_{n+1} = \frac{1}{2} \left(x_n + \frac{b}{x_n} \right), \quad n = 1, 2, \dots$$

Show that $(x_n)_{n \geq 1}$ converges to \sqrt{b} .

A sequence of this type is called a **Heron sequence**. The case $b = 2$ has been investigated in Proposition 5.3.5. Moreover, the result remains valid with any starting point $x_1 > 0$.

Exercise 5.3.2. Evaluate

$$\lim_{n \rightarrow \infty} \left(1 - \frac{1}{n}\right)^n \quad \text{and} \quad \lim_{n \rightarrow \infty} \left(1 + \frac{1}{n}\right)^{3n}.$$

Exercise 5.3.3. Evaluate the following limits or prove that they do not exist.

$$(1) \quad \lim_{n \rightarrow \infty} \sqrt[n]{n^2} \quad \text{and} \quad \lim_{n \rightarrow \infty} \sqrt[n]{(2n)!}.$$

$$(2) \quad \lim_{n \rightarrow \infty} n^3 2^{-n} \quad \text{and} \quad \lim_{n \rightarrow \infty} \frac{n^3}{n!}.$$

$$(3) \quad \lim_{n \rightarrow \infty} \left(1 + \frac{1}{n^2}\right)^n \quad \text{and} \quad \lim_{n \rightarrow \infty} \left(\frac{1}{n}\right)^{1/n}.$$

Exercise 5.3.4. Show that

$$\left(1 + \frac{1}{n}\right)^n \leq \sum_{k=0}^n \frac{1}{k!}.$$

Exercise 5.3.5. Let $n \in \mathbb{N}_0$. Prove that $e \geq \sum_{k=0}^n \frac{1}{k!}$.

Exercise 5.3.6. Which of the two following numbers is greater:

$$99^{100} \quad \text{or} \quad 100^{99}?$$

In which relation are 1000^{999} and 999^{1000} ? Justify your answers.

Exercise 5.3.7. Let $n \in \mathbb{N}$. Determine which inequality is true:

$$n^{n+1} < (n+1)^n \quad \text{or} \quad (n+1)^n < n^{n+1}?$$

Exercise 5.3.8. Let $k \in \mathbb{N}$ and a_1, \dots, a_k be some positive numbers. Suppose that $M = \max(a_1, \dots, a_k)$. Prove that $\lim_{n \rightarrow \infty} \sqrt[n]{a_1^n + \dots + a_k^n} = M$.

Exercise 5.3.9. Let $A \subseteq \mathbb{R}$ be a nonempty subset bounded above. Furthermore, assume that $\sup(A)$ is **not** attained. Prove that under these conditions the following is true: There exist $x_1 < x_2 < x_3 < \dots$ in A such that

$$\lim_{n \rightarrow \infty} x_n = \sup(A).$$

Why does this become false without the assumption that $\sup(A)$ is not attained?

Exercise 5.3.10. Let $(x_n)_{n \geq 1}$ be a **bounded** sequence of real numbers. Define the **upper limit**⁸ of the x_n 's by

$$\limsup_{n \rightarrow \infty} x_n := \lim_{n \rightarrow \infty} \left[\sup_{m \geq n} x_m \right].$$

- (1) Why does this upper limit always exist, even in the case of divergent sequences?
- (2) Show that the upper limit coincides with the limit in the case of convergent sequences.
- (3) Let $(x_n)_{n \geq 1}$ and $(y_n)_{n \geq 1}$ be two bounded sequences. Prove that

$$\limsup_{n \rightarrow \infty} (x_n + y_n) \leq \limsup_{n \rightarrow \infty} x_n + \limsup_{n \rightarrow \infty} y_n.$$

Give an example that shows that the left-hand side of this estimate may be strictly less than the right hand one.

- (4) The **lower limit** (or limit inferior) of the x_n 's is defined by

$$\liminf_{n \rightarrow \infty} x_n := \lim_{n \rightarrow \infty} \left[\inf_{m \geq n} x_m \right].$$

Show that this lower limit always exists and that

$$\liminf_{n \rightarrow \infty} x_n \leq \limsup_{n \rightarrow \infty} x_n.$$

⁸also called **limit superior**.

Moreover, for two bounded sequences $(x_n)_{n \geq 1}$ and $(y_n)_{n \geq 1}$ one has

$$\liminf_{n \rightarrow \infty} (x_n + y_n) \geq \liminf_{n \rightarrow \infty} x_n + \liminf_{n \rightarrow \infty} y_n.$$

- (5) Evaluate the upper and the lower limit of the sequence $\left((-1)^n \frac{n}{n+1}\right)_{n \geq 1}$.

5.4. Subsequences

Suppose we are given a sequence $(x_n)_{n \geq 1}$. What happens if we single out infinitely many elements of the sequence. For example, we may choose the first one, the tenth, the hundredth, the thousandth, and so on? Doing so we will get a new sequence which can be considered as a part of the old one. It is common to say that this new sequence is a subsequence of the given one.

Example 5.4.1. If $x_n = \frac{1}{n}$, $n \geq 1$, and we single out those elements at position 10^k for some $k \geq 1$, then the new sequence is

$$1, \frac{1}{10}, \frac{1}{100}, \frac{1}{1000}, \dots$$

Another example is as follows: Take the sequence $0, -\frac{1}{2}, \frac{2}{3}, -\frac{3}{4}, \frac{4}{5}, \dots$. That is

$$x_n = \begin{cases} \frac{n-1}{n} & : n \text{ odd} \\ -\frac{n-1}{n} & : n \text{ even.} \end{cases}$$

The sequence $(x_n)_{n \geq 1}$ is divergent but the two subsequences x_1, x_3, x_5, \dots and x_2, x_4, x_6, \dots converge to 1 and -1 , respectively. Of course, we may also choose other convergent subsequences as for example $x_2, x_4, x_8, x_{16}, \dots$

One of the basic questions treated in this section is whether **any** sequence contains a convergent subsequence. The answer is negative. Consider the sequence $1, 2, 3, \dots$. Then any of its subsequences tends to infinity as well. So there do not exist convergent subsequences. But how about bounded sequences. Is it in this case always possible to extract convergent subsequence?

Before proceeding further let us first fix in a precise way what it means that a sequence is a subsequence of a given $(x_n)_{n \geq 1}$.

Definition 5.4.1. Let $(x_n)_{n \geq 1}$ be a sequence of real numbers. Given an increasing sequence $n_1 < n_2 < \dots$ of natural numbers, the sequence x_{n_1}, x_{n_2}, \dots , shorter written as $(x_{n_k})_{k \geq 1}$, is called a **subsequence** of $(x_n)_{n \geq 1}$. In other words: If $(y_k)_{k \geq 1}$ is a sequence of real numbers, then we say

$$((y_k)_{k \geq 1} \text{ is a subsequence of } (x_n)_{n \geq 1}) \Leftrightarrow (\exists n_k \in \mathbb{N}, n_k \nearrow \infty)(\forall k \geq 1)(y_k = x_{n_k}).$$

Example 5.4.2. If $x_n = \frac{1}{n}$, then the sequence $\{1, \frac{1}{5}, \frac{1}{10}, \frac{1}{15}, \dots\}$ is a subsequence of $(x_n)_{n \geq 1}$. Another subsequence would be $\{1, \frac{1}{4}, \frac{1}{9}, \frac{1}{16}, \dots\}$ or one could choose $(1/n!)_{n \geq 1}$ as subsequence.

The following result is obvious.

Proposition 5.4.1. *Let $(x_{n_k})_{k \geq 1}$ be a subsequence of $(x_n)_{n \geq 1}$. If $\lim_{n \rightarrow \infty} x_n = L$ for some $L \in \mathbb{R}$, then also $\lim_{k \rightarrow \infty} x_{n_k} = L$.*

Proof: Given $\varepsilon > 0$, choose $N \geq 1$ such that $|x_n - L| < \varepsilon$ if $n \geq N$. Take any $K \geq 1$ for which $n_K \geq N$. Then $k \geq K$ implies $n_k \geq N$, hence

$$|x_{n_k} - L| < \varepsilon \quad \text{if } k \geq K.$$

This completes the proof. ■

Remark 5.4.1. Proposition 5.4.1 remains valid also in the case that the given sequence diverges properly. That is, if $\lim_{n \rightarrow \infty} x_n = \infty$, then this is so also for any subsequence of $(x_n)_{n \geq 1}$. The proof follows exactly by the same arguments as we used to verify Proposition 5.4.1.

We prove now an important property of real sequences.

Proposition 5.4.2. *Let $(x_n)_{n \geq 1}$ be an arbitrary sequence of real numbers. Then it contains a monotone subsequence.*

Proof: Fix a sequence $(x_n)_{n \geq 1}$. A number $r \geq 1$ is called a peak point of $(x_n)_{n \geq 1}$ if

$$x_r > x_n \quad \text{for all } n > r.$$

To visualize this definition one may think of the following. Suppose at each $n \geq 1$ there is a mountain of height x_n . Then r is a peak point of $(x_n)_{n \geq 1}$, if standing on the summit of the mountain at r and looking towards the direction of larger n , there is no mountain blocking the view to the horizon.

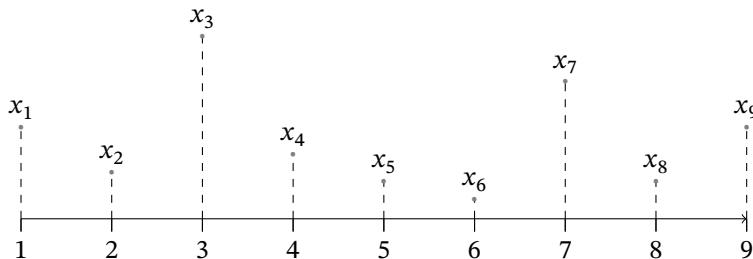


Figure 5.4.1. Peak points at $r = 3$, $r = 7$ and $r = 9$.

Now we distinguish two different cases:

Case 1: There are infinitely many peak points $n_1 < n_2 < n_3 < \dots$.

By the definition of peak points,

$$x_{n_1} > x_{n_2} > x_{n_3} > \dots$$

Consequently, we found a decreasing, hence monotone, subsequence of $(x_n)_{n \geq 1}$.

Case 2: The number of peak points is finite. Say there are $m \geq 0$ peak points r_1, \dots, r_m .

Then there is a $N \geq 1$ such that $r_j < N$, $j = 1, \dots, m$. Consequently, all $n \geq N$ are no peak points. Choose some $n_1 \geq N$. Then n_1 cannot be a peak point (all peak points are smaller than N) i.e., there is an $n_2 > n_1$ such that $x_{n_2} \geq x_{n_1}$. Why? Because, if such $n_2 > n_1$ did not exist, then n_1 would be a peak point contradicting $n_1 \geq N$.

Now $n_2 > n_1 \geq N$, hence n_2 is also not a peak point. By the same argument as for n_1 , there is an $n_3 > n_2$ such that $x_{n_3} \geq x_{n_2}$. Proceeding in this way we obtain a subsequence $(x_{n_k})_{k \geq 1}$ such that

$$x_{n_1} \leq x_{n_2} \leq \dots .$$

Thus, we found a nondecreasing subsequence. This completes the proof. ■

Example 5.4.3. If $x_n = (-1)^n/n$, then the peak points are $\{2, 4, 6, \dots\}$. On the other hand, the sequence $x_n = (-1)^n \frac{n}{n+1}$ does not possess any peak point.

Combining Theorem 5.3.1 with Proposition 5.4.2 leads to the following crucial result.

Theorem 5.4.3 (Bolzano–Weierstrass Theorem⁹). *Each bounded sequence of real numbers contains a convergent subsequence.*

$$(\sup_{n \geq 1} |x_n| < \infty) \Rightarrow (\exists (n_k)_{k \geq 1} \text{ increasing}) (\lim_{k \rightarrow \infty} x_{n_k} \text{ exists})$$

In order to get a deeper insight into the structure of sequences, our next aim is to reformulate the preceding theorem. To this end the following notation is useful.

Definition 5.4.2. Let $(x_n)_{n \geq 1}$ be an arbitrary sequence of real numbers. A point $L \in \mathbb{R}$ is said to be a **cluster point** or **accumulation point** of $(x_n)_{n \geq 1}$ provided that there is a subsequence $(x_{n_k})_{k \geq 1}$ of $(x_n)_{n \geq 1}$ such that

$$\lim_{k \rightarrow \infty} x_{n_k} = L .$$

Example 5.4.4. The sequence $1, 0, 1, 0, \dots$ possesses exactly two cluster points, namely 0 and 1. But note that the set of its values is the finite set $\{0, 1\}$ which does not cluster at all. Therefore, one has to distinguish carefully between cluster points of sequences and of sets.

In the notation of Definition 5.4.2 the previous theorem can now be formulated as follows:

Theorem 5.4.4. *Any bounded sequence possesses at least one cluster point.*

The next result gives a characterization of cluster points.

⁹It bears the names of Bernhard Bolzano (1781–1848) and of Karl Weierstrass (1815–1897). The result was first verified by Bernhard Bolzano in 1817 as a lemma in the proof of the intermediate value theorem. Later on the result was identified as significant in its own right, and reproved by Karl Weierstrass. Nowadays, it is one of the most essential theorems of analysis.

Proposition 5.4.5. A point $L \in \mathbb{R}$ is a cluster point of a sequence $(x_n)_{n \geq 1}$ if and only if any ε -neighborhood

$$U_\varepsilon(L) = \{x \in \mathbb{R} : |x - L| < \varepsilon\} = (L - \varepsilon, L + \varepsilon)$$

of L contains infinitely many of the x_n 's.

$$(L \text{ cluster point of } (x_n)_{n \geq 1}) \Leftrightarrow (\forall \varepsilon > 0)(|\{n \geq 1 : x_n \in U_\varepsilon(L)\}| = \infty).$$

Proof: Suppose first that L is a cluster point of the sequence $(x_n)_{n \geq 1}$. By definition there exists a subsequence $(x_{n_k})_{k \geq 1}$ which converges to L . Thus, given an $\varepsilon > 0$, there exists a $K \in \mathbb{N}$ such that

$$x_{n_k} \in U_\varepsilon(L) \quad \text{provided that } k \geq K.$$

In other words,

$$\{n_k \in \mathbb{N} : k \geq K\} \subseteq \{n \geq 1 : x_n \in U_\varepsilon(L)\}.$$

Of course, this implies

$$|\{n \geq 1 : x_n \in U_\varepsilon(L)\}| = \infty.$$

So we see that every ε -neighborhood of L contains infinitely many elements of the given sequence.

To prove the converse let us assume that every ε -neighborhood of L contains infinitely many of the x_n 's. Our aim is to construct a subsequence of $(x_n)_{n \geq 1}$ converging to L .

So we start with $\varepsilon = 1$ and choose an n_1 for which $x_{n_1} \in U_1(L)$. Such an n_1 exists because there are infinitely many x_n in $U_1(L)$.

Next take $\varepsilon = 1/2$. Since there are infinitely many x_n in $U_{1/2}(L)$, there is $n_2 > n_1$ such that $x_{n_2} \in U_{1/2}(L)$.

Proceeding further in this way for each $k \geq 1$ there exist integers n_k with $n_1 < n_2 < n_3 < \dots$ for which $x_{n_k} \in U_{1/k}(L)$. We claim now that the subsequence $(x_{n_k})_{k \geq 1}$ converges to L . Thus, take an arbitrary $\varepsilon > 0$ and choose $K \geq 1$ with $1/K < \varepsilon$. If $k \geq K$, then

$$U_{1/k}(L) \subseteq U_\varepsilon(L).$$

So for $k \geq K$ from $x_{n_k} \in U_{1/k}(L)$ we derive $x_{n_k} \in U_\varepsilon(L)$. This being true for any $\varepsilon > 0$ proves the convergence of $(x_{n_k})_{k \geq 1}$ to L and completes the proof. ■

Remark 5.4.2. Proposition 5.4.5 sheds some new light onto Theorem 5.4.4. It says that any bounded sequence has at least one point around which infinitely many of the x_n 's are concentrated. This may be exactly one point in the case of convergent sequence (cf. Exercise 5.4.5) but also many points. It is not difficult to construct sequences with finitely many cluster points, but there even exist those sequences with infinitely many cluster points. One may even have that all $L \in [0, 1]$ are cluster points of a single sequence (difficult to imagine, but not too complicated to construct).

Theorem 5.4.3 and its equivalent formulation Theorem 5.4.4 have many important consequences in Calculus. In particular, it is a very helpful tool in the investigation of continuous functions. We do not treat this topic in the context of the book. To show the power of Theorem 5.4.3 let us state one topological application, the so-called

covering theorem due to Heine and Borel¹⁰. It asserts the following: Suppose we cover a closed interval $[a, b]$ by open intervals (α_j, β_j) , $j = 1, 2, \dots$. Then these open intervals have necessarily some overlap (if they do not possess an overlap, then between them there exist points not covered by these intervals). So one may guess that in fact not all intervals are needed to cover $[a, b]$. And, indeed, this is so how the next theorem shows.

Its proof is a little more involved.

Theorem 5.4.6 (Heine-Borel Covering Theorem). *Let a closed interval $[a, b]$ be covered by open intervals (α_j, β_j) , $j = 1, 2, \dots$. That is*

$$[a, b] \subseteq \bigcup_{j=1}^{\infty} (\alpha_j, \beta_j).$$

Then $[a, b]$ is already covered by finitely many of these intervals. In other words, for some $n \geq 1$ it follows that

$$[a, b] \subseteq \bigcup_{j=1}^n (\alpha_j, \beta_j).$$

$$\left([a, b] \subseteq \bigcup_{j=1}^{\infty} (\alpha_j, \beta_j) \right) \Rightarrow (\exists n \in \mathbb{N}) \left([a, b] \subseteq \bigcup_{j=1}^n (\alpha_j, \beta_j) \right).$$

Proof: Assume the result is false, i.e., finitely many of the intervals (α_j, β_j) never cover $[a, b]$. Thus, if we define sets G_n by

$$G_n = \bigcup_{j=1}^n (\alpha_j, \beta_j), \quad n = 1, 2, \dots,$$

then for all $n \geq 1$ it follows that $[a, b] \not\subseteq G_n$. Otherwise, G_n would be a finite covering of $[a, b]$, which is impossible by the assumption.

Hence, there exist $x_n \in [a, b]$ with $x_n \notin G_n$ for all $n \geq 1$. By the construction of the sets G_n this implies that

$$(5.4.1) \quad x_n \notin (\alpha_j, \beta_j), \quad 1 \leq j \leq n, \quad n = 1, 2, \dots .$$

Combining Proposition 5.4.5 with Theorem 5.4.4 implies the following: There exists a cluster point $L \in [a, b]$ of the sequence $(x_n)_{n \geq 1}$, i.e., every ε -neighborhood of L contains infinitely many elements of the sequence. But the open intervals cover $[a, b]$. Hence, at least one of these intervals covers the cluster point L . That is,

$$\exists j_0 \in \mathbb{N} \text{ such that } \alpha_{j_0} < L < \beta_{j_0} .$$

¹⁰In 1852 Peter Gustav Lejeune Dirichlet proved in his lectures in Berlin that continuous functions on closed bounded intervals are even uniformly continuous. To this purpose he used implicitly the fact that each open covering of a bounded closed interval contains a finite subcover. So he was the first who used an argument similar to the Heine-Borel covering theorem. His name does not appear in the denomination of the theorem because his notes were published only in 1904. Eduard Heine later used a similar technique, and finally, in 1895, Émile Borel stated and proved the covering theorem as it is formulated in Theorem 5.4.6. Later on, Pierre Cousin (1895), Henri Lebesgue (1896) and Arthur Moritz Schönflies (1900) generalized the result to arbitrary open covers (not necessarily countably many open intervals or subsets of \mathbb{R}) and to more general subsets of \mathbb{R} .

If $\varepsilon < \min\{L - \alpha_{j_0}, \beta_{j_0} - L\}$, then this implies that

$$U_\varepsilon(L) \subseteq (\alpha_{j_0}, \beta_{j_0}).$$

Consequently, since L is chosen as cluster point, it follows that

$$|\{n \geq 1 : \alpha_{j_0} < x_n < \beta_{j_0}\}| = \infty.$$

In particular, there are certain $n \geq j_0$ with $\alpha_{j_0} < x_n < \beta_{j_0}$. But this contradicts property (5.4.1). Thus, our assumption that finitely many intervals never cover $[a, b]$ was false. And this completes the proof. ■

Remark 5.4.3. If one analyzes the proof carefully, it becomes clear that we never used that the covering sets are intervals. The only property we applied was the following: If an element $x \in \mathbb{R}$ (in our case $x = L$) belongs to one of the covering sets, then there is an $\varepsilon > 0$ such that it even contains $U_\varepsilon(x) = (x - \varepsilon, x + \varepsilon)$. Sets possessing these properties are called open sets¹¹. Thus, a generalization of the previous theorem is that whenever infinitely many open sets cover a closed finite interval, then this is already achieved by finitely many of them.

Example 5.4.5. The crucial assumption in Theorem 5.4.6 is that the covered interval is bounded and closed; The result fails if one of these two properties is violated. For instance, consider the coverings of $(0, 1)$ or $[1, \infty)$ given by

$$(0, 1] \subseteq \bigcup_{j=1}^{\infty} \left(\frac{1}{j+1}, \frac{2}{j} \right) = \left(\frac{1}{2}, 2 \right) \cup \left(\frac{1}{3}, 1 \right) \cup \left(\frac{1}{4}, \frac{2}{3} \right) \cup \left(\frac{1}{5}, \frac{1}{2} \right) \cup \dots$$

or

$$[1, \infty) \subseteq \bigcup_{j=1}^{\infty} (j-1, j+1).$$

In both cases finitely many of the covering intervals never suffice to cover all of $(0, 1]$ or $[1, \infty)$, respectively.

Exercise 5.4.1. Let $(x_n)_{n \geq 1}$ be a bounded sequence. Define the upper limit of this sequence as in Exercise 5.3.10. Show that this upper limit is a cluster point. Moreover, if¹²

$$S := \{L \in \mathbb{R} : L \text{ is cluster point of } (x_n)_{n \geq 1}\},$$

then it follows that

$$\max(S) = \limsup_{n \rightarrow \infty} x_n.$$

Exercise 5.4.2. Describe all convergent subsequences of $1, -1, 1, -1, \dots$. Note that there are infinitely many, although there are only two limits which such subsequences can have.

Exercise 5.4.3. Let $(x_n)_{n \geq 1}$ be a sequence of real numbers such that the limits

$$\lim_{n \rightarrow \infty} x_{2n} = \lim_{n \rightarrow \infty} x_{2n+1} := L$$

¹¹see Exercise 5.8.18 for a precise definition

¹²The set S is usually called the cluster set of $(x_n)_{n \geq 1}$.

exist and coincide. Show that then $(x_n)_{n \geq 1}$ is convergent with

$$\lim_{n \rightarrow \infty} x_n = L.$$

Exercise 5.4.4. Suppose a sequence $(x_n)_{n \geq 1}$ is unbounded above, i.e.,

$$\sup_{n \geq 1} x_n = \infty.$$

Show that there exists a subsequence $(x_{n_k})_{k \geq 1}$ of $(x_n)_{n \geq 1}$ such that

$$\lim_{k \rightarrow \infty} x_{n_k} = \infty.$$

Exercise 5.4.5. Show that a sequence $(x_n)_{n \geq 1}$ is convergent if and only if it is bounded and, moreover, it possesses exactly one cluster point.

Exercise 5.4.6. Give an example of an (unbounded) divergent sequence with exactly one cluster point.

Exercise 5.4.7. Construct a sequence of real numbers for which each natural number is a cluster point.

Exercise 5.4.8. Prove the following result:

Proposition 5.4.7 (Nested Interval Theorem). *Let $[a_n, b_n]$, $n = 1, 2, \dots$ be nonempty closed intervals with $[a_1, b_1] \supseteq [a_2, b_2] \supseteq \dots$. Then*

$$\bigcap_{n=1}^{\infty} [a_n, b_n] \neq \emptyset.$$

Moreover, this infinite intersection of the intervals consists of a unique single point if and only if $\lim_{n \rightarrow \infty} (b_n - a_n) = 0$.

Exercise 5.4.9. Give an example that shows that Proposition 5.4.7 becomes false in the case of a nested sequence of **open** intervals. That is, find a sequence of nonempty open intervals (a_n, b_n) satisfying $(a_1, b_1) \supseteq (a_2, b_2) \supseteq \dots$ such that

$$\bigcap_{n=1}^{\infty} (a_n, b_n) = \emptyset.$$

Exercise 5.4.10. As observed in Example 4.7.3, the Cantor set \mathcal{C} consists of two different types of numbers: Endpoints (of withdrawn intervals) and infinite points (those which admit an infinite ternary expansion without digit 1). Show that for any infinite point $x \in \mathcal{C}$ there exists a sequence e_1, e_2, \dots of endpoints with $e_n \xrightarrow{n \rightarrow \infty} x$.

5.5. Cauchy Sequences

Say we are given a sequence $(x_n)_{n \geq 1}$ of real numbers. Is there some way to decide whether this sequence is converging or divergent? Of course, one could take **all** $L \in \mathbb{R}$ and check whether $x_n \xrightarrow{n \rightarrow \infty} L$. It is immediately clear that in many cases this is an unsolvable problem. Therefore, one is interested in an intrinsic property of $(x_n)_{n \geq 1}$ which characterizes convergent sequences.

Definition 5.5.1. A sequence $(x_n)_{n \geq 1}$ is said to be a **Cauchy sequence** if for all $\varepsilon > 0$ there is an $N = N(\varepsilon) \geq 1$ such that

$$|x_n - x_m| < \varepsilon \quad \text{whenever } n, m \geq N.$$

$$((x_n)_{n \geq 1} \text{ Cauchy sequence}) \Leftrightarrow (\forall \varepsilon > 0)(\exists N \in \mathbb{N})(\forall n, m \geq N)(|x_n - x_m| < \varepsilon)$$

Remark 5.5.1. It is important to mention that it depends only on the sequence $(x_n)_{n \geq 1}$ whether it is a Cauchy sequence. No external $L \in \mathbb{R}$ as possible limit nor any other condition independent of the investigated sequence occur.

Roughly spoken, $(x_n)_{n \geq 1}$ is a Cauchy sequence if its elements concentrate around themselves. Given an $\varepsilon > 0$, there is an $N(\varepsilon) \geq 1$ such that the distance between all possible x_n 's, $n \geq N(\varepsilon)$, becomes smaller than ε . Think of tiny particles whose distance is getting smaller and smaller over time. The basic question is whether there is a center (a limit) which they approach. Or, maybe, there is a hole at the position one expects their concentration.

Proposition 5.5.1. *Each convergent sequence is a Cauchy sequence.*

Proof: Suppose a sequence $(x_n)_{n \geq 1}$ converges to some $L \in \mathbb{R}$. Given $\varepsilon > 0$ there is an $N \geq 1$ such that $|x_n - L| < \varepsilon/2$ whenever $n \geq N$. Consequently, if $n, m \geq N$, then

$$|x_n - x_m| = |(x_n - L) + (L - x_m)| \leq |x_n - L| + |L - x_m| < \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon.$$

Since $\varepsilon > 0$ was arbitrary, this shows that $(x_n)_{n \geq 1}$ is a Cauchy sequence. ■

Lemma 5.5.2. *Let $(x_n)_{n \geq 1}$ be a Cauchy sequence. If this sequence contains a convergent subsequence, then it is convergent.*

Proof: Let $(x_n)_{n \geq 1}$ be a Cauchy sequence and suppose that its subsequence $(x_{n_k})_{k \geq 1}$ converges to some $L \in \mathbb{R}$. That is, given $\varepsilon > 0$, there is a $K \geq 1$ such that

$$|x_{n_k} - L| < \varepsilon/2 \quad \text{whenever } k \geq K.$$

Since the sequence $(x_n)_{n \geq 1}$ is Cauchy, there is an $N \geq 1$ such that

$$|x_n - x_m| < \varepsilon/2 \quad \text{provided that } n, m \geq N.$$

We choose $k \geq K$ satisfying $n_k \geq N$. An application of the previous two estimates with $m = n_k$ then leads to

$$|L - x_n| \leq |L - x_{n_k}| + |x_n - x_{n_k}| < \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon.$$

Recall that we chose the $k \geq 1$ such that at the same time $k \geq K$ and $n_k \geq N$. Since $\varepsilon > 0$ was arbitrary, this completes the proof. ■

The proof of the next result is almost identical with that of Proposition 5.2.6. Therefore, we only sketch its main steps.

Proposition 5.5.3. *Each Cauchy sequence is bounded.*

Proof: Choose an $N \geq 1$ such that $|x_n - x_m| < 1$ whenever $n, m \geq N$. In particular,

$$(5.5.1) \quad |x_n - x_N| < 1, \quad \text{which implies} \quad |x_n| < |x_N| + 1, \quad n \geq N.$$

Setting

$$c := \max\{|x_1|, |x_2|, \dots, |x_{N-1}|, |x_N| + 1\},$$

then estimate (5.5.1) and the definition of $c > 0$ yield $|x_n| \leq c$ for all $n \geq 1$. ■

We are now able to prove one of the most important results about the field of real numbers. The theorem characterizes convergent sequences in an intrinsic way, a way without knowing that an (external) limit exists. Furthermore, it sheds some new light onto the completeness condition of \mathbb{R} .

Theorem 5.5.4. *A sequence in \mathbb{R} is convergent if and only if it is a Cauchy sequence.*

$$((x_n)_{n \geq 1} \text{ is convergent}) \Leftrightarrow ((x_n)_{n \geq 1} \text{ is a Cauchy sequence})$$

Proof: First note that we proved already in Proposition 5.5.1 that each convergent sequence satisfies the Cauchy condition. Consequently, we have to prove only the converse direction, i.e., that each Cauchy sequence is convergent.

Thus, let $(x_n)_{n \geq 1}$ be an arbitrary Cauchy sequence of real numbers. By Proposition 5.5.3 it is bounded and Theorem 5.4.3 applies. Consequently, there is a subsequence $(x_{n_k})_{k \geq 1}$ of the x_n 's such that

$$x_{n_k} \xrightarrow{k \rightarrow \infty} L$$

for some $L \in \mathbb{R}$. Now Lemma 5.5.2 implies then that also $x_n \xrightarrow{n \rightarrow \infty} L$. This completes the proof. ■

Remark 5.5.2. If one analyzes the proof of Theorem 5.5.4, then it turns out that the basic ingredient is Theorem 5.4.3 which mainly rests on Theorem 5.3.1. Recall that this theorem asserts that bounded monotone sequences converge. To verify this it was essential that for each bounded set $\sup(A)$ or $\inf(A)$ exist in \mathbb{R} . Hence, the previous theorem is more or less a direct consequence of the (order) completeness of \mathbb{R} .

On the other hand, it can be proved (the proof is not too complicated) that conversely the existence of suprema follows by the assertion of Theorem 5.5.4. That is, if one assumes that each Cauchy sequence converges in \mathbb{R} , then also $\sup(A)$ exists for all nonempty subsets $A \subseteq \mathbb{R}$ bounded above. Thus, the completeness of \mathbb{R} may either be postulated by the existence of lowest upper bounds or, equivalently, by the fact that Cauchy sequences converge.

Example 5.5.1. Let $(x_n)_{n \geq 1}$ be the sequence investigated in Proposition 5.3.5. Recall that it was defined by

$$x_1 = 2 \quad \text{and} \quad x_{n+1} = \frac{x_n}{2} + \frac{1}{x_n}, \quad n = 1, 2, \dots$$

We proved that this sequence converges to $\sqrt{2}$ in \mathbb{R} . Consequently, it is a Cauchy sequence in \mathbb{R} . On the other hand, by construction the x_n 's are rational numbers. Thus,

$(x_n)_{n \geq 1}$ is also a Cauchy sequence in \mathbb{Q} which does **not** converge in \mathbb{Q} . Hereby a Cauchy sequence in \mathbb{Q} means that it is a sequence $(x_n)_{n \geq 1}$ of rational numbers which satisfies

$$(\forall \varepsilon > 0)(\exists N \in \mathbb{N})(n, m \geq N \Rightarrow |x_n - x_m| < \varepsilon).$$

Why is this sequence divergent in \mathbb{Q} ? If $(x_n)_{n \geq 1}$ converges in \mathbb{Q} , say to some $L' \in \mathbb{Q}$, then it would also converge to L' in \mathbb{R} , and by the uniqueness of the limit it follows that $L' = \sqrt{2}$. But as we know, $\sqrt{2}$ does not belong to \mathbb{Q} . Therefore, there does not exist a limit of these x_n 's inside the set of rational numbers.

Remark 5.5.3. Example 5.5.1 gives a hint how to use Cauchy sequences to complete \mathbb{Q} by adding irrational numbers. The basic idea, due of Bernard Bolzano and Augustin-Louis Cauchy, is to investigate Cauchy sequences of rational numbers, so-called rational Cauchy sequences. If such a sequence converges in \mathbb{Q} , nothing new occurs. But if it does not, we may assign to this sequence a new object which we add to the field \mathbb{Q} . In this way irrational numbers are defined as limits of rational Cauchy sequences which do **not** converge in \mathbb{Q} . But one technical problem arises: There are many Cauchy sequences which correspond to the same (nonexisting in \mathbb{Q}) limit. For example, the sequences $(x_n)_{n \geq 1}$ investigated in Propositions 5.3.5 and 5.3.6, respectively, both may be identified with $\sqrt{2}$. But any subsequence of these sequences also approaches $\sqrt{2}$. Which of all these rational Cauchy sequences should we choose for the definition of $\sqrt{2}$? And there are many more such sequences. For example, we may also choose the sequence

$$1, 1.4, 1.41, 1.414, 1.4142, 1.41421, \dots$$

which corresponds to the decimal series representation of $\sqrt{2}$. To overcome this difficulty one has to identify Cauchy sequences converging to the same (nonexisting) limit. This is done in the following way: One defines a relation \sim between rational Cauchy sequences as follows:

$$(5.5.2) \quad (x_n)_{n \geq 1} \sim (y_n)_{n \geq 1} \Leftrightarrow \lim_{n \rightarrow \infty} |x_n - y_n| = 0.$$

It is not difficult to see that this is an equivalence relation on the set of all rational Cauchy sequences. In this setting rational numbers $q \in \mathbb{Q}$ may be identified with the set of rational Cauchy sequences converging to q . Consequently, the set of equivalence classes of rational Cauchy sequences may be taken as a model for the real numbers. So, for example, $\sqrt{2}$ may be identified with the set of rational Cauchy sequences which are equivalent to the sequence $(x_n)_{n \geq 1}$ either investigated in Proposition 5.3.5 or in Proposition 5.3.6, respectively.

Exercise 5.5.1. Why does every Cauchy sequence with values in \mathbb{Z} converge?

Exercise 5.5.2. Show directly (without using its convergence) that $(1/n)_{n \geq 1}$ is a Cauchy sequence.

Exercise 5.5.3. For $n \geq 1$, let $a_n = \sum_{k=1}^n \frac{1}{k^2}$. Show that $(a_n)_{n \geq 1}$ is a Cauchy sequence.

Exercise 5.5.4. Prove that the sum of two Cauchy sequences is a Cauchy sequence as well. Give a direct proof and another one via Theorem 5.5.4.

Exercise 5.5.5. Is the previous exercise true for the product of two Cauchy sequences?

Exercise 5.5.6. Show that any subsequence of a Cauchy sequence is also a Cauchy sequence. Give two different proofs: One direct one via the definition of Cauchy sequences and another one via Theorem 5.5.4.

Exercise 5.5.7. Let $(x_n)_{n \geq 1}$, $(y_n)_{n \geq 1}$, and $(z_n)_{n \geq 1}$ be three sequences of real numbers such that $x_n \leq y_n \leq z_n$ for any $n \geq 1$. If $(x_n)_{n \geq 1}$ and $(z_n)_{n \geq 1}$ are Cauchy sequences, does this imply that $(y_n)_{n \geq 1}$ is also a Cauchy sequence? If this is not so, what can be said if, furthermore, $\lim_{n \rightarrow \infty} |z_n - x_n| = 0$?

Exercise 5.5.8. Prove that a sequence $(x_n)_{n \geq 1}$ is a Cauchy sequence if and only if the following is satisfied:

$$(\forall \varepsilon > 0)(\exists N \in \mathbb{N})(\forall n > N)(|x_n - x_N| < \varepsilon).$$

Exercise 5.5.9. Let $(x_n)_{n \geq 1}$ be a sequence of real numbers. Show that the following properties are equivalent.

- (1) The sequence $(x_n)_{n \geq 1}$ is **not** a Cauchy sequence.
- (2) There is some $\varepsilon_0 > 0$ such that

$$(\forall N \in \mathbb{N})(\exists n > N)(|x_n - x_N| \geq \varepsilon_0).$$

- (3) There are $\varepsilon_0 > 0$ and natural numbers $n_1 < n_2 < \dots$ with

$$|x_{n_{j+1}} - x_{n_j}| \geq \varepsilon_0, \quad j = 1, 2, \dots.$$

Exercise 5.5.10. Show that the relation \sim defined in (5.5.2) is indeed an equivalence relation on the set of rational Cauchy sequences.

Exercise 5.5.11. Suppose the rational Cauchy sequences satisfy

$$(x_n)_{n \geq 1} \sim (y_n)_{n \geq 1} \quad \text{as well as} \quad (u_n)_{n \geq 1} \sim (v_n)_{n \geq 1}.$$

Show that then

$$(x_n + u_n)_{n \geq 1} \sim (y_n + v_n)_{n \geq 1} \quad \text{and} \quad (x_n \cdot u_n)_{n \geq 1} \sim (y_n \cdot v_n)_{n \geq 1}.$$

5.6. Infinite Series

Suppose we are given a sequence $(x_n)_{n \geq 1}$ of real numbers. How is the infinite sum

$$x_1 + x_2 + \dots = \sum_{n=1}^{\infty} x_n$$

defined? To add one after the other infinitely many of the x_n 's does not make much sense. We would never be finished in finite time. For example, what are

$$\begin{aligned} 1 + \frac{1}{2} + \frac{1}{3} + \dots &= \sum_{n=1}^{\infty} \frac{1}{n} \quad \text{or} \\ 1 - \frac{1}{2} + \frac{1}{3} - \frac{1}{4} \pm \dots &= \sum_{n=1}^{\infty} \frac{(-1)^{n+1}}{n} \quad \text{or} \\ 1 + \frac{1}{2} + \frac{1}{4} + \frac{1}{8} + \dots &= \sum_{n=0}^{\infty} \frac{1}{2^n} ? \end{aligned}$$

While all these sums can be given an exact meaning, it is far from clear how

$$1 - 1 + 1 - 1 \pm \dots = \sum_{n=0}^{\infty} (-1)^n$$

should be defined. If one adds these elements successively, then one gets either 1 or 0 in dependence of adding an odd or an even number of elements in the sequence.

The mathematical exact way to introduce the infinite sum of real numbers is as follows.

Definition 5.6.1. Let $(x_n)_{n \geq 1}$ be an infinite sequence of real numbers. Its *n*th **partial sum** s_n is defined by

$$s_n = x_1 + \dots + x_n = \sum_{j=1}^n x_j, \quad n \geq 1.$$

The sequence $(s_n)_{n \geq 1}$ is called **summable** provided that $\lim_{n \rightarrow \infty} s_n$ exists. The limit is denoted by

$$\sum_{n=1}^{\infty} x_n := \lim_{n \rightarrow \infty} s_n = \lim_{n \rightarrow \infty} \sum_{i=1}^n x_i.$$

Another way to express this is to say that the infinite sum $\sum_{n=1}^{\infty} x_n$ **converges** or **exists**.

If the sequence $(s_n)_{n \geq 1}$ of partial sums diverges properly, we write

$$\sum_{n=1}^{\infty} x_n = \infty \quad \text{or} \quad \sum_{n=1}^{\infty} x_n = -\infty,$$

in dependence of $s_n \xrightarrow{n \rightarrow \infty} \infty$ or $s_n \xrightarrow{n \rightarrow \infty} -\infty$, respectively.

Summary:

Let $s_n = x_1 + \dots + x_n$ be the *n*th partial sum of x_1, x_2, \dots

$$(x_n)_{n \geq 1} \text{ summable} \Leftrightarrow \sum_{n=1}^{\infty} x_n \text{ conv.} \Leftrightarrow (s_n)_{n \geq 1} \text{ conv.}$$

$$\sum_{n=1}^{\infty} x_n = \lim_{n \rightarrow \infty} s_n = \lim_{n \rightarrow \infty} \sum_{i=1}^n x_i$$

$$\sum_{n=1}^{\infty} x_n = \pm\infty \Leftrightarrow \lim_{n \rightarrow \infty} s_n = \lim_{n \rightarrow \infty} \sum_{i=1}^n x_i = \pm\infty$$

Remark 5.6.1. If the sequence $(s_n)_{n \geq 1}$ does not converge, then this is often expressed by saying that the infinite sum $\sum_{n=1}^{\infty} x_n$ does not converge or does not exist. This is mathematically not quite precise because in this case the occurring infinite sum is not defined properly. For example, if $x_n = (-1)^n$, $n \geq 0$, then the sequence of partial sums is $(1, 0, 1, 0, 1, \dots)$ which, of course, does not converge. Thus, to state that $\sum_{n=0}^{\infty} (-1)^n$

does not converge is somehow misleading because this infinite sum does not make sense. Nevertheless, it is common to say that

$$\sum_{n=0}^{\infty} (-1)^n \quad \text{does not converge or does not exist.}$$

It is important to rewrite Definition 5.6.1 into the language of convergent sequences.

Proposition 5.6.1. *The sequence $(x_n)_{n \geq 1}$ is summable if and only if there is some $S \in \mathbb{R}$ such that the following holds. Given $\varepsilon > 0$ there is an $N = N(\varepsilon) \geq 1$ such that*

$$\left| S - \sum_{i=1}^n x_i \right| < \varepsilon \quad \text{whenever } n \geq N.$$

The limit $S \in \mathbb{R}$ is then denoted by $\sum_{n=1}^{\infty} x_n$.

$$((x_n)_{n \geq 1} \text{ summable}) \Leftrightarrow (\exists S \in \mathbb{R})(\forall \varepsilon > 0)(\exists N \geq 1)(\forall n \geq N) \left(\left| S - \sum_{i=1}^n x_i \right| < \varepsilon \right)$$

Example 5.6.1 (Geometric Series). Let q be a given real number. If $x_n = q^n$, $n \geq 0$, by (1.3.1) it follows that

$$s_n = 1 + q + \cdots + q^n = \begin{cases} \frac{q^{n+1}-1}{q-1} & : q \neq 1 \\ n+1 & : q = 1 \end{cases}$$

In Proposition 5.3.1 we proved that $q^n \xrightarrow{n \rightarrow \infty} 0$ if $|q| < 1$ while $q^n \xrightarrow{n \rightarrow \infty} \infty$ for $1 < q < \infty$.

Thus, $s_n \xrightarrow{n \rightarrow \infty} \frac{1}{1-q}$ if $|q| < 1$ and $s_n \xrightarrow{n \rightarrow \infty} \infty$ if $q \geq 1$. That is,

$$\sum_{n=0}^{\infty} q^n = \frac{1}{1-q}, \quad |q| < 1, \quad \text{while} \quad \sum_{n=0}^{\infty} q^n = \infty, \quad q \geq 1.$$

Note that the sum $\sum_{n=0}^{\infty} q^n$ diverges (does not exist) whenever $-\infty < q \leq -1$.

From Proposition 5.2.8 one immediately gets the following property of infinite sums.

Proposition 5.6.2. *Let $(x_n)_{n \geq 1}$ and $(y_n)_{n \geq 1}$ be two summable sequences. Then for all real numbers α and β also $(\alpha x_n + \beta y_n)_{n \geq 1}$ is summable and*

$$(5.6.1) \quad \sum_{n=1}^{\infty} (\alpha x_n + \beta y_n) = \alpha \sum_{n=1}^{\infty} x_n + \beta \sum_{n=1}^{\infty} y_n.$$

Proof: It easily follows by Proposition 5.2.8, assertion (1), that

$$\begin{aligned} \lim_{n \rightarrow \infty} \sum_{i=1}^n (\alpha x_i + \beta y_i) &= \lim_{n \rightarrow \infty} \left[\alpha \sum_{i=1}^n x_i + \beta \sum_{i=1}^n y_i \right] \\ &= \alpha \lim_{n \rightarrow \infty} \sum_{i=1}^n x_i + \beta \lim_{n \rightarrow \infty} \sum_{i=1}^n y_i = \alpha \sum_{n=1}^{\infty} x_n + \beta \sum_{n=1}^{\infty} y_n. \end{aligned}$$

Thus, the left-hand limit exists, hence the sequence $(\alpha x_n + \beta y_n)_{n \geq 1}$ is summable and (5.6.1) follows by the definition of infinite sums. ■

There exist many criteria which ensure the existence of $\sum_{n=1}^{\infty} x_n$. Here we state and prove a first very important one.

Proposition 5.6.3 (Cauchy Criterion). *A sequence $(x_n)_{n \geq 1}$ is summable if and only if the tails of the sum become arbitrarily small. That is, given $\varepsilon > 0$ there is an $N = N(\varepsilon) \geq 1$ such that*

$$\left| \sum_{i=n+1}^m x_i \right| < \varepsilon \quad \text{for all } N \leq n < m.$$

$$\left(\sum_{n=1}^{\infty} x_n \text{ exists} \right) \Leftrightarrow \left[(\forall \varepsilon > 0)(\exists N \geq 1)(\forall m > n \geq N) \left(\left| \sum_{i=n+1}^m x_i \right| < \varepsilon \right) \right]$$

Proof: By Theorem 5.5.4 the partial sums s_n converge if and only if $(s_n)_{n \geq 1}$ is a Cauchy sequence. That is, given $\varepsilon > 0$ there is an $N = N(\varepsilon)$ such that

$$|s_n - s_m| < \varepsilon \quad \text{if } n, m \geq N.$$

Suppose $n < m$, then

$$s_m - s_n = \sum_{i=n+1}^m x_i.$$

From this these two observations the result follows directly. ■

Corollary 5.6.4. *If $(x_n)_{n \geq 1}$ is summable, then*

$$\lim_{n \rightarrow \infty} \sum_{i=n+1}^{\infty} x_i = 0.$$

Proof: Given $\varepsilon > 0$ there is an $N \geq 1$ such that

$$|s_m - s_n| < \varepsilon/2 \quad \text{if } N \leq n < m.$$

Fix $n \geq N$. Then for all $m > n$

$$-\varepsilon/2 \leq s_m - s_n \leq \varepsilon/2$$

Now

$$\lim_{m \rightarrow \infty} [s_m - s_n] = \sum_{i=n+1}^{\infty} x_i.$$

By the assertion in Exercise 5.2.3 it follows that

$$-\varepsilon/2 \leq \sum_{i=n+1}^{\infty} x_i \leq \varepsilon/2,$$

and consequently,

$$\left| \sum_{i=n+1}^{\infty} x_i \right| < \varepsilon \quad \text{whenever } n \geq N.$$

This completes the proof. ■

Example 5.6.2. Since $\sum_{n=0}^{\infty} \frac{1}{2^n}$ exists, it follows that

$$\lim_{n \rightarrow \infty} \sum_{i=n+1}^{\infty} \frac{1}{2^i} = 0.$$

In this case this also follows directly because of

$$\sum_{i=n+1}^{\infty} \frac{1}{2^i} = \frac{1}{2^{n+1}} \cdot \sum_{i=n+1}^{\infty} \frac{1}{2^{i-n-1}} = \frac{1}{2^{n+1}} \cdot \sum_{i=0}^{\infty} \frac{1}{2^i} = \frac{2}{2^{n+1}} = \frac{1}{2^n}.$$

Corollary 5.6.5 (Vanishing Condition). *In order that $(x_n)_{n \geq 1}$ is summable it is necessary that $\lim_{n \rightarrow \infty} x_n = 0$.*

$$\left((x_n)_{n \geq 1} \text{ summable} \right) \Rightarrow \left(\lim_{n \rightarrow \infty} x_n = 0 \right)$$

Proof: This easily follows from Corollary 5.6.4 by

$$\lim_{n \rightarrow \infty} x_n = \lim_{n \rightarrow \infty} \left[\sum_{i=n}^{\infty} x_i - \sum_{i=n+1}^{\infty} x_i \right] = \lim_{n \rightarrow \infty} \sum_{i=n}^{\infty} x_i - \lim_{n \rightarrow \infty} \sum_{i=n+1}^{\infty} x_i = 0 - 0 = 0.$$

■

The next (very important) example shows that the converse of Corollary 5.6.5 is not valid. That is, there exist sequences tending to zero which are not summable.

Example 5.6.3 (Harmonic Series). Let $x_n = \frac{1}{n}$, $n = 1, 2, \dots$. We claim that this series is not summable. To see this write

$$s_{2^n} = 1 + \left(\frac{1}{2} \right) + \left(\frac{1}{3} + \frac{1}{4} \right) + \cdots + \left(\frac{1}{2^{n-1}+1} + \cdots + \frac{1}{2^n} \right), \quad n \geq 1.$$

Observe that

$$\frac{1}{2^{n-1}+1} + \cdots + \frac{1}{2^n} \geq \underbrace{\frac{1}{2^n} + \cdots + \frac{1}{2^n}}_{2^{n-1}} \geq \frac{1}{2}.$$

Consequently,

$$s_{2^n} \geq 1 + n \cdot \frac{1}{2} > \frac{n}{2}$$

which implies $\lim_{n \rightarrow \infty} s_{2^n} = \infty$. But $s_1 < s_2 < \cdots$, hence $s_n \xrightarrow[n \rightarrow \infty]{} \infty$. That is¹³,

$$\sum_{k=1}^{\infty} \frac{1}{k} = \lim_{n \rightarrow \infty} \left(1 + \frac{1}{2} + \frac{1}{3} + \cdots + \frac{1}{n} \right) = \infty.$$

To get an impression how slowly the partial sums $s_n = \sum_{k=1}^n \frac{1}{k}$ tend to ∞ , here are some values:

$$s_{100} = 5.18738, \quad s_{500} = 6.79282, \quad s_{1000} = 7.48547, \quad s_{10000} = 9.78761.$$

¹³The first proof for the divergence of the harmonic series is due to the French mathematician Nicole Oresme (1320–1382) in 1350.

Remark 5.6.2. A much stronger result than the preceding one about the divergence of the harmonic series was already proved in 1737 by Leonhard Euler. He showed that the sum over the reciprocals of prime numbers diverges:

$$\frac{1}{2} + \frac{1}{3} + \frac{1}{5} + \frac{1}{7} + \frac{1}{11} + \frac{1}{13} + \frac{1}{17} + \dots = \sum_{p \text{ prime}} \frac{1}{p} = \infty.$$

In particular, this implies that there are infinitely many prime numbers.

In contrast to that, the Norwegian mathematician Viggo Brun (1885–1975) proved in 1915 that the sum over the reciprocals of twin primes is finite. That is,

$$\begin{aligned} & \left[\frac{1}{3} + \frac{1}{5} \right] + \left[\frac{1}{5} + \frac{1}{7} \right] + \left[\frac{1}{11} + \frac{1}{13} \right] + \left[\frac{1}{17} + \frac{1}{19} \right] + \left[\frac{1}{29} + \frac{1}{31} \right] + \dots \\ &= \sum_{p, p+2 \text{ prime}} \left[\frac{1}{p} + \frac{1}{p+2} \right] := C_B < \infty. \end{aligned}$$

Note that this does not answer the question about the existence of finitely or infinitely many twin primes. The constant C_B is called Brun's constant. Its value is known to be about 1.902160583104 (cf. OEIS A065421).

Let us treat a special case now. Suppose the sequence $(x_n)_{n \geq 1}$ consists of nonnegative real numbers. Then this implies

$$s_1 \leq s_2 \leq \dots$$

Theorem 5.3.1 lets us conclude that the partial sums $(s_n)_{n \geq 1}$ either converge or diverge improperly to ∞ . That is, the following is true.

Proposition 5.6.6 (Boundedness Criterion). *Let $(x_n)_{n \geq 1}$ be a sequence of **nonnegative** real numbers. Then either*

$$\sum_{n=1}^{\infty} x_n \quad \text{exists or} \quad \sum_{n=1}^{\infty} x_n = \infty.$$

The former happens if and only if the sequence of partial sums is bounded, i.e., if we have $\sup_{n \geq 1} s_n < \infty$.

$$(x_n \geq 0, n \geq 1) \Rightarrow \left[\text{Either } (x_n)_{n \geq 1} \text{ is summable or } \sum_{n=1}^{\infty} x_n = \infty \right]$$

Remark 5.6.3. It is common to write $\sum_{n=1}^{\infty} x_n < \infty$ in the case that $\sum_{n=1}^{\infty} x_n$ exists. Recall that this happens if and only if

$$\sup_{n \geq 1} s_n = \sup_{n \geq 1} \sum_{i=1}^n x_i < \infty.$$

$$[(\forall n \geq 1)(x_n \geq 0)] \Rightarrow \left[((x_n)_{n \geq 1} \text{ is summable}) \Leftrightarrow \left(\sum_{n=1}^{\infty} x_n < \infty \right) \right]$$

It is important to mention that only in the case of nonnegative x_n 's one may express the existence of $\sum_{n=1}^{\infty} x_n$ by writing $\sum_{n=1}^{\infty} x_n < \infty$. It does **not** make sense for arbitrary real numbers x_n . For example, if $x_n = 1$ for even $n \geq 1$ and $x_n = -1$ for odd ones, then $\sup_{n \geq 1} s_n < \infty$, but this sequence is not summable.

Definition 5.6.2. A sequence of real numbers is said to be **absolutely summable** if $(|x_n|)_{n \geq 1}$ is summable.

$$[(x_n)_{n \geq 1} \text{ absolutely summable}] \Leftrightarrow \sum_{n=1}^{\infty} |x_n| < \infty$$

Proposition 5.6.7. Each absolutely summable sequence is summable.

$$\left(\sum_{n=1}^{\infty} |x_n| < \infty \right) \Rightarrow \left(\sum_{n=1}^{\infty} x_n \text{ exists} \right)$$

Proof: We will apply Proposition 5.6.3 twice. Since $\sum_{n=1}^{\infty} |x_n|$ exists, given $\varepsilon > 0$ by Proposition 5.6.3 there is an $N \in \mathbb{N}$ such that

$$\sum_{i=n+1}^m |x_i| < \varepsilon \quad \text{whenever } m > n \geq N.$$

But

$$\left| \sum_{i=n+1}^m x_i \right| \leq \sum_{i=n+1}^m |x_i|,$$

hence

$$|s_m - s_n| = \left| \sum_{i=n+1}^m x_i \right| < \varepsilon \quad \text{whenever } m > n \geq N.$$

Since $\varepsilon > 0$ was arbitrary, another application of Proposition 5.6.3 implies that the partial sums converge, i.e., $(x_n)_{n \geq 1}$ is summable. ■

Example 5.6.4. Suppose we are given a number $a \in \mathbb{R}$ with $|a| < 1$ and some $\alpha > 0$. Does the sum

$$\sum_{n=1}^{\infty} \frac{a^n}{n^{\alpha}}$$

exist? The answer is affirmative. Because

$$\left| \frac{a^k}{k^{\alpha}} \right| = \frac{|a|^k}{k^{\alpha}} \leq |a|^k,$$

by virtue of Example 5.6.1 it follows that

$$\sum_{k=1}^n \left| \frac{a^k}{k^{\alpha}} \right| \leq \sum_{k=0}^n |a|^k \leq \sum_{k=0}^{\infty} |a|^k = \frac{1}{1 - |a|}.$$

Consequently, the partial sums of $(|a|^n/n^{\alpha})_{n \geq 1}$ are bounded above, that is

$$\sum_{k=1}^{\infty} \left| \frac{a^k}{k^{\alpha}} \right| < \infty.$$

This tells us, that the sequence $(a^n/n^{\alpha})_{n \geq 1}$ is absolutely summable. Thus, for any $|a| < 1$ and any $\alpha > 0$ the infinite sum

$$\sum_{n=1}^{\infty} \frac{a^n}{n^{\alpha}}$$

exists. It is even absolutely summable.

Remark 5.6.4. As we will see below, the sequence $((-1)^n/n)_{n \geq 1}$ is summable, but not absolutely summable as we observed in Example 5.6.3. Thus, the converse implication of Proposition 5.6.7 is not valid.

There exist many tests or criteria to check whether a given sequence is summable. We already have proved some of them, as the Cauchy and the boundedness criterion or the vanishing condition. Yet there exist many more. To state and to prove all of them would go beyond the scope of the present book. Therefore, we will only mention the most important tests and show with a few examples how these tests apply. Those who want to know more facts about infinite series are referred to any textbook about Calculus, as e.g. to [32].

Proposition 5.6.8 (Comparison Test). *Let $(x_n)_{n \geq 1}$ and $(y_n)_{n \geq 1}$ be two sequences of positive real numbers. If*

$$\sup_{n \geq 1} \frac{x_n}{y_n} < \infty,$$

then $\sum_{n=1}^{\infty} x_n$ exists if $\sum_{n=1}^{\infty} y_n$ does so. Conversely, if

$$\inf_{n \geq 1} \frac{x_n}{y_n} > 0,$$

then $(y_n)_{n \geq 1}$ is summable if $(x_n)_{n \geq 1}$ is so.

In particular, if the limit

$$L := \lim_{n \rightarrow \infty} \frac{x_n}{y_n}$$

exists, then $L > 0$ implies that $(x_n)_{n \geq 1}$ is summable if and only if $(y_n)_{n \geq 1}$ is so.

$$(\forall x_n, y_n > 0) \left(\lim_{n \rightarrow \infty} \frac{x_n}{y_n} > 0 \right) \Rightarrow \left[\left(\sum_{n=1}^{\infty} x_n \text{ exists} \right) \Leftrightarrow \left(\sum_{n=1}^{\infty} y_n \text{ exists} \right) \right].$$

Proof: We only prove the first assertion. The second one can be proved by similar methods while the last one follows by known properties of limits. Use, for example, that convergent sequences are bounded and that $\lim_{n \rightarrow \infty} x_n/y_n = L > 0$ implies $x_n/y_n \geq L/2 > 0$ provided that $n \geq N$ for a suitable $N \geq 1$ (compare Exercise 5.2.4).

Thus, suppose that $\sup_{n \geq 1} \frac{x_n}{y_n} < \infty$. Then there is a number $c > 0$ for which $x_n/y_n \leq c$ for all $n \geq 1$, i.e., $x_n \leq c y_n$. If

$$s_n = x_1 + \cdots + x_n \quad \text{and} \quad t_n = y_1 + \cdots + y_n,$$

the previous estimate implies $s_n \leq c t_n$ for all $n \geq 1$. Thus, if $(y_n)_{n \geq 1}$ is summable, from $\sup_{n \geq 1} t_n < \infty$ we derive $\sup_{n \geq 1} s_n < \infty$. The boundedness Criterion (Proposition 5.6.6) lets us conclude (recall that we assumed $x_n > 0$) that $(x_n)_{n \geq 1}$ is summable. ■

Example 5.6.5. Let

$$x_n = \frac{n^2 + 1}{2n^3 - n^2} \quad \text{and} \quad y_n = \frac{1}{n}, \quad n \geq 1.$$

Since

$$\lim_{n \rightarrow \infty} \frac{x_n}{y_n} = \frac{1}{2} > 0,$$

by Example 5.6.3 the comparison test implies that

$$\sum_{n=1}^{\infty} y_n = \infty \quad \Rightarrow \quad \sum_{n=1}^{\infty} x_n = \infty.$$

On the other hand, if

$$x_n = \frac{n^2 + 1}{2n^4 - n^2} \quad \text{and} \quad y_n = \frac{1}{n^2}, \quad n \geq 1,$$

then also

$$\lim_{n \rightarrow \infty} \frac{x_n}{y_n} = \frac{1}{2} > 0.$$

But this time, as we will see below, the sequence $(y_n)_{n \geq 1}$ is summable. Hence, now the comparison test yields the existence of

$$\sum_{n=1}^{\infty} \frac{n^2 + 1}{2n^4 - n^2}.$$

Proposition 5.6.9 (Ratio Test). *Let $(x_n)_{n \geq 1}$ be a sequence of positive numbers. Suppose that the limit*

$$(5.6.2) \quad r := \lim_{n \rightarrow \infty} \frac{x_{n+1}}{x_n}$$

exists. If $r < 1$ the sequence $(x_n)_{n \geq 1}$ is summable while for $r > 1$ the partial sums diverge.

$$\left(r = \lim_{n \rightarrow \infty} \frac{x_{n+1}}{x_n} \text{ exists} \right) \Rightarrow \begin{cases} (x_n)_{n \geq 1} \text{ summable} & : 0 \leq r < 1 \\ (x_n)_{n \geq 1} \text{ not summable} & : 1 < r \leq \infty \end{cases}$$

Proof: We first assume $r < 1$. Given a number $q > 0$ with $r < q < 1$, there is an $N \in \mathbb{N}$ (cf. Exercise 5.2.5) such that

$$\frac{x_{n+1}}{x_n} \leq q \quad \text{provided that } n \geq N.$$

In particular,

$$\frac{x_{N+1}}{x_N} \leq q, \quad \frac{x_{N+2}}{x_{N+1}} \leq q, \quad \frac{x_{N+3}}{x_{N+2}} \leq q, \dots.$$

Iterating these estimates it follows that

$$x_{N+k} \leq q^k x_N, \quad k = 0, 1, 2, \dots.$$

From this we derive

$$\frac{x_{N+k}}{q^{N+k}} = q^{-N} \frac{x_{N+k}}{q^k} \leq \frac{x_N}{q^N}, \quad k = 0, 1, 2, \dots,$$

which easily implies

$$\sup_{n \geq 1} \frac{x_n}{q^n} \leq \max \left\{ \frac{x_1}{q}, \frac{x_2}{q^2}, \dots, \frac{x_{N-1}}{q^{N-1}}, \frac{x_N}{q^N} \right\} < \infty.$$

Since $\sum_{n=1}^{\infty} q^n < \infty$ (recall $q < 1$) an application of the first part in Proposition 5.6.8 with $y_n = q^n$ proves that the x_n 's are summable.

If $r > 1$, we choose now a number q with $1 < q < r$. Use Exercise 5.2.5 to prove that there is an $N \geq 1$ with

$$\frac{x_{n+1}}{x_n} \geq q, \quad n \geq N.$$

As above an iteration of these estimates leads to $x_{N+k} \geq q^k x_N$, $k = 0, 1, 2, \dots$. But $q^k \xrightarrow{k \rightarrow \infty} \infty$, implying $x_n \xrightarrow{n \rightarrow \infty} \infty$. By the Vanishing Condition (Proposition 5.6.5) the sequence $(x_n)_{n \geq 1}$ cannot be summable. This completes the proof. ■

Remark 5.6.5. If the limit r in (5.6.2) equals 1, then, in general, nothing can be said about the existence or nonexistence of $\sum_{n=1}^{\infty} x_n$. For example, if either $x_n = 1/n$ or $x_n = 1/n^2$, then in both cases it follows that $r = 1$. But as we saw in Example 5.6.3, $\sum_{n=1}^{\infty} \frac{1}{n} = \infty$ while, as we will prove in Example 5.6.6 below, the infinite sum $\sum_{n=1}^{\infty} \frac{1}{n^2}$ exists.

Let us state a first important application of Proposition 5.6.9.

Proposition 5.6.10. *For any real number $x \in \mathbb{R}$ the infinite sum*

$$(5.6.3) \quad \sum_{n=0}^{\infty} \frac{x^n}{n!}$$

exists.

Proof: We are going to prove that

$$\sum_{n=0}^{\infty} \left| \frac{x^n}{n!} \right| = \sum_{n=0}^{\infty} \frac{|x|^n}{n!} < \infty.$$

Then $(x^n/n!)_{n \geq 1}$ is absolutely summable, hence by Proposition 5.6.7 also summable. Thus, if

$$x_n = \frac{|x|^n}{n!},$$

then it follows that

$$\frac{x_{n+1}}{x_n} = \frac{|x|^{n+1} n!}{|x|^n (n+1)!} = \frac{|x|}{n+1} \xrightarrow{n \rightarrow \infty} 0.$$

So we get that the limit r occurring in (5.6.2) equals 0, hence in view of Proposition 5.6.9 this implies the (even absolute) convergence of the infinite sum. ■

Let us evaluate the value of the infinite sun (5.6.3) in the case $x = 1$. Here we get the following result.

Proposition 5.6.11. *We have*

$$\sum_{k=0}^{\infty} \frac{1}{k!} = e$$

with Euler number e introduced in Definition 5.3.1.

Proof: We use the notations and results of Theorem 5.3.7. Recall that the x_n s defined by

$$x_n = \left(1 + \frac{1}{n}\right)^n$$

form an increasing sequence which, as $n \rightarrow \infty$, converges to the Euler number e . Next we apply the binomial theorem 1.7.6 to the n th power of $1 + \frac{1}{n}$ and obtain

$$(5.6.4) \quad x_n = \sum_{k=0}^n \binom{n}{k} \frac{1}{n^k} = 1 + \sum_{k=1}^n \left[\frac{n}{n} \cdot \frac{n-1}{n} \cdots \frac{n-k+1}{n} \right] \frac{1}{k!}.$$

If we estimate each of the n terms inside the bracket by 1, then this implies that

$$x_n \leq \sum_{k=0}^n \frac{1}{k!} \leq \sum_{k=0}^{\infty} \frac{1}{k!}.$$

Finally, taking the limit $n \rightarrow \infty$ it follows that

$$(5.6.5) \quad e = \lim_{n \rightarrow \infty} x_n \leq \sum_{k=0}^{\infty} \frac{1}{k!}.$$

Here we used that the right-hand side is a finite number and that $x_n \leq b$ for some $b \in \mathbb{R}$ also implies $\lim_{n \rightarrow \infty} x_n \leq b$.

The proof of the reversed estimate is a bit more complicated. Suppose we have two integers $m \leq n$. Since all elements in the sum (5.6.4) are positive, it becomes smaller if we add up less elements, i.e., it follows that

$$(5.6.6) \quad x_n = 1 + \sum_{k=1}^n \left[\frac{n}{n} \dots \frac{n-k+1}{n} \right] \frac{1}{k!} \geq 1 + \sum_{k=1}^m \left[\frac{n}{n} \dots \frac{n-k+1}{n} \right] \frac{1}{k!}.$$

Next we estimate the product in the bracket as follows: For each $1 \leq k \leq m$ we have

$$\frac{n}{n} \dots \frac{n-k+1}{n} \geq \frac{n}{n} \dots \frac{n-k+1}{n} \dots \frac{n-m+1}{n} \geq \left(\frac{n-m+1}{n} \right)^m.$$

Plugging this into (5.6.6) leads to

$$x_n \geq 1 + \left(\frac{n-m+1}{n} \right)^m \cdot \sum_{k=1}^m \frac{1}{k!}, \quad 1 \leq m \leq n.$$

Fix $m \in \mathbb{N}$ now and take the limit $n \rightarrow \infty$. Because

$$\lim_{n \rightarrow \infty} \left(\frac{n-m+1}{n} \right)^m = 1,$$

the last estimate implies that

$$e = \lim_{n \rightarrow \infty} x_n \geq 1 + \lim_{n \rightarrow \infty} \left(\frac{n-m+1}{n} \right)^m \cdot \sum_{k=1}^m \frac{1}{k!} = \sum_{k=0}^m \frac{1}{k!}.$$

This is valid for any $m \geq 1$, hence we may take the limit $m \rightarrow \infty$ and get

$$e \geq \sum_{k=0}^{\infty} \frac{1}{k!}.$$

Together with (5.6.5) this completes the proof. ■

Another formulation of Proposition 5.6.11 is

$$\lim_{n \rightarrow \infty} \sum_{k=0}^n \frac{1}{k!} = e.$$

For application it may be of interest how fast the left-hand sum converges to e . So let

$$e_n := \sum_{k=0}^n \frac{1}{k!}, \quad n = 0, 1, 2, \dots$$

Then on one hand we have

$$e_0 < e_1 < e_2 < \dots < e$$

and on the other hand Proposition 5.6.11 implies

$$\lim_{n \rightarrow \infty} e_n = e.$$

Proposition 5.6.12. *For all $n \geq 0$*

$$0 < e - e_n \leq \frac{2}{(n+1)!}.$$

Proof: Fix some $0 \leq n < m$. Then

$$e_m - e_n = \sum_{k=n+1}^m \frac{1}{k!} = \frac{1}{n!} \left[\frac{1}{n+1} + \frac{1}{(n+1)(n+2)} + \dots + \frac{1}{(n+1)(n+2)\dots m} \right].$$

Next we enlarge the second, third, and so on, fraction by taking only the two last factors in the product. Furthermore, we use

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

and obtain that

$$\begin{aligned} e_m - e_n &\leq \frac{1}{n!} \left[\frac{1}{n+1} + \frac{1}{(n+1)(n+2)} + \frac{1}{(n+2)(n+3)} + \dots + \frac{1}{(m-1)m} \right] \\ &= \frac{1}{n!} \left[\frac{1}{n+1} + \left(\frac{1}{n+1} - \frac{1}{n+2} \right) + \dots + \left(\frac{1}{m-1} - \frac{1}{m} \right) \right] = \frac{1}{n!} \left[\frac{2}{n+1} - \frac{1}{m} \right] \\ &\leq \frac{1}{n!} \cdot \frac{2}{n+1} = \frac{2}{(n+1)!}. \end{aligned}$$

So we end up with

$$e_m - e_n \leq \frac{2}{(n+1)!}.$$

The right-hand side is independent of m . Thus, we may take on the left-hand side the limit $m \rightarrow \infty$ and obtain

$$e - e_n = \lim_{m \rightarrow \infty} e_m - e_n \leq \frac{2}{(n+1)!}.$$

This completes the proof. ■

So we see that the sequence $(e_n)_{n \geq 1}$ converges to e very quickly. For example, the first values of this sequence are

$$e_0 = 1, e_1 = 2, e_2 = 2.5, e_3 = 2.6\bar{6}, e_4 = 2.70833, e_5 = 2.71667, \text{ and } e_6 = 2.71806$$

Recall that $e = 2.71828 \dots$.

Proposition 5.6.12 has a very interesting consequence.

Corollary 5.6.13. *The Euler number e is irrational¹⁴.*

¹⁴In 1873, the French mathematician Charles Hermite (1822–1901) proved that e is even transcendental. The original proof is quite complicated. In 1893, the German mathematician David Hilbert (1862–1943) gave an easier proof of this fact, not only for e , but also for π . See Remark 4.7.5 for the definition of transcendental numbers.

Proof: Assume the contrary, that is that e is rational. Then there are integers n and m with $n \geq 2$ such that

$$e = \frac{m}{n}.$$

Using Proposition 5.6.12, this implies

$$0 < \frac{m}{n} - e_n = e - e_n \leq \frac{2}{(n+1)!}.$$

If we multiply these estimate by $n!$ we get

$$0 < m(n-1)! - \sum_{k=0}^n \frac{n!}{k!} \leq \frac{2}{n+1} < 1.$$

Note that $n!/k!$ is an integer whenever $0 \leq k \leq n$, hence so is

$$\ell := m(n-1)! - \sum_{k=0}^n \frac{n!}{k!}.$$

So we found an integer ℓ satisfying $0 < \ell < 1$. Of course, this is impossible, and our assumption about e was wrong. This completes the proof. ■

Remark 5.6.6. Much more than Proposition 5.6.11 is valid. Using techniques from Calculus one may show the following result:

Proposition 5.6.14. *For any real number $x \in \mathbb{R}$ it follows that*

$$\lim_{n \rightarrow \infty} \left(1 + \frac{x}{n}\right)^n = \sum_{k=0}^{\infty} \frac{x^k}{k!} = e^x.$$

In the case $x = 1$ we rediscover Proposition 5.6.11.

The next criterion for the convergence of infinite series is probably the most powerful one. But it uses facts from Calculus not presented in this textbook. Nevertheless, because of its importance we want to state it here and refer to suitable textbooks about Integration, for example to [32].

Proposition 5.6.15 (Integral Test). *Let $f : [1, \infty) \rightarrow (0, \infty)$ be a nonincreasing function. Suppose $x_n = f(n)$, $n = 1, 2, \dots$. Then the x_n 's are summable if and only if $\int_1^\infty f(s) ds < \infty$.*

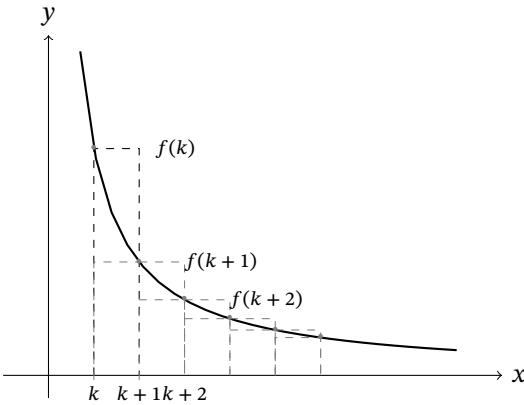
$$\left(\sum_{n=1}^{\infty} f(n) < \infty \right) \Leftrightarrow \left(\int_1^{\infty} f(s) ds < \infty \right).$$

Proof: Since f is assumed to be nonincreasing, it follows that for each $k \geq 1$

$$x_{k+1} = f(k+1) \leq f(u) \leq f(k) = x_k, \quad k \leq u \leq k+1,$$

which implies

$$x_{k+1} \leq \int_k^{k+1} f(u) du \leq x_k, \quad k = 1, 2, \dots.$$



As before, let $s_n = x_1 + \dots + x_n$ be the n th partial sum. If $n \geq 2$, the previous estimates lead to

$$\begin{aligned} s_n - x_1 &= \sum_{k=2}^n x_k = \sum_{k=1}^{n-1} x_{k+1} \leq \sum_{k=1}^{n-1} \int_k^{k+1} f(u) du = \int_1^n f(u) du \\ &= \sum_{k=1}^{n-1} \int_k^{k+1} f(u) du \leq \sum_{k=1}^{n-1} x_k = s_{n-1}. \end{aligned}$$

These estimates show that $\sup_{n \geq 1} s_n < \infty$ if and only if $\sup_{n \geq 1} \int_1^n f(u) du < \infty$. But $f(u) \geq 0$ for all $u \geq 1$, hence $n \mapsto \int_1^n f(u) du$ is nondecreasing and

$$\int_1^\infty f(u) du = \lim_{n \rightarrow \infty} \int_1^n f(u) du = \sup \int_1^n f(u) du.$$

Furthermore, $\sup_{n \geq 1} s_n = \lim_{n \rightarrow \infty} s_n = \sum_{n=1}^\infty x_n$. These two observations complete the proof. ■

Example 5.6.6. Given some $\alpha > 0$, we apply the previous Integral Test to the function f on $[1, \infty)$ defined by $f(u) = u^{-\alpha}$. Because of

$$\int_1^n u^{-\alpha} du = \begin{cases} \frac{n^{-\alpha+1}-1}{1-\alpha} & : \alpha \neq 1 \\ \log n & : \alpha = 1 \end{cases}$$

it follows $\int_1^\infty u^{-\alpha} du < \infty$ if and only if $1 < \alpha < \infty$. Note that

$$n^{-\alpha+1} \xrightarrow[n \rightarrow \infty]{} \begin{cases} \infty & : 0 < \alpha < 1 \\ 0 & : 1 < \alpha < \infty. \end{cases}$$

From Proposition 5.6.15 we derive $\sum_{n=1}^\infty n^{-\alpha} < \infty$ if $\alpha > 1$ while $\sum_{n=1}^\infty n^{-\alpha} = \infty$ in the case $0 < \alpha < 1$. The case $\alpha = 1$ was already treated in Example 5.6.3, but the result follows also from Proposition 5.6.15 because of $\log n \xrightarrow[n \rightarrow \infty]{} \infty$.

$$\sum_{n=1}^\infty \frac{1}{n^\alpha} \begin{cases} < \infty & : 1 < \alpha < \infty \\ = \infty & : 0 < \alpha \leq 1 \end{cases}$$

A final well-known result investigates the case of alternating sequences.

Proposition 5.6.16 (Leibniz's Theorem¹⁵ about Alternating Sequences). *Suppose $(a_n)_{n \geq 0}$ is a nonincreasing sequence of positive real numbers tending to zero. Then the alternating sequence $((-1)^n a_n)_{n \geq 0}$ is summable.*

$$(a_n \searrow 0) \Rightarrow \sum_{n=0}^{\infty} (-1)^n a_n \text{ exists}$$

Proof: Let us investigate the behavior of the partial sums s_n for even n . It holds

$$s_{2n+2} = s_{2n} + \underbrace{a_{2n+2} - a_{2n+1}}_{\leq 0} \leq s_{2n},$$

hence $s_0 \geq s_2 \geq s_4 \geq \dots$.

Furthermore, for all $n \geq 0$ it follows that

$$s_{2n} = a_0 - a_1 + \underbrace{a_2 - a_3}_{\geq 0} + \dots + \underbrace{a_{2n-2} - a_{2n-1}}_{\geq 0} + a_{2n} \geq a_0 - a_1 = s_1,$$

thus, s_1 is a lower bound for $\{s_{2n} : n \geq 0\}$. Consequently, by Proposition 5.3.1 the limit

$$S := \lim_{n \rightarrow \infty} s_{2n}$$

exists.

It remains to prove that $s_n \xrightarrow{n \rightarrow \infty} S$. To this end, given $\varepsilon > 0$, we choose an $N \geq 1$ for which

$$a_n < \frac{\varepsilon}{2}, \quad n \geq N, \quad \text{and} \quad |s_n - S| < \frac{\varepsilon}{2} \quad n \geq N, \quad n \text{ even}.$$

Of course, if n is even, by the choice of N it follows that $|s_n - S| < \frac{\varepsilon}{2} < \varepsilon$. But for odd $n \geq N$ we use that $n+1$ is even and obtain

$$|s_n - S| \leq |s_n - s_{n+1}| + |s_{n+1} - S| = a_{n+1} + |s_{n+1} - S| < \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon.$$

This proves $|s_n - S|$ for **all** $n \geq N$, hence $s_n \xrightarrow{n \rightarrow \infty} S$ as asserted. Thus, the proof is completed. ■

Remark 5.6.7. Both assumptions about the sequence $(a_n)_{n \geq 0}$ (nonincreasing and tending to zero) are crucial and cannot be omitted. For example, if $a_n = 1$ for all $n \geq 0$, then, of course, $((-1)^n a_n)_{n \geq 1}$ is not summable.

Furthermore, if we consider the sequence $(a_n)_{n \geq 0}$ given by

$$\frac{1}{\sqrt{2}+1}, \frac{1}{\sqrt{2}-1}, \frac{1}{\sqrt{3}+1}, \frac{1}{\sqrt{3}-1}, \frac{1}{\sqrt{4}+1}, \frac{1}{\sqrt{4}-1}, \dots,$$

¹⁵Gottfried Wilhelm Leibniz (1646–1716) was a German mathematician who developed differential and integral calculus independently of Isaac Newton.

then $a_n \xrightarrow{n \rightarrow \infty} 0$, but $\sum_{n=0}^{\infty} (-1)^n a_n$ does not exist. Why? The partial sums for odd indices equal

$$\begin{aligned}s_{2n+1} &= \frac{1}{\sqrt{2}+1} - \frac{1}{\sqrt{2}-1} + \frac{1}{\sqrt{3}+1} - \frac{1}{\sqrt{3}-1} + \cdots + \frac{1}{\sqrt{n+2}+1} - \frac{1}{\sqrt{n+2}-1} \\ &= 2 + 1 + \frac{2}{3} + \cdots + \frac{2}{n+1} = 2 \left(1 + \frac{1}{2} + \cdots + \frac{1}{n+1} \right).\end{aligned}$$

The results presented in Example 5.6.3 imply $\lim_{n \rightarrow \infty} s_{2n+1} = \infty$, hence, in particular, the sequence of partial sums is not bounded above. By Proposition 5.2.6 we know that convergent sequences are necessarily bounded. Consequently, $(s_n)_{n \geq 0}$ cannot be convergent and $((-1)^n a_n)_{n \geq 0}$ is not summable.

Question: Why does the previous example not contradict Proposition 5.6.16?

Example 5.6.7. For any $\alpha > 0$ the infinite sum

$$\sum_{n=1}^{\infty} \frac{(-1)^{n+1}}{n^{\alpha}}$$

exists. In particular, the limit

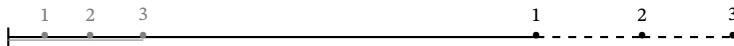
$$\lim_{n \rightarrow \infty} \left(1 - \frac{1}{2} + \frac{1}{3} - \frac{1}{4} \pm \cdots \pm \frac{1}{n} \right)$$

exists. It is interesting¹⁶ to compare this with the result presented in Example 5.6.3.

Next we present a more elaborated example. We decided to include it into this textbook because it demonstrates the usefulness of infinite series even in cases where one does not expect it.

Example 5.6.8. A rubber band has length 1 m (meter). At its left end sits a snail. Every morning the snail moves 2 cm (centimeter) to the right. In the evening the band is stretched by 4 cm. Thus, after one day the length of the band is 1.04 m, the next day 1.08 m and so on.

The question is now whether the snail arrives at the right end of the band within finite time. And if the answer is positive, one may ask how many days does it take for the snail to reach the finish.



The left-hand bullets are the positions of the snail after its move at days 1, 2, and 3 while the right-hand bullets show the length of the band before stretching it that day.

To answer these questions we argue as follows: At the first day the snail crawls a $2/100$ part of the band. Next day the proportion is $2/104$ of the band, and so on. So, at day n the progress of the snail is the

$$\frac{1}{2} \cdot \frac{1}{24+n}, \quad n = 1, 2, \dots$$

part of the band.

¹⁶A fascinating problem in probability theory is what happens if one chooses the signs of $1/n$ randomly, for example, by tossing a fair coin labeled with +1 and -1. It turns out, that then the sum exists for almost all choices of signs. Thus, the divergence of the harmonic series is the exception, the convergence of the alternating sum is the normal case.

Summing up the proportions of all days, after move n , the snail reaches the

$$S_n := \frac{1}{2} \left(\frac{1}{25} + \frac{1}{26} + \frac{1}{27} + \cdots + \frac{1}{24+n} \right) = \frac{1}{2} \left[\sum_{k=1}^{24+n} \frac{1}{k} - \sum_{k=1}^{24} \frac{1}{k} \right]$$

part of the band.

The snail arrives at the right-hand end of the band at day n provided that $S_{n-1} < 1$, yet $S_n \geq 1$. At this point infinite series come into the game. We know by Example 5.6.3 that the Harmonic Series diverges, i.e.,

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

Of course, this implies (only finitely many terms of the Harmonic Series are missing) that

$$\frac{1}{25} + \frac{1}{26} + \frac{1}{27} + \cdots + \frac{1}{24+n} \xrightarrow{n \rightarrow \infty} \infty ,$$

hence $S_n \xrightarrow{n \rightarrow \infty} \infty$. Consequently, there is a smallest $n \geq 1$ with $S_n \geq 1$. This tells us that the snail arrives at the right-hand end after finitely many steps.

If one asks now how many days it takes for the snail to succeed, one has to look for the smallest number $n \in \mathbb{N}$ satisfying

$$S_n \geq 1 \Leftrightarrow \frac{1}{2} \left[\sum_{k=1}^{24+n} \frac{1}{k} - \sum_{k=1}^{24} \frac{1}{k} \right] \geq 1 .$$

Numerical calculations give

$$S_{156} = 0,998495 \quad \text{and} \quad S_{157} = 1,00126 ,$$

hence the snail arrives at the right-hand end on day 157. During this time the band was stretched 156 times. Thus, its length at the time of arrival is

$$100 + 4 \cdot 156 = 724 \text{ cm} = 7,24 \text{ m} .$$

One may ask now what happens if the snail moves only one cm per day while the band is elongated by one meter per day. Does the snail succeed in this extreme situation as well?

The answer is affirmative. Here we get

$$S_n = \frac{1}{100} + \cdots + \frac{1}{100n} = \frac{1}{100} \cdot \sum_{k=1}^n \frac{1}{k} \xrightarrow{n \rightarrow \infty} \infty .$$

By the same arguments as before there is a smallest $n \geq 1$ for which $S_n \geq 1$. The only difference is that now we have to apply asymptotic methods to get an approximate solution for the day of arrival. As it is known by methods from Calculus,

$$\lim_{n \rightarrow \infty} \left[\sum_{k=1}^n \frac{1}{k} - \log n \right] = \gamma$$

with Euler–Mascheroni constant $\gamma \approx 0.5577216$. Hence, we may replace S_n by

$$\tilde{S}_n = \frac{1}{100} \cdot (\log(n) + \gamma) .$$

Doing so we get that $\tilde{S}_n = 1$ if n is approximately

$$e^{100-\gamma} \approx e^{100} \cdot 0.561459 \approx 1.50927 \cdot 10^{42}.$$

So we see, under these conditions the snail succeeds as well, it only takes a little more time.

Remark 5.6.8. The model of the preceding example is maybe not very realistic. In general, the snail will not move step by step but continually. And also the band may be stretched in the same way. So let us suppose now that the speed of the snail is v_S while the band is elongated with speed v_B . Methods from Calculus then imply that the snail reaches the right-hand end of the band at time $T > 0$ where

$$T = \frac{L}{v_B} (e^{v_B/v_S} - 1).$$

Here $L > 0$ denotes the length of the band at time 0. For instance, if as in the first example $v_S = 2$ cm per day, $v_B = 4$ cm per day and $L = 100$ cm, then

$$T = \frac{100}{4} (e^{4/2} - 1) = 25 \cdot (e^2 - 1) \approx 159.72 \text{ days.}$$

This is not so much different from the 157 days in the case of noncontinual moving and stretching.

Our final goal is to transform the results in Section 4.7 into the language of infinite series. So let us recall Definition 4.7.1. Let $b \geq 2$ be a fixed integer, the base of the representation. A real number $x \in [0, 1)$ admits the representation

$$x =_b 0.x_1x_2\cdots \quad \text{with some } x_j \in \{0, \dots, b-1\}$$

provided that for each $n \geq 1$ the following estimate is satisfied:

$$(5.6.7) \quad \sum_{j=1}^n \frac{x_j}{b^j} \leq x < \sum_{j=1}^n \frac{x_j}{b^j} + \frac{1}{b^n}.$$

Setting

$$S_n := \sum_{j=1}^n \frac{x_j}{b^j}, \quad n = 1, 2, \dots,$$

then condition (5.6.7) may be rewritten as

$$(5.6.8) \quad S_n \leq x < S_n + \frac{1}{b^n}, \quad n = 1, 2, \dots.$$

Next observe that S_n is nothing else as the n th partial sum of the sequence $\frac{x_1}{b}, \frac{x_2}{b^2}, \dots$. Moreover, by $x_j \leq b-1$ it follows that

$$S_n = \sum_{j=1}^n \frac{x_j}{b^j} \leq (b-1) \sum_{j=1}^n \frac{1}{b^j} = 1 - \frac{1}{b^n} \leq 1.$$

Since $x_j/b^j \geq 0$, Proposition 5.6.6 applies. Hence, the infinite sum

$$S = \sum_{j=1}^{\infty} \frac{x_j}{b^j} = \lim_{n \rightarrow \infty} S_n$$

exists for any choice of the x_j 's in $\{0, \dots, b-1\}$. Moreover, $S \leq 1$. After this preparation we are now able to prove a result which relates representations of real numbers as b -fractions with infinite series.

Proposition 5.6.17. Suppose a number $x \in [0, 1)$ can be represented as

$$x =_b 0.x_1x_2\dots$$

for certain $x_j \in \{0, \dots, b-1\}$. Then this implies

$$x = \sum_{j=1}^{\infty} \frac{x_j}{b^j} = \lim_{n \rightarrow \infty} \sum_{j=1}^n \frac{x_j}{b^j}.$$

$$x \in [0, 1) \Rightarrow x =_b 0.x_1x_2x_3\dots \Rightarrow x = \sum_{j=1}^{\infty} \frac{x_j}{b^j}.$$

Proof: Let as before

$$S_n = \sum_{j=1}^n \frac{x_j}{b^j}, \quad n = 1, 2, \dots$$

Because of $b^{-n} \underset{n \rightarrow \infty}{\rightarrow} 0$ it follows

$$S = \lim_{n \rightarrow \infty} S_n = \lim_{n \rightarrow \infty} \left[S_n + \frac{1}{b^n} \right].$$

Since $x =_b 0.x_1x_2\dots$, by definition, estimate (5.6.8) has to be satisfied for all $n \geq 1$. That is, for all $n \geq 1$ we have

$$S_n \leq x < S_n + \frac{1}{b^n}.$$

Now Proposition 5.2.7 applies and leads to $x = S$ as asserted. ■

Example 5.6.9. We have

$$\frac{1}{7} = 0.\overline{142857} \Rightarrow \frac{1}{7} = \frac{1}{10} + \frac{4}{10^2} + \dots + \frac{7}{10^6} + \frac{1}{10^7} + \frac{4}{10^8} + \dots$$

or

$$0.3 =_2 0.0\overline{1001} \Rightarrow 0.3 = \frac{1}{4} + \frac{1}{32} + \frac{1}{64} + \frac{1}{2^9} + \frac{1}{2^{10}} + \frac{1}{2^{13}} + \frac{1}{2^{14}} + \dots$$

One may ask now whether Proposition 5.6.17 can be reversed. In Proposition 5.6.17 we start with a given $x \in [0, 1)$, write it as b -fraction $x =_b 0.x_1x_2\dots$, build the infinite sum

$$\sum_{j=1}^{\infty} \frac{x_j}{b^j}$$

and get back the number x we started with.

The reverse problem may be formulated as follows. One starts with an arbitrary sequence x_1, x_2, \dots of integers between 0 and $b-1$ and with this sequence one builds the infinite sum

$$S := \sum_{j=1}^{\infty} \frac{x_j}{b^j}.$$

Then, as shown above, $0 \leq S \leq 1$. Next we represent S as b -fraction and the question is whether

$$S =_b 0.x_1x_2\cdots,$$

i.e., whether we get back the sequence we started with. Unfortunately, the answer is negative.

Example 5.6.10. Choose $b = 2$ and start with the sequence $0, 0, 1, 1, 1, \dots$. Then

$$S = \sum_{j=3}^{\infty} \frac{1}{2^j} = \frac{1}{8} \cdot \sum_{j=0}^{\infty} \frac{1}{2^j} = \frac{1}{4} =_2 0.01 = 0.01000\cdots.$$

So we started with the infinite sequence $0, 0, 1, 1, \dots$ and ended up with the sequence $0, 1, 0, 0, 0, \dots$

To make the following investigation clearer, let us start with an elementary lemma which is an interesting consequence of the summation formula for the geometric series as presented in Example 5.6.1.

Lemma 5.6.18. Let $(x_j)_{j \geq 1}$ be a sequence of integers with $0 \leq x_j \leq b - 1$ for each $j \geq 1$. Then for every integer $k \geq 0$

$$(5.6.9) \quad 0 \leq \sum_{j=k+1}^{\infty} \frac{x_j}{b^j} \leq \frac{1}{b^k}.$$

Moreover, we have equality on the right-hand side, i.e.,

$$\sum_{j=k+1}^{\infty} \frac{x_j}{b^j} = \frac{1}{b^k},$$

if and only if $x_{k+1} = x_{k+2} = \dots = b - 1$.

$$(\forall k \geq 0) \left(0 \leq \sum_{j=k+1}^{\infty} \frac{x_j}{b^j} \leq \frac{1}{b^k} \right) \text{ and } \left(\sum_{j=k+1}^{\infty} \frac{x_j}{b^j} = \frac{1}{b^k} \right) \Leftrightarrow (x_{k+1} = x_{k+2} = \dots = b - 1).$$

Proof: Since the x_j 's and $1/b^j$ are nonnegative, the left-hand inequality in (5.6.9) is obvious.

The right-hand inequality is an easy consequence of

$$\sum_{j=k+1}^{\infty} \frac{x_j}{b^j} \leq (b-1) \sum_{j=k+1}^{\infty} \frac{1}{b^j} = \frac{b-1}{b^k} \sum_{j=k+1}^{\infty} \frac{1}{b^{j-k}}$$

and

$$\sum_{j=k+1}^{\infty} \frac{1}{b^{j-k}} = \sum_{j=1}^{\infty} \frac{1}{b^j} = \frac{1}{1-1/b} - 1 = \frac{1}{b-1}.$$

If there is a $j_0 \geq k + 1$ with $x_{j_0} < b - 1$, then this implies (cf. Exercise 5.6.1) that

$$\sum_{j=k+1}^{\infty} \frac{x_j}{b^j} < (b-1) \sum_{j=k+1}^{\infty} \frac{1}{b^j} = \frac{1}{b^k}.$$

Hence, it follows

$$\sum_{j=k+1}^{\infty} \frac{x_j}{b^j} = \frac{1}{b^k}$$

if and only if $x_j = b - 1$ for all $j \geq k + 1$. This completes the proof. ■

Proposition 5.6.19. Let $(x_j)_{j \geq 1}$ be a sequence of integers in $\{0, \dots, b - 1\}$. Define a real number $x \in [0, 1]$ by

$$x := \sum_{j=1}^{\infty} \frac{x_j}{b^j}.$$

- (1) If there is a $k \geq 1$ with $x_k < b - 1$ and $x_{k+1} = x_{k+2} = \dots = b - 1$, then x is b -rational with (finite) representation

$$x =_b 0.x_1 \cdots x_{k-1}(x_k + 1).$$

- (2) If infinitely many of the x_j 's are different from $b - 1$, then x may be represented as

$$x =_b 0.x_1 x_2 \cdots$$

Proof: The first case is a direct consequence of Lemma 5.6.18. Indeed, here it follows that

$$x = \sum_{j=1}^{\infty} \frac{x_j}{b^j} = \sum_{j=1}^k \frac{x_j}{b^j} + \sum_{j=k+1}^{\infty} \frac{b-1}{b^j} = \sum_{j=1}^{k-1} \frac{x_j}{b^j} + \frac{x_k + 1}{b^k},$$

hence

$$x =_b 0.x_1 \cdots x_{k-1}(x_k + 1)$$

as asserted.

To prove the second part we assume that infinitely many of the x_j 's are different from $b - 1$. Then Lemma 5.6.18 applies and leads to

$$(5.6.10) \quad x - \sum_{j=1}^n \frac{x_j}{b^j} = \sum_{j=1}^{\infty} \frac{x_j}{b^j} - \sum_{j=1}^n \frac{x_j}{b^j} = \sum_{j=n+1}^{\infty} \frac{x_j}{b^j} < \frac{1}{b^n}.$$

Of course, we have

$$\sum_{j=1}^n \frac{x_j}{b^j} \leq \sum_{j=1}^{\infty} \frac{x_j}{b^j} = x.$$

Thus, if we combine this with (5.6.10) we obtain

$$\sum_{j=1}^n \frac{x_j}{b^j} \leq x < \sum_{j=1}^n \frac{x_j}{b^j} + \frac{1}{b^n}.$$

This being true for all $n \geq 1$ means that

$$x =_b 0.x_1 x_2 \cdots,$$

as claimed. ■

Summary: Let $x \in [0, 1)$. Then it follows that

$$\begin{aligned} x \text{ is } b\text{-rational} &\Leftrightarrow x = \sum_{j=1}^n \frac{x_j}{b^j} \text{ for } n \geq 1, x_j \in \{0, \dots, b-1\}, x_n \neq 0 \\ &\Leftrightarrow x =_b 0.x_1 \dots x_n = 0.x_1 \dots x_{n-1}(x_n - 1)(b-1) \dots \\ x \text{ is not } b\text{-rational} &\Leftrightarrow x = \sum_{j=1}^{\infty} \frac{x_j}{b^j} \text{ with } x_j < b-1 \text{ infinitely often} \\ &\Leftrightarrow x =_b 0.x_1 x_2 \dots \forall k \geq 1 \exists j \geq k, x_j \neq 0 \end{aligned}$$

Example 5.6.11. (1) We have

$$\begin{aligned} x = \frac{1}{2} &\Rightarrow x = 0.5 \quad \text{or} \quad x = \frac{4}{10} + \sum_{j=2}^{\infty} \frac{9}{10^j} = 0.4999\dots, \\ x = \frac{3}{4} &\Rightarrow x =_2 0.11 \quad \text{or} \quad x = \frac{1}{2} + \frac{0}{4} + \sum_{j=3}^{\infty} \frac{1}{2^j} =_2 0.101111\dots, \\ x = \frac{5}{8} &\Rightarrow x =_{16} 0.A \quad \text{or} \quad x = \frac{9}{16} + \sum_{j=2}^{\infty} \frac{F}{16^j} =_{16} 0.9FFFF\dots. \end{aligned}$$

(2) Examples of non- b -rational numbers are

$$\begin{aligned} x = \frac{8}{11} &\Rightarrow x = 0.\overline{72} = \frac{7}{10} + \frac{2}{10^2} + \frac{7}{10^3} + \dots, \\ x = \frac{5}{7} &\Rightarrow x =_2 0.\overline{101} = \frac{1}{2} + \frac{1}{8} + \frac{1}{16} + \frac{1}{64} + \dots, \\ \pi = 3.14159\dots &\Rightarrow \pi = 3 + \frac{1}{10} + \frac{4}{10^2} + \frac{1}{10^3} + \frac{5}{10^4} + \frac{9}{10^5} + \dots. \end{aligned}$$

Finally, let us answer a question raised in Remark 4.7.6. Recall that we introduced \mathbb{R} by an axiomatic approach, hence it is natural to ask whether \mathbb{R} is a minimal extension of \mathbb{Q} . Theoretically, there could exist elements in \mathbb{R} which are not needed to fill the holes in \mathbb{Q} .

Proposition 5.6.20. *For each real number $x \in \mathbb{R}$ there exists a sequence $(q_n)_{n \geq 1}$ of rational numbers such that*

$$\lim_{n \rightarrow \infty} q_n = x.$$

Consequently, \mathbb{R} is the minimal set which contains \mathbb{Q} and which is closed under taking limits of rational numbers¹⁷.

Proof: There are many ways to verify this result. Let us choose here an approach via the decimal expansion of real numbers.

¹⁷In the language of analysis one would say that \mathbb{R} is the closed hull (or the closure) of \mathbb{Q} .

Let $x \in \mathbb{R}$ be an arbitrary real number. An application of Proposition 5.6.17 combined with Theorem 4.7.3 implies the existence of an integer x_0 as well as of $x_j \in \{0, \dots, 9\}$ such that

$$x = x_0 + 0.x_1x_2\dots = x_0 + \sum_{j=1}^{\infty} \frac{x_j}{10^j}.$$

Now set

$$q_n := x_0 + \sum_{j=1}^n \frac{x_j}{10^j}, \quad n = 1, 2, \dots$$

Of course, the q_n are rational numbers converging to x by the definition of infinite sums.

The second part of the proposition is an immediate consequence of the first one. Indeed, whenever a set A contains \mathbb{Q} as well as limits of elements in \mathbb{Q} , then by the first part the set A contains also each real number, saying $\mathbb{R} \subseteq A$. Thus, \mathbb{R} contains all elements which one needs in order to complete \mathbb{Q} and nothing more. ■

Exercise 5.6.1. Let $(x_j)_{j \geq 1}$ and $(y_j)_{j \geq 1}$ be two summable sequences of real numbers such that $x_j \leq y_j$, $j = 1, 2, \dots$ Why does this imply

$$\sum_{j=1}^{\infty} x_j \leq \sum_{j=1}^{\infty} y_j ?$$

Moreover, it follows

$$\sum_{j=1}^{\infty} x_j = \sum_{j=1}^{\infty} y_j \Leftrightarrow x_1 = y_1, x_2 = y_2, \dots$$

Exercise 5.6.2. Prove that the series $\sum_{n=0}^{\infty} \frac{1}{2^n+1}$ converges.

Exercise 5.6.3. Evaluate $\sum_{n=1}^{\infty} \frac{1}{n^2+n}$.

Exercise 5.6.4. Verify¹⁸ that $\sum_{n=1}^{\infty} \frac{1}{n^2} \leq 2$.

Exercise 5.6.5. Show that

$$\begin{array}{lll} a) \quad \sum_{n=0}^{\infty} \frac{(-1)^n}{2^n} = \frac{2}{3} & b) \quad \sum_{n=2}^{\infty} \frac{1}{3^{n-1}} = \frac{1}{2} & c) \quad \sum_{n=1}^{\infty} \frac{3}{4^n} = 1 \\ d) \quad \sum_{n=0}^{\infty} \frac{(-3)^n}{4^n} = \frac{4}{7} & e) \quad \sum_{n=1}^{\infty} \frac{1}{4n^2-1} = \frac{1}{2}. & \end{array}$$

Exercise 5.6.6. Evaluate for $|q| < 1$ and natural numbers $N < M$,

$$(a) \quad \sum_{n=0}^{\infty} q^{2n} \quad (b) \quad \sum_{n=N+1}^{\infty} q^n \quad (c) \quad \sum_{n=N+1}^M q^n \quad (d) \quad \sum_{n=1}^{\infty} n q^n$$

as well as the infinite sums

$$(e) \quad \sum_{n=0}^{\infty} [3^{-2n} + 4 \cdot 5^{-n}] \quad (f) \quad \sum_{n=1}^{\infty} \frac{3}{n(n+1)}.$$

¹⁸The value of this sum equals $\pi^2/6$ (known as the Basel Problem and solved by Leonhard Euler in 1734).

Exercise 5.6.7. Prove the following **root test** for infinite series.

Proposition 5.6.21. Let $(x_n)_{n \geq 1}$ be a sequence of real numbers such that

$$\lim_{n \rightarrow \infty} \sqrt[n]{|x_n|} := r$$

exists. If $0 \leq r < 1$, then $(x_n)_{n \geq 1}$ is absolutely summable. In contrast to that, $r > 1$ implies that $(x_n)_{n \geq 1}$ is **not** summable. In the case $r = 1$ the sequence $(x_n)_{n \geq 1}$ may or may not be summable,

Exercise 5.6.8. Do the following infinite series converge or diverge? Justify your answer.

- | | | |
|--|--------------------------------------|---|
| (a) | (b) | (c) |
| $\sum_{n=0}^{\infty} \frac{1}{2n+1}$ | $\sum_{n=1}^{\infty} \frac{n!}{n^n}$ | $\sum_{n=1}^{\infty} \frac{n+1}{n^2+1}$ |
| (d) | (e) | |
| $\sum_{n=2}^{\infty} \frac{n^2 - 2n + 5}{n^4 - 3n^2 + 2n}$ | $\sum_{n=1}^{\infty} n^2 2^{-n}$ | |

Exercise 5.6.9. Determine $x \in \mathbb{R}$ such that the subsequent series converge.

$$(a) \quad \sum_{n=1}^{\infty} \frac{(n!)^2}{(2n)!} x^n, \quad \text{respectively} \quad (b) \quad \sum_{n=1}^{\infty} \frac{x^n}{\sqrt{n}} ?$$

Exercise 5.6.10. Prove the last estimate left open in the proof of Proposition 4.7.8. That is, show the following: Given $\varepsilon > 0$, there is an $N \geq 1$ such that for all $k \geq 1$ we have $|1 - y_{k,N}| < \varepsilon$ where

$$y_{k,N} =_3 0.222\dots 2 \underbrace{1}_{N+k} \underbrace{1}_{N+k+1} 000\dots$$

Exercise 5.6.11. In Proposition 5.6.7 we proved that every absolutely summable sequence is summable. The proof of this fact is based upon Theorem 5.5.4 which asserts that any Cauchy sequence in \mathbb{R} is convergent. Show conversely that the assertion of Theorem 5.5.4, every real Cauchy sequence converges, may be derived from the fact that any absolutely summable sequence is summable.

Hint: Choose a Cauchy sequence and show that it has a convergent subsequence. Then the convergence of the Cauchy sequence follows by Lemma 5.5.2.

5.7. Infinite Continued Fractions

In the previous section, we discussed the representation of real numbers as finite or infinite, periodic or nonperiodic, general b -fractions for some base $b \geq 2$. We observed that any real number x can be written in a unique way as

$$x = a_0 + \sum_{k=1}^{\infty} \frac{a_k}{10^k} = a_0 + \frac{a_1}{10} + \frac{a_2}{10^2} + \frac{a_3}{10^3} + \dots = a_0.a_1a_2a_3\dots$$

with integers $a_0 \in \mathbb{Z}$ and $0 \leq a_j \leq 9$ if $j \geq 1$.

The aim of this section is to study another representation of real numbers which is a natural extension of the results of Section 3.4 on finite continued fractions. In that

section, we investigated finite continued fractions defined as follows. Given $n \in \mathbb{N}_0$, $a_0 \in \mathbb{Z}$ and $a_1, \dots, a_n \in \mathbb{N}$, define

$$[a_0; a_1, \dots, a_n] = a_0 + \cfrac{1}{a_1 + \cfrac{1}{\ddots + \cfrac{1}{a_{n-1} + \cfrac{1}{a_n}}}}.$$

In Section 3.4, we proved that any rational number has a unique representation as a finite continued fraction (with $a_n > 1$) and any finite continued fraction is a rational number.

Example 5.7.1. Note that

$$\frac{43}{30} = 1 + \cfrac{1}{2 + \cfrac{1}{3 + \cfrac{1}{4}}} = [1; 2, 3, 4].$$

We now introduce the notion of an infinite continued fraction.

Definition 5.7.1. Let $a_0 \in \mathbb{Z}$ and an infinite sequence of natural numbers $(a_n)_{n \geq 1}$. If the limit $\lim_{n \rightarrow \infty} [a_0; a_1, \dots, a_n]$ exists and equals x , then we write that

$$x = [a_0; a_1, a_2, \dots]$$

and we call this equation a representation of the number x as **infinite continued fraction**. For each $m \in \mathbb{N}_0$, **the m -th convergent** of x is defined as

$$r_m := [a_0; a_1, \dots, a_m].$$

The sequence $(r_m)_{m \geq 0}$ is called the sequence of **convergents** of x or of the fraction $[a_0; a_1, a_2, \dots]$, respectively.

Example 5.7.2. In Example 3.4.6, we noticed that

$$[1; \underbrace{1, \dots, 1}_{n \text{ terms}}] = F_{n+2}/F_{n+1}$$

where $(F_n)_{n \geq 0}$ is the sequence of Fibonacci numbers. And we argued (without using the notion of limit at that time), that

$$\lim_{n \rightarrow \infty} \frac{F_{n+1}}{F_n} = \frac{1 + \sqrt{5}}{2} = \varphi.$$

Later when limits of sequences were available, we gave an exact proof of this fact in Proposition 5.3.4.

Thus, the golden ratio φ can be represented as infinite continued fraction as follows:

$$\varphi = [1; 1, 1, 1, \dots] = 1 + \cfrac{1}{1 + \cfrac{1}{1 + \cfrac{1}{1 + \cfrac{1}{1 + \dots}}}}$$

Let us still state another example of interest.

Example 5.7.3. In Example 3.4.7 we investigated the continued fraction

$$[1; \underbrace{2, \dots, 2}_{n \text{ terms}}],$$

which, as shown in Proposition 5.3.6, tends to $\sqrt{2}$. Consequently, the representation of $\sqrt{2}$ as infinite continued fraction is

$$\sqrt{2} = [1; 2, 2, 2, \dots] = 1 + \cfrac{1}{2 + \cfrac{1}{2 + \cfrac{1}{2 + \cfrac{1}{2 + \dots}}}}$$

Now the following basic questions arise:

- (1) For which $a_0 \in \mathbb{Z}$ and $a_j \in \mathbb{N}$ does

$$\lim_{n \rightarrow \infty} [a_0; a_1, \dots, a_n]$$

exist?

- (2) Which real numbers $x \in \mathbb{R}$ admit a representation as infinite continued fractions? That is, which $x \in \mathbb{R}$ may be written as

$$x = [a_0; a_1, a_2, \dots]$$

for suitable integer $a_0 \in \mathbb{Z}$ and some $a_j \in \mathbb{N}$?

- (3) How about uniqueness of representations of infinite continued fractions? In other words, does

$$[a_0; a_1, a_2, \dots] = [b_0; b_1, b_2, \dots]$$

imply $a_0 = b_0, a_1 = b_1$ and so on?

- (4) A natural question is which $x \in \mathbb{R}$ may be represented as an infinite continued fraction with a periodic sequence a_1, a_2, \dots of natural numbers? We know from Section 3.4 that rational numbers correspond to finite continued fractions, hence a new class of real numbers has to show up when considering infinite continued fractions.

Let us answer now some of these questions.

Theorem 5.7.1. For any $a_0 \in \mathbb{Z}$ and all natural numbers a_1, a_2, \dots the limit

$$\lim_{n \rightarrow \infty} [a_0; a_1, \dots, a_n] := [a_0; a_1, a_2, \dots]$$

exists in \mathbb{R} . Moreover, this limit is always an irrational number.

Proof: Without loss of generality, let assume $a_0 \geq 0$. The proof rests on the results in Section 3.4. Let us summarize the basic facts about finite continued fractions proved there.

Let $(a_n)_{n \geq 0}$ be a sequence of integers with $a_0 \geq 0$ and $a_j > 0$ if $j \geq 1$. For $n \geq 0$, define

$$(5.7.1) \quad r_n = [a_0; a_1, a_2, \dots, a_n] = \frac{b_n}{c_n}$$

with $b_n, c_n \in \mathbb{N}_0$ and $\gcd(b_n, c_n) = 1$. From Section 3.4, we get that the sequence $(c_n)_{n \geq 0}$ is strictly increasing and therefore,

$$(5.7.2) \quad \lim_{n \rightarrow \infty} c_n = \infty.$$

Secondly, we have that the sequence of even convergents $(r_{2n})_{n \geq 0}$ is strictly increasing and the sequence of odd convergents $(r_{2n+1})_{n \geq 0}$ is strictly decreasing. Other important properties of the b_n and c_n are that for $n \geq 1$,

$$(5.7.3) \quad b_n c_{n-1} - b_{n-1} c_n = (-1)^{n-1}$$

and

$$(5.7.4) \quad |r_n - r_{n-1}| = \left| \frac{(-1)^{n-1}}{c_n c_{n-1}} \right| = \frac{1}{c_n c_{n-1}}.$$

We now prove the convergence of the sequence $(r_n)_{n \geq 0}$. In view of (5.7.3) the monotone convergence theorem (Theorem 5.3.1) implies the existence of the limits

$$A := \lim_{n \rightarrow \infty} r_{2n} \quad \text{and} \quad B := \lim_{n \rightarrow \infty} r_{2n-1}.$$

Moreover, using (5.7.2) and (5.7.4) one gets

$$|r_{2n} - r_{2n-1}| = \frac{1}{c_{2n} c_{2n-1}} \xrightarrow{n \rightarrow \infty} 0.$$

This implies that

$$A - B = \lim_{n \rightarrow \infty} [r_{2n} - r_{2n-1}] = 0.$$

So we obtain $A = B$ and, therefore (compare Exercise 5.4.3), the limit

$$\lim_{n \rightarrow \infty} r_n = \lim_{n \rightarrow \infty} [a_0; a_1, \dots, a_n]$$

exists for all $a_0 \in \mathbb{Z}$ and $a_j \in \mathbb{N}$, $j \geq 1$. This proves the first assertion.

Our next goal is to show that a real number x is irrational whenever it can be represented as

$$x = [a_0; a_1, a_2, \dots] = \lim_{n \rightarrow \infty} [a_0; a_1, \dots, a_n].$$

Because the sequences $(r_{2n})_{n \geq 0}$ and $(r_{2n+1})_{n \geq 0}$ are monotone, we get that

$$\frac{b_{2n}}{c_{2n}} = r_{2n} < x < r_{2n-1} = \frac{b_{2n-1}}{c_{2n-1}}, \quad n \geq 1,$$

which by (5.7.4) implies

$$0 < \left| x - \frac{b_{2n}}{c_{2n}} \right| < |r_{2n} - r_{2n-1}| = \frac{1}{c_{2n}c_{2n-1}}.$$

Multiplying this estimate by c_{2n} leads to

$$(5.7.5) \quad 0 < |xc_{2n} - b_{2n}| < \frac{1}{c_{2n-1}}.$$

Assume now that x is rational, i.e., $x = p/q$ for some integers p and $q > 0$. Plugging this into (5.7.5) and multiplying by $q \neq 0$ implies

$$0 < |pc_{2n} - qb_{2n}| < \frac{q}{c_{2n-1}} \xrightarrow{n \rightarrow \infty} 0.$$

In particular, if we choose n large enough, then $q/c_{2n-1} < 1$, which leads for those n to the estimate

$$0 < |pc_{2n} - qb_{2n}| < 1.$$

But $pc_{2n} - qb_{2n} \in \mathbb{Z}$, thus its absolute value can never belong to $(0, 1)$. This contradiction shows, that x cannot be written as fraction p/q , hence it has to be irrational.

The next objective is to verify the uniqueness of the representation as infinite continued fraction. Suppose we have that

$$x = [a_0; a_1, a_2, \dots] = [b_0; b_1, b_2, \dots]$$

for some $a_0, b_0 \in \mathbb{Z}$ and $a_j, b_j \in \mathbb{N}$ if $j \geq 1$. In a first step we have that

$$a_0 = r_0 < x < r_1 = a_0 + \frac{1}{a_1} \quad \text{as well as} \quad b_0 = r_0 < x < r_1 = b_0 + \frac{1}{b_1},$$

which implies $a_0 = \lfloor x \rfloor = b_0$, hence $a_0 = b_0$. Since

$$\begin{aligned} \lim_{n \rightarrow \infty} [a_0; a_1, \dots, a_n] &= \lim_{n \rightarrow \infty} \left[a_0 + \frac{1}{[a_1; a_2, \dots, a_n]} \right] = a_0 + \frac{1}{[a_1; a_2, a_3, \dots]} \\ \lim_{n \rightarrow \infty} [b_0; b_1, \dots, b_n] &= \lim_{n \rightarrow \infty} \left[b_0 + \frac{1}{[b_1; b_2, \dots, b_n]} \right] = b_0 + \frac{1}{[b_1; b_2, b_3, \dots]}, \end{aligned}$$

by $a_0 = b_0$ necessarily follows

$$[a_1; a_2, a_3, \dots] = [b_1; b_2, b_3, \dots].$$

An application of the first step to these two infinite continued fractions yields $a_1 = b_1$. Proceeding further in this way leads to $a_2 = b_2$ and so on. By induction, we finally get $a_j = b_j$ for all $j \geq 0$. This proves the uniqueness of the representation. ■

Our next goal is to answer the second question. Here we show that every irrational number admits a representation as an infinite continued fraction. To verify this, we need the following lemma.

Lemma 5.7.2. *Let $a_0 \in \mathbb{Z}$ and $a_j \in \mathbb{N}$, $j \geq 1$. Define the coprime integers b_n and c_n by (5.7.1). Moreover, for any $\alpha > 0$ set*

$$[a_0; a_1, \dots, a_{n-1}, \alpha] = a_0 + \cfrac{1}{a_1 + \cfrac{1}{\ddots + \cfrac{1}{a_{n-1} + \cfrac{1}{\alpha}}}}.$$

Then for all $n \geq 2$ it follows that

$$(5.7.6) \quad [a_0; a_1, \dots, a_{n-1}, \alpha] = \frac{b_{n-1}\alpha + b_{n-2}}{c_{n-1}\alpha + c_{n-2}}.$$

Proof: We prove the lemma by induction. Start with the base case $n = 2$. Then on one hand we have

$$[a_0; a_1, \alpha] = a_0 + \frac{1}{a_1 + \frac{1}{\alpha}} = \frac{a_0 a_1 \alpha + \alpha + a_0}{a_1 \alpha + 1}.$$

On the other hand, as we have shown in the proof of Proposition 3.4.5,

$$b_0 = a_0, \quad b_1 = a_0 a_1 + 1, \quad c_0 = 1, \quad \text{and} \quad c_1 = a_1.$$

This leads to

$$\frac{b_1 \alpha + b_0}{c_1 \alpha + c_0} = \frac{a_0 a_1 \alpha + \alpha + a_0}{a_1 \alpha + 1}.$$

Consequently, equation (5.7.6) is true for $n = 2$.

Suppose now that for some $n \geq 2$ and all $\alpha > 0$ the equation

$$(5.7.7) \quad [a_0; a_1, \dots, a_{n-1}, \alpha] = \frac{b_{n-1}\alpha + b_{n-2}}{c_{n-1}\alpha + c_{n-2}}$$

is satisfied. Our aim is to show that it is also true for $n + 1$. Thus, take any $\alpha > 0$. Then

$$[a_0; a_1, \dots, a_n, \alpha] = [a_0; a_1, \dots, a_n + \frac{1}{\alpha}]$$

allows us to apply (5.7.7) with $a_n + 1/\alpha$ instead of α . Doing so we obtain

$$\begin{aligned} [a_0; a_1, \dots, a_n, \alpha] &= \frac{b_{n-1}(a_n + \frac{1}{\alpha}) + b_{n-2}}{c_{n-1}(a_n + \frac{1}{\alpha}) + c_{n-2}} = \frac{b_{n-1}(a_n \alpha + 1) + b_{n-2}\alpha}{c_{n-1}(a_n \alpha + 1) + c_{n-2}\alpha} \\ &= \frac{(b_{n-1}a_n + b_{n-2})\alpha + b_{n-1}}{(c_{n-1}a_n + c_{n-2})\alpha + c_{n-1}} = \frac{b_n\alpha + b_{n-1}}{c_n\alpha + c_{n-1}}. \end{aligned}$$

Here in the last step we used the recursive formulas

$$b_n = a_n b_{n-1} + b_{n-2} \quad \text{and} \quad c_n = a_n c_{n-1} + c_{n-2}$$

as stated and proved in Proposition 3.4.5. Thus, equation (5.7.6) is also valid for $n + 1$, and this ends the proof of the lemma. \blacksquare

Now we are in position to complete the answer to question two.

Theorem 5.7.3. *Any irrational $x \in \mathbb{R}$ admits a unique representation as infinite continued fraction.*

$$(\forall x \notin \mathbb{Q})(\exists! a_0 \in \mathbb{Z}, a_1, a_2, \dots \in \mathbb{N})(x = [a_0; a_1, a_2, \dots])$$

Proof: Since we already proved the uniqueness of the representation, it remains to verify its existence for irrational numbers.

Take any irrational number $x = x_0 \in \mathbb{R}$. We start the construction of the integers a_0, a_1, \dots so that $x = [a_0; a_1, a_2, \dots]$ in the following way.

Set $a_0 = \lfloor x_0 \rfloor$. Since x_0 is irrational, it follows that $x_1 := \frac{1}{x_0 - a_0}$ is a well-defined positive irrational number with $x_1 > 1$. Hence, $a_1 = \lfloor x_1 \rfloor \in \mathbb{N}$.

Suppose now we already constructed x_0, \dots, x_n and a_0, a_1, \dots, a_n possessing following properties: The x_j s are irrational, $a_0 \in \mathbb{Z}$ and $a_j \in \mathbb{N}$, $1 \leq j \leq n$, and with $a_j = \lfloor x_j \rfloor$.

In the next step we set

$$x_{n+1} = \frac{1}{x_n - a_n} \quad \text{and} \quad a_{n+1} = \lfloor x_{n+1} \rfloor.$$

Since x_n is irrational, so is x_{n+1} , and because of $0 < x_n - a_n < 1$ it follows that $x_{n+1} > 1$, which implies $a_{n+1} \in \mathbb{N}$.

We claim now that

$$(5.7.8) \quad \lim_{n \rightarrow \infty} [a_0; a_1, \dots, a_n] = x.$$

By the construction we have $x_j = a_j + 1/x_{j+1}$, hence using the notation introduced in Lemma 5.7.2 we conclude that

$$(5.7.9) \quad x = x_0 = a_0 + \frac{1}{x_1} = [a_0; x_1] = [a_0; a_1, x_2] = \dots = [a_0; a_1, \dots, a_{n-1}, x_n].$$

This representation of x makes (5.7.8) very likely, but does not prove it.

To finish the proof we have to show that the convergents $r_n = b_n/c_n$ approach the irrational number x . An application of Lemma 5.7.2 and (5.7.9) leads to

$$\begin{aligned} x - r_{n-1} &= [a_0; a_1, \dots, a_{n-1}, x_n] - \frac{b_{n-1}}{c_{n-1}} = \frac{b_{n-1}x_n + b_{n-2}}{c_{n-1}x_n + c_{n-2}} - \frac{b_{n-1}}{c_{n-1}} \\ &= \frac{c_{n-1}b_{n-1}x_n + c_{n-1}b_{n-2} - c_{n-1}b_{n-1}x_n - b_{n-1}c_{n-2}}{c_{n-1}(c_{n-1}x_n + c_{n-2})} \\ &= \frac{c_{n-1}b_{n-2} - b_{n-1}c_{n-2}}{c_{n-1}(c_{n-1}x_n + c_{n-2})} = \frac{-(-1)^{n-2}}{c_{n-1}(c_{n-1}x_n + c_{n-2})} \end{aligned}$$

where in the last step we used (5.7.3) with n replaced by $n - 1$. So we arrive at

$$(5.7.10) \quad |x - r_{n-1}| = \frac{1}{c_{n-1}(c_{n-1}x_n + c_{n-2})}.$$

Next observe that $x_n > 0$, hence

$$c_{n-1}(c_{n-1}x_n + c_{n-2}) > c_{n-1}c_{n-2} \xrightarrow{n \rightarrow \infty} \infty,$$

which implies

$$\lim_{n \rightarrow \infty} \frac{1}{c_{n-1}(c_{n-1}x_n + c_{n-2})} = 0.$$

By (5.7.10) we finally arrive at

$$\lim_{n \rightarrow \infty} [x - r_{n-1}] = 0,$$

proving $x = [a_0; a_1, \dots]$ as asserted. ■

The estimates in the proof of Theorem 5.7.4 imply the following bounds for the distance of an irrational number to its convergents.

Corollary 5.7.4. Suppose that $x = [a_0; a_1, a_2, \dots]$. If

$$[a_0; a_1, \dots, a_n] = \frac{b_n}{c_n}$$

for some integers b_n, c_n , then

$$(5.7.11) \quad \left| x - \frac{b_n}{c_n} \right| \leq \frac{1}{c_n c_{n-1}}.$$

Remark 5.7.1. The proof of Theorem 5.7.3 is constructive. Indeed, let $x = x_0$ be an arbitrary irrational number. In a first step one sets $a_0 = \lfloor x_0 \rfloor$, then

$$x_1 = \frac{1}{x_0 - a_0} \quad \text{and} \quad a_1 = \lfloor x_1 \rfloor.$$

If a_0, \dots, a_n are already defined in this way, in a next step one chooses

$$x_{n+1} = \frac{1}{x_n - a_n} \quad \text{and} \quad a_{n+1} = \lfloor x_{n+1} \rfloor.$$

And as shown in Theorem 5.7.3, by this construction

$$x = \lim_{n \rightarrow \infty} [a_0; a_1, \dots, a_n] = a_0 + \cfrac{1}{a_1 + \cfrac{1}{\ddots + \cfrac{1}{a_n + \dots}}}.$$

Example 5.7.4. Let us evaluate the first entries of the infinite continued fraction of $\sqrt{3}$.

$$\begin{aligned} x_0 &= \sqrt{3} \approx 1.73205 & \Rightarrow a_0 &= 1 \\ x_1 &= \frac{1}{x_0 - a_0} = \frac{1}{0.73205} = 1.36603 & \Rightarrow a_1 &= 1 \\ x_2 &= \frac{1}{y_1 - a_1} = \frac{1}{0.36603} = 2.73205 & \Rightarrow a_2 &= 2 \\ x_3 &= \frac{1}{y_2 - a_2} = \frac{1}{0.73205} = 1.3660 & \Rightarrow a_3 &= 1 \end{aligned}$$

This suggests that $\sqrt{3} = [1; 1, 2, 1, 2, \dots]$. Why is this so? Set $x := [1; 1, 2, 1, 2, \dots]$. In view of Theorem 5.7.1 this is a well-defined positive irrational number satisfying

$$x = 1 + \cfrac{1}{[1; 2, 1, 2, \dots]} = 1 + \cfrac{1}{1 + \cfrac{1}{[2; 1, 2, 1, \dots]}} = 1 + \cfrac{1}{1 + \cfrac{1}{1 + x}}.$$

Solving this equation for x gives

$$x = \frac{3 + 2x}{2 + x} \quad \Rightarrow \quad x^2 = 3 \quad \Rightarrow \quad x = \sqrt{3}$$

because $x > 0$. So, we see that $[1; 1, 2, 1, 2, \dots]$ is indeed the representation of $\sqrt{3}$ as infinite continued fraction.

Example 5.7.5. Let us evaluate the first integers a_0, a_1, \dots of the representation of the Euler number e.

$$\begin{aligned} x_0 &= e \approx 2.71828 & \Rightarrow a_0 &= 2 \\ x_1 &= \frac{1}{x_0 - a_0} = \frac{1}{0.71828} = 1.39221 & \Rightarrow a_1 &= 1 \\ x_2 &= \frac{1}{x_1 - a_1} = \frac{1}{0.39221} = 2.54964 & \Rightarrow a_2 &= 2 \\ x_3 &= \frac{1}{x_2 - a_2} = \frac{1}{0.54964} = 1.81935 & \Rightarrow a_3 &= 1 \\ x_4 &= \frac{1}{x_3 - a_3} = \frac{1}{0.81935} = 1.22048 & \Rightarrow a_4 &= 1 \\ x_5 &= \frac{1}{x_4 - a_4} = \frac{1}{0.22048} = 4.53557 & \Rightarrow a_5 &= 4 \end{aligned}$$

The following representation of e was already known to Euler (for a new approach see [5]).

$$e = [2; 1, 2, 1, 1, 4, 1, 1, 6, 1, 1, 8, 1, 1, 10, 1, 1, 12, 1, 1, 14, 1, 1, 16, 1, 1, 18, 1, 1, 20, 1, 1, 22, 1, 1, \dots, 1, 1, 2n, 1, 1, 2n + 2, 1, 1, 2n + 4, \dots].$$

There is a nice and surprising pattern in the representation of the Euler number e.

Example 5.7.6. Another example of interest is the number π . Here it is known (cf. sequence A001203 in the OEIS) that

$$\begin{aligned} \pi = [3; 7, 15, 1, 292, 1, 1, 1, 2, 1, 3, 1, 14, 2, 1, 1, 2, 2, 2, 1, 84, 2, 1, 1, 15, 3, 13, 1, 4, 2, 6, 6, 99, 1, 2, 2, 6, 3, 5, 1, 1, 6, 8, 1, 7, 1, 2, 3, 7, 1, 2, 1, 1, 12, 1, 1, 1, 3, 1, 1, 8, 1, 1, 2, 1, 6, 1, 1, 5, 2, 2, 3, 1, 2, 4, 4, 16, 1, 161, 45, 1, 22, 1, 2, 2, 1, 4, 1, 2, 24, 1, 2, 1, 3, 1, 2, 1, \dots]. \end{aligned}$$

No pattern is known for the representation of π .

Epilogue:

There are some more interesting properties of infinite continued fractions. Let us mention the most important ones. For further reading we refer to [25].

(1) Given an irrational number $x = [a_0; a_1, \dots]$, then its convergents

$$[a_0; a_1, \dots, a_n] = \frac{b_n}{c_n} \quad \text{with} \quad \gcd(b_n, c_n) = 1,$$

approximate the number x . Note that these convergents are rational numbers. The first result tells us that this approximation becomes better and better when n increases. That is, for all $n \geq 1$ it follows that

$$\left| x - \frac{b_n}{c_n} \right| < \left| x - \frac{b_{n-1}}{c_{n-1}} \right|.$$

(2) The approximation of the irrational x by its convergents is in the following sense the best possible by rational numbers. If for some integers p and $q > 0$ we have

$$\left| x - \frac{p}{q} \right| < \left| x - \frac{b_n}{c_n} \right|,$$

then necessarily $q > c_n$. Thus, if we bound the size of the denominator, then the approximation of x by its convergents is the optimal one.

For example, the first entries of the expansion of π are $[3; 7, 15, 1, \dots]$. Thus, a possible approximation is the well-known fraction

$$[3; 7] = 3 + \frac{1}{7} = \frac{22}{7} \approx 3.1425.$$

A more precise approximation of π is given by its second convergent

$$[3; 7, 15] = \frac{333}{106} = 3.14150943.$$

Here the error is about 8.32×10^{-5} . Moreover, this is the best approximation of π by a rational number p/q with positive denominator $q \leq 106$.

(3) An irrational number x is said to possess a **periodic representation** as infinite continued fraction if $x = [a_0; a_1, a_2, \dots]$ with $a_{k+n} = a_k$ for some $n \geq 1$ and k sufficiently large.

Theorem 5.7.5. *A real number x has a periodic infinite continued fraction if and only if x is a quadratic irrational number. That is, x is irrational and may be written as*

$$x = \frac{a \pm \sqrt{b}}{c}$$

for integers $a, b > 0$ and $c \neq 0$.

Exercise 5.7.1. Argue why for all infinite continued fractions one has

$$[a_0; a_1, a_2, \dots] = a_0 + \frac{1}{[a_1; a_2, a_3, \dots]}.$$

Exercise 5.7.2. Give a direct proof for $\sqrt{2} = [1; 2, 2, 2, \dots]$.

Exercise 5.7.3. Show that $[2; 3, 2, 3, 2, 3, \dots] = \frac{3+\sqrt{15}}{3}$.

Exercise 5.7.4. Prove that $[4; 1, 2, 3, 2, 3, 2, \dots] = \frac{29+\sqrt{15}}{7}$.

Exercise 5.7.5. Suppose an irrational number $x > 1$ possesses the representation $x = [a_0; a_1, a_2, \dots]$. Show that

$$\frac{1}{x} = [0; a_0, a_1, \dots].$$

Give the expansions of $1/\sqrt{2}$ and of $1/\sqrt{3}$.

Exercise 5.7.6. Expand $\sqrt{5}$ as infinite continued fraction. Verify your result.

Exercise 5.7.7. Which real numbers are represented as

- (a) $[2; 1, 1, 1, 1, \dots]$, (b) $[2; 3, 1, 1, 1, 1, \dots]$,
- (c) $[2; 2, 2, 2, 2, \dots]$, (d) $[1; 2, 1, 2, 1, 2, \dots]$,
- (e) $[2; 1, 2, 1, 2, \dots]$, (f) $[1; 3, 1, 2, 1, 2, \dots]$?

Exercise 5.7.8. Use (5.7.11) to estimate how precise the convergents

$$[2; 1], [2; 1, 2], [2; 1, 2, 1], \text{ and } [2; 1, 2, 1, 1]$$

approximate the Euler number e?

Exercise 5.7.9. As we saw, the second convergent r_2 of π equals

$$[3; 7, 15] = \frac{333}{106} \approx 3.141509.$$

Show in the same way that the third and the fourth convergent of π are

$$\frac{355}{113} \text{ and } \frac{103993}{33102}.$$

Use a calculator to estimate the error when replacing π by these numbers.

Exercise 5.7.10. Say a real number x has the expansion

$$x = [0; a, b, a, b, a, b, a, \dots] = [0; \overline{a, b}]$$

for certain natural numbers a and b . Express x in dependence on a and b .

5.8. More Exercises

Exercise 5.8.1. Start with an equilateral triangle F_1 whose side has length 1. Divide each side into three segments of equal length $1/3$ and draw an equilateral triangle on the middle segment of each side as a base towards the exterior of F_1 . Delete the three middle segments and the resulting figure F_2 is a polygon with twelve sides each of length $1/3$. Divide each side of F_2 into three equal segments of length $1/9$ and for each side, build an equilateral triangle with the middle segment as a base in the exterior of F_2 . Deleting the twelve middle segments, we obtain a polygon F_3 with 48 sides, each of length $1/9$ (see Figure 5.8.1 for a picture of F_1, F_2 and F_3). Continue this process and denote by F_n the polygon obtained after $n - 1$ steps. For each $n \geq 1$, determine with proof the following

- (1) the number of sides of F_n ,
- (2) the length of each side of F_n ,
- (3) the perimeter of F_n ,
- (4) the area of F_n .

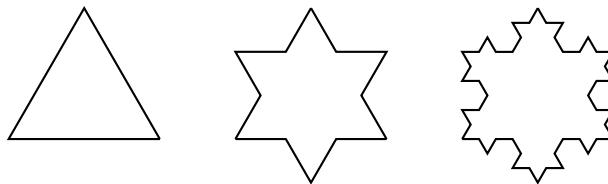


Figure 5.8.1. The polygons F_1, F_2, F_3 .

The limit of F_n as n goes to infinity is called the **Koch snowflake**¹⁹ and is one of the first examples of fractals.

¹⁹Helge von Koch (1870–1924) was a Swedish mathematician.

Exercise 5.8.2. Let $(x_n)_{n \geq 1}$ be a sequence of real numbers. Which of the following four properties describes the fact that the sequence does **not** converge to a given $L \in \mathbb{R}$?

- (1) For each $\varepsilon > 0$ there is $N \in \mathbb{N}$ such that for all $n \geq N$

$$|x_n - L| > \varepsilon.$$

- (2) There is $\varepsilon > 0$ such that for each $N \geq 1$ there is $n \geq N$ such that

$$|x_n - L| > \varepsilon.$$

- (3) There is $\varepsilon > 0$ such that for all $n \in \mathbb{N}$

$$|x_n - L| > \varepsilon.$$

- (4) There is $\varepsilon > 0$ such that for each $N \geq 1$ there is $n \geq N$ such that

$$|x_n - L| \geq \varepsilon.$$

Exercise 5.8.3. Let $(x_n)_{n \geq 1}$ be a sequence of positive numbers with $x_n \xrightarrow{n \rightarrow \infty} 0$. Show that then $\max_{n \geq 1} x_n$ exists. In other words, there is some $n_0 \in \mathbb{N}$ with

$$x_n \leq x_{n_0}, \quad n = 1, 2, \dots$$

Exercise 5.8.4. Prove that

$$\lim_{n \rightarrow \infty} [(n+1)^\alpha - n^\alpha] = \begin{cases} 0 & : 0 < \alpha < 1 \\ 1 & : \alpha = 1 \\ \infty & : 1 < \alpha < \infty. \end{cases}$$

Exercise 5.8.5. Evaluate

$$\lim_{n \rightarrow \infty} (\sqrt{n+2} - \sqrt{n}) \quad \text{and} \quad \lim_{n \rightarrow \infty} n^2 \cdot \left(\frac{1}{n} - \frac{1}{n+1} \right).$$

Exercise 5.8.6. Let $b > 1$ and $\alpha > 0$ be real numbers. Prove that $\lim_{n \rightarrow \infty} \frac{n^\alpha}{b^n} = 0$.

Exercise 5.8.7. Let $(x_n)_{n \geq 1}$ be a convergent sequence of real numbers. Show that for any given natural number k , the sequence $(x_{n+k} - x_n)_{n \geq 1}$ converges to zero.

Exercise 5.8.8. Is the converse of the previous exercise true? That is, does $(x_n)_{n \geq 1}$ converge provided that the sequence $(x_{n+k} - x_n)_{n \geq 1}$ converges to zero for some $k \in \mathbb{N}$?

Exercise 5.8.9. Let $(x_n)_{n \geq 1}$ be a convergent sequence of nonzero real numbers. Show that for any given natural number k , the sequence $(x_{n+k}/x_n)_{n \geq 1}$ converges to one.

Exercise 5.8.10. Is the converse of the previous exercise true? In other words, does $(x_n)_{n \geq 1}$ converge provided that the sequence $(x_{n+k}/x_n)_{n \geq 1}$ converges to one for some $k \in \mathbb{N}$?

Exercise 5.8.11. Let $(x_n)_{n \geq 1}$ be a convergent sequence of real numbers. Define

$$c_n = \frac{x_1 + \dots + x_n}{n}, \quad n \in \mathbb{N}.$$

Show that the sequence $(c_n)_{n \geq 1}$ is convergent and has the same limit as the sequence $(x_n)_{n \geq 1}$.

Give an example where the c_n 's converge²⁰ for diverging x_n 's.

²⁰The limit of the c_n 's, whenever it exists, is called Cesáro limit of the x_n 's. Named after the Italian mathematician Ernesto Cesáro (1859–1906). To investigate the Cesáro limit is a useful and quite often applied tool to assign (generalized) limits to diverging sequences.

Exercise 5.8.12. Suppose that a sequence $(x_n)_{n \geq 1}$ of positive numbers converges to some $L \in \mathbb{R}$. Show that this implies

$$\lim_{n \rightarrow \infty} \sqrt{x_n} = \sqrt{L}.$$

Hint: Investigate the cases $L > 0$ and $L = 0$ separately.

Exercise 5.8.13. Show that

$$\lim_{n \rightarrow \infty} \left(1 + \frac{1}{2n}\right)^n = e^{1/2}.$$

Exercise 5.8.14. Let $(x_n)_{n \geq 1}$ and $(y_n)_{n \geq 1}$ be two sequences of real numbers with

$$x_n \xrightarrow[n \rightarrow \infty]{} L \quad \text{and} \quad y_n \xrightarrow[n \rightarrow \infty]{} M$$

for some real numbers L and M . Show that then

$$\min\{x_n, y_n\} \xrightarrow[n \rightarrow \infty]{} \min\{L, M\} \quad \text{and} \quad \max\{x_n, y_n\} \xrightarrow[n \rightarrow \infty]{} \max\{L, M\}.$$

Hint: Use Exercise 4.3.11.

Exercise 5.8.15. Why does

$$x_n \xrightarrow[n \rightarrow \infty]{} L$$

imply

$$|x_n| \xrightarrow[n \rightarrow \infty]{} |L|?$$

Exercise 5.8.16. Verify the following: For each $x \in \mathbb{R}$ there exist a decreasing sequence $(q_n)_{n \geq 1}$, $q_n \in \mathbb{Q}$, as well as an increasing sequence $(r_n)_{n \geq 1}$, $r_n \in \mathbb{Q}$, such that

$$\lim_{n \rightarrow \infty} q_n = \lim_{n \rightarrow \infty} r_n = x.$$

Exercise 5.8.17. In view of Theorem A.4.8 is the set of rational numbers countably infinite, thus may be written as

$$\mathbb{Q} = \{q_1, q_2, \dots\}.$$

Describe all cluster points of the sequence $(q_j)_{j \geq 1}$. Recall that cluster points of a sequence were introduced in Definition 5.4.2.

Exercise 5.8.18. A subset $G \subseteq \mathbb{R}$ is said to be **open** provided the following property is satisfied:

$$(\forall x \in G)(\exists \varepsilon > 0)(U_\varepsilon(x) \subseteq G).$$

Recall that

$$U_\varepsilon(x) = (x - \varepsilon, x + \varepsilon) = \{y \in \mathbb{R} : |x - y| < \varepsilon\}$$

denotes the (open) ε -neighborhood of the element $x \in \mathbb{R}$. By definition the empty set $\emptyset \subseteq \mathbb{R}$ is supposed to be open.

Prove the following properties of open sets:

- (1) Any open interval (bounded or unbounded) is an open subset of \mathbb{R} .
- (2) The union of arbitrarily many open sets is open as well. Similarly, the intersection of **finitely** many open sets is also open. Give an example which shows that this becomes false for infinitely many open sets.

Exercise 5.8.19. A subset $A \subseteq \mathbb{R}$ is said to be **closed** if it satisfies the following property:

$$(\forall x_n \in A) (\lim_{n \rightarrow \infty} x_n = L) \Rightarrow (L \in A)$$

In other words, the set A is closed whenever it is closed by taking limits. Again we suppose that $\emptyset \subseteq \mathbb{R}$ is closed.

Prove the following properties of closed sets:

- (1) Any closed interval, bounded or unbounded, is closed.
- (2) The union of **finitely** many closed sets is closed. Moreover, the intersection of arbitrarily many closed sets is closed. Give an example of infinitely many closed sets for which their union is not closed.
- (3) (\star) Prove the following result:

Proposition 5.8.1. *A subset $G \subseteq \mathbb{R}$ is open if and only if its complement G^c is closed.*

Complex Numbers \mathbb{C}

The beauty of mathematics only shows itself to more patient followers.

Maryam Mirzakhani¹

6.1. Basic Properties

A common belief is that complex numbers arose from the need to solve quadratic equations as, e.g., $x^2 + 1 = 0$. But this is not so; equations of this type were accepted for centuries to be unsolvable. In fact, the investigation of cubic equations demanded extracting roots of negative numbers. Around 1545, the Italian mathematician² Gerolamo Cardano (1501–1576) found a formula for the solution of cubic equations of the type $x^3 = ax + b$ for given $a, b \in \mathbb{R}_+$. This formula (now known as Cardano's formula) is

$$x = \sqrt[3]{\frac{b}{2} + \sqrt{\left(\frac{b}{2}\right)^2 - \left(\frac{a}{3}\right)^3}} + \sqrt[3]{\frac{b}{2} - \sqrt{\left(\frac{b}{2}\right)^2 - \left(\frac{a}{3}\right)^3}}$$

for one solution x . For example, the equation $x^3 = 9x + 2$ has three real solutions³, but the above formula leads to

$$(6.1.1) \quad x = \sqrt[3]{1 + \sqrt{-26}} + \sqrt[3]{1 - \sqrt{-26}}.$$

Or $x^3 = 15x + 4$ has the three solutions $x_1 = 4$, $x_{2/3} = -2 \pm \sqrt{3}$. In this case Cardano's formula gives

$$(6.1.2) \quad x = \sqrt[3]{2 + \sqrt{-121}} + \sqrt[3]{2 - \sqrt{-121}}.$$

¹Maryam Mirzakhani (1977–2017) was an Iranian mathematician who in 2014 became the first woman to win the Fields Medal.

²He was also a physician, biologist, physicist, chemist, astrologer, astronomer, philosopher, writer, and gambler.

³Using Newton's Method it follows that the three real solutions of $x^3 = 9x + 2$ are approximately $x_1 = -2.882021$, $x_2 = -0.223462$ and $x_3 = 3.105483$.

But what to do with such formulas as stated in (6.1.1) and (6.1.2), respectively? In 1572, another Italian mathematician Rafael Bombelli (1526–1572) found a way how to deal with those expressions. He conjectured that there exist real numbers α and β for which

$$\sqrt[3]{2 + \sqrt{-121}} = \alpha + \beta\sqrt{-1} \quad \text{and} \quad \sqrt[3]{2 - \sqrt{-121}} = \alpha - \beta\sqrt{-1}.$$

And applying the rules for addition and multiplication of real numbers he finally deduced $\alpha = 2$ and $\beta = 1$, that is

$$\sqrt[3]{2 + \sqrt{-121}} = 2 + \sqrt{-1} \quad \text{while} \quad \sqrt[3]{2 - \sqrt{-121}} = 2 - \sqrt{-1}.$$

Hence, $x = (2 + \sqrt{-1}) + (2 - \sqrt{-1}) = 4$, which solves $x^3 = 15x + 4$. Thus, he was the first who gave a meaning to something meaningless.

For the next two and a half centuries Bombelli complex numbers were extensively and very successfully used. For example, we should mention the theoretical work of the French mathematician René Descartes (1596–1650) who coined the term *imaginary number*. Before these numbers were called *sophisticated* or *subtle*. Other contributions are due to another French mathematician Abraham de Moivre (1667–1754) who evaluated $(\cos \theta + i \sin \theta)^n$. Important work was done by the Swiss mathematician Leonard Euler (1707–1783) who introduced the letter i for $\sqrt{-1}$ and who found a formula which relates the exponential function with the trigonometric ones.

Nevertheless, complex numbers remained mysterious. For example, Leonhard Euler stated the following:⁴

Because all conceivable numbers are either greater than zero, less than zero or equal to zero, then it is clear that the square roots of negative numbers cannot be included among the possible numbers. Consequently, we must say that these are impossible numbers. And this circumstance leads us to the concept of such numbers, which by their nature are impossible, and ordinarily are called imaginary or fancied numbers, because they exist only in the imagination.

In 1797, the German mathematician Carl Friedrich Gauss (1777–1855) used complex numbers for his proof of the fundamental theorem of algebra. Even then they remained mysterious for him. So he wrote in 1825 that *the true metaphysics of $\sqrt{-1}$ is elusive*. Things changed, when Gauss published in 1831 his scheme for the geometric representation of complex numbers as points in the plane⁵. After that mathematicians used complex numbers very successfully, and nowadays mathematics or physics without complex numbers is inconceivable.

Definition 6.1.1. The **complex numbers** (or **complex plane**) \mathbb{C} consists of ordered pairs (a, b) with $a, b \in \mathbb{R}$, equipped with addition and multiplication defined as follows:

$$(a_1, b_1) + (a_2, b_2) = (a_1 + a_2, b_1 + b_2), \quad (a_1, b_1), (a_2, b_2) \in \mathbb{C},$$

$$(a_1, b_1) \cdot (a_2, b_2) = (a_1 a_2 - b_1 b_2, a_1 b_2 + a_2 b_1), \quad (a_1, b_1), (a_2, b_2) \in \mathbb{C}.$$

⁴See [19], p. 594.

⁵Similar representations by the Norwegian mathematician Caspar Wessel (1745–1818) in 1797 or by the French mathematician Jean-Robert Argand (1768–1822) in 1806, had gone largely unnoticed.

Notation: Let $i = (0, 1) \in \mathbb{C}$ and, given $a \in \mathbb{R}$, write $a \in \mathbb{C}$ instead of $(a, 0) \in \mathbb{C}$. Then every complex number $z = (a, b)$ can be represented as

$$(a, b) = (a, 0) + (0, b) = (a, 0) + (b, 0)(0, 1) = a + bi = a + ib, \quad (a, b) \in \mathbb{C}.$$

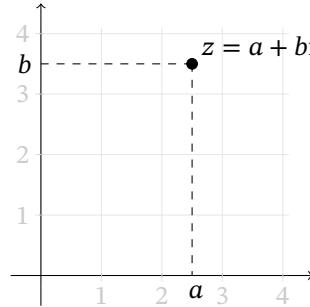


Figure 6.1.1. Representation of $z = a + bi$ in the plane.

In other words, we consider \mathbb{R} as subset of \mathbb{C} by identifying $a \in \mathbb{R}$ with $(a, 0) \in \mathbb{C}$, i.e., $\mathbb{R} \cong \{(a, 0) : a \in \mathbb{R}\} \subset \mathbb{C}$. Thereby addition and multiplication in \mathbb{R} coincide with the restriction of these binary operations from \mathbb{C} to \mathbb{R} . That is,

$$(a_1, 0) + (a_2, 0) = (a_1 + a_2, 0) \quad \text{and} \quad (a_1, 0) \cdot (a_2, 0) = (a_1 \cdot a_2, 0).$$

$$\mathbb{C} = \{a + bi : a, b \in \mathbb{R}\} \cong \{(a, b) : a, b \in \mathbb{R}\}$$

Definition 6.1.2. Let P be a point in the Cartesian plane of coordinates (a, b) . We call $a + bi$ the complex number associated to P , and we denote it by z_P .

Closely related to complex numbers is the notion of a vector. Informally, a vector is a directed segment or arrow, consisting of a segment in the Euclidean plane that is directed from one endpoint to the other. Vectors can be used to represent physical concepts such as force or velocity that consist of a certain length and a direction. The word *vector* translates as *carrier* from Latin.

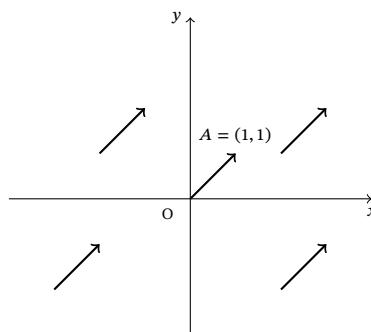


Figure 6.1.2. Representations of the same vector.

Definition 6.1.3. Let A, B, A', B' be four points in the Euclidean plane. We say that the ordered pair (A, B) is equivalent to the ordered pair (A', B') provided that $ABB'A'$ is a parallelogram.

One can show that this relation is an equivalence relation (see Exercise 6.1.6). The equivalence classes of this relation are what we call vectors. Given two points A and B in the Euclidean plane, we can think of the vector \vec{AB} as the segment AB directed or oriented from the starting point A to the terminal point B . Figure 6.1.2 gives the representation of the same vector with different starting points. For every vector \vec{AB} , there is a unique vector \vec{OD} such that $\vec{AB} = \vec{OD}$, where O is the origin.

Example 6.1.1. If O represents the origin, A is the point $(1, 1) = 1 + i$, B is $(-2, 1) = -2 + i$, and $C = (-1, 2) = -1 + 2i$, then $\vec{OA} = \vec{BC}$.

Definition 6.1.4. The length or magnitude of the vector \vec{AB} is the length of the segment AB , and we will denote it by $|AB|$ or $|\vec{AB}|$.

Example 6.1.2. If A is as in the previous example, then the length of the vector \vec{OA} is $\sqrt{2}$.

Definition 6.1.5. When $A = B$, the vector \vec{AB} is called the zero vector and is denoted by \vec{O} .

Algebraically, the addition and multiplication of complex numbers can be written as follows:

$$(6.1.3) \quad z_1 + z_2 = (a_1 + b_1i) + (a_2 + b_2i) = a_1 + a_2 + (b_1 + b_2)i,$$

$$(6.1.4) \quad z_1 \cdot z_2 = (a_1 + b_1i) \cdot (a_2 + b_2i) = a_1a_2 - b_1b_2 + (a_1b_2 + a_2b_1)i.$$

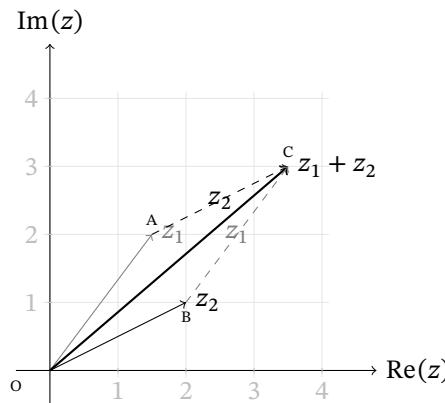


Figure 6.1.3. The sum of the complex numbers z_1 and z_2 .

Geometrically, the addition of complex numbers or of their corresponding vectors is done via the **parallelogram rule** (see Figure 6.1.3). By a remark above, for any vector \vec{AB} in the plane, there is a vector starting at the origin that equals \vec{AB} .

Definition 6.1.6. The sum of two vectors \overrightarrow{OA} and \overrightarrow{OB} is defined as the vector \overrightarrow{OC} , where C is the unique point such that $OACB$ is a parallelogram.

In other words, \overrightarrow{OC} is the diagonal of the parallelogram containing O, A , and B and having AB as the other diagonal.

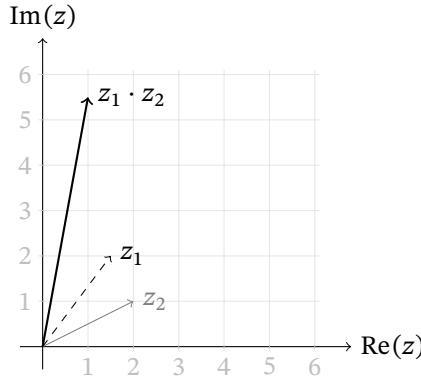


Figure 6.1.4. The multiplication of the complex numbers z_1 and z_2 .

Remark 6.1.1. Since

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

we see that the multiplication of complex numbers follows the well-known rules for the multiplication of bilinear forms of real numbers

$$(a_1 + b_1i) \cdot (a_2 + b_2i) = a_1a_2 + a_1b_2i + a_2b_1i + b_1b_2i^2.$$

Definition 6.1.7. Given a complex number $z = a + bi$, the number $a \in \mathbb{R}$ is called the **real part** of z while b is its **imaginary part**:

$$a = \operatorname{Re}(z) \quad \text{and} \quad b = \operatorname{Im}(z), \quad z = a + bi \in \mathbb{C}.$$

A complex number z is said to be **real** if $\operatorname{Im}(z) = 0$ and (purely) **imaginary** provided that $\operatorname{Re}(z) = 0$.

Note that $z_1 = z_2$ if and only if $\operatorname{Re}(z_1) = \operatorname{Re}(z_2)$ as well as $\operatorname{Im}(z_1) = \operatorname{Im}(z_2)$.

Theorem 6.1.1. The set \mathbb{C} of complex numbers with addition and multiplication defined in (6.1.3) and (6.1.4) is a field⁶, denoted by $(\mathbb{C}, +, \cdot)$.

Proof: (1) In a first step we observe that $(\mathbb{C}, +)$ is the Cartesian product of $(\mathbb{R}, +)$ with itself. Since $(\mathbb{R}, +)$ is a commutative group, this property is also valid for their direct product, hence for $(\mathbb{C}, +)$ (compare Exercise A.8.14). The unit element is $0 = 0 + 0i$ and the inverse of some $z \in \mathbb{C}$ equals $-z = -a - bi$ for $z = a + bi$.

(2) The product of two complex numbers is associative, and we leave the details as an exercise. The unit is $1 = (1, 0) = 1 + 0i$. Now take some nonzero $z = a + bi$. Set

⁶That is, the set \mathbb{C} endowed with the binary operations $+$ and \cdot fulfills the properties (I a), (I b), and (I c) as stated at the beginning of Section 4.2.

$w := \frac{a-bi}{a^2+b^2}$. We claim that w is inverse to z . First note that w is well-defined because of $z \neq 0$, hence $a^2 + b^2 > 0$. Furthermore,

$$\begin{aligned} z \cdot w &= (a+bi) \left(\frac{a}{a^2+b^2} - \frac{b}{a^2+b^2}i \right) \\ &= \frac{a^2}{a^2+b^2} + \frac{b^2}{a^2+b^2} + i \left(\frac{-ab}{a^2+b^2} + \frac{ab}{a^2+b^2} \right) = 1. \end{aligned}$$

Consequently, if $z = a+bi \neq 0$, then

$$(6.1.5) \quad z^{-1} = \frac{a-bi}{a^2+b^2} = \frac{a}{a^2+b^2} - \frac{bi}{a^2+b^2}.$$

Hence, $(\mathbb{C} \setminus \{0\}, \cdot)$ is an abelian group.

(3) It remains to prove the distributive law. Choose complex numbers z_1, z_2 , and z_3 with $z_j = a_j + b_ji$, $j = 1, 2, 3$. Then

$$\begin{aligned} (z_1 + z_2)z_3 &= [(a_1 + a_2 + i(b_1 + b_2)) [a_3 + b_3i]] \\ &= (a_1 + a_2)a_3 - (b_1 + b_2)b_3 + i((a_1 + a_2)b_3 + (b_1 + b_2)a_3). \end{aligned}$$

On the other hand, by the distributive law for addition and multiplication of real numbers it follows that

$$\begin{aligned} z_1z_3 + z_2z_3 &= a_1a_3 - b_1b_3 + i(a_1b_3 + b_1a_3) + a_2a_3 - b_2b_3 + i(a_2b_3 + b_2a_3) \\ &= a_1a_3 + a_2a_3 - b_1b_3 - b_2b_3 + i(a_1b_3 + b_1a_3 + a_2b_3 + b_2a_3) \\ &= (a_1 + a_2)a_3 - (b_1 + b_2)b_3 + i((a_1 + a_2)b_3 + (b_1 + b_2)a_3). \end{aligned}$$

Thus, we have proved that

$$(z_1 + z_2)z_3 = z_1z_3 + z_2z_3 \quad \text{for all } z_1, z_2, z_3 \in \mathbb{C},$$

that is, the distributive law is valid, and $(\mathbb{C}, +, \cdot)$ is a field. ■

Basic properties of addition and multiplication on $\mathbb{C} = \{a+bi : a, b \in \mathbb{R}\}$ are

- (1) $(a_1 + b_1i) + (a_2 + b_2i) = (a_1 + a_2) + (b_1 + b_2)i$,
- (2) $(a_1 + b_1i) \cdot (a_2 + b_2i) = (a_1a_2 - b_1b_2) + (a_1b_2 + a_2b_1)i$,
- (3) $\frac{1}{a+bi} = \frac{a}{a^2+b^2} - \frac{bi}{a^2+b^2}$, $a+bi \neq 0$,
- (4) $\frac{a_1 + b_1i}{a_2 + b_2i} = (a_1 + b_1i) \cdot \left(\frac{1}{a_2 + b_2i} \right)$, $a_2 + b_2i \neq 0$.

Example 6.1.3. The reader should verify the results below.

$$(1) \quad (1 + 3i) + (2 - i) = 3 + 2i \text{ and } (1 + 3i) \cdot (2 - i) = (2 + 3) + (-1 + 6)i = 5 + 5i.$$

$$(2) \quad \frac{1}{2-i} = \frac{2+i}{5} = \frac{2}{5} + \frac{i}{5}.$$

$$(3) \quad \frac{1+3i}{2-i} = (1+3i) \cdot \left(\frac{2}{5} + \frac{i}{5}\right) = -\frac{1}{5} + \frac{7}{5}i.$$

$$(4) \quad i^2 = -1, \quad i^3 = -i, \quad i^4 = 1, \quad \frac{1}{i} = \frac{i}{i^2} = \frac{i}{-1} = -i.$$

$$(5) \quad (1+i)^4 = 1 + \binom{4}{1}i + \binom{4}{2}i^2 + \binom{4}{3}i^3 + i^4 = 1 + 4i - 6 - 4i + 1 = -4.$$

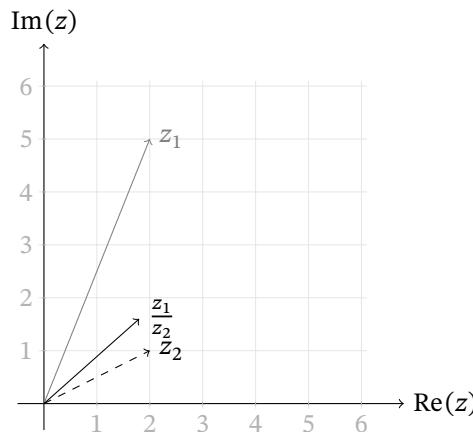


Figure 6.1.5. The quotient $\frac{z_1}{z_2} = \frac{9}{5} + \frac{8}{5}i$ of $z_1 = 2 + 5i$ and $z_2 = 2 + i$.

Remark 6.1.2. As we proved, the addition and multiplication of complex numbers satisfy properties (I a), (I b), and (I c) stated in Section 4.2. Thus, all algebraic rules of real addition and multiplication are also valid in the complex case⁷. For example, the summation formula of geometric sequences (as in \mathbb{R} we let $z^0 = 1$ for complex z)

$$(6.1.6) \quad \sum_{k=0}^n z^k = \frac{z^{n+1} - 1}{z - 1}, \quad z \neq 1.$$

as well as the binomial formula

$$(6.1.7) \quad (z_1 + z_2)^n = \sum_{k=0}^n \binom{n}{k} z_1^k z_2^{n-k}, \quad z_1, z_2 \in \mathbb{C}, \quad n \geq 1,$$

or

$$z_1^n - z_2^n = (z_1 - z_2)(z_1^{n-1} + z_1^{n-2}z_2 + \cdots + z_2^{n-1}), \quad z_1, z_2 \in \mathbb{C}, \quad n \geq 1,$$

remain valid in the case of complex numbers.

⁷This is one of the main advantages in modern mathematics: A result once proved in a general setting can later on be applied universally.

Remark 6.1.3. In contrast to the real line, the field $(\mathbb{C}, +, \cdot)$ of complex numbers **cannot** be ordered in a suitable way. One might think of

$$\mathbb{C}_+ := \{z \in \mathbb{C} : \operatorname{Re}(z) > 0 \text{ and } \operatorname{Im}(z) > 0\}$$

as a cone of positive elements. But this set does not satisfy the second and the third of the properties (4.3.1) of \mathbb{R}_+ . Since $z_1, z_2 \in \mathbb{C}_+$ implies $z_1 + z_2 \in \mathbb{C}_+$, the set \mathbb{C}_+ is closed under addition. But this is no longer valid for the multiplication. Note that $1 + i \in \mathbb{C}_+$, but $(1 + i)(1 + i) = 2i \notin \mathbb{C}_+$. Moreover, also the third condition in (4.3.1) is violated: There are nonzero elements z in \mathbb{C} which neither belong to \mathbb{C}_+ nor do we have $-z \in \mathbb{C}_+$. Take $z = 1 - i$ as such an example.

Exercise 6.1.1. Evaluate

$$\begin{array}{lll} (a) & (3 + 2i) - (8 - 5i) & (b) \quad (4 - 2i)(1 - 5i) \quad (c) \quad \frac{-2 - 4i}{-1} \\ (d) & \frac{-3 + 3i}{3 - 6i} & (e) \quad (2 + i)^3 \quad (f) \quad \frac{1}{(1 + i)^5} \\ (g) & (1 - i)^4 & (h) \quad (1 - i)^{-4}. \end{array}$$

Exercise 6.1.2. Solve the following equations for unknowns $x, y \in \mathbb{R}$.

$$\begin{array}{ll} (i) & x^2 + ix + 1 = 0. \\ (ii) & x^4 + x^2 - 1 = 0. \\ (iii) & x^2 + 2ix - 1 = 0. \\ (iv) & ix - (1 + i)y = 3 \quad \text{and} \quad (2 + i)x + iy = 4.. \end{array}$$

Exercise 6.1.3. Find all $z \in \mathbb{C}$ for which $z^2 = -1 + 2\sqrt{6}i$.

Exercise 6.1.4. Characterize complex numbers $z \in \mathbb{C}$ satisfying $z^2 \in \mathbb{R}$. For which $z \in \mathbb{C}$ do we have $\operatorname{Re}(z^2) = 0$?

Exercise 6.1.5. Draw a rough sketch of the following sets of complex numbers:

$$\begin{aligned} & \{z \in \mathbb{C} : \operatorname{Re}(z) \geq 2\operatorname{Im}(z)\}, \quad \{z \in \mathbb{C} : \operatorname{Re}(z)^2 + \operatorname{Im}(z)^2 \leq 4\}, \\ & \{z \in \mathbb{C} : \operatorname{Re}(z) = -\operatorname{Im}(z)\}, \quad \{z \in \mathbb{C} : \operatorname{Re}(z) \cdot \operatorname{Im}(z) \leq 1\}. \end{aligned}$$

Exercise 6.1.6. Let n be a natural number. Evaluate

$$\sum_{k=0}^{n-1} i^k \quad \text{and} \quad \sum_{k=0}^{n-1} i^{-k}$$

in dependence on n modulo 4.

Exercise 6.1.7. Prove that the binary relation in Definition 6.1.3 is an equivalence relation.

Exercise 6.1.8. Let $n \geq 2$ be a natural number. If A_1, \dots, A_n are points in the Euclidean plane, show that

$$\overrightarrow{A_1A_2} + \dots + \overrightarrow{A_{n-1}A_n} + \overrightarrow{A_nA_1} = \vec{O}.$$

Exercise 6.1.9. Let A, B , and C be three points in the Euclidean plane, not all on the same line. A median of the triangle ABC is a line passing through a vertex and the midpoint of the opposite side. Prove that the three medians of the triangle ABC are concurrent, meaning they intersect in one point.

Exercise 6.1.10. Let A, B , and C be three points in the Euclidean plane, not all on the same line. Denote by G the intersection point of the medians of the triangle ABC . Show that

$$\overrightarrow{GA} + \overrightarrow{GB} + \overrightarrow{GC} = \vec{O}.$$

The point G is called **the centroid** of the triangle ABC .

Exercise 6.1.11. Let A, B and C be three points in the Euclidean plane, not all on the same line. Denote by G the intersection point of the medians of the triangle ABC . If z_K denotes the complex number associated to the point K , prove that

$$z_G = \frac{z_A + z_B + z_C}{3}.$$

6.2. The Conjugate and the Absolute Value

We start with the definitions of the main concepts of this section.

Definition 6.2.1. The **absolute value** or the **modulus** of a complex number $z = a+bi$ is defined as

$$|z| := (a^2 + b^2)^{1/2}.$$

Remark 6.2.1. The absolute value of $|z|$ of the complex number $z = a+bi$ is the length of the segment whose endpoints are the origin and the point of coordinates (a, b) (see Figure 6.2.1).

Definition 6.2.2. The **conjugate** \bar{z} of a complex number $z = a+bi$ is defined as

$$\bar{z} = a - bi.$$

Remark 6.2.2. The conjugate \bar{z} of the complex number z is obtained by reflecting the complex number z or the point of coordinates (a, b) with respect to the horizontal axis (see Figure 6.2.1).

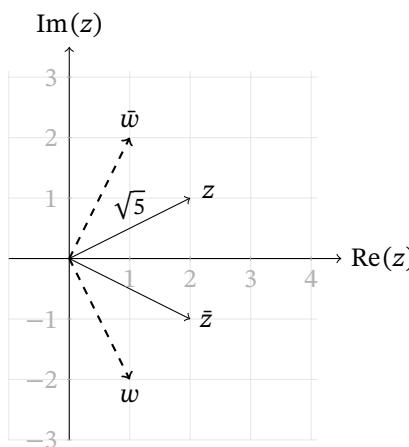


Figure 6.2.1. The complex numbers $w = 1 - 2i$ and $\bar{w} = 1 + 2i$ and $z = 2 + i$ and $\bar{z} = 2 - i$.

We now state and prove some properties of the absolute value and of the conjugate complex number.

Proposition 6.2.1.

(1) For all $z_1, z_2 \in \mathbb{C}$,

$$\overline{z_1 + z_2} = \bar{z}_1 + \bar{z}_2 \quad \text{and} \quad \overline{z_1 \cdot z_2} = \bar{z}_1 \cdot \bar{z}_2.$$

(2) For any $z \in \mathbb{C}$,

$$z + \bar{z} = 2 \operatorname{Re}(z) \quad \text{and} \quad \frac{z - \bar{z}}{i} = 2 \operatorname{Im}(z).$$

(3) If $z \in \mathbb{C}$, then

$$|\operatorname{Re}(z)| \leq |z| \quad \text{and} \quad |\operatorname{Im}(z)| \leq |z|.$$

(4) For all $z \in \mathbb{C}$

$$|z|^2 = z \cdot \bar{z} \quad \text{and if } z \text{ is nonzero, then} \quad z^{-1} = \frac{\bar{z}}{|z|^2}.$$

(5) For every $z \in \mathbb{C}$, $|z| \geq 0$ and $|z| = 0$ if and only if $z = 0$.

(6) For all $z_1, z_2 \in \mathbb{C}$

$$|z_1 \cdot z_2| = |z_1| \cdot |z_2|.$$

In particular, $|z^n| = |z|^n$ for $z \in \mathbb{C}$ and $n \geq 1$.

Proof: (1) Let $z_1 = a_1 + b_1 i$ and $z_2 = a_2 + b_2 i$. Then the first assertion easily follows by

$$\overline{z_1 + z_2} = \overline{a_1 + a_2 + (b_1 + b_2)i} = a_1 + a_2 - (b_1 + b_2)i = \bar{z}_1 + \bar{z}_2.$$

To prove the second part, we use

$$\overline{z_1 \cdot z_2} = \overline{a_1 a_2 - b_1 b_2 + i(a_1 b_2 + a_2 b_1)} = a_1 a_2 - b_1 b_2 - i(a_1 b_2 + a_2 b_1)$$

and

$$\bar{z}_1 \cdot \bar{z}_2 = (a_1 - b_1 i)(a_2 - b_2 i) = a_1 a_2 - (-b_1)(-b_2) + i(a_1(-b_2) + a_2(-b_1)),$$

which shows that $\overline{z_1 \cdot z_2} = \bar{z}_1 \cdot \bar{z}_2$.

(2) If $z = a + bi$, then

$$z + \bar{z} = (a + bi) + (a - bi) = 2a \quad \text{and} \quad z - \bar{z} = (a + bi) - (a - bi) = 2bi.$$

This completes the proof of (2).

(3) Assume $z = a + bi$. Then we get

$$|\operatorname{Re}(z)| = |a| = \sqrt{a^2} \leq \sqrt{a^2 + b^2} = |z|,$$

which proves the estimate for the real part of z . The second estimate for the imaginary part can be proved similarly.

(4) This is an easy consequence of

$$z \cdot \bar{z} = (a + bi)(a - bi) = a^2 + b^2 + i(-ab + ab) = |z|^2.$$

The second part is nothing other than a reformulation of equation (6.1.5).

(5) This follows directly from the definition of the absolute value, and we leave the details as an exercise.

(6) To prove (6) we use properties (1) and (4). Doing so, we get

$$|z_1 \cdot z_2|^2 = (z_1 z_2)(\overline{z_1 z_2}) = (z_1 z_2)(\bar{z}_1 \cdot \bar{z}_2) = (z_1 \bar{z}_1)(z_2 \bar{z}_2) = |z_1|^2 |z_2|^2.$$

Since $|z_1 \cdot z_2| \geq 0$ and $|z_1| |z_2| \geq 0$, the equality of the squares implies $|z_1 \cdot z_2| = |z_1| \cdot |z_2|$ as asserted. \blacksquare

Corollary 6.2.2.

(1) *The complex number z is real if and only if $z = \bar{z}$.*

The complex number z is purely imaginary if and only if $\bar{z} = -z$.

(2) *If $z \in \mathbb{C}$ is nonzero, then for all $n \geq 1$*

$$\overline{z^{-n}} = \bar{z}^{-n} \quad \text{and} \quad |z^{-n}| = |z|^{-n}.$$

Thus, if $z_2 \neq 0$, then $|z_1/z_2| = |z_1|/|z_2|$.

(3) *If $z_1, \dots, z_n \in \mathbb{C}$, then*

$$\overline{\sum_{k=1}^n z_k} = \sum_{k=1}^n \bar{z}_k \quad \text{and} \quad \overline{\prod_{k=1}^n z_k} = \prod_{k=1}^n \bar{z}_k.$$

Proof: (1) This is a direct consequence of property (2) in Proposition 6.2.1.

(2) Since

$$1 = \bar{1} = \overline{\bar{z}^n \cdot z^{-n}} = \overline{\bar{z}^n} \cdot \overline{z^{-n}} = \bar{z}^n \cdot \overline{z^{-n}}$$

it follows that the inverse of \bar{z}^n is $\overline{z^{-n}}$. But, $(\bar{z}^n)^{-1} = \bar{z}^{-n}$. This proves the first part. Then the second one follows by

$$|z^{-n}|^2 = z^{-n} \overline{z^{-n}} = z^{-n} \cdot \bar{z}^{-n} = (z^{-1} \bar{z}^{-1})^n = |z^{-1}|^{2n}.$$

Let us still give an alternative proof of the second assertion in (2): Using $|z| = |\bar{z}|$, this is a consequence of

$$|z^{-n}| = |(z^{-1})^n| = |z^{-1}|^n = \left| \frac{\bar{z}}{|z|^2} \right|^n = \left(\frac{|\bar{z}|}{|z|^2} \right)^n = |z|^{-n}.$$

(3) These assertions follow by iterative applications of properties (1), and (6) of Proposition 6.2.1, respectively. \blacksquare

Theorem 6.2.3 (Triangle Inequality). *If $z_1, z_2 \in \mathbb{C}$, then*

$$(6.2.1) \quad |z_1 + z_2| \leq |z_1| + |z_2|.$$

Equality happens if and only if $z_2 = 0$ or $z_2 \neq 0$ and z_1/z_2 is a real nonnegative number.

Proof: Geometrically, we use Figure 6.1.3 to get an intuition. If A and B are the points corresponding to z_1 and z_2 , respectively, then $|z_1| = |OA|$, $|z_2| = |OB|$ and $|z_1 + z_2| = |OC|$, where C is the complex number corresponding to $z_1 + z_2$. Because $OACB$ is a parallelogram, $|OB| = |AC|$. In any triangle, the length of one side is smaller than or equal to the sum of the lengths of the other two sides and therefore, $|OC| \leq |OA| + |AC|$. This gives the desired inequality. Equality happens if and only if O , A , and C are collinear with A between O and C . This is the same as our claim for our equality case, and we leave the verification as an exercise.

Let us give now a complex, nongeometric proof of the triangle inequality. We start with the following special case of (6.2.1). Given $z \in \mathbb{C}$, then

$$(6.2.2) \quad |1+z| \leq 1+|z|, \quad \text{or, equivalently,} \quad |1+z|^2 \leq (1+|z|)^2.$$

To verify (6.2.2) take an arbitrary $z = a+bi$. Since $\operatorname{Re}(z) \leq |\operatorname{Re}(z)|$, estimate (3) implies that

$$|1+z|^2 = (1+a)^2 + b^2 = 1+2a+a^2+b^2 = 1+2\operatorname{Re}(z)+|z|^2 \leq 1+2|z|+|z|^2 = (1+|z|)^2.$$

Thus we proved (6.2.1) with $z_1 = 1$ and for arbitrary $z_2 \in \mathbb{C}$. Note that equality happens above if and only if $\operatorname{Re}(z) = |z|$ which is equivalent with z being a nonnegative real number.

We deduce now the general case. If $z_2 = 0$, there is nothing to show. Thus, we may assume $z_2 \neq 0$. Property (6) and (6.2.2) with $z = z_1/z_2$ let us conclude that

$$|z_1 + z_2| = |z_2| \cdot |1 + (z_1/z_2)| \leq |z_2|(1 + |z_1/z_2|) = |z_1| + |z_2|.$$

Equality happens if and only if z_1/z_2 is a nonnegative real number. This completes the proof of (6.2.1), hence that of the proposition. ■

Exercise 6.2.1. Give a geometric description of the set of all complex numbers z such that $|z - 3| = 5$.

Exercise 6.2.2 (Reverse Triangle Inequality). If $z_1, z_2 \in \mathbb{C}$, then prove that

$$(6.2.3) \quad ||z_1| - |z_2|| \leq |z_1 - z_2|.$$

Characterize the equality case.

Exercise 6.2.3. Find $z \in \mathbb{C}$ satisfying

$$z + 3\bar{z} = 5 - 6i, \quad \text{those such that} \quad z\bar{z} = 25 \quad \text{and those with} \quad \operatorname{Re} z + \operatorname{Im} z = 7.$$

Exercise 6.2.4. Describe the set of all complex numbers $z \in \mathbb{C}$ which satisfy the following:

- (i) $\bar{z} = -z$
- (ii) $\bar{z} = z^{-1}$
- (iii) $|z - 5| = |z + 3|$
- (iv) $|z + 2| + |z - 4| = 7$
- (v) $|z| < 1 - \operatorname{Re} z$.

Exercise 6.2.5. Show that there is no complex number $z \in \mathbb{C}$ such that

$$|z| - z = i.$$

Exercise 6.2.6. Find all complex numbers $z \in \mathbb{C}$ satisfying one of the following properties:

$$(a) \quad \bar{z} = i(z - 1) \quad (b) \quad z^2 \cdot \bar{z} = z \quad (c) \quad |z + 3i| = 3|z|.$$

Exercise 6.2.7. Prove that for all $z \in \mathbb{C}$

$$z^2 + \bar{z}^2 = 2 \cdot \operatorname{Re}(z)^2 - 2 \cdot \operatorname{Im}(z)^2.$$

Exercise 6.2.8. Let $n \in \mathbb{N}$. If z_1, \dots, z_n are complex numbers, then

$$|z_1 + \dots + z_n| \leq |z_1| + \dots + |z_n|.$$

When does equality happen?

Exercise 6.2.9. Let $z \in \mathbb{C} \setminus \{-1, 1\}$. Prove that $|z| = 1$ if and only if $\frac{z-1}{z+1}$ is purely imaginary.

Exercise 6.2.10. Determine all the real numbers x such that $|x + 1/x| = 2$. Find all the complex numbers z such that $|z + 1/z| = 2$.

6.3. Polar Representation of Complex Numbers

Let $z = a + bi$ be a nonzero complex number. Then

$$\frac{z}{|z|} = x + yi \quad \text{where} \quad x = \frac{a}{\sqrt{a^2 + b^2}} \quad \text{and} \quad y = \frac{b}{\sqrt{a^2 + b^2}}.$$

The numbers x and y satisfy $x^2 + y^2 = 1$, that is, the point (x, y) lies on a circle of radius 1 in \mathbb{R}^2 . From elementary trigonometry is known that there is an angle θ such that

$$x = \cos \theta \quad \text{and} \quad y = \sin \theta.$$

Consequently, given a nonzero $z \in \mathbb{C}$, there exist a number $r > 0$ (choose $r = |z|$) and an angle $\theta \in \mathbb{R}$ such that

$$(6.3.1) \quad z = r(\cos \theta + i \sin \theta).$$

Definition 6.3.1. We call (6.3.1) a **polar representation** of $z \in \mathbb{C}$.

Is the polar representation of a complex number unique? If $z = r(\cos \theta + i \sin \theta)$ for some $r > 0$ and $\theta \in \mathbb{R}$, then

$$|z|^2 = r^2 \cos^2 \theta + r^2 \sin^2 \theta = r^2(\cos^2 \theta + \sin^2 \theta) = r^2.$$

This implies, that $r = |z|$, and $r > 0$ is unique. How about the angle θ ? Since for each integer $k \in \mathbb{Z}$

$$\cos(\theta + 2k\pi) = \cos \theta \quad \text{and} \quad \sin(\theta + 2k\pi) = \sin \theta,$$

there are infinitely many angles θ representing the same complex number. To get uniqueness, one has to restrict the range of the allowed angles. In the literature, one requires either that $-\pi \leq \theta < \pi$ or that $0 \leq \theta < 2\pi$. We will use the latter convention. Elementary trigonometry theorems imply the following result.

Proposition 6.3.1. For each nonzero complex number $z = a + bi$ there are a unique $r > 0$ and a unique angle $\theta \in [0, 2\pi)$ such that

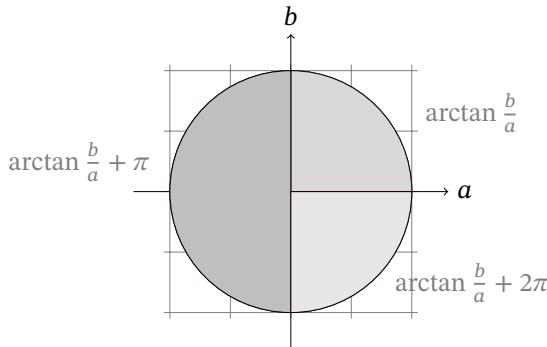
$$z = r(\cos \theta + i \sin \theta).$$

Then

$$r = |z| = \sqrt{a^2 + b^2}$$

and

$$(6.3.2) \quad \theta = \begin{cases} \arctan \frac{b}{a} & : a > 0, b \geq 0 \\ \arctan \frac{b}{a} + 2\pi & : a > 0, b < 0 \\ \arctan \frac{b}{a} + \pi & : a < 0 \\ \frac{\pi}{2} & : a = 0, b > 0 \\ \frac{3\pi}{2} & : a = 0, b < 0 \end{cases}$$



Remark 6.3.1. There are two reasons why the shifts by π or 2π are necessary in the cases $a < 0$ and $a > 0, b < 0$, respectively.

- (1) Recall that the tangent of an angle θ is defined by

$$\tan \theta = \frac{\sin \theta}{\cos \theta}, \quad \theta \notin \left\{ \frac{(2k+1)\pi}{2} : k \in \mathbb{Z} \right\}.$$

It is an one-to-one mapping from $(-\pi/2, \pi/2)$ onto \mathbb{R} . Thus, its inverse arctan (sometimes also denoted by \tan^{-1}) maps \mathbb{R} onto the interval $(-\pi/2, \pi/2)$. Hence, in order to get $0 \leq \theta < 2\pi$ a shift is necessary whenever $\arctan \frac{b}{a} < 0$.

- (2) Since $\arctan\left(\frac{b}{a}\right) = \arctan\left(\frac{-b}{-a}\right)$, without shift the complex numbers z and $-z$ would possess the same radial representation.

Summary: Each nonzero complex number $z = a + bi$ admits the unique polar representation

$$(6.3.3) \quad z = r(\cos \theta + i \sin \theta), \quad r > 0, \quad 0 \leq \theta < 2\pi.$$

Here $r = |z| = \sqrt{a^2 + b^2}$ and θ is evaluated by (6.3.2). Moreover, then for each $k \in \mathbb{Z}$ one also has

$$z = r(\cos(\theta + 2k\pi) + i \sin(\theta + 2k\pi)).$$

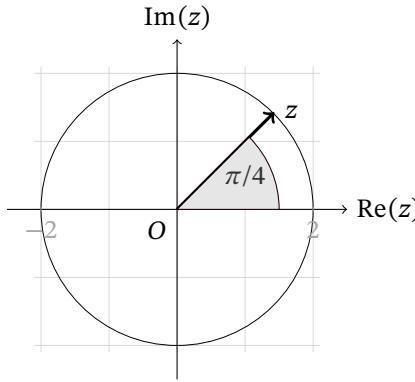


Figure 6.3.1. The complex number $z = \sqrt{2}(1 + i)$ in polar representation with $|z| = 2$ and $\arg(z) = \pi/4$.

Definition 6.3.2. Let $z \in \mathbb{C}$ be written in polar representation (6.3.3). The angle $\theta \in [0, 2\pi)$ is called the **argument** of z . It is denoted by $\arg(z)$,

$$\arg(z) = \theta \quad \text{if } z = r(\cos \theta + i \sin \theta), \quad r > 0, \quad 0 \leq \theta < 2\pi.$$

Example 6.3.1.

$$\begin{aligned} 1 &= \cos 0 + i \sin 0 \\ i &= \cos(\pi/2) + i \sin(\pi/2) \\ -1 &= \cos \pi + i \sin \pi \\ -i &= \cos(3\pi/2) + i \sin(3\pi/2) \\ 1+i &= \sqrt{2}(\cos(\pi/4) + i \sin(\pi/4)) \\ -1+i &= \sqrt{2}(\cos(3\pi/4) + i \sin(3\pi/4)) \\ -1-i &= \sqrt{2}(\cos(5\pi/4) + i \sin(5\pi/4)) \\ 1-i &= \sqrt{2}(\cos(7\pi/4) + i \sin(7\pi/4)). \end{aligned}$$

Hence,

$$\arg(1) = 0, \quad \arg(i) = \frac{\pi}{2}, \quad \arg(-1) = \pi, \quad \arg(-i) = \frac{3\pi}{2}, \quad \arg(1-i) = \frac{7\pi}{4}.$$

Proposition 6.3.2. Let $z_1 = r_1(\cos \theta_1 + i \sin \theta_1)$ and $z_2 = r_2(\cos \theta_2 + i \sin \theta_2)$ be two complex numbers in polar representation. Then their product has the representation

$$(6.3.4) \quad z_1 \cdot z_2 = (r_1 r_2)(\cos(\theta_1 + \theta_2) + i \sin(\theta_1 + \theta_2)).$$

Proof: Multiplying z_1 and z_2 by the rules for the multiplication of complex numbers gives

$$(6.3.5) \quad z_1 \cdot z_2 = (r_1 r_2)(\cos \theta_1 \cos \theta_2 - \sin \theta_1 \sin \theta_2 + i(\cos \theta_1 \sin \theta_2 + \sin \theta_1 \cos \theta_2)).$$

The trigonometric addition formulas yield

$$\cos(\alpha + \beta) = \cos \alpha \cos \beta - \sin \alpha \sin \beta \quad \text{and} \quad \sin(\alpha + \beta) = \cos \alpha \sin \beta + \sin \alpha \cos \beta$$

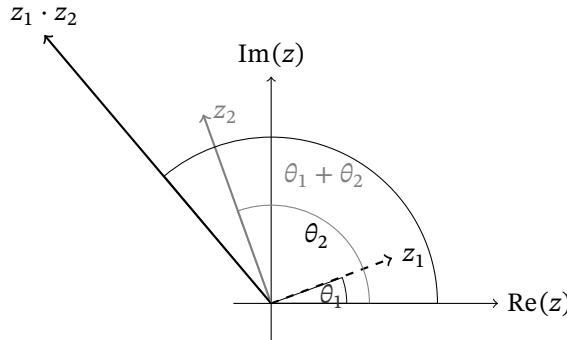


Figure 6.3.2. The multiplication of two complex numbers z_1 and z_2 .

for any two angles α and β . An application of these formulas to (6.3.5) with $\alpha = \theta_1$ and $\beta = \theta_2$ implies

$$z_1 \cdot z_2 = (r_1 r_2)(\cos(\theta_1 + \theta_2) + i \sin(\theta_1 + \theta_2)).$$

This completes the proof. ■

Remark 6.3.2. Formula (6.3.4) implies that

$$\arg(z_1 \cdot z_2) = \begin{cases} \arg(z_1) + \arg(z_2) & : \arg(z_1) + \arg(z_2) < 2\pi \\ \arg(z_1) + \arg(z_2) - 2\pi & : \arg(z_1) + \arg(z_2) \geq 2\pi. \end{cases}$$

Another way to express this is:

$$\arg(z_1 \cdot z_2) \equiv \arg(z_1) + \arg(z_2) \pmod{2\pi}.$$

Is there a similar formula as (6.3.4) for the division of two complex numbers? Let us treat first a special case.

Proposition 6.3.3. *If $r > 0$, then*

$$\frac{1}{r(\cos \theta + i \sin \theta)} = r^{-1}(\cos(-\theta) + i \sin(-\theta)) = r^{-1}(\cos \theta - i \sin \theta).$$

Proof: Setting $z = r(\cos \theta + i \sin \theta)$, its conjugate complex number is given by $\bar{z} = r(\cos \theta - i \sin \theta)$. Consequently, the assertion follows by

$$\begin{aligned} \frac{1}{r(\cos \theta + i \sin \theta)} &= \frac{1}{z} = \frac{\bar{z}}{|z|^2} = \frac{r(\cos \theta - i \sin \theta)}{r^2} = r^{-1}(\cos(-\theta) + i \sin(-\theta)) \\ &= r^{-1}(\cos(\theta) - i \sin(\theta)). \end{aligned}$$

This proves the result. ■

Combining Proposition 6.3.2 and Proposition 6.3.3, we get the following corollary.

Corollary 6.3.4. *If $r_2 > 0$, then*

$$(6.3.6) \quad \frac{r_1(\cos \theta_1 + i \sin \theta_1)}{r_2(\cos \theta_2 + i \sin \theta_2)} = \frac{r_1}{r_2}(\cos(\theta_1 - \theta_2) + i \sin(\theta_1 - \theta_2)).$$

Remark 6.3.3. Another way to express (6.3.6) is

$$\left| \frac{z_1}{z_2} \right| = \frac{|z_1|}{|z_2|} \quad \text{and} \quad \arg\left(\frac{z_1}{z_2}\right) \equiv \arg(z_1) - \arg(z_2) \pmod{2\pi}.$$

Proposition 6.3.5 (de Moivre's Theorem). *If $r > 0$ and $\theta \in \mathbb{R}$, then for any $n \in \mathbb{Z}$:*

$$(6.3.7) \quad (r(\cos \theta + i \sin \theta))^n = r^n(\cos(n\theta) + i \sin(n\theta)).$$

Proof: In a first step we treat the case of nonnegative powers n . In this case we verify the assertion by induction over n . If $n = 0$, the left-hand side of (6.3.7) equals 1 and so does the right-hand side by $\cos 0 = 1$ and $\sin 0 = 0$.

Suppose now (6.3.7) is satisfied for some $n \geq 0$. Then Proposition 6.3.2 implies

$$\begin{aligned} (r(\cos \theta + i \sin \theta))^{n+1} &= (r(\cos \theta + i \sin \theta))^n \cdot [r(\cos \theta + i \sin \theta)] \\ &= [r^n(\cos(n\theta) + i \sin(n\theta))] \cdot [r(\cos \theta + i \sin \theta)] \\ &= r^{n+1}(\cos((n+1)\theta) + i \sin((n+1)\theta)), \end{aligned}$$

and (6.3.7) is true for $n + 1$, thus for all $n \geq 0$.

If $n < 0$, then we apply (6.3.7) with $-n$ and $[r(\cos \theta + i \sin \theta)]^{-1}$. Doing so, an application of Proposition 6.3.3 and of (6.3.7) with $-n > 0$ and with $-\theta$ leads to

$$\begin{aligned} (r(\cos \theta + i \sin \theta))^n &= \left[\frac{1}{r(\cos \theta + i \sin \theta)} \right]^{-n} \\ &= [r^{-1}(\cos(-\theta) + i \sin(-\theta))]^{-n} \\ &= r^n(\cos((-n)(-\theta)) + i \sin((-n)(-\theta))) \\ &= r^n(\cos(n\theta) + i \sin(n\theta)). \end{aligned}$$

Hence, (6.3.7) is also true for negative powers, and this completes the proof. ■

Example 6.3.2. We want to evaluate $(1 + i)^{10}$. To this end we use the polar representation

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

Thus, an application of de Moivre's theorem implies

$$(1 + i)^{10} = 2^5(\cos(5\pi/2) + i \sin(5\pi/2)) = 32(\cos(\pi/2) + i \sin(\pi/2)) = 32i.$$

Example 6.3.3. Let $z_1 = \frac{\sqrt{3}}{2} + \frac{1}{2}i$ and $z_2 = 1 + i$. Our aim is to evaluate z_1/z_2 . Since

$$z_1 = \cos(\pi/6) + i \sin(\pi/6) \quad \text{and} \quad z_2 = \sqrt{2}(\cos(\pi/4) + i \sin(\pi/4))$$

it follows that

$$\begin{aligned} \frac{z_1}{z_2} &= \frac{1}{\sqrt{2}}(\cos(\pi/6 - \pi/4) + i \sin(\pi/6 - \pi/4)) \\ &= \frac{1}{\sqrt{2}}(\cos(-\pi/12) + i \sin(-\pi/12)) \\ &= \frac{1}{\sqrt{2}}(\cos(23\pi/12) + i \sin(23\pi/12)) \approx 0.683013 - 0.183013i. \end{aligned}$$

Example 6.3.4. We want to determine $(1 - i)^{-8}$.

By (6.3.7) it follows that

$$\begin{aligned}(1 - i)^{-8} &= \left[\sqrt{2}(\cos(7\pi/4) + i \sin(7\pi/4)) \right]^{-8} \\ &= 2^{-4}(\cos(-14\pi) + i \sin(-14\pi)) = 2^{-4}.\end{aligned}$$

Exercise 6.3.1. Find the absolute value and argument of each of the following numbers: $3 + 4i$, $(3 + 4i)^{-1}$, $(1 + i)^5$, and $|3 + 4i|$.

Exercise 6.3.2. Evaluate the following complex numbers:

$$(1 - i)^8, \quad (1 + i)^{-6}, \quad \text{and} \quad \frac{(1 + i)^2}{(1 - i)^4}.$$

Exercise 6.3.3. Let $\theta \in \mathbb{R}$.

- (1) Write $\cos(2\theta)$, $\cos(3\theta)$, $\cos(4\theta)$ in terms of $\cos \theta$.
- (2) If $\sin \theta \neq 0$, write $\frac{\sin(2\theta)}{\sin \theta}$, $\frac{\sin(3\theta)}{\sin \theta}$, $\frac{\sin(4\theta)}{\sin \theta}$ in terms of $\cos \theta$.

Exercise 6.3.4. Let $z \in \mathbb{C}$.

- (1) The number z is a nonnegative real number if and only if $\arg(z) = 0$.
- (2) The number z is a negative real number if and only if $\arg(z) = \pi$.

Exercise 6.3.5. Let A and B be two points in the Cartesian plane with O denoting the origin.

- (1) The points O , A , and B are collinear with O between A and B if and only if $|\arg(z_A) - \arg(z_B)| = \pi$.
- (2) The points O , A , and B are collinear with O not between A and B if and only if $\arg(z_A) = \arg(z_B)$.

Exercise 6.3.6. Let $z_1, z_2 \in \mathbb{C}$ be nonzero complex numbers.

- (1) Prove that $\operatorname{Re}(z_1/z_2) \geq 0$ if and only if $z_1\bar{z}_2 + \bar{z}_1z_2 \geq 0$.
- (2) Show that $\operatorname{Im}(z_1/z_2) \geq 0$ if and only if $i(\bar{z}_1z_2 - z_1\bar{z}_2) \geq 0$.

Exercise 6.3.7. Use de Moivre's theorem and the binomial formula to prove that for any angle θ and any natural number n :

$$\begin{aligned}\sin(n\theta) &= \sum_{\substack{k=0 \\ k \text{ odd}}}^n (-1)^{\frac{k-1}{2}} \binom{n}{k} \cos^{n-k} \theta \sin^k \theta \quad \text{and} \\ \cos(n\theta) &= \sum_{\substack{k=0 \\ k \text{ even}}}^n (-1)^{\frac{k}{2}} \binom{n}{k} \cos^{n-k} \theta \sin^k \theta.\end{aligned}$$

Exercise 6.3.8. Let $z_1, z_2 \in \mathbb{C}$ satisfy $|z_1| = |z_2| > 0$. Prove that $\frac{z_1+z_2}{|z_1z_2|+z_1z_2}$ is a real number.

Exercise 6.3.9. Let z be a nonzero complex number.

- (1) Prove that $\frac{z}{|z|} + \frac{|z|}{z}$ is a real number.
- (2) Show that there exist $a, b \in \mathbb{R}$ such that $z^2 = az + b$.
- (3) Show that for any natural number n , there are $a_n, b_n \in \mathbb{R}$ with $z^n = a_n z + b_n$.

Exercise 6.3.10. If $z \in \mathbb{C}$ such that $|z| = 1$, show that $\frac{z^{2n}+1}{z^n}$ is a real number for any natural number n .

Exercise 6.3.11. Let $z \in \mathbb{C}$ such that $|z - 1| \leq 1$ and $|z + 1| \leq 1$. Prove that $z = 0$.

6.4. Roots of Complex Numbers

The main motivation for the introduction of complex numbers was to extract roots of arbitrary numbers, including negative ones. We still do not know whether this is so, i.e., whether for each complex number its n th root exists. And if this is so, which properties do these roots possess? The next example shows that the answers to these questions may not be completely trivial. It has disconcerted mathematicians for several centuries.

Example 6.4.1. Since $i^2 = i \cdot i = -1$ it follows that $i = \sqrt{-1}$. Applying the usual rules for square roots implies

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

Something is wrong. But what? We will answer this question in Remark 6.4.5 below.

Proposition 6.4.1. Let $z \in \mathbb{C}$ be represented as $z = r(\cos \theta + i \sin \theta)$ for some $r > 0$ and $\theta \in [0, 2\pi)$. Given $n \geq 2$ we define n complex numbers w_0, w_1, \dots, w_{n-1} by

$$(6.4.1) \quad w_k = r^{1/n} \left(\cos \left(\frac{\theta}{n} + \frac{2k\pi}{n} \right) + i \sin \left(\frac{\theta}{n} + \frac{2k\pi}{n} \right) \right), \quad k = 0, \dots, n-1.$$

Then for all $k \in \{0, \dots, n-1\}$ it follows that $w_k^n = z$. Conversely, if w is a complex number with $w^n = z$, then there is a $k \leq n-1$ such that $w = w_k$.

Proof: Given $z = r(\cos \theta + i \sin \theta)$, set

$$\theta_k := \frac{\theta}{n} + \frac{2k\pi}{n}, \quad k = 0, \dots, n-1.$$

With this notation, $\arg(w_k) = \theta_k$ where w_k is as in (6.4.1). Now de Moivre's theorem implies

$$w_k^n = r(\cos(n\theta_k) + i \sin(n\theta_k)) = r(\cos(\theta + 2k\pi) + i \sin(\theta + 2k\pi)) = r(\cos \theta + i \sin \theta) = z.$$

This proves the first part of the proposition.

Suppose now $w = |w|(\cos \varphi + i \sin \varphi)$, $0 \leq \varphi < 2\pi$, satisfies $w^n = z$ for the complex number $z \neq 0$. If $z = r(\cos \theta + i \sin \theta)$, then (6.3.7) implies

$$|w|^n (\cos(n\varphi) + i \sin(n\varphi)) = r(\cos \theta + i \sin \theta).$$

Hence, it follows $|w| = r^{1/n}$ and, there exists an integer $k \in \mathbb{Z}$ for which $n\varphi - 2k\pi = \theta$. Since $0 \leq \varphi, \theta < 2\pi$, the integer k has to be in⁸ $\{0, 1, \dots, n-1\}$, and we conclude that

$$\varphi = \frac{\theta}{n} + \frac{2k\pi}{n}.$$

Consequently, $w = w_k$ which completes the proof. ■

Definition 6.4.1. Let z be a complex number and $n \in \mathbb{N}$. The complex number w is said to be an *n th root* of z if $w^n = z$.

$$(w \in \mathbb{C} \text{ is an } n\text{th root of } z \in \mathbb{C}) \Leftrightarrow (w^n = z).$$

Remark 6.4.1. Proposition 6.4.1 may now be formulated as follows: Each nonzero complex number z possesses exactly n roots of order n . Moreover, these n roots are w_0, \dots, w_{n-1} as introduced in (6.4.1).

Remark 6.4.2. A special role plays the case $n = 2$. Here $w_1 = -w_0$, thus in this case it is common to write for the two square roots

$$\pm\sqrt{z}, \quad z \in \mathbb{C}.$$

This looks like the formula for the square root of positive real numbers. But there is a big difference. In the real case, one root is positive, the other negative. Thus, it is always possible to select one of the roots, mostly the positive one, and to define it as the square root of the positive real number. In contrast to that, there is no natural choice for \sqrt{z} . Should we always take w_0 or better $w_1 = -w_0$?

Example 6.4.2. The n roots of $1 \in \mathbb{C}$ of order n are 1 and the $n-1$ complex numbers with absolute value 1 and with arguments

$$\frac{2\pi}{n}, \frac{4\pi}{n}, \dots, \frac{2(n-1)\pi}{n}.$$

In other words, these n roots are

$$w_k = \cos\left(\frac{2k\pi}{n}\right) + i \sin\left(\frac{2k\pi}{n}\right), \quad k = 0, \dots, n-1.$$

These n numbers are called **roots of unity** (of order n). For example, if $n = 4$, the 4 roots of unity are $1, i, -1, -i$. Or if $n = 8$, the eight roots of unity equal

$$1, \frac{1+i}{\sqrt{2}}, i, \frac{-1+i}{\sqrt{2}}, -1, \frac{-1-i}{\sqrt{2}}, -i, \frac{1-i}{\sqrt{2}}.$$

Remark 6.4.3. If $n \geq 2$, let $E_n = \{w_0, w_1, \dots, w_{n-1}\} \subseteq \mathbb{C}$ be the set of roots of unity. Since $\arg(w_k) = \frac{2k\pi}{n}$, $k = 0, \dots, n-1$, Proposition 6.3.2 implies that

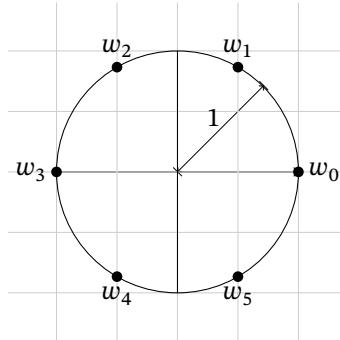
$$w_k \cdot w_\ell = w_m \quad \text{where} \quad m \equiv k + \ell \pmod{n}.$$

Example 6.4.3.

(1) The n roots of $1 + i$ are $2^{1/2n}(\cos \theta_k + i \sin \theta_k)$ with

$$\theta_k = \frac{\pi}{4n} + \frac{2\pi k}{n}, \quad k = 0, \dots, n-1.$$

⁸Observe that $2k\pi = n\varphi - \theta \leq n\varphi < 2n\pi$ and $2k\pi \geq -\theta > -2\pi$.

**Figure 6.4.1.** Roots of unity of order 6.

- (2) The two square roots w_0 and w_1 of i have absolute value 1 and arguments $\pi/4$ and $5\pi/4$. They equal $w_0 = \frac{1+i}{\sqrt{2}}$ and $w_1 = -\frac{1+i}{\sqrt{2}}$, which also can be checked directly by evaluating w_0^2 and w_1^2 . So we got

$$\sqrt{i} = \pm \frac{1+i}{\sqrt{2}}.$$

- (3) The n roots of $z = -1$ have absolute value 1 and arguments

$$\left\{ \frac{\pi}{n}, \frac{\pi}{n} + \frac{2\pi}{n}, \dots, \frac{\pi}{n} + \frac{2(n-1)\pi}{n} \right\} = \left\{ \frac{\pi}{n}, \frac{3\pi}{n}, \dots, \frac{(2n-1)\pi}{n} \right\}.$$

For example, if $n = 5$, the five roots of -1 have the arguments

$$\pi/5, 3\pi/5, \pi, 7\pi/5 \text{ and } 9\pi/5.$$

In degrees these are $36^\circ, 108^\circ, 180^\circ, 252^\circ$, and 324° .

Remark 6.4.4. The fact that a complex number has n different roots of order n is not too surprising. A positive real numbers x has two different real square roots, namely \sqrt{x} and $-\sqrt{x}$. The new phenomenon occurs when one investigates n th roots with $n \geq 3$ or when considering square roots of negative numbers. There is exactly one (real) solution x of $x^3 = 1$. In contrast to that there are three different complex roots of unity, only one of them is real. Any negative number y has two complex square roots, namely $i\sqrt{-y}$, and $-i\sqrt{-y}$.

Let us formulate some properties of the roots of a given complex number.

Proposition 6.4.2.

- (1) Suppose that certain w and w' are both n th roots of z and z' , respectively. Then ww' is an n th root of zz' .
- (2) Let w be an n th root of z . If $m \geq 1$, then the number w^m is an n th root of z^m .
- (3) If w is an n th root of z , then $1/w$ is an n th root of $1/z$.

Proof:

(1) By assumption $w^n = z$ and $(w')^n = z'$. This implies

$$(ww')^n = w^n \cdot (w')^n = zz'.$$

This shows that $w w'$ is an n th root of zz' .

(2) This easily follows from $(w^m)^n = w^{m \cdot n} = (w^n)^m = z^m$.

(3) Note that

$$\left(\frac{1}{w}\right)^n = \frac{1}{w^n} = \frac{1}{z}.$$

But this says nothing else as that $1/w$ is an n th root of $1/z$. ■

Remark 6.4.5. Now we may answer the question of why the calculations in Example 6.4.1 led to the wrong result $-1 = 1$. The deeper reason is that a notation $\sqrt[n]{z}$ does not make sense. There are n different roots, and the question is, which one⁹ has to be chosen as $\sqrt[n]{z}$. If w_0, \dots, w_{n-1} are the roots defined in (6.4.1), one could conjecture that w_0 is the best choice for the n th root of z . Assume, we follow this idea and define for all nonzero complex numbers z its n th root by

$$\sqrt[n]{z} = w_0 = |z|^{1/n} \left(\cos\left(\frac{\theta}{n}\right) + i \sin\left(\frac{\theta}{n}\right) \right).$$

If we do so, then it may happen that $\sqrt[n]{z_1 z_2} \neq \sqrt[n]{z_1} \cdot \sqrt[n]{z_2}$. Why? As we proved in (1) of Proposition 6.4.2, the product $\sqrt[n]{z_1} \cdot \sqrt[n]{z_2}$ is an n th root of $z_1 z_2$, but not necessarily the one with $k = 0$, i.e., not the one we had chosen as $\sqrt[n]{z_1 z_2}$. And this exactly happens in Example 6.4.1. There are two square roots of $z = 1$, namely $w_0 = 1$ and $w_1 = -1$. Since $i \in \mathbb{C}$ is a root of -1 , property (1) of Proposition 6.4.2 implies that $i^2 = -1$ is a root of $1^2 = 1$. This is correct, but the square root -1 of 1 is w_1 , not w_0 . Hence,

$$\sqrt{(-1)(-1)} \neq \sqrt{-1} \cdot \sqrt{-1}.$$

It would be correct if we choose on the left-hand side the root with $k = 1$ and on the right-hand side the roots with $k = 0$. As a consequence we see that the notation $i = \sqrt{-1}$ is dangerous. It suggests that the rules of square roots, valid in the real case, are also true in the complex plane.

Summing up, there is **no reasonable way** to define $\sqrt[n]{z}$ satisfying

$$\sqrt[n]{z_1 z_2} = \sqrt[n]{z_1} \cdot \sqrt[n]{z_2}$$

for all $z_1, z_2 \in \mathbb{C}$. We only know that the right-hand side is an n th root of $z_1 z_2$, but we do not know which of the n ones. It may be the first one w_0 , or maybe the second one w_1 , or it could even be the last one w_{n-1} .

⁹Some authors consider $\sqrt[n]{\cdot}$ as set valued *function*. It maps any $z \in \mathbb{C}$ to the set of its n roots. This can be done, but has to be handled with extreme care. For example, how to define then the sum or the product of roots?

Summary: For each complex number $z \neq 0$ and any $n \geq 1$ there exist n different roots w_0, w_1, \dots, w_{n-1} of z . These roots are

$$w_k = |z|^{1/n} \left(\cos\left(\frac{\theta + 2k\pi}{n}\right) + i \sin\left(\frac{\theta + 2k\pi}{n}\right) \right), \quad k = 0, \dots, n-1,$$

where $\theta \in [0, 2\pi)$ denotes the argument of z . If $n = 2$, then

$$w_0 = |z|^{1/2} \left(\cos\left(\frac{\theta}{2}\right) + i \sin\left(\frac{\theta}{2}\right) \right) = -w_1.$$

The n numbers w_0, w_1, \dots, w_{n-1} with

$$w_k^n = 1, \quad k = 0, \dots, n-1,$$

are called roots of unity of order n . Their representation is

$$w_k = \cos\left(\frac{2k\pi}{n}\right) + i \sin\left(\frac{2k\pi}{n}\right), \quad k = 0, \dots, n-1.$$

We conclude this section with a fundamental result about zeroes of complex polynomials. It asserts the following

Theorem 6.4.3 (Fundamental Theorem of Algebra¹⁰). *Let $p : \mathbb{C} \rightarrow \mathbb{C}$ be a complex polynomial of degree n . That is, there are complex numbers c_0, \dots, c_n with $c_n \neq 0$ such that*

$$(6.4.2) \quad p(z) = c_n z^n + c_{n-1} z^{n-1} + \dots + c_1 z + c_0, \quad z \in \mathbb{C}.$$

Then there are $m \leq n$ complex numbers z_1, \dots, z_m and positive integers k_1, \dots, k_m satisfying $k_1 + \dots + k_m = n$ so that the polynomial p may be represented as

$$p(z) = c_n (z - z_1)^{k_1} \cdots (z - z_m)^{k_m}, \quad z \in \mathbb{C}.$$

In other words, the numbers z_1, \dots, z_m are the zeroes of p with corresponding multiplicities k_1, \dots, k_m .

Remark 6.4.6. Nowadays, there exist several approaches for the proof of the fundamental theorem. But all of them use facts which exceed the scope of this book. We refer interested readers to [32], Chapter 25, Theorem 2, for a proof.

Example 6.4.4.

- (1) The polynomial $p(z) = z^2 + 1$ can be written as $p(z) = (z - i)(z + i)$.
- (2) The polynomial $p(z) = z^4 - 4z + 3$ can be represented as

$$p(z) = (z - 1)^2(z + 1 + i\sqrt{2})(z + 1 - i\sqrt{2}), \quad z \in \mathbb{C}.$$

- (3) The polynomial $p(z) = z^3 - 1$ has the zeroes $1, -\frac{1}{2} + \frac{\sqrt{3}}{2}i$ and $-\frac{1}{2} - \frac{\sqrt{3}}{2}i$.

¹⁰This theorem has a long and interesting history. First conjectures about the number of zeroes of a polynomial were already published in the 17th century. There were several mathematicians who tried to prove that a polynomial of degree n has exactly n zeroes counted according to their multiplicity. Among them were Leonhard Euler, Joseph-Louis Lagrange, and Pierre-Simon Laplace. But either their attempts failed or their proofs were incorrect. In 1799, Carl Friedrich Gauss gave in his dissertation a proof of the fundamental theorem which was accepted at that time. But later on it turned out that Gauss's proof also contained a gap which was filled in 1920 by the Ukrainian mathematician Alexander Ostrowski (1893–1986). The first rigorous proof of the fundamental theorem was found in 1806 by Jean-Robert Argand, an amateur mathematician from Geneva who later worked in a bookstore in Paris. In 1821, Argand's proof appeared in a textbook by the French mathematician Augustin-Louis Cauchy (1789–1857), but Argand is not credited for it.

In the case of real coefficients Theorem 6.4.3 can be refined. To do so we need the following result.

Proposition 6.4.4. *Suppose the polynomial*

$$p(z) = c_n z^n + c_{n-1} z^{n-1} + \cdots + c_1 z + c_0, \quad z \in \mathbb{C}.$$

has **real** coefficients c_0, \dots, c_n . Then a number $z \in \mathbb{C}$ is a zero of p if and only if its conjugate number \bar{z} is a zero as well.

Proof: This easily follows by the properties of the conjugate complex number as stated in property (3) of Corollary 6.2.2. Using that the coefficients c_j are real, hence $\overline{c_j} = c_j$, it follows that

$$\begin{aligned} p(\bar{z}) &= c_n \bar{z}^n + c_{n-1} \bar{z}^{n-1} + \cdots + c_1 \bar{z} + c_0 \\ &= \overline{c_n z^n} + \cdots + \overline{c_1 z} + \overline{c_0} = \overline{c_n z^n + \cdots + c_1 z + c_0} \\ &= \overline{p(z)} = \bar{0} = 0. \end{aligned}$$

This proves our assertion. ■

The preceding proposition says that the complex zeroes of polynomials with real coefficients always occur pairwise. Moreover, it is not difficult to show, but needs some more facts about polynomials, that also the multiplicities of the zeroes z and \bar{z} coincide.

Thus, in the case of real coefficients Theorem 6.4.3 implies the following.

Theorem 6.4.5. *Let p as in (6.4.2) be a polynomial of degree n with **real coefficients**. Then there exist real numbers x_1, \dots, x_ℓ and nonreal complex numbers z_1, \dots, z_m such that for certain integers k_1, \dots, k_ℓ and r_1, \dots, r_m it follows that*

$$p(z) = c_n (z - x_1)^{k_1} \cdots (z - x_\ell)^{k_\ell} \cdot (z - z_1)^{r_1} \cdot (z - \bar{z}_1)^{r_1} \cdots (z - z_m)^{r_m} \cdot (z - \bar{z}_m)^{r_m}.$$

The multiplicities satisfy

$$k_1 + \cdots + k_\ell + 2r_1 + \cdots + 2r_m = n.$$

In other words, x_1, \dots, x_ℓ are the real zeroes of the polynomial p with multiplicities k_1, \dots, k_ℓ while $z_1, \bar{z}_1, \dots, z_m, \bar{z}_m$ denote its complex roots with multiplicities $r_1, r_1, \dots, r_m, r_m$.

Example 6.4.5. For example, the polynomial

$$p(z) = z^6 - 1, \quad z \in \mathbb{C},$$

can be represented as

$$p(z) = (z - 1)(z + 1)(z - w_1)(z - w_2)(z - w_4)(z - w_5)$$

where w_1, w_2, w_4 , and w_5 are the nonreal unit roots of order 6 (cf. Figure 6.4.1). All zeroes have multiplicity 1. It is worthwhile to mention that $\bar{w}_1 = w_5$ and $\bar{w}_2 = w_4$.

Another example is the polynomial

$$p(z) = (z - 1)^3(z - 2)^2(z - i)^2(z + i)^2, \quad z \in \mathbb{C}.$$

It has the two real zeroes 1 and 2 with multiplicities 3 and 2, respectively, as well as the complex zeroes i and $-i$, each with multiplicity 2. All together we see that p has degree 9.

Remark 6.4.7. There is another way to formulate the representation in Theorem 6.4.5. Using that

$$(z - z_j)(z - \bar{z}_j) = z^2 - z(z_j + \bar{z}_j) + z_j\bar{z}_j = z^2 - 2 \operatorname{Re}(z_j)z + |z_j|^2$$

we get in the case of real coefficients that

$$p(z) = c_n (z - x_1)^{k_1} \cdots (z - x_\ell)^{k_\ell} \cdot (z^2 - 2 \operatorname{Re}(z_1)z + |z_1|^2)^{r_1} \cdots (z^2 - 2 \operatorname{Re}(z_m)z + |z_m|^2)^{r_m}$$

where

$$(6.4.3) \quad k_1 + \cdots + k_\ell + 2r_1 + \cdots + 2r_m = n.$$

An alternative way to state the previous formula is as follows:

$$p(z) = c_n (z - x_1)^{k_1} \cdots (z - x_\ell)^{k_\ell} p_1(z)^{r_1} \cdots p_m(z)^{r_m}$$

where p_1, \dots, p_m are quadratic polynomials without real zero. Thereby, the multiplicities k_i and r_j satisfy condition (6.4.3) as before.

Example 6.4.6. The polynomial

$$p(z) = z^4 + 5z^2 + 4$$

of degree 4 has the complex zeroes $i, -i, 2i$ and $-2i$. Therefore,

$$p(z) = (z - i)(z + i)(z - 2i)(z + 2i) = (z^2 + 1)(z^2 + 4).$$

Here the quadratic polynomials $p_1(z) = z^2 + 1$ and $p_2(z) = z^2 + 4$ do not possess a real zero.

Let us come back to the Fundamental Theorem 6.4.3 in its general form. Given a complex polynomial, an important question is how to find its zeroes z_1, \dots, z_m as well as their multiplicities k_1, \dots, k_m . The answer is maybe discouraging: The so-called Abel–Ruffini theorem (also named Abel’s impossibility theorem) asserts that there are no algebraic formulas for the evaluation of the zeroes of a general polynomial p in the case that the degree of p is greater than or equal 5. One should mention that there exist quite complicated formulas for the zeroes of polynomials of degree 3 (Cardano’s formula mentioned in Section 6.1) and of degree 4, the so-called Ferrari formula.

Much easier and well-known for centuries is the case $n = 2$, i.e., the case of quadratic polynomials¹¹.

Proposition 6.4.6. Let the complex polynomial p of degree 2 be defined by

$$p(z) = az^2 + bz + c, \quad z \in \mathbb{C},$$

where the coefficients a, b , and c are complex numbers with $a \neq 0$. Define the (complex) discriminant D of p by

$$(6.4.4) \quad D = \frac{b^2}{4a^2} - \frac{c}{a} = \frac{b^2 - 4ac}{4a^2}.$$

¹¹The problem of solving quadratic equations has a long history. Dating back to 2000–1600 BCE Babylonians investigated certain sum-product problems of the type presented in Example 1.1.1. Greek mathematicians tried to solve quadratic equations by algebraic and geometrical methods. The first formulas for the solution of quadratic equations appeared at 825 CE in a book by Al-Khwarizmi. Because at that time neither negative numbers nor zero were available, he only treated quadratic equations which could be written in the form $ax^2 + bx = c$ with positive a, b , and c . So, necessarily some cases had to be left open. In Europe successive proofs for solution of quadratic equations appeared, from François Viète in 1579 to Simon Stevin in 1585. Later on, in 1637 René Descartes stated and proved the formula for the solution of quadratic equations in the way we know it today and as it is mostly taught at school.

If $D = 0$, then $z_0 = -b/2$ is the only zero of p with multiplicity 2. Otherwise, if $D \neq 0$, then the two zeroes are

(6.4.5)

$$z_1 = -\frac{b}{2a} + \sqrt{D} = \frac{-b + \sqrt{b^2 - 4ac}}{2a} \quad \text{and} \quad z_2 = -\frac{b}{2a} - \sqrt{D} = \frac{-b - \sqrt{b^2 - 4ac}}{2a}.$$

Remark 6.4.8. The square root \sqrt{D} denotes one of the two roots of D . The result does not depend on the special choice of \sqrt{D} . Interchanging the two roots of D leads only to an interchange of the complex zeroes z_1 and z_2 .

Proof: If we divide p by $a \neq 0$, then we may suppose that the leading coefficient of p is normalized to 1. That is, the polynomial is given by

$$p(z) = z^2 + bz + c, \quad z \in \mathbb{C}.$$

Then the discriminant is the complex number

$$D = \frac{b^2}{4} - c,$$

while the complex numbers z_1, z_2 in (6.4.5) may now be written as

$$z_1 = -\frac{b}{2} + \sqrt{D} = \frac{-b + \sqrt{b^2 - 4c}}{2} \quad \text{and} \quad z_2 = -\frac{b}{2} - \sqrt{D} = \frac{-b - \sqrt{b^2 - 4c}}{2}.$$

Let us first assume $D \neq 0$ which implies $z_1 \neq z_2$. Given an arbitrary $z \in \mathbb{C}$, an application of the distributive law leads to

$$\begin{aligned} (z - z_1)(z - z_2) &= z^2 - z(z_1 + z_2) + z_1 z_2 = z^2 + z b + \left(\frac{b}{2} + \sqrt{D}\right)\left(\frac{b}{2} - \sqrt{D}\right) \\ &= z^2 + z b + \left(\frac{b}{2}\right)^2 - (\sqrt{D})^2 = z^2 + z b + \frac{b^2}{4} - D \\ &= z^2 + bz + c = p(z). \end{aligned}$$

Thus, z_1 and z_2 are the two (different) zeroes of p .

If $D = 0$ or, equivalently, $(b/2)^2 = c$, then the assertion follows by

$$(z - z_0)^2 = \left(z + \frac{b}{2}\right)^2 = z^2 + bz + \left(\frac{b}{2}\right)^2 = z^2 + bz + c = p(z).$$

This concludes our proof. ■

Example 6.4.7. If $p(z) = z^2 + 2iz - 1 + i$, then

$$b = 2i, \quad \text{thus} \quad \frac{b^2}{4} = -1, \quad c = -1 + i, \quad \text{hence} \quad D = -i.$$

Now

$$\sqrt{D} = \sqrt{-i} = \frac{-1+i}{\sqrt{2}},$$

which implies

$$z_{1/2} = -i \pm \sqrt{-i} = -i \pm \frac{-1+i}{\sqrt{2}}.$$

In the case of **real** coefficients Proposition 6.4.6 leads to the following result.

Proposition 6.4.7. Let p be a quadratic polynomial with **real** coefficients $a \neq 0$, b and c . Define the (real) discriminant D as in (6.4.4). In dependence of D the following is valid:

(1) $D > 0$: Then p has exactly two real zeroes

$$x_{1/2} = -\frac{b}{2a} \pm \sqrt{\frac{b^2 - 4ac}{4a^2}} = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}.$$

(2) $D = 0$: Then $x_0 = -\frac{b}{2a}$ is the only (real) zero with multiplicity 2

(3) $D < 0$: In this case has p the two complex zeroes

$$z_{1/2} = -\frac{b}{2a} \pm i\sqrt{-D} = \frac{-b \pm i\sqrt{4ac - b^2}}{2a},$$

Example 6.4.8. If $p(z) = z^2 + 2z - 3$, then $D = 4 > 0$. Hence, $x_1 = 1$ as well as $x_2 = -3$ are the two real zeroes of p .

Example 6.4.9. For $p(z) = z^2 + 2z + 5$, $D = -4$. Thus, the complex numbers $z_1 = -1 + 2i$ and $z_2 = \bar{z}_1 = -1 - 2i$ are the distinct complex zeroes of p .

In Section 1.4, we promised that we would return to determining a general formula for the terms of a linear recurrence relation when the characteristic quadratic equation does not possess real roots. More precisely, given two numbers α and β , consider the sequence of numbers $(x_n)_{n \geq 0}$, whose first two values x_0 and x_1 are known and which satisfies the linear recurrence relation

$$(6.4.6) \quad x_{n+1} = \alpha x_n + \beta x_{n-1},$$

for $n \geq 1$. When $\alpha^2 - 4\beta \geq 0$, we described finding a general formula for x_n in Section 1.4, page 41.

If $\alpha^2 - 4\beta < 0$, then the characteristic equation $r^2 - \alpha r - \beta = 0$ does not have real solutions. However, according to case (3) in Proposition 6.4.7 it has the complex solutions:

$$(6.4.7) \quad r_1 = \frac{\alpha - i\sqrt{4\beta - \alpha^2}}{2} \quad \text{and} \quad r_2 = \frac{\alpha + i\sqrt{4\beta - \alpha^2}}{2}.$$

Thus, $r_j^2 = \alpha r_j + \beta$ and consequently, $r_j^{n+1} = \alpha r_j^n + \beta r_j^{n-1}$ for $1 \leq j \leq 2$ and any $n \geq 1$. Hence, each of the sequences $(r_1^n)_{n \geq 0}$ and $(r_2^n)_{n \geq 0}$ satisfies the recurrence relation (6.4.6). It turns out that x_n will have the following form:

$$x_n = ar_1^n + br_2^n,$$

for some complex numbers a and b to be determined. The sequence $(ar_1^n + br_2^n)_{n \geq 0}$ satisfies the same recurrence relation as the sequence $(x_n)_{n \geq 0}$. In order for the two sequences to match, we need to find a and b in \mathbb{C} such that the sequences are the same on the first two values:

$$\begin{aligned} a + b &= x_0 \\ ar_1 + br_2 &= x_1. \end{aligned}$$

Because $r_1 \neq r_2$, this system of equations has a unique solution (a, b) . Note that a and b are complex numbers. Take $b = x_0 - a$ from the first equation and substituting into the

second equation, we get that $ar_1 + (x_0 - a)r_2 = x_1$. This leads to $a(r_1 - r_2) = x_1 - x_0r_2$ and therefore,

$$(6.4.8) \quad a = \frac{x_1 - x_0r_2}{r_1 - r_2}, \quad \text{hence} \quad b = x_0 - a = \frac{x_0r_1 - x_1}{r_1 - r_2}.$$

One does not need to memorized these formulas as they can be deduced when needed.

The task of computing r_1^n and r_2^n may seem difficult when r_1 and r_2 are complex numbers, but this is where writing r_1 and r_2 in polar coordinates simplifies things greatly since r_1 and r_2 are conjugates as seen from equation (6.4.7).

Example 6.4.10. Consider the sequence $(x_n)_{n \geq 0}$ defined recursively as follows: $x_0 = 1$, $x_1 = -1$ and

$$x_{n+1} = x_n - x_{n-1},$$

for $n \geq 1$.

The characteristic equation is $r^2 - r + 1 = 0$ and, therefore, its discriminant equals $(-1)^2 - 4 = -3 < 0$. Thus, the solutions of this equation are

$$r_1 = \frac{1}{2} + \frac{\sqrt{3}}{2}i = \cos\left(\frac{\pi}{3}\right) + i \sin\left(\frac{\pi}{3}\right)$$

and

$$r_2 = \frac{1}{2} - \frac{\sqrt{3}}{2}i = \cos\left(\frac{\pi}{3}\right) - i \sin\left(\frac{\pi}{3}\right).$$

Note that

$$|r_1| = |r_2| = 1 \quad \text{and as always} \quad r_2 = \bar{r}_1.$$

Moreover, de Moivre's theorem (cf. Proposition 6.3.5) implies that

$$(6.4.9) \quad r_1^n = \cos\left(\frac{n\pi}{3}\right) + i \sin\left(\frac{n\pi}{3}\right) \quad \text{and} \quad r_2^n = \cos\left(\frac{n\pi}{3}\right) - i \sin\left(\frac{n\pi}{3}\right), \quad n \geq 1.$$

In order to find complex numbers a and b satisfying

$$(6.4.10) \quad a + b = 1 \quad \text{and} \quad ar_1 + br_2 = -1$$

we could use equations (6.4.8). But in this special case we prefer a more direct approach. We claim that

$$(6.4.11) \quad r_1 + r_2 = 1 \quad \text{and} \quad r_1 \cdot r_1 + r_2 \cdot r_2 = -1.$$

In other words, we are going to prove that (6.4.10) is valid with $a = r_1$ and $b = r_2$. The first equation in (6.4.11) is true because $r_1 = \bar{r}_2$ and $\operatorname{Re}(r_1) = 1/2$. For the second equation in (6.4.11), note (cf. Exercise 6.2.7) that for any $z \in \mathbb{C}$

$$z^2 + \bar{z}^2 = 2 \operatorname{Re}(z)^2 - 2 \operatorname{Im}(z)^2.$$

If $z = r_1$, we get

$$r_1^2 + r_2^2 = 2 \cdot \frac{1}{4} - 2 \cdot \frac{3}{4} = \frac{1}{2} - \frac{3}{2} = -1.$$

as claimed above.

An application of (6.4.9) with $n + 1$ finally leads to

$$\begin{aligned}x_n &= r_1 \cdot r_1^n + r_2 \cdot r_2^n = r_1^{n+1} + r_2^{n+1} \\&= \cos\left(\frac{(n+1)\pi}{3}\right) + i \sin\left(\frac{(n+1)\pi}{3}\right) + \cos\left(\frac{(n+1)\pi}{3}\right) - i \sin\left(\frac{(n+1)\pi}{3}\right) \\&= 2 \cdot \cos\left(\frac{(n+1)\pi}{3}\right).\end{aligned}$$

The representation of the x_n 's as

$$x_n = 2 \cdot \cos\left(\frac{(n+1)\pi}{3}\right), \quad n = 0, 1, 2, \dots$$

tells us that the x_n occur periodically with period 6, that is

$$x_{n+6k} = x_n, \quad n, k \geq 0.$$

So, only the values x_0, \dots, x_5 are of interest, and they are

$$x_0 = 1, \quad x_1 = -1, \quad x_2 = -2, \quad x_3 = -1, \quad x_4 = 1, \text{ and} \quad x_5 = 2.$$

Of course, the formula for the evaluation of the x_n could also be verified by strong induction. But to do so, one has to have a guess how the sequence of the x_n 's looks like.

Exercise 6.4.1. Find all roots of order n of $1 - i$.

Exercise 6.4.2. Evaluate the following complex roots:

$$\sqrt[4]{i}, \quad \sqrt{-i}, \quad \sqrt{1+i} \quad \text{and} \quad \sqrt[4]{1-i}.$$

Exercise 6.4.3. Let E_n be the set of all n th roots of unity. Define a binary operation on E_n by $w * z = w \cdot z$, $w, z \in E_n$. Show that $(E_n, *)$ is a commutative group in the sense of the definition given in Exercise A.8.6? Establish a bijection φ from E_n to $(\mathbb{Z}_n, +)$ such that

$$(6.4.12) \quad \varphi(w * z) = \varphi(w) + \varphi(z) \pmod{n}, \quad w, z \in E_n.$$

Exercise 6.4.4. Given $n \in \mathbb{N}$, an n th root of unity is said to be a **primitive**¹² root whenever $z^j \neq 1$ for all $1 \leq j < n$, i.e., n is the smallest number for which $z^n = 1$.

- (1) Find all roots of unity of order 4, order 6 and order 10. Which of them are primitive?
- (2) Let z be an n th root of unity. Show that then \bar{z} as well as z^k , $k \in \mathbb{Z}$, are n th roots of unity.
- (3) Verify the following: If k and ℓ are two integers with $k \equiv \ell \pmod{n}$, then this implies $z^k = z^\ell$ for all n th roots $z \in \mathbb{C}$.
- (4) Suppose that z is a primitive root of unity. Why does this imply that the set $\{z^0, z^1, \dots, z^{n-1}\}$ coincides with the set E_n of **all** n th roots of unity?
- (5) (*) Show that there are exactly $\phi(n)$ roots of unity of order n which are primitive. Here ϕ denotes Euler's totient function introduced in Definition 2.7.1. Why does this imply that in the case of prime n all n th roots $z \neq 1$ are primitive?

¹²Primitive roots may be introduced and investigated in the much more general context of so-called cyclic groups. For example, the primitive roots in Definition 2.7.3 are those in the group (\mathbb{Z}_n^*, \cdot) while a root of unity is one in the group in $(E_n, *)$ or, equivalently, in view (6.4.12), in $(\mathbb{Z}_n, +)$.

Exercise 6.4.5. Let $z \neq 1$ be an n th root of unity. Show that

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

Exercise 6.4.6. Recall the following definition. If c_0, \dots, c_n are $n+1$ complex numbers with $c_n \neq 0$, then the function $p : \mathbb{C} \rightarrow \mathbb{C}$ with

$$p(z) = c_n z^n + \cdots + c_1 z + c_0, \quad z \in \mathbb{C},$$

is said to be a (complex) polynomial of degree n (written $\deg(p) = n$). Note that polynomials of degree 0 are constant functions, those of degree 1 are linear ones and that polynomials of degree 2 are called quadratic polynomials.

Prove the two following assertions: If p and q are two complex polynomials, then

$$(6.4.13) \quad \deg(p + q) \leq \max\{\deg(p), \deg(q)\} \text{ and}$$

$$(6.4.14) \quad \deg(p \cdot q) = \deg(p) + \deg(q).$$

Here the sum and the product of two polynomials are defined by

$$(p + q)(z) = p(z) + q(z) \text{ and } (p \cdot q)(z) = p(z) \cdot q(z), \quad z \in \mathbb{C}.$$

Give an example which shows that the left-hand side in (6.4.13) may be strictly smaller than the right-hand one.

Why does (6.4.14) imply $\deg(\alpha p) = \deg(p)$ for all complex $\alpha \neq 0$?

Exercise 6.4.7. (\star) Mimic the Euclidean division in Proposition 1.5.1 to prove the following result.

Proposition 6.4.8. For all nonzero complex polynomials p and q there are unique polynomials a and r such that

$$(6.4.15) \quad p(z) = q(z)a(z) + r(z), \quad z \in \mathbb{C},$$

and with $\deg(r) < \deg(q)$. The polynomial a is called quotient of the division of p by q while r is said to be the remainder of the division.

Hint: First observe that the result is trivially true if $\deg(p) < \deg(q)$. Why? So one may assume $\deg(p) \geq \deg(q)$. In a first step represent p as

$$p(z) = q(z)a_1(z) + r_1(z), \quad z \in \mathbb{C},$$

where $\deg(r_1) < \deg(p)$. If $\deg(r_1) < \deg(q)$, then (6.4.15) is valid with $a = a_1$ and with $r = r_1$. Otherwise, repeat the procedure with p replaced by r_1 . Proceed in this way until the degree of the remainder is less than the degree of q .

Exercise 6.4.8. Suppose the polynomials p and q are defined as follows:

$$p(z) = 3z^5 - 3z^4 - 2z^2 + z - 1 \quad \text{and} \quad q(z) = z^2 + 1, \quad z \in \mathbb{C}.$$

Find polynomials a and r with $\deg(r) < 2 = \deg(q)$ such that

$$p(z) = q(z)a(z) + r(z), \quad z \in \mathbb{C}.$$

Exercise 6.4.9. Use Proposition 6.4.8 to prove the following:

Proposition 6.4.9. Let p be a complex polynomial. Then the number $z_0 \in \mathbb{C}$ is a zero of p , i.e., $p(z_0) = 0$, if and only if there is a polynomial a such that

$$p(z) = (z - z_0)a(z), \quad z \in \mathbb{C}.$$

Equivalently one may say, if and only if the polynomial $q(z) = z - z_0$ divides p .

Exercise 6.4.10. In order to prove the fundamental theorem of algebra (Theorem 6.4.3) it suffices to prove the following:

Theorem 6.4.10. Each complex polynomial of degree greater than or equal 1 has at least one complex zero.

Justify why it is enough to verify this (apparently weaker) assertion.

Exercise 6.4.11. Use Theorem 6.4.3 to prove the following result about complex polynomials:

Proposition 6.4.11. Let p and q be two complex polynomials of degree less than or equal some $n \geq 1$. Suppose that there are $m > n$ different complex numbers z_1, \dots, z_m such that

$$p(z_1) = q(z_1), \dots, p(z_m) = q(z_m).$$

Then necessarily $p(z) = q(z)$ for all $z \in \mathbb{C}$.

Exercise 6.4.12. Argue why any polynomial with real coefficients and of odd degree has at least one real zero.

Exercise 6.4.13. Find the zeroes of the polynomial

$$p(z) = z^2 - 2z + i, \quad z \in \mathbb{C}.$$

Exercise 6.4.14. Determine all zeroes of the two degree 4 polynomials

$$p_1(z) = z^4 - 2z^2 + 1 \quad \text{and} \quad p_2(z) = z^4 + 2z^2 + 1.$$

Exercise 6.4.15. Describe the sequence $(x_n)_{n \geq 0}$ satisfying $x_0 = 2$, $x_1 = 0$ and

$$x_{n+1} = 2x_n - 2x_{n-1}, \quad n \geq 1.$$

Exercise 6.4.16. Let $n \geq 2$ be a natural number and denote by $\omega_0 = 1, \omega_1, \dots, \omega_{n-1}$ the n th roots of unity. Prove that

$$\sum_{k=0}^{n-1} \omega_k = 0 \quad \text{and} \quad \sum_{0 \leq j < k \leq n-1} \omega_j \omega_k = 0.$$

6.5. Geometric Applications

The geometric interpretation of complex numbers leads naturally to their use in geometry problems. In this section, we present some applications of complex numbers in this area.

Proposition 6.5.1. Let A, B , and C be three distinct points in the Euclidean plane whose coordinates are given by the complex numbers z_A, z_B , and z_C , respectively. The points A, B , and C are collinear if and only if $\frac{z_B - z_A}{z_C - z_A}$ is a real number.

Proof: Note that A, B , and C are collinear if and only if the points O, D, E are collinear, where D and E are the points corresponding to the complex numbers $z_B - z_A$ and $z_C - z_A$, respectively. The result follows by using Exercise 6.3.5. ■

Proposition 6.5.2. *The midpoints of the four sides of a quadrilateral form a parallelogram.*

Proof: Consider a quadrilateral whose vertices are A_1, A_2, A_3, A_4 whose coordinates correspond to the complex numbers z_1, z_2, z_3, z_4 , respectively. Whenever $1 \leq i < j \leq 4$, let M_{ij} denote the midpoint of the segment A_iA_j (see Figure 6.5.1). Each point M_{ij} corresponds to the complex number $\frac{z_i+z_j}{2}$. Hence, the midpoint of the segment $M_{12}M_{34}$ corresponds to the complex number $\frac{z_1+z_2+z_3+z_4}{4}$ and the same can be said about the midpoint of the segment $M_{23}M_{14}$. Therefore, these points are the same implying that $M_{12}M_{23}M_{34}M_{14}$ is a parallelogram. ■

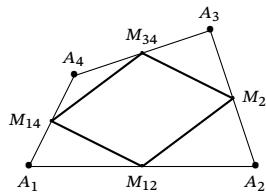


Figure 6.5.1. The midpoints of the sides of a quadrilateral.

Proposition 6.5.3. *Let A, B, C , and D be four distinct points. These points lie on a circle in the order A, B, C , and D if and only if $\frac{z_A-z_C}{z_A-z_D} \cdot \frac{z_C-z_D}{z_B-z_C}$ is a real nonnegative number.*

Proof: The measure $m(\angle BAD)$ of the angle $\angle BAD$ equals $\arg\left(\frac{z_B-z_A}{z_D-z_A}\right)$. The measure $m(\angle BCD)$ of the angle $\angle BCD$ equals $\arg\left(\frac{z_B-z_C}{z_D-z_C}\right)$.

If $\frac{z_A-z_C}{z_A-z_D} \cdot \frac{z_C-z_D}{z_B-z_C}$ is a real nonnegative number, then we can first rewrite it as $\frac{z_A-z_C}{z_A-z_D} \cdot \frac{z_C-z_D}{z_C-z_B} \cdot (-1)$. Using $\arg(-1) = \pi$, we deduce that

$$m(\angle BAD) + m(\angle BCD) + \pi = 2n\pi,$$

for some natural number n . The only possible solution is that

$$m(\angle BAD) + m(\angle BCD) = \pi.$$

This implies that A, B, C , and D are on a circle in this order. We leave the converse for the reader to complete. ■

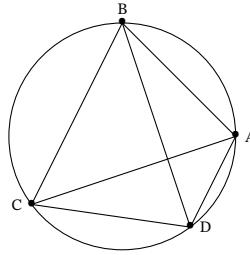


Figure 6.5.2. Four points on a circle.

The following result is known as the Ptolemy's inequality and is due to the Greek mathematician Ptolemy (100–170 CE).

Theorem 6.5.4. *If A, B, C , and D are four points in the Cartesian plane, then*

$$(6.5.1) \quad |AC| \cdot |BD| \leq |AB| \cdot |CD| + |AC| \cdot |BD|.$$

Equality happens if and only if A, B, C, D are on a circle in this order.

Proof: We recall Exercise 4.2.10 which states that

$$(x_1 - x_2)(x_3 - x_4) + (x_1 - x_3)(x_2 - x_4) + (x_1 - x_4)(x_3 - x_2) = 0,$$

for any real numbers x_1, x_2, x_3, x_4 . It turns out that this identity is true for complex numbers as well, and we leave the verification of this fact to the reader. Applying this result to the complex numbers associated to our four points, we get that

$$(z_C - z_A)(z_B - z_D) = (z_A - z_B)(z_C - z_D) + (z_A - z_D)(z_C - z_B).$$

Using the triangle inequality, it follows that

$$\begin{aligned} |(z_C - z_A)| \cdot |(z_B - z_D)| &= |(z_A - z_B)(z_C - z_D) + (z_A - z_D)(z_C - z_B)| \\ &\leq |z_A - z_B| \cdot |z_C - z_D| + |z_A - z_D| \cdot |z_C - z_B|. \end{aligned}$$

This proves the desired inequality.

Equality happens in equation 6.5.1 if and only if we have equality in the inequality above. This is the same as the ratio $\frac{(z_A - z_B)(z_C - z_D)}{(z_A - z_D)(z_C - z_B)}$ being real and nonnegative (see Theorem 6.2.3). Using Proposition 6.5.3, the result follows. ■

Exercise 6.5.1. Prove that for any two complex numbers $z, w \in \mathbb{C}$

$$|z + w|^2 + |z - w|^2 = 2(|z|^2 + |w|^2).$$

Interpret this result geometrically.

Exercise 6.5.2. Consider the three points in the plane corresponding to the complex numbers $2 + i$, $-3 + 2i$, and $-1 - i$. If these points are three vertices of a parallelogram, find the possible complex numbers corresponding to the fourth vertex.

Exercise 6.5.3. Let A, B , and C be three points. If $z_A \overline{z_B} + z_B \overline{z_C} + z_C \overline{z_A} = 0$, then prove that A, B , and C are collinear.

Exercise 6.5.4. Let A, B, C , and D be four distinct points. Show that the lines AB and CD are parallel if and only if $(z_A - z_B)\overline{z_C - z_D}$ is real which is also equivalent to $\frac{z_A - z_B}{z_C - z_D}$ being real.

Exercise 6.5.5. Let A and B be two points in the Cartesian plane with origin O . Show that OA and OB are perpendicular if and only if $z_A \overline{z_B}$ is purely imaginary which is equivalent to z_A/z_B being purely imaginary.

Exercise 6.5.6. Let A, B, C , and D be four distinct points. Show that the lines AB and CD are perpendicular if and only if $(z_A - z_B)\overline{z_C - z_D}$ is purely imaginary which is the same as $\frac{z_A - z_B}{z_C - z_D}$ being purely imaginary.

Exercise 6.5.7. Let A, B, C , and D be four distinct points. These points lie on a circle if and only if $\frac{z_A - z_C}{z_A - z_D} \cdot \frac{z_C - z_D}{z_B - z_C}$ is a real number.

Exercise 6.5.8. Let A be a point in the Cartesian plane and $\theta \in \mathbb{R}$. Prove that the point corresponding to the complex number $z_A e^{i\theta}$ is obtained from A by a counterclockwise rotation with angle θ around the origin.

Exercise 6.5.9. Let A, B , and C be three points in the Cartesian plane such that the triangle ABC is equilateral. Prove that $z_A = z_B + \omega(z_C - z_B)$, where $\omega = e^{2\pi i/3}$.

Exercise 6.5.10. Let A, B , and C be three points in the Cartesian plane. Show that the triangle ABC is equilateral if and only if

$$z_A + \omega z_B + \omega^2 z_C = 0,$$

where $\omega = e^{\frac{2\pi i}{3}} = \cos\left(\frac{2\pi}{3}\right) + i \sin\left(\frac{2\pi}{3}\right)$ is a third root of unity.

Exercise 6.5.11. Give a geometric proof of the Ptolemy's theorem.

Exercise 6.5.12. Let A, B , and C be three points, not all on a line. The **altitude** AA' from A in the triangle ABC is the line segment joining A to the point A' on the line BC such that the line supporting AA' and the line BC are perpendicular or orthogonal. The altitudes BB' and CC' can be defined similarly. Prove that the altitudes AA' , BB' , and CC' intersect in one point H which is called the **orthocenter** of the triangle ABC .

Exercise 6.5.13. Prove another result of Ptolemy, namely given four distinct points A, B, C , and D that are lying on the same circle in this order, then

$$(6.5.2) \quad \frac{|AC|}{|BD|} = \frac{|AB| \cdot |AD| + |CB| \cdot |CD|}{|BA| \cdot |BC| + |DA| \cdot |DC|}.$$

6.6. Sequences of Complex Numbers

Definition 6.6.1. A sequence $(z_n)_{n \geq 1}$ of complex numbers is a function $z : \mathbb{N} \rightarrow \mathbb{C}$, where, as in the real case, one writes z_n instead of $z(n)$, i.e.,

$$z_n = z(n), \quad n = 1, 2, \dots$$

In this sense, z_1, z_2, \dots are the values of z in its natural order.

$$(z_n)_{n \geq 1} \text{ sequence of complex numbers} \Leftrightarrow (\exists z : \mathbb{N} \rightarrow \mathbb{C})(\forall n \in \mathbb{N})(z(n) = z_n).$$

Example 6.6.1. The complex sequence $i, -\frac{1}{2}, -\frac{i}{3}, \frac{1}{4}, \frac{i}{5}, -\frac{1}{6}, \dots$ is defined by the function $z : \mathbb{N} \rightarrow \mathbb{C}$ with $z(n) = \frac{i^n}{n}$. If $z : \mathbb{N} \rightarrow \mathbb{C}$ is given by $z(n) = \frac{1}{(1+i)^n}$, then the sequence of complex numbers is $\frac{1}{1+i}, \frac{1}{(1+i)^2}, \dots$

Next we investigate the convergence of complex sequences.

Definition 6.6.2. Let $(z_n)_{n \geq 1}$ be a sequence of complex numbers. Then $(z_n)_{n \geq 1}$ converges to some $z \in \mathbb{C}$ if

$$(6.6.1) \quad \lim_{n \rightarrow \infty} |z - z_n| = 0.$$

If this happens, we write $z_n \xrightarrow{n \rightarrow \infty} z$ or $\lim_{n \rightarrow \infty} z_n = z$.

A sequence $(z_n)_{n \geq 1}$ of complex numbers is said to be **convergent** provided that there is some $z \in \mathbb{C}$ with $\lim_{n \rightarrow \infty} z_n = z$. A sequence that is not convergent is called **divergent**.

Remark 6.6.1. By the definition of limits in \mathbb{R} , condition (6.6.1) may be reformulated as follows.

$$\lim_{n \rightarrow \infty} z_n = z \Leftrightarrow (\forall \varepsilon > 0)(\exists N = N(\varepsilon))(\forall n \geq N)(|z - z_n| < \varepsilon).$$

What does this mean? Let $U_\varepsilon(z) := \{w \in \mathbb{C} : |z - w| < \varepsilon\}$ be an ε -neighborhood of $z \in \mathbb{C}$. That is, $U_\varepsilon(z) \subseteq \mathbb{C}$ is an open sphere of radius $\varepsilon > 0$ centered at $z \in \mathbb{C}$. In this setting we have $z_n \xrightarrow{n \rightarrow \infty} z$ if and only if there is an integer N such that $z_n \in U_\varepsilon(z)$ for all $n \geq N$. And this has to be satisfied for all $\varepsilon > 0$. Compare Figure 6.6.1.

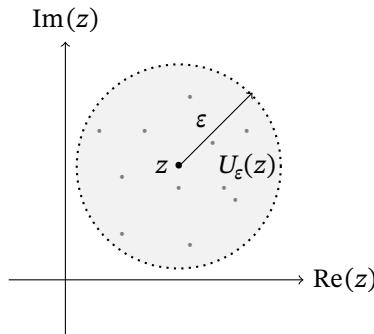


Figure 6.6.1. Convergence of a sequence to $z \in \mathbb{C}$.

The following result shows that the convergence of complex sequences can be reduced to the case of real ones.

Proposition 6.6.1. A sequence $(z_n)_{n \geq 1}$ of complex numbers is convergent if and only if $(\operatorname{Re}(z_n))_{n \geq 1}$ and $(\operatorname{Im}(z_n))_{n \geq 1}$ are convergent (real) sequences. Moreover, then

$$(\lim_{n \rightarrow \infty} z_n = z) \Leftrightarrow (\lim_{n \rightarrow \infty} \operatorname{Re}(z_n) = \operatorname{Re}(z) \text{ and } \lim_{n \rightarrow \infty} \operatorname{Im}(z_n) = \operatorname{Im}(z)).$$

Proof: Suppose first $\lim_{n \rightarrow \infty} z_n = z$. If $z_n = a_n + b_n i$ and if $z = a + bi$, by (3) in Proposition 6.2.1 we get

$$0 \leq |a - a_n| = |\operatorname{Re}(z - z_n)| \leq |z - z_n| \xrightarrow{n \rightarrow \infty} 0.$$

Hence, by the sandwich theorem this implies $|a - a_n| \xrightarrow{n \rightarrow \infty} 0$, i.e., $\lim_{n \rightarrow \infty} a_n = a$. The proof for the imaginary part follows exactly in the same way.

To prove the converse implication suppose that the real parts $a_n = \operatorname{Re}(z_n)$ and the imaginary parts $b_n = \operatorname{Im}(z_n)$ converge to some $a \in \mathbb{R}$ and some $b \in \mathbb{R}$, respectively. Consequently, given $\varepsilon > 0$ there are integers N_1 and N_2 such that

$$|a - a_n| < \frac{\varepsilon}{\sqrt{2}} \quad \text{if } n \geq N_1 \quad \text{and} \quad |b - b_n| < \frac{\varepsilon}{\sqrt{2}} \quad \text{if } n \geq N_2.$$

Set $N = \max\{N_1, N_2\}$. Then, if $n \geq N$, both estimates are satisfied, and we conclude that

$$|z - z_n| = \sqrt{|a - a_n|^2 + |b - b_n|^2} < \sqrt{\frac{\varepsilon^2}{2} + \frac{\varepsilon^2}{2}} = \varepsilon.$$

This being true for all $\varepsilon > 0$ means that, as asserted, $\lim_{n \rightarrow \infty} z_n = z$. Thus, the proof is completed. \blacksquare

Corollary 6.6.2. *The limit of a convergent sequence in \mathbb{C} is uniquely determined.*

Proof: Suppose the sequence $z_n = a_n + b_n i$ converges to $z = a + bi$ and at the same time to $z' = a' + b'i$. By Proposition 6.6.1 this implies $a_n \xrightarrow{n \rightarrow \infty} a$ as well as $a_n \xrightarrow{n \rightarrow \infty} a'$. Since the limit of convergent real sequences is unique, we conclude that $a = a'$. The same arguments lead to $b = b'$, hence $z = z'$ and the limit is unique. \blacksquare

Proposition 6.6.1 allows us to transfer the main properties of convergent sequences from the real to the complex case.

Proposition 6.6.3. *Let $(w_n)_{n \geq 1}$ and $(z_n)_{n \geq 1}$ be two sequences of complex numbers. Then the following are valid.*

- (1) $z_n \xrightarrow{n \rightarrow \infty} z$ implies $\bar{z}_n \xrightarrow{n \rightarrow \infty} \bar{z}$ and $|z_n| \xrightarrow{n \rightarrow \infty} |z|$.
- (2) If $c, d \in \mathbb{C}$, $w_n \xrightarrow{n \rightarrow \infty} w$ and $z_n \xrightarrow{n \rightarrow \infty} z$, then $\lim_{n \rightarrow \infty} (cw_n + dz_n) = cw + dz$.
- (3) $w_n \xrightarrow{n \rightarrow \infty} w$ and $z_n \xrightarrow{n \rightarrow \infty} z$ leads to $\lim_{n \rightarrow \infty} (w_n z_n) = w \cdot z$.
- (4) $w_n \xrightarrow{n \rightarrow \infty} w$ and $z_n \xrightarrow{n \rightarrow \infty} z$ with $z \neq 0$ implies $\lim_{n \rightarrow \infty} \frac{w_n}{z_n} = \frac{w}{z}$.

Proof: (1) Since $\operatorname{Re}(\bar{z}_n) = \operatorname{Re}(z_n)$ and $\operatorname{Im}(\bar{z}_n) = -\operatorname{Im}(z_n)$, Proposition 6.6.1 easily implies that $\bar{z}_n \xrightarrow{n \rightarrow \infty} \bar{z}$ whenever $z_n \xrightarrow{n \rightarrow \infty} z$.

To verify the second part of (1) we use inequality (6.2.3). Then we get

$$0 \leq ||z| - |z_n|| \leq |z - z_n| \xrightarrow{n \rightarrow \infty} 0.$$

This proves $|z_n| \xrightarrow{n \rightarrow \infty} |z|$ as asserted.

As already mentioned, via Proposition 6.6.1 all remaining results follow easily from the corresponding properties for real sequences. Let us shortly sketch how this is done to prove (3) and (4).

(3) Because of $\operatorname{Re}(w_n z_n) = \operatorname{Re}(w_n) \operatorname{Re}(z_n) - \operatorname{Im}(w_n) \operatorname{Im}(z_n)$ the convergence of the real and imaginary parts of the w_n 's and z_n 's implies the convergence of the real parts of $w_n \cdot z_n$. A similar argument proves the convergence of the imaginary parts of $w_n \cdot z_n$, hence, by Proposition 6.6.1 $\lim_{n \rightarrow \infty} (w_n z_n)$ exists and, moreover, as can be seen easily, equals $[\lim_{n \rightarrow \infty} w_n] \cdot [\lim_{n \rightarrow \infty} z_n]$.

(4) In view of property (3) it suffices to prove $\lim_{n \rightarrow \infty} (z_n^{-1}) = \frac{1}{z}$ whenever $z_n \xrightarrow{n \rightarrow \infty} z$ and $z \neq 0$. Suppose $z_n = a_n + b_n i$ and $z = a + bi$. Then, because of $a_n \xrightarrow{n \rightarrow \infty} a$, $b_n \xrightarrow{n \rightarrow \infty} b$ and $|z_n| \xrightarrow{n \rightarrow \infty} |z|$, by $|z| \neq 0$ we obtain

$$\frac{1}{z_n} = \frac{\bar{z}_n}{|z_n|^2} = \frac{a_n}{|z_n|^2} - \frac{b_n}{|z_n|^2} i \xrightarrow{n \rightarrow \infty} \frac{a}{|z|^2} - \frac{b}{|z|^2} i = \frac{1}{z}.$$

This completes the proof of (4). ■

Example 6.6.2.

(1) Let $z \in \mathbb{C}$ be a fixed complex number with $|z| < 1$. We claim that then $z^n \xrightarrow{n \rightarrow \infty} 0$.

This follows by $|z^n| = |z|^n$ and by $q^n \xrightarrow{n \rightarrow \infty} 0$ for any $0 \leq q < 1$.

(2) If $z_n = \frac{1}{n}(1 + i^n)$, then $|z_n| = \frac{\sqrt{2}}{n} \xrightarrow{n \rightarrow \infty} 0$, hence $z_n \xrightarrow{n \rightarrow \infty} 0$.

(3) Let $z_n = \frac{1+i}{1+ni}$. Does $(z_n)_{n \geq 1}$ converge in \mathbb{C} ? Since

$$|z_n| = \frac{|1+i|}{|1+ni|} = \frac{\sqrt{2}}{\sqrt{1+n^2}} \xrightarrow{n \rightarrow \infty} 0,$$

the sequence $(z_n)_{n \geq 1}$ converges to zero.

(4) Let us consider the sequence $(z_n)_{n \geq 1}$ with $z_n = 1 + \frac{1}{n} + \frac{(1+n)i}{n}$. Then the real parts $\operatorname{Re}(z_n) = 1 + \frac{1}{n}$ converge to 1 and because of $\operatorname{Im}(z_n) = \frac{1+n}{n} \xrightarrow{n \rightarrow \infty} 1$, it follows that $z_n \xrightarrow{n \rightarrow \infty} 1 + i$.

Remark 6.6.2. Although almost all results about convergent sequences transfer from the real to the complex case, at least two significant differences exist.

- (1) There is no analogue of sandwich theorem for complex sequences. The reason is that there is no order on \mathbb{C} compatible with the algebraic operations. The only possible statement is that $z_n \xrightarrow{n \rightarrow \infty} 0$ and $|w_n| \leq |z_n|$ imply $w_n \xrightarrow{n \rightarrow \infty} 0$. But this is obvious and mostly not very helpful.
- (2) For sequences in \mathbb{C} there is no direct way to define the proper divergence. There exists some substitute, namely if $|z_n| \xrightarrow{n \rightarrow \infty} \infty$, but this does not give any information about the behavior of the real and/or the imaginary parts of the sequence. Compare also Exercise 6.6.1 below.

Finally, we ask for conditions which ensure that a given sequence in \mathbb{C} converges.

Definition 6.6.3. A sequence $(z_n)_{n \geq 1}$ of complex numbers is said to be a **Cauchy sequence** if it satisfies the following condition: For each $\varepsilon > 0$ there is an integer $N = N(\varepsilon)$ such that for all $n, m \geq N$ follows that

$$|z_n - z_m| < \varepsilon.$$

$$(z_n)_{n \geq 1} \text{ Cauchy sequence} \Leftrightarrow (\forall \varepsilon > 0)(\exists N = N(\varepsilon))(\forall n, m \geq N)(|z_n - z_m| < \varepsilon).$$

Theorem 6.6.4. *A sequence of complex numbers is convergent if and only if it is a Cauchy sequence.*

Proof: Suppose first that $(z_n)_{n \geq 1}$ is a convergent sequence of complex numbers. Using exactly the same arguments as in the real case (in fact we used only the triangle inequality in \mathbb{R} , which, as we saw, is also valid in \mathbb{C}), it follows that $(z_n)_{n \geq 1}$ is a Cauchy sequence.

Choose now an arbitrary Cauchy sequence $(z_n)_{n \geq 1}$ in \mathbb{C} . If we can show that then the real $(a_n)_{n \geq 1}$ and the imaginary parts $(b_n)_{n \geq 1}$ of the z_n 's are Cauchy sequences (in \mathbb{R}), then we are done. Indeed, then by Theorem 5.5.4 the a_n 's and b_n 's converge, which by Proposition 6.6.1 implies the convergence of $(z_n)_{n \geq 1}$.

Thus, it remains to investigate the real sequences $(a_n)_{n \geq 1}$ and $(b_n)_{n \geq 1}$. By assumption, given $\varepsilon > 0$, there is an integer N for which $|z_n - z_m| < \varepsilon$ provided that $n, m \geq N$. But by estimate (3) in Proposition 6.2.1 this implies

$$|a_n - a_m| = |\operatorname{Re}(z_n - z_m)| \leq |z_n - z_m| < \varepsilon$$

whenever $n, m \geq N$. This being true for all $\varepsilon > 0$, the sequence $(a_n)_{n \geq 1}$ is a Cauchy sequence in \mathbb{R} . In the same way one shows that $(b_n)_{n \geq 1}$ is a Cauchy sequence, and as we observed above, this completes the proof. ■

Exercise 6.6.1. Let $z_n = a_n + ib_n$, $n \geq 1$, be a sequence of complex numbers such that

$$\lim_{n \rightarrow \infty} |z_n| = \infty.$$

Does this imply that at least one of the sequences $(a_n)_{n \geq 1}$ or $(b_n)_{n \geq 1}$ converges properly? Give a proof or a counterexample.

Exercise 6.6.2. Define $z_n = \left(\frac{1}{\sqrt{2}} + i \frac{1}{\sqrt{2}} \right)^n$ for $n \in \mathbb{N}$. Prove that the sequence $(z_n)_{n \geq 1}$ is divergent.

Exercise 6.6.3. Give examples of a divergent sequence $(z_n)_{n \geq 1}$ of complex numbers such that $(|z_n|)_{n \geq 1}$ is a convergent nonconstant sequence.

Exercise 6.6.4. Give examples of a divergent sequence $(z_n)_{n \geq 1}$ of complex numbers such that $(\arg(z_n))_{n \geq 1}$ is a convergent sequence.

Exercise 6.6.5. Let $(z_n)_{n \geq 1}$ be a convergent sequence of complex numbers. Prove that the sequence $(z_n)_{n \geq 1}$ is bounded.

Exercise 6.6.6. Show that if $(z_n)_{n \geq 1}$ is a bounded sequence of complex numbers, then it has a convergent subsequence.

Exercise 6.6.7. Let z_n and z be nonzero complex numbers written in polar representation as

$$z_n = r_n(\cos \theta_n + i \sin \theta_n) \quad \text{and} \quad z = r(\cos \theta + i \sin \theta)$$

for some $r_n, r > 0$ and $0 \leq \theta_n, \theta < 2\pi$. Prove that

$$(6.6.2) \quad \lim_{n \rightarrow \infty} r_n = r \text{ and } \lim_{n \rightarrow \infty} \theta_n = \theta \Rightarrow \lim_{n \rightarrow \infty} z_n = z.$$

Hint: Use that sine and cosine are continuous functions. That is, for all sequences $(x_n)_{n \geq 1}$ of real numbers one has

$$\lim_{n \rightarrow \infty} x_n = x \Rightarrow \lim_{n \rightarrow \infty} \sin x_n = \sin x \text{ and } \lim_{n \rightarrow \infty} \cos x_n = \cos x.$$

Give an example that shows that the converse implication in (6.6.2) is not valid. Although (prove this)

$$\lim_{n \rightarrow \infty} z_n = z \Rightarrow \lim_{n \rightarrow \infty} r_n = r,$$

in general we do **not** have

$$\lim_{n \rightarrow \infty} z_n = z \Rightarrow \lim_{n \rightarrow \infty} \theta_n = \theta.$$

Exercise 6.6.8. Check whether $\lim_{n \rightarrow \infty} \frac{2+i}{1+ni}$ exists. If so, find it.

Exercise 6.6.9. Determine whether $\lim_{n \rightarrow \infty} \frac{(1+in)^4 - n^4}{(1+in)^3}$ exists and if it does, then calculate it.

Exercise 6.6.10. Same question as above for $\lim_{n \rightarrow \infty} \frac{1-n^2i}{(1+ni)^2}$.

Exercise 6.6.11. Let $(w_n)_{n \geq 1}$ and $(z_n)_{n \geq 1}$ be two sequences of complex numbers with

$$\lim_{n \rightarrow \infty} [w_n - z_n] = 0.$$

Why does $w_n \xrightarrow[n \rightarrow \infty]{} w$ imply $z_n \xrightarrow[n \rightarrow \infty]{} w$?

6.7. Infinite Series of Complex Numbers

Definition 6.7.1. Let $(z_n)_{n \geq 1}$ be an infinite sequence of complex numbers. Its *n*th **partial sum** s_n is defined by

$$s_n = z_1 + \cdots + z_n, \quad n \geq 1.$$

Then $(z_n)_{n \geq 1}$ is said to be a **summable** sequence provided that $\lim_{n \rightarrow \infty} s_n$ exists in \mathbb{C} . The limit is denoted by

$$\sum_{n=1}^{\infty} z_n := \lim_{n \rightarrow \infty} s_n = \lim_{n \rightarrow \infty} \sum_{j=1}^n z_j.$$

If $z_n = a_n + b_ni$ for real numbers a_n and b_n , then the partial sums can be written as

$$s_n = u_n + v_ni \quad \text{where} \quad u_n = a_1 + \cdots + a_n \text{ and } v_n = b_1 + \cdots + b_n.$$

Consequently, Proposition 6.6.1 implies the following result.

Proposition 6.7.1. Let $z_n = a_n + b_n i$ be a sequence of complex numbers. Then $(z_n)_{n \geq 1}$ is summable if and only if $\sum_{n=1}^{\infty} a_n$ and $\sum_{n=1}^{\infty} b_n$ exist. Moreover, then

$$\sum_{n=1}^{\infty} z_n = \sum_{n=1}^{\infty} a_n + i \sum_{n=1}^{\infty} b_n.$$

A very useful criterion for the summability of a sequence is the following consequence of Theorem 6.6.4.

Proposition 6.7.2 (Cauchy Criterion). A sequence $(z_n)_{n \geq 1}$ of complex numbers is summable if and only if for each $\varepsilon > 0$ there exists an integer $N \geq 1$ such that

$$\left| \sum_{k=n+1}^m z_k \right| < \varepsilon,$$

whenever $N < n < m$.

Definition 6.7.2. A sequence $(z_n)_{n \geq 1}$ of complex numbers is said to be **absolutely summable** if $\sum_{n=1}^{\infty} |z_n| < \infty$.

Proposition 6.7.3. Each absolutely summable sequence of complex numbers is summable.

Proof: Let $z_n = a_n + b_n i$ be an absolutely summable sequence. By estimate (3) in Proposition 6.2.1 get $|a_n| \leq |z_n|$ and $|b_n| \leq |z_n|$. Hence

$$\sum_{n=1}^{\infty} |a_n| \leq \sum_{n=1}^{\infty} |z_n| < \infty \quad \text{and} \quad \sum_{n=1}^{\infty} |b_n| \leq \sum_{n=1}^{\infty} |z_n| < \infty.$$

Thus, $(a_n)_{n \geq 1}$ and $(b_n)_{n \geq 1}$ are both absolutely summable sequences of real numbers. As shown in Proposition 5.6.7 they are also summable. Finally, Proposition 6.7.1 implies that $(z_n)_{n \geq 1}$ is summable, which completes the proof.

Let us give still another proof of the assertion. Since $\sum_{n=1}^{\infty} |z_n| < \infty$, by Proposition 5.6.3 the following is satisfied: For each $\varepsilon > 0$ there is an $N \geq 1$ such that

$$\sum_{k=n+1}^m |z_k| < \varepsilon,$$

whenever $N < n < m$. Now, the triangle inequality for complex numbers yields

$$\left| \sum_{k=n+1}^m z_k \right| \leq \sum_{k=n+1}^m |z_k| < \varepsilon$$

as soon as $N < n < m$. Hence, by Proposition 6.7.2 the sequence $(z_n)_{n \geq 1}$ is summable. This completes the alternative proof of the assertion. \blacksquare

Remark 6.7.1. As in the real case there exist summable sequences which are not absolutely summable. For example, the sequence $(i^n/n)_{n \geq 1}$ (see Exercise 6.7.1) is summable, but

$$\sum_{n=1}^{\infty} \left| \frac{i^n}{n} \right| = \sum_{n=1}^{\infty} \frac{1}{n} = \infty.$$

Example 6.7.1. Let $z \in \mathbb{C}$ be a complex number with $|z| < 1$. Example 6.6.2 implies $z^n \xrightarrow[n \rightarrow \infty]{} 0$, hence by (6.1.6) it follows that

$$s_n = 1 + z + z^2 + \cdots + z^n = \frac{z^{n+1} - 1}{z - 1} \xrightarrow[n \rightarrow \infty]{} \frac{1}{1 - z}.$$

In other words, if $|z| < 1$, then

$$\sum_{n=0}^{\infty} z^n = \frac{1}{1 - z}.$$

In particular, if we apply this formula with $z = it$ for some real number t , $|t| < 1$, then we obtain

$$\begin{aligned} \frac{1}{1 - it} &= \sum_{n=0}^{\infty} (it)^n = 1 + it - t^2 - it^3 + t^4 + it^5 \pm \cdots \\ &= \sum_{n=0}^{\infty} (-1)^n t^{2n} + i \sum_{n=0}^{\infty} (-1)^n t^{2n+1}. \end{aligned}$$

Example 6.7.2. We claim that for any $z \in \mathbb{C}$ the sum

$$\sum_{n=0}^{\infty} \frac{z^n}{n!} \text{ exists.}$$

By Proposition 5.6.10 we know that for any $x \in \mathbb{R}$ the sequence $x^n/n!$ is summable. Hence, in view of $|z^n/n!| = |z|^n/n!$, with $x = |z|$ this implies

$$\sum_{n=0}^{\infty} \left| \frac{z^n}{n!} \right| = \sum_{n=0}^{\infty} \frac{|z|^n}{n!} = \sum_{n=0}^{\infty} \frac{x^n}{n!} < \infty.$$

Consequently, the sequence $(z_n)_{n \geq 0}$ is absolutely summable, hence summable as claimed.

Our final goal is to investigate the product of two infinite series of complex numbers. So suppose we are given two summable sequences $(w_n)_{n \geq 0}$ and $(z_n)_{n \geq 0}$ of complex numbers. Let

$$(6.7.1) \quad s_n = \sum_{k=0}^n w_k \quad \text{and} \quad t_n = \sum_{k=0}^n z_k$$

be their partial sums. They converge to

$$\sum_{n=0}^{\infty} w_n \quad \text{and} \quad \sum_{n=0}^{\infty} z_n,$$

respectively. By Proposition 6.6.3 this implies that

$$\lim_{n \rightarrow \infty} (t_n \cdot s_n) = \left(\sum_{n=0}^{\infty} w_n \right) \cdot \left(\sum_{n=0}^{\infty} z_n \right).$$

Now, the following is very likely, but, unfortunately, wrong in general. Instead of summing over all pairs of integers (k, ℓ) in $\mathbb{N}_0 \times \mathbb{N}_0$, in a first step we sum the entries along

all diagonals from $(0, n)$ to $(n, 0)$, and then we take the sum over all $n \geq 0$. In other words, the hope is that we may rewrite the product of the two infinite sums as follows:

$$\left(\sum_{n=0}^{\infty} w_n \right) \cdot \left(\sum_{n=0}^{\infty} z_n \right) = \sum_{(k,\ell) \in \mathbb{N}_0 \times \mathbb{N}_0} w_k z_\ell = \sum_{n=0}^{\infty} \sum_{k+\ell=n} w_k z_\ell = \sum_{n=0}^{\infty} \sum_{k=0}^n w_k z_{n-k}.$$

As already said, in general it is not possible to change the order of summation as suggested above. Fortunately, the previous calculations are correct in the case that both sums converge absolutely.

Theorem 6.7.4. *Let $(w_n)_{n \geq 0}$ and $(z_n)_{n \geq 0}$ be two **absolutely summable** sequences of complex numbers. For each $n \geq 0$ let*

$$c_n := \sum_{k=0}^n w_k z_{n-k}$$

be the sum of $w_k z_\ell$ along the diagonal starting at $(0, n)$ and terminating at $(n, 0)$. Then the sequence $(c_n)_{n \geq 0}$ is absolutely summable and, moreover,

$$\sum_{n=0}^{\infty} c_n = \sum_{n=0}^{\infty} \sum_{k=0}^n w_k z_{n-k} = \left[\sum_{n=0}^{\infty} w_n \right] \cdot \left[\sum_{n=0}^{\infty} z_n \right].$$

Proof: Let us introduce the following sets of pairs of integers:

$$A_n := \{(k, \ell) \in \mathbb{N}_0 \times \mathbb{N}_0 : 0 \leq k, \ell \leq n\} \quad \text{and} \quad B_n := \{(k, \ell) \in \mathbb{N}_0 \times \mathbb{N}_0 : 0 \leq k + \ell \leq n\}.$$

It is immediately clear (see Figure 6.7.1) that these sets are related by

$$(6.7.2) \quad A_{\lfloor n/2 \rfloor} \subseteq B_n \subseteq A_n \quad \text{and}$$

$$(6.7.3) \quad A_n \setminus A_{\lfloor n/2 \rfloor} \subseteq \{(k, \ell) : \ell \leq n, \lfloor n/2 \rfloor < k \leq n \text{ or } k \leq n, \lfloor n/2 \rfloor < \ell \leq n\}.$$

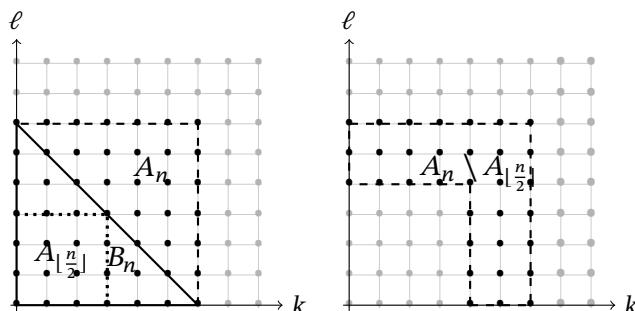


Figure 6.7.1. $A_n := \{(k, \ell) : 0 \leq k, \ell \leq n\}$, $B_n := \{(k, \ell) : 0 \leq k + \ell \leq n\}$.

Moreover, if the s_n and t_n are defined as in (6.7.1), then by the construction of A_n and B_n we get

$$(6.7.4) \quad \sum_{(k,\ell) \in A_n} w_k z_\ell = s_n t_n \quad \text{while} \quad \sum_{(k,\ell) \in B_n} w_k z_\ell = u_n,$$

where u_n denotes the n th partial sum of the c_n 's, i.e.,

$$u_n = \sum_{k=0}^n c_k = \sum_{k=0}^n \sum_{\ell=0}^k w_\ell z_{k-\ell}.$$

The triangle inequality for complex numbers combined with properties (6.7.4), (6.7.2) and (6.7.3) lets us conclude that

$$\begin{aligned} |s_n t_n - u_n| &= \left| \sum_{(k,\ell) \in A_n} w_k z_\ell - \sum_{(k,\ell) \in B_n} w_k z_\ell \right| = \left| \sum_{(k,\ell) \in A_n \setminus B_n} w_k z_\ell \right| \\ &\leq \sum_{(k,\ell) \in A_n \setminus B_n} |w_k| |z_\ell| \leq \sum_{(k,\ell) \in A_n \setminus A_{\lfloor n/2 \rfloor}} |w_k| |z_\ell| \\ &\leq \left(\sum_{k=\lfloor n/2 \rfloor + 1}^n |w_k| \right) \cdot \left(\sum_{\ell=0}^n |z_\ell| \right) + \left(\sum_{k=0}^{\lfloor n/2 \rfloor} |w_k| \right) \cdot \left(\sum_{\ell=\lfloor n/2 \rfloor + 1}^n |z_\ell| \right) \\ (6.7.5) \quad &\leq \left(\sum_{k=\lfloor n/2 \rfloor + 1}^n |w_k| \right) \cdot T + \left(\sum_{\ell=\lfloor n/2 \rfloor + 1}^n |z_\ell| \right) \cdot S \end{aligned}$$

where

$$S = \sum_{n=0}^{\infty} |w_n| < \infty \quad \text{and} \quad T = \sum_{n=0}^{\infty} |z_n| < \infty.$$

Since both, $(w_n)_{n \geq 0}$ and $(z_n)_{n \geq 0}$ are assumed to be absolutely summable, given $\varepsilon > 0$, by Proposition 6.7.2 there exists an integer $N \geq 1$ such that

$$\left(\sum_{k=\lfloor n/2 \rfloor + 1}^n |w_k| \right) \cdot T < \frac{\varepsilon}{2} \quad \text{as well as} \quad \left(\sum_{\ell=\lfloor n/2 \rfloor + 1}^n |z_\ell| \right) \cdot S < \frac{\varepsilon}{2}$$

whenever $\lfloor n/2 \rfloor > N$, which surely is satisfied if $n > 2N + 2$. Thus, for those integers n estimate 6.7.5 implies

$$|s_n t_n - u_n| \leq \left(\sum_{k=\lfloor n/2 \rfloor + 1}^n |w_k| \right) \cdot T + \left(\sum_{\ell=\lfloor n/2 \rfloor + 1}^n |z_\ell| \right) \cdot S < \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon.$$

Since $\varepsilon > 0$ was arbitrary, it follows that $\lim_{n \rightarrow \infty} [s_n t_n - u_n] = 0$, hence, as asserted,

$$\left(\sum_{n=0}^{\infty} w_n \right) \cdot \left(\sum_{n=0}^{\infty} z_n \right) = (\lim_{n \rightarrow \infty} s_n) \cdot (\lim_{n \rightarrow \infty} t_n) = \lim_{n \rightarrow \infty} [s_n t_n] = \lim_{n \rightarrow \infty} u_n = \sum_{n=0}^{\infty} c_n.$$

It remains to prove that the c_n 's are absolutely summable. Using

$$|c_k| = \left| \sum_{j=0}^k w_j z_{k-j} \right| \leq \sum_{j=0}^k |w_j| |z_{k-j}|,$$

this follows from

$$\begin{aligned} \sum_{k=0}^n |c_k| &\leq \sum_{(k,\ell) \in B_n} |w_k||z_\ell| \leq \sum_{(k,\ell) \in A_n} |w_k||z_\ell| \\ &= \left[\sum_{k=0}^n |w_k| \right] \left[\sum_{\ell=0}^n |z_\ell| \right] \leq \left[\sum_{k=0}^{\infty} |w_k| \right] \left[\sum_{\ell=0}^{\infty} |z_\ell| \right] = S \cdot T < \infty \end{aligned}$$

for all $n \geq 1$. ■

Definition 6.7.3. The infinite sum

$$\sum_{n=0}^{\infty} c_n \quad \text{with} \quad c_n = \sum_{k=0}^n w_k z_{n-k}$$

is called the **Cauchy product** of the two series

$$\sum_{n=0}^{\infty} w_n \quad \text{and} \quad \sum_{n=0}^{\infty} z_n .$$

Example 6.7.3. The aim of this example is showing that Theorem 6.7.4 is false for arbitrary summable sequences. Choose

$$x_n = y_n = \frac{(-1)^n}{\sqrt{n+1}}, \quad n = 0, 1, 2, \dots$$

Due to the Leibniz test, Proposition 5.6.16, both sequences are summable. Summing their product over the n th diagonal gives in that case

$$c_n = \sum_{k=0}^n x_k y_{n-k} = \sum_{k=0}^n \frac{(-1)^k}{\sqrt{k+1}} \cdot \frac{(-1)^{n-k}}{\sqrt{n-k+1}} = (-1)^n \sum_{k=0}^n \frac{1}{\sqrt{k+1} \cdot \sqrt{n-k+1}} .$$

We claim now that $|c_n| \geq 1$ for all $n \geq 0$. To verify this we use the estimate stated in Proposition 1.1.4. If we take the square root of this inequality and invert it, we get for all positive numbers a and b that

$$(6.7.6) \quad \frac{1}{\sqrt{ab}} \geq \frac{2}{a+b} .$$

An application of (6.7.6) with $a = k+1$ and $b = n-k+1$, where $k = 0, \dots, n$, implies

$$\frac{1}{\sqrt{k+1} \cdot \sqrt{n-k+1}} \geq \frac{2}{n+2} ,$$

hence

$$|c_n| = |(-1)^n| \sum_{k=0}^n \frac{1}{\sqrt{(k+1)(n-k+1)}} \geq \sum_{k=0}^n \frac{2}{n+2} = 2 \frac{n+1}{n+2} \geq 1 .$$

So we got $|c_n| \geq 1$ for all $n \geq 0$. Corollary 5.6.5 implies now that the sequence $(c_n)_{n \geq 0}$ of the sums over the diagonals cannot be summable. So we found two summable sequences for which their Cauchy product does not exist.

Remark 6.7.2. The deeper reason for the phenomenon in the preceding example is that the sequences $(x_n)_{n \geq 0}$ and $(y_n)_{n \geq 0}$ are both not absolutely summable. Let us mention that by refined methods one may show that Theorem 6.7.4 remains valid under the (weaker) condition that at least one of the two sequences is absolutely summable.

Example 6.7.4. Suppose $w_n = z_n = z^n$ for some complex z with $|z| < 1$. Then

$$c_n = \sum_{k=0}^n z^k z^{n-k} = \sum_{k=0}^n z^n = (n+1) \cdot z^n.$$

Consequently, an application of Example 6.7.1 leads to

$$\sum_{n=0}^{\infty} (n+1) z^n = \left(\sum_{n=0}^{\infty} z^n \right)^2 = \frac{1}{(1-z)^2}.$$

In Example 6.7.2 we proved that the infinite sum $\sum_{n=0}^{\infty} z^n/n!$ exists for all $z \in \mathbb{C}$. In accordance with Proposition 5.6.14 we define

$$e^z = \exp(z) := \sum_{n=0}^{\infty} \frac{z^n}{n!}, \quad z \in \mathbb{C}.$$

Then we see that $\exp(x) = e^x$ for real numbers x , hence “exp” may be considered as extension of the (real) exponential function $x \mapsto e^x$ to the complex plane. This justifies the notation e^z for nonreal $z \in \mathbb{C}$.

Which properties does the function $z \mapsto \exp(z)$ possess?

Proposition 6.7.5. *The complex exponential function possesses the following properties.*

(1) *For all $w, z \in \mathbb{C}$ it follows*

$$e^{w+z} = e^w \cdot e^z.$$

(2) *For all $z \in \mathbb{C}$ one has*

$$e^{\bar{z}} = \overline{e^z}.$$

(3) *If $t \in \mathbb{R}$, then $|e^{it}| = 1$. More general, for all $z \in \mathbb{C}$ we have*

$$|e^z| = e^{\operatorname{Re}(z)}.$$

In particular, $e^z \neq 0$ for all $z \in \mathbb{C}$.

(4) *For all $t \in \mathbb{R}$ the following formula (called **Euler's formula**) holds:*

$$e^{it} = \cos t + i \sin t.$$

(5) *Any $z \neq 0$ can uniquely represented as*

$$z = |z| e^{i\theta} \quad \text{where} \quad 0 \leq \theta < 2\pi.$$

(6) *For all $z \in \mathbb{C}$ and all $k \in \mathbb{Z}$ it follows that*

$$e^{z+2k\pi i} = e^z.$$

Proof: (1) By definition

$$e^w \cdot e^z = \left(\sum_{n=0}^{\infty} \frac{w^n}{n!} \right) \cdot \left(\sum_{n=0}^{\infty} \frac{z^n}{n!} \right),$$

hence Theorem 6.7.4 applies and leads to

$$\begin{aligned} e^w \cdot e^z &= \sum_{n=0}^{\infty} \sum_{k=0}^n \frac{w^k}{k!} \frac{z^{n-k}}{(n-k)!} = \sum_{n=0}^{\infty} \frac{1}{n!} \sum_{k=0}^n \frac{n!}{k!(n-k)!} w^k z^{n-k} \\ &= \sum_{n=0}^{\infty} \frac{1}{n!} \sum_{k=0}^n \binom{n}{k} w^k z^{n-k} = \sum_{n=0}^{\infty} \frac{(w+z)^n}{n!} = e^{w+z}. \end{aligned}$$

The last step of the proof rests upon the binomial theorem which, as we observed in (6.1.7), is also valid in the case of complex numbers.

The proof of property (2) is left as exercise (see Exercise 6.7.5 below).

Assertion (3) is a direct consequence of (1) and (2). Given $t \in \mathbb{R}$, by $i\bar{t} = -it$ it follows that

$$|e^{it}|^2 = e^{it} \overline{e^{it}} = e^{it} e^{-it} = e^{it} \cdot e^{-it} = e^{it-it} = e^0 = 1.$$

Since the absolute value is always nonnegative, we finally get $|e^{it}| = 1$.

For the proof of the second part of (3) suppose that $z = a+ib$ for some real numbers a and b . Then by (1) the assertion is a direct consequence of

$$|e^z| = |e^{a+ib}| = |e^a \cdot e^{ib}| = |e^a| |e^{ib}| = |e^a| \cdot 1 = e^{\operatorname{Re}(z)}.$$

Furthermore, since $e^t > 0$ for all $t \in \mathbb{R}$, this implies $|e^z| > 0$ for all $z \in \mathbb{C}$, hence it follows that $e^z \neq 0$.

To verify (4) we need a result from Calculus not contained in the present book. We refer to [32], Chapter IV, Section 22, for the following result: For all $t \in \mathbb{R}$ one has

$$\cos t = \sum_{n=0}^{\infty} (-1)^n \frac{t^{2n}}{(2n)!} \quad \text{and} \quad \sin t = \sum_{n=0}^{\infty} (-1)^n \frac{t^{2n+1}}{(2n+1)!}.$$

This enables us to express e^{it} as follows:

$$\begin{aligned} e^{it} &= \sum_{n=0}^{\infty} \frac{(it)^n}{n!} = \sum_{m=0}^{\infty} \frac{i^{2m} t^{2m}}{(2m)!} + \sum_{m=0}^{\infty} \frac{i^{2m+1} t^{2m+1}}{(2m+1)!} \\ &= \sum_{m=0}^{\infty} (-1)^m \frac{t^{2m}}{(2m)!} + i \left(\sum_{m=0}^{\infty} (-1)^m \frac{t^{2m+1}}{(2m+1)!} \right) = \cos t + i \sin t. \end{aligned}$$

This completes the proof of (4).

Property (5) is a direct consequence of (4) combined with Proposition 6.3.1. Given a complex number $z \neq 0$ by Proposition 6.3.1 there is a unique angle $0 \leq \theta < 2\pi$ such that

$$z = |z|(\cos \theta + i \sin \theta).$$

But as we have shown, $\cos \theta + i \sin \theta = e^{i\theta}$, hence $z \neq 0$ has the representation as

$$z = |z|e^{i\theta}$$

which is unique as long as we restrict θ to be in $[0, 2\pi)$.

The periodicity of the exponential function follows easily by (1) and (4). Use that

$$e^{z+2k\pi i} = e^z e^{2k\pi i} = e^z (\cos 2k\pi + i \sin 2k\pi) = e^z.$$

Thus, (6) is valid. ■

Let us state some examples and applications of the previous formulas.

(a) For all $t \in \mathbb{R}$ one has

$$(6.7.7) \quad \sin t = \frac{e^{it} - e^{-it}}{2i} \quad \text{and} \quad \cos t = \frac{e^{it} + e^{-it}}{2}.$$

(b) For example, the exponential function attains the following values:

$$e^{i\pi/2} = i, \quad e^{i\pi} = -1 \quad \text{and} \quad e^{i\pi/4} = \frac{1+i}{\sqrt{2}}.$$

Writing the second identity differently leads to one of the most beautiful¹³ mathematical formulas:

$$e^{i\pi} + 1 = 0.$$

(c) Property (6) of Proposition 6.7.5 tells us that the complex exponential function is **not** injective. Yet the following weaker property is true: The function “exp” is bijective as mapping from $\{z \in \mathbb{C} : 0 \leq \operatorname{Im}(z) < 2\pi\}$ to $\mathbb{C} \setminus \{0\}$.

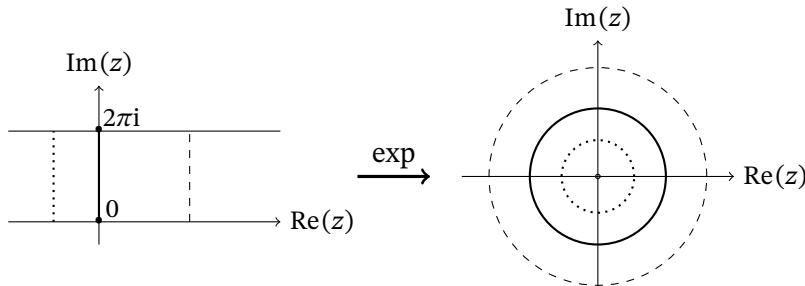


Figure 6.7.2. The function \exp maps vertical lines from x to $x + 2\pi i$ into circles with radius e^x .

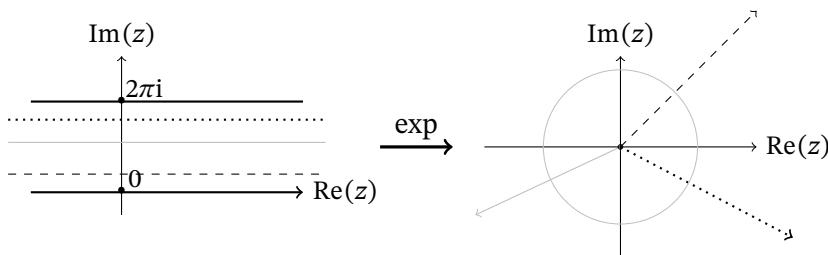


Figure 6.7.3. The function \exp maps the horizontal line $\{x + i\theta : x \in \mathbb{R}\}$ to the infinite ray $\{e^x(\cos \theta + i \sin \theta) : x \in \mathbb{R}\}$.

¹³It combines the 5 most important numbers in mathematics, namely 0, 1, e, π , and i .

The inverse of the function

$$\exp : \{z \in \mathbb{C} : 0 \leq \operatorname{Im}(z) < 2\pi\} \rightarrow \mathbb{C} \setminus \{0\}$$

is denoted by “Log” and is called the **main branch of the complex logarithm**. It is a bijection from $\mathbb{C} \setminus \{0\}$ to $\{z \in \mathbb{C} : 0 \leq \operatorname{Im}(z) < 2\pi\}$ and applies as follows: If $z \neq 0$ equals in polar representation $z = |z|(\cos \theta + i \sin \theta)$ with $0 \leq \theta < 2\pi$, then one gets

$$\operatorname{Log}(z) = \log |z| + i \theta = \log |z| + i \arg(z).$$

Here “log” denotes the inverse function of the (ordinary) real exponential function $x \mapsto e^x$. So, for example, we obtain

$$\operatorname{Log}(-1) = i\pi, \quad \operatorname{Log}(i) = \frac{i\pi}{2}, \quad \operatorname{Log}(1+i) = \frac{\log 2}{2} + \frac{i\pi}{4}.$$

Exercise 6.7.1. Argue why the infinite sum

$$\sum_{n=1}^{\infty} \frac{i^n}{n}$$

exists.

Exercise 6.7.2. Let c_0, c_1, \dots be a sequence of complex numbers such that

$$\lim_{n \rightarrow \infty} \sqrt[n]{|c_n|} = r.$$

Show that for all $z \in \mathbb{C}$ with $|z| < 1/r$ the infinite sum

$$(6.7.8) \quad \sum_{n=0}^{\infty} c_n z^n$$

converges absolutely. If $r = 0$, then the sum (6.7.8) exists even for all $z \in \mathbb{C}$.

Exercise 6.7.3. Given $z \in \mathbb{C}$ with $|z| < 1$, evaluate the Cauchy product of the two infinite series

$$\sum_{n=0}^{\infty} z^n \quad \text{and} \quad \sum_{n=0}^{\infty} (n+1) z^n.$$

Exercise 6.7.4. Suppose $\sum_{k=1}^{\infty} z_k$ exists for some sequence $(z_n)_{n \geq 1}$ of complex numbers. Show that then the sequence of conjugate complex numbers $(\bar{z}_n)_{n \geq 1}$ is also summable and, moreover,

$$(6.7.9) \quad \sum_{k=1}^{\infty} \bar{z}_k = \overline{\sum_{k=1}^{\infty} z_k}.$$

Exercise 6.7.5. Use property (6.7.9) to prove assertion (2) in Proposition 6.7.5. That is, show that

$$e^{\bar{z}} = \overline{e^z}$$

for all $z \in \mathbb{C}$.

Exercise 6.7.6. Verify that for all $z \in \mathbb{C}$ and all $k \in \mathbb{Z}$ always

$$(6.7.10) \quad e^{kz} = (e^z)^k.$$

In particular, it follows that

$$e^{-z} = \frac{1}{e^z} = (e^z)^{-1}, \quad z \in \mathbb{C}.$$

How is (6.7.10) related to de Moivre's formula as stated in Proposition 6.3.5?

Exercise 6.7.7. Evaluate the following values of the complex exponential function:

$$e^{2+i\pi/2}, e^{3-i\pi/2} \quad \text{and} \quad e^{-3\pi/4}.$$

Exercise 6.7.8. Use (6.7.7) and properties of the exponential function to verify the following key relationships between sine and cosine of sums and differences: If $x, y \in \mathbb{R}$, then prove that

$$\begin{aligned} \sin(x \pm y) &= \sin x \cos y \pm \cos x \sin y, \\ \cos(x \pm y) &= \cos x \cos y \mp \sin x \sin y, \\ \sin(2x) &= 2 \sin x \cos x. \end{aligned}$$

Exercise 6.7.9. Sketch the set $\{t e^{it} : 0 \leq t \leq 4\pi\}$ of complex numbers.

Exercise 6.7.10. Give a rigorous proof of the fact that \exp is a bijection from $\{z \in \mathbb{C} : 0 \leq \operatorname{Im}(z) < 2\pi\}$ to $\mathbb{C} \setminus \{0\}$.

Exercise 6.7.11. Do we have

$$\operatorname{Log}(w \cdot z) = \operatorname{Log}(w) + \operatorname{Log}(z),$$

for all nonzero complex numbers w and z ? Give a proof or a counterexample.

Exercise 6.7.12. Show that for all $z \in \mathbb{C}$

$$(6.7.11) \quad |e^z - 1| \leq |z| e^{|z|}.$$

Hint: Note that

$$e^z - 1 = \sum_{n=1}^{\infty} \frac{z^n}{n!} = z \cdot \sum_{n=0}^{\infty} \frac{z^n}{(n+1)!}.$$

Prove that (6.7.11) implies the following:

$$\lim_{n \rightarrow \infty} z_n = 0 \quad \Rightarrow \quad \lim_{n \rightarrow \infty} e^{z_n} = 1.$$

Show that this yields the following more general assertion¹⁴: For all $z \in \mathbb{C}$ one has

$$\lim_{n \rightarrow \infty} z_n = 0 \quad \Rightarrow \quad \lim_{n \rightarrow \infty} e^{z+z_n} = e^z.$$

¹⁴In the language of calculus this property means: The complex exponential function is continuous at every point $z \in \mathbb{C}$.

6.8. More Exercises

Exercise 6.8.1. Calculate and simplify the following expressions:

$$(4 + 3i)(2 - i) + 3 - 5i, \frac{4 + 3i}{2 - i} - 3 + 5i \quad \text{and} \quad (1 + i)^{20}.$$

Exercise 6.8.2. Solve the equation $z^{20} = -1024(1 + i)$.

Exercise 6.8.3. Find all complex numbers z such that $z^6 = 8(1 - i)$.

Exercise 6.8.4. Solve the equation

$$z^7 + z^6 + z^5 + z^4 + z^3 + z^2 + z + 1 = 0.$$

Exercise 6.8.5. Find all complex number z such that

$$z^6 - z^5 + z^4 - z^3 + z^2 - z^2 + z - 1 = 0.$$

Exercise 6.8.6. Let $z = \cos \alpha + i \sin \alpha$ be a nonzero complex number of absolute value one. Prove that for any $n \in \mathbb{N}$:

$$\cos(n\alpha) = \frac{z^{2n} + 1}{z^n} \quad \text{and} \quad \sin(n\alpha) = \frac{z^{2n} - 1}{2iz^n}.$$

Exercise 6.8.7. Let $z_1, z_2 \in \mathbb{C}$. Show that $|z_1 - z_2| = |z_1 + z_2|$ if and only if we have $|z_1 - z_2|^2 = |z_1|^2 + |z_2|^2$.

Exercise 6.8.8. Let $z_1, z_2, z_3 \in \mathbb{C}$. Prove that

$$3(|z_1|^2 + |z_2|^2 + |z_3|^2) = |z_1 + z_2 + z_3|^2 + |z_1 - z_2|^2 + |z_2 - z_3|^2 + |z_3 - z_1|^2.$$

Exercise 6.8.9. Let $n \in \mathbb{N}$. If $z_1, \dots, z_n \in \mathbb{C}$, then prove that

$$(n-2) \sum_{j=1}^n z_j^2 + \left| \sum_{j=1}^n z_j \right|^2 = \sum_{1 \leq k < \ell \leq n} |z_k + z_\ell|^2.$$

Exercise 6.8.10. Let $z \in \mathbb{C}, z \neq 0$ and $\alpha \in \mathbb{R}$. If $z + \frac{1}{z} = 2 \cos \alpha$, then prove that $z^n + \frac{1}{z^n} = 2 \cos(n\alpha)$, for any $n \in \mathbb{N}$.

Exercise 6.8.11. Determine all the solutions to $x^2 + y^2 = 0$ with x and y real. Determine all solutions to $z^2 + w^2 = 0$ with z and w complex.

Exercise 6.8.12. Let A and B be two points in the Euclidean plane. Show that a point C lies on the segment AB if and only if $z_C = \alpha z_A + (1 - \alpha) z_B$, for some $\alpha \in [0, 1]$.

Exercise 6.8.13. Let A, B , and C be three points in the Euclidean plane. Show that a point D lies in the interior or on the sides of the triangle ABC if and only if

$$z_D = \alpha z_A + \beta z_B + \gamma z_C,$$

for some nonnegative real numbers α, β , and γ with $\alpha + \beta + \gamma = 1$.

Exercise 6.8.14. Let $z_1, z_2, z_3 \in \mathbb{C}$ such that $|z_j| = |z_k + z_\ell|$, for any j, k, ℓ such that $\{j, k, \ell\} = \{1, 2, 3\}$. Show that $z_1 + z_2 + z_3 = 0$.

Exercise 6.8.15. Let $n \geq 2$ be a natural number. Show that if w_1, \dots, w_{n-1} are the n th roots of unity not equal to one, then

$$\prod_{k=1}^{n-1} (1 - w_k) = n.$$

Exercise 6.8.16. Let $n \geq 2$ be a natural number. Prove that

$$\prod_{k=1}^{n-1} \sin\left(\frac{k\pi}{n}\right) = \frac{n}{2^{n-1}}.$$

Exercise 6.8.17. Let $\theta \in (0, 2\pi)$. Prove that for any natural number n :

$$\sum_{k=0}^n \cos(k\theta) = \frac{\cos\left(\frac{n\theta}{2}\right) \sin\left(\frac{(n+1)\theta}{2}\right)}{\sin\left(\frac{\theta}{2}\right)} \quad \text{and} \quad \sum_{k=0}^n \sin(k\theta) = \frac{\sin\left(\frac{n\theta}{2}\right) \sin\left(\frac{(n+1)\theta}{2}\right)}{\sin\left(\frac{\theta}{2}\right)}.$$

Exercise 6.8.18. Calculate and simplify $(2 + i)(3 + i)$. Use your work to prove that $\arctan(1/2) + \arctan(1/3) = \pi/4$.

Exercise 6.8.19. Calculate and simplify $(5 + i)^4(239 - i)$. Use your work to show that $4 \arctan(1/5) - \arctan(1/239) = \pi/4$.

Exercise 6.8.20. Let ABC be a triangle in the Cartesian plane such that the origin O is the center of the circle passing through A, B , and C . If H is the orthocenter of ABC , prove that $z_H = z_A + z_B + z_C$. If G is the centroid of ABC , prove that O, G , and H are collinear with G between O and H and that $|OH| = 3|OG|$.

Exercise 6.8.21. Define the complex sine and cosine functions as follows:

$$\text{Sin } z = \frac{e^{iz} - e^{-iz}}{2i} \quad \text{and} \quad \text{Cos } z = \frac{e^{iz} + e^{-iz}}{2}, \quad z \in \mathbb{C}.$$

Due to formula (6.7.7) these functions are extensions of the real sine and cosine functions, i.e., for real x it follows

$$\text{Sin } x = \sin x \quad \text{and} \quad \text{Cos } x = \cos x.$$

(1) Argue why

$$\text{Cos } z = \sum_{n=0}^{\infty} (-1)^n \frac{z^{2n}}{(2n)!} \quad \text{and} \quad \text{Sin } z = \sum_{n=0}^{\infty} (-1)^n \frac{z^{2n+1}}{(2n+1)!}, \quad z \in \mathbb{C}.$$

(2) Show that

$$\text{Sin}(-z) = -\text{Sin } z \quad \text{and} \quad \text{Cos}(-z) = \text{Cos } z.$$

(3) Verify that for all $k \in \mathbb{Z}$ and $z \in \mathbb{C}$ one has

$$\text{Sin}(z + 2k\pi) = \text{Sin } z \quad \text{and} \quad \text{Cos}(z + 2k\pi) = \text{Cos } z.$$

(4) Evaluate $\text{Sin}(i)$ and $\text{Cos}(i)$.

(5) Prove that $\text{Sin}^2 z + \text{Cos}^2 z = 1$, $z \in \mathbb{C}$.

- (6) Find $z_n \in \mathbb{C}$ so that

$$\lim_{n \rightarrow \infty} |\operatorname{Sin} z_n| = \infty \quad \text{and} \quad \lim_{n \rightarrow \infty} |\operatorname{Cos} z_n| = \infty.$$

Thus, in contrast to the real case, complex sine and cosine are unbounded functions.

- (7) (\star) Show that the only zeroes of the complex sine and cosine functions are those which are already known for the real functions. In other words, for all $z \in \mathbb{C}$ with $\operatorname{Im}(z) \neq 0$ follows that

$$\operatorname{Sin} z \neq 0 \quad \text{and} \quad \operatorname{Cos} z \neq 0.$$

- (8) (\star) Verify that Sin and Cos are surjective functions from \mathbb{C} to \mathbb{C} .

Epilogue

After we finished our excursion through the world of numbers, let us shortly analyze what we saw: Every time when we stepped from one system of numbers to the next one, there was either a cultural, an economical or sometimes also only a mathematical (triggered by the curiosity of mathematicians) need to extend the old system and to switch to a new one. In other words, when we started at the set of natural numbers and ended at those of complex ones, then at each step there was an urgent need to extend the existing scale. For example, the step from \mathbb{N} to \mathbb{Z} was necessary to subtract two integers. In order to divide two integers one had to go from \mathbb{Z} to \mathbb{Q} . Since \mathbb{Q} was not complete, it had to be extended to \mathbb{R} . And finally, the investigation of polynomial equations enforced the introduction of the complex numbers. Every time important new operations between numbers were possible, but also some useful properties got lost. For example, after extending \mathbb{N} to \mathbb{Z} , the “Least Value Principle” was no longer valid. Or after extending \mathbb{R} to \mathbb{C} , there was no longer a suitable way to compare the size of two numbers. So let us finally summarize the main advantages and drawbacks of each of the number systems:

| | Advantage | Disadvantage |
|--------------|---|--|
| \mathbb{N} | is well-ordered, induction principle | $(\mathbb{N}, +)$ is not a group |
| \mathbb{Z} | $(\mathbb{Z}, +)$ is a group | $(\mathbb{Z}, +, \cdot)$ is not a field |
| \mathbb{Q} | $(\mathbb{Q}, +, \cdot)$ is an ordered field | not order complete, $\sqrt{2} \notin \mathbb{Q}$ |
| \mathbb{R} | is a complete and ordered field, $\exists \sqrt[n]{x}, x > 0$ | $x^2 + 1 = 0$ is not solvable in \mathbb{R} |
| \mathbb{C} | is a field, $p(z) = 0$ always solvable | not ordered, $\sqrt[n]{z}$ has n values |

Sets, Functions, and Relations

Ignoring difficulties is a poor way of solving them.
Raymond Greene

The aim of this chapter is to present basic definitions and properties involving some fundamental mathematical objects: logic, sets, functions, and relations. We also describe some basic proof methods. And finally, in the last section we shortly sketch how to construct \mathbb{N} and \mathbb{Z} in a mathematical precise way.

A.1. Logic

We start with the following questionnaire that was given to some students. It starts with the following definitions.

Definition A.1.1. A **fact** is something that can be proven to be true or false. An **opinion** is someone's feelings about a particular topic.

It then proceeds to ask to label each sentence below as a fact or an opinion.

- (1) Sunday is the best day of the week.
- (2) George Washington was born in August.
- (3) Thanksgiving is celebrated in autumn in the US.
- (4) This has been a good week.
- (5) Learning math is hard work.
- (6) Monday, Wednesday, and Friday are weekdays.
- (7) The first day of school is scary.

- (8) Spring is the most beautiful season of the year.
- (9) Soccer is a great game.
- (10) Your birthday comes only one day a year.

Sunday may be the best day of the week for most people, but it is likely that there are folks who prefer other days. This makes the first statement an opinion. For the second statement, even without knowing this date of birth, we know that the statement is true or false. It turns out that George Washington was born on February 23, 1732, and therefore, the statement is false. Thanksgiving is celebrated in many countries and in US, it falls on the fourth Thursday in November which is in autumn and thus, (3) is true. While many people may agree with statements (4), (5), (7), (8), and (9), those sentences only express opinions. The answer to question (10) depends on whether you were born on February 29 or not.

Mathematics is quantitative common sense. It deals with mathematical statements which are statements that can only be true or false. For example, the statement $7 > 9$ is a mathematical statement, and it is false. The statement $1 + 1 = 2$ is also a mathematical statement, and it is true. The statements *There is no largest natural number*, or *I woke up at 6am today* are also mathematical statements as they can be assigned a truth value of true or false. For convenience, we can denote mathematical statements by letters as $p : 7 > 9$ or $q : 1 + 1 = 2$. This allows us to create compound statements such as the ones we describe below.

- (1) **not p , $\neg p$ - the negation of p** The negation $\neg p$ is true when p is false and is false when p is true. This definition can be represented by the following truth table:

| p | $\neg p$ |
|-----|----------|
| T | F |
| F | T |

The negation of a statement may contain words such as *not*.

Example A.1.1. If p is the statement *7 is a natural number*, then its negation $\neg p$ is *7 is not a natural number*.

Example A.1.2. If q is the statement *$\sqrt{5}$ is a rational number*, then its negation $\neg q$ is *$\sqrt{5}$ is an irrational number*.

The double negation $\neg(\neg p)$ has the same truth value as p .

- (2) **p and q , $p \wedge q$**

The proposition p and q , also written $p \wedge q$ is true when both p and q are true and is false otherwise.

| p | q | $p \wedge q$ |
|-----|-----|--------------|
| T | T | T |
| F | T | F |
| T | F | F |
| F | F | F |

A tell sign for $p \wedge q$ is the word *and*.

Example A.1.3. The sentence $\sqrt{2}$ is a positive and irrational number is of the form $p \wedge q$, where p is $\sqrt{2}$ is a positive number and q is $\sqrt{2}$ is an irrational number.

Example A.1.4. The sentence the function $f : \mathbb{R} \rightarrow \mathbb{R}$, $f(x) = 2x + 1$ is injective and surjective is also of the form $p \wedge q$, where p denotes the function $f : \mathbb{R} \rightarrow \mathbb{R}$, $f(x) = 2x + 1$ is injective and q is the function $f : \mathbb{R} \rightarrow \mathbb{R}$, $f(x) = 2x + 1$ is surjective.

- (3) **p or q , $p \vee q$**

The proposition p or q , also written $p \vee q$, is true when at least one of p or q is true and is false otherwise.

| p | q | $p \vee q$ |
|-----|-----|------------|
| T | T | T |
| F | T | T |
| T | F | T |
| F | F | F |

A statement of the form $p \vee q$ often contains words such as *or*.

Example A.1.5. The statement *The number 7 is prime or composite* is of the form $p \vee q$, where p is *The number 7 is prime* and q is *The number 7 is composite*.

Example A.1.6. The statement *The number $\frac{2}{3}$ is rational or negative* is of the form $p \vee q$, where p is *The number $\frac{2}{3}$ is rational* and q is *The number $\frac{2}{3}$ is negative*.

- (4) **p implies q , $p \rightarrow q$, $p \Rightarrow q$**

The implication p implies q , also written $p \rightarrow q$ or $p \Rightarrow q$ is true when both p and q are true or when p is false. Another way to read this is *if p , then q* and this justifies calling p the hypothesis of the implication $p \rightarrow q$ and calling q the conclusion of $p \rightarrow q$.

| p | q | $p \Rightarrow q$ |
|-----|-----|-------------------|
| T | T | T |
| F | T | T |
| T | F | F |
| F | F | T |

Some tell signs of an implication are the words *if* and *then*.

Example A.1.7. The implication *If $7 > 9$, then $1 + 1 = 2$* is true since the hypothesis $7 > 9$ is false.

Each implication $p \Rightarrow q$ has a converse $q \Rightarrow p$. Note that the truth value of $p \Rightarrow q$ is not related to the truth value of its converse $q \Rightarrow p$ in general (see the table on page 424).

Example A.1.8. The converse implication *If $1 + 1 = 2$, then $7 > 9$* is false because the hypothesis $1 + 1 = 2$ is true, but the conclusion $7 > 9$ is false.

(5) **p is equivalent to q , $p \leftrightarrow q$, $p \Leftrightarrow q$**

The equivalence $p \leftrightarrow q$ is true when both p and q are true or when both of p and q are false. The truth table below shows that $p \leftrightarrow q$ has the same truth value as $(p \Rightarrow q) \wedge (q \Rightarrow p)$. Note that $p \leftrightarrow q$ has the same truth table as $q \leftrightarrow p$.

| p | q | $p \Rightarrow q$ | $q \Rightarrow p$ | $(p \Rightarrow q) \wedge (q \Rightarrow p)$ | $p \leftrightarrow q$ | $q \leftrightarrow p$ |
|-----|-----|-------------------|-------------------|--|-----------------------|-----------------------|
| T | T | T | T | T | T | T |
| F | T | T | F | F | F | F |
| T | F | F | T | F | F | F |
| F | F | T | T | T | T | T |

Example A.1.9. The equivalence $(0 = 1) \Leftrightarrow (1 = 5)$ is true because both sentences $0 = 1$ and $1 = 5$ are false.

(6) **$The contrapositive of the implication $p \Rightarrow q$ is that $(\neg q) \Rightarrow (\neg p)$.$**

The truth table below shows that any implication $p \rightarrow q$ and its contrapositive $(\neg q) \rightarrow (\neg p)$ are equivalent.

| p | q | $p \Rightarrow q$ | $\neg q$ | $\neg p$ | $(\neg q) \Rightarrow (\neg p)$ |
|-----|-----|-------------------|----------|----------|---------------------------------|
| T | T | T | F | F | T |
| F | T | T | F | T | T |
| T | F | F | T | F | F |
| F | F | T | T | T | T |

Example A.1.10. The contrapositive of the statement *If I study, then I am happy* is *If I am not happy, then I do not study*. The converse of the statement *If I study, then I am happy* is *If I am happy, then I study*.

Definition A.1.2. A mathematical statement that is always true is called a **tautology**. A mathematical statement that is always false is called a **contradiction**.

Example A.1.11. If p is any proposition, then $p \vee (\neg p)$ is a tautology and $p \wedge (\neg p)$ is a contradiction. These facts can be checked using the truth tables.

Using truth tables, we can prove various logical formulas such as De Morgan's laws¹:

$$(A.1.1) \quad \neg(p \vee q) \Leftrightarrow (\neg p) \wedge (\neg q) \text{ and } \neg(p \wedge q) \Leftrightarrow (\neg p) \vee (\neg q)$$

or transitive law:

$$[(p \Rightarrow q) \wedge (q \Rightarrow r)] \Rightarrow (p \Rightarrow r)$$

or associative laws

$$(A.1.2) \quad [p \vee (q \vee r)] \Leftrightarrow [(p \vee q) \vee r] \text{ and } [p \wedge (q \wedge r)] \Leftrightarrow [(p \wedge q) \wedge r]$$

or distributive laws

$$(A.1.3) \quad [p \wedge (q \vee r)] \Leftrightarrow [(p \wedge q) \vee (p \wedge r)] \text{ and } [p \vee (q \wedge r)] \Leftrightarrow [(p \vee q) \wedge (p \vee r)].$$

¹These laws/rules bear the name of the English mathematician Augustus De Morgan (1806–1871).

It is important to understand how to negate implications correctly. Using truth tables, we can deduce that $p \Rightarrow q$ has the same truth value as $(\neg p) \vee q$:

| p | q | $p \Rightarrow q$ | $\neg p$ | $(\neg p) \vee q$ |
|-----|-----|-------------------|----------|-------------------|
| T | T | T | F | T |
| F | T | T | F | T |
| T | F | F | T | F |
| F | F | T | T | T |

Combining this fact with the De Morgan's laws, we get that

$$\neg(p \Rightarrow q) \Leftrightarrow \neg[(\neg p) \vee q] \Leftrightarrow p \wedge (\neg q).$$

This is the basis of the method of proof by contradiction.

In mathematics, we try to answer general questions about patterns in numbers, shapes, sets and so on. These questions can be phrased in words as follows

The square of every real number is nonnegative.

or

The last digit of any power of 5 is 0 or 5.

or

Any even number is either a prime or the sum of two primes.

or

Among every 6 people, there are either 3 friends or 3 non-friends.

In many situations, it is useful to write such statements in equivalent form using mathematical quantifiers. There are two types of quantifiers that we will use: *universal* \forall and *existential* \exists .

Definition A.1.3. The **universal quantifier** is usually denoted by \forall and it can be translated as *for any, for all, or for every*.

The simplest form of a statement involving this quantifier is

$$(\forall x)P(x),$$

where $P(x)$ is a mathematical statement (sometimes called a *predicate*) that depends on the variable x . Of course, there are statements that may depend on several variables and one may have study statements of the form

$$(\forall x)(\forall y)Q(x, y) \text{ or } (\forall x)(\forall y)(\forall z)R(x, y, z) \text{ and so on.}$$

Example A.1.12. The statement

The square of any real number is nonnegative

can be also written as

$$(\forall x \text{ real number})(x^2 \geq 0).$$

For the readers familiar with sets (see the next section for a refresher), the previous statement is the same as

$$(\forall x \in \mathbb{R})(x^2 \geq 0).$$

For now, let us try to prove or disprove the statement above: the square of every real number is nonnegative. To do that, it might be helpful to try a few values.

$$\begin{aligned} 0^2 &= 0 \geq 0, 1^2 = 1 \geq 0, (-1)^2 = 1 \geq 0 \\ 2^2 &= 4 \geq 0, 3^2 = 9 \geq 0, (-2)^2 = 4 \geq 0. \end{aligned}$$

While this might seem convincing to indicate the statement is true, it does not constitute a complete proof (see Example 1.7.6 for a reminder that patterns can be deceiving sometimes). Our statement refers to any real number and while we could continue to pick more values and check their squares, we will not be able to complete the proof this way as there are infinitely many real numbers. If the statement was

the square of any integer between -2 and 2 is nonnegative,

then we would be done.

The way we deal with the general nature of the statement, namely the part involving *any real number* is to *pick* or *consider* or *take* a real number and try to prove the nonnegativity of its square. We will not discuss here the axiomatic nature of this process, and we will just assume that it can be done. Now how do you do this? In mathematics, we can just say

Let x be a real number.

The notation x confers its generality as we do not say *Let 5 be a real number*. Now we have x and reading our statement again, we realize that we are required to prove something about the square of x . Studying our pattern above might give us the following idea to proceed further:

If $x \geq 0$, then multiplying both sides of $x \geq 0$ by x will not change the direction of the inequality as $x \geq 0$. Thus, we get $x^2 \geq 0 \cdot x = 0$. If $x < 0$, then multiplying both sides of the inequality $x < 0$ by x changes the direction of the inequality as $x < 0$. Thus, we get $x^2 > 0 \cdot x = 0$. In either case, $x^2 \geq 0$ and we are done.

This is an example of **proof by cases**. To summarize, we split our argument in two cases: $x \geq 0$ and $x < 0$. In each case, we proceeded in slightly different ways towards the desired result $x^2 \geq 0$.

Example A.1.13. Another problem that can be solved using proof by cases involves the parity (odd or even) of the product of two consecutive integers. Testing a few examples: $0, 1$ or $5, 6$ or $2021, 2022$ or $-8, -7$ or $-199, -198$, we noticed that in each case,

the product of the respective consecutive numbers is even.

This statement can be written as

$$(\forall a \in \mathbb{Z})[a(a + 1) \text{ is even}].$$

As before, we cannot consider each pair of consecutive integers one at a time, so we need a general procedure or recipe that would work in each situation. So let us take a pair of consecutive integers: a and $a + 1$. Since our conjecture is that our product

$a(a + 1)$ is even, it makes sense to see what happens when a is even. In such case, $a(a + 1)$ will be the product of an even number a by the integer $a + 1$ and it is clear that the product $a(a + 1)$ should be even. Formally, one could write it as:

If a is even, then $a = 2k$ where k is some integer. Consequently, we obtain $a(a + 1) = 2k(a + 1) = 2[k(a + 1)]$. Since $k(a + 1)$ is an integer, we get that $a(a + 1)$ is even.

The proof is not complete as the case a even does not cover all possible situations. Of course, a could be odd, and we have to consider this case separately. If a is odd, then $a + 1$ must be even. Then $a(a + 1)$ is the product of an even number $a + 1$ by the integer a and must be even. Formally, one could write this as follows:

If a is odd, then $a = 2t + 1$ where t is some integer. Therefore, we get that $a(a + 1) = a(2t + 1 + 1) = 2[a(t + 1)]$. Since $a(t + 1)$ is an integer, $a(a + 1)$ is even.

Combining these two paragraphs, we get a valid proof of the statement that the product of any two consecutive integers is even.

In some situations, the statement $P(x)$ is an implication, and we are required to prove statements of the form $(\forall x)(Q(x) \Rightarrow R(x))$. To prove such statement, we would have to start by considering an arbitrary x in our universe that satisfies $Q(x)$. Then we would have to work our way to the conclusion that $R(x)$ is true.

Example A.1.14. Let's try our hand with $(\forall x \in \mathbb{R})[(x^2 < 9) \Rightarrow (-3 < x < 3)]$. Here $Q(x)$ is $x^2 < 9$ and $R(x)$ is $-3 < x < 3$. Let $x \in \mathbb{R}$ such that $x^2 < 9$. Subtracting 9 from both sides, we deduce that $x^2 - 9 < 0$. We can rewrite the left-hand side as $(x + 3)(x - 3) = x^2 - 9$ and we have that $(x + 3)(x - 3) < 0$. Because the product of $x + 3$ and $x - 3$ is negative, it follows that one of these numbers is positive and the other is negative. Given that $x + 3 > x - 3$, it must be that $x + 3 > 0 > x - 3$. Therefore, $-3 < x < 3$ which finishes our proof.

Definition A.1.4. The **existential quantifier** is usually denoted by \exists and it can be translated as *there is*, *there are*, or *there exist(s)*.

Example A.1.15. The statement

There is a natural number greater than 2021

can be written as

$\exists n$ natural number, $n > 2021$ or $(\exists n \in \mathbb{N})(n > 2021)$.

To provide this statement, one needs to *describe* or *show the existence* of a natural number strictly greater than 2021. In this case, an explicit description is possible and one can write something like

Consider $n = 2022$. Then n is a natural number and $n > 2021$. Thus, the statement above is true.

The quantifiers may be combined to create mathematical statements.

Example A.1.16. Goldbach's conjecture from number theory stating that every even natural number greater other than 2, is the sum of two primes, can be written as follows

$$(\forall k \geq 2)(\exists p, q \text{ primes})(2k = p + q).$$

If one would like to prove this conjecture, then such proof would start with an arbitrary natural number $k \geq 2$. The next step which is the most important one, is determining or proving the existence of two prime numbers p and q such that $2k = p + q$. So far, nobody has been able to complete this part.

On the other hand, disproving Goldbach's conjecture reduces to showing that

$$\neg[(\forall k \geq 2)(\exists p, q \text{ primes})(2k = p + q)],$$

which is the same as

$$(\exists k \geq 2)(\forall p, q \text{ primes})(2k \neq p + q).$$

Thus, disproving this conjecture reduces to determining or showing the existence of a natural number $k \geq 2$ such that $2k$ is not the sum of two primes. At present time, the Goldbach's conjecture has been verified for all even numbers that are at most $4 \cdot 10^{18}$, meaning that for any such even number, at least one way of writing it as a sum of two primes is known, but it is not known if the conjecture is true or false in general.

Whenever we use the existential quantifier \exists , then it expresses the existence of at least one object which satisfies the desired property. In general, there may be more than one object with that property.

Example A.1.17. If we write

$$(\exists \text{ natural number } k)(2 \leq k \leq 5),$$

then this statement is true for $k = 2, 3, 4$ and $k = 5$.

In order to express that a given property is valid for exactly one object, it is common to write $\exists!$ which is translated as *there is/exists a unique/exactly one*

$$\exists! x \Leftrightarrow \text{exists a unique or exists exactly one } x.$$

Example A.1.18. The statement

$$(\forall x \in \mathbb{R})(\exists! k \in \mathbb{N})(k \leq x < k + 1),$$

is the same as stating that

For any real number x , there exists a unique integer k that $k \leq x < k + 1$.

Example A.1.19. Writing that

$$(\forall x > 0)(\exists! y > 0)(y^2 = x),$$

means that

for any positive number x , there is a unique positive y such that $y^2 = x$.

This statement is true.

An important concept to understand when dealing with statements involving quantifiers is how to construct their negation. Since we live in the realm of true and false, we have to figure out if a statement is true or if it is false which is the same thing as

its negation being true. The general rules for determining the negation of a statement involving quantifiers are below:

$\neg[(\forall x)P(x)]$ is the same as $(\exists x)(\neg P(x))$

$\neg[(\exists x)P(x)]$ is the same as $(\forall x)(\neg P(x))$.

Example A.1.20. We write down the negations of some statements below:

$\neg(\forall x \in \mathbb{R})(x^2 > 1)$ is $(\exists x \in \mathbb{R})(x^2 \leq 1)$,

$\neg(\exists x \in \mathbb{R})(x^2 > 1)$ is $(\forall x \in \mathbb{R})(x^2 \leq 1)$.

In general, a predicate $P(x)$ may be more complicated. For example, it may be an implication $Q(x) \Rightarrow R(x)$. How does one determine the negation of $(\forall x)(P(x))$ in such situation? The same rules as above apply and the negation is $(\exists x)(\neg P(x))$. Now since $P(x)$ is $Q(x) \Rightarrow R(x)$ (which by Exercise A.1.2 is the same as the statement $(\neg Q(x)) \vee R(x)$). Hence, $\neg(Q(x) \Rightarrow R(x))$ is $\neg((\neg Q(x)) \vee R(x))$ which is the same as $Q(x) \wedge (\neg R(x))$. Hence, the negation of $(\forall x)(Q(x) \Rightarrow R(x))$ is $(\exists x)(Q(x) \wedge (\neg R(x)))$.

Exercise A.1.1. Let p , q , and r be mathematical statements. Prove that the following propositions are tautologies.

- (1) $[(p \vee q) \wedge (\neg p)] \Rightarrow q$,
- (2) $[p \wedge (p \Rightarrow q)] \Rightarrow q$,
- (3) $[(p \Rightarrow q) \wedge (\neg q)] \Rightarrow (\neg p)$,
- (4) $[(p \Rightarrow q) \wedge (q \Rightarrow r)] \Rightarrow (p \Rightarrow r)$.

Exercise A.1.2. Let p and q be two mathematical statements.

- (1) Prove that $p \Rightarrow q$ is equivalent to $\neg p \vee q$.
- (2) Show that $\neg(p \Rightarrow q)$ is equivalent to $p \wedge (\neg q)$.

Exercise A.1.3. Prove that $(p \Rightarrow q) \wedge ((\neg p) \Rightarrow q)$ and q are equivalent.

Exercise A.1.4. Prove De Morgan's laws (A.1.1), the associativity laws (A.1.2), and the distributivity laws (A.1.3).

Exercise A.1.5. Write down the negation of each of the following statements.

- (1) Every student at our university has at least one cell phone.
- (2) Every soccer ball is round.
- (3) If you don't study hard, you will not understand the material.
- (4) For any math problem, there is at least one person who cannot solve it.
- (5) Any person can solve at least one math problem.

Exercise A.1.6. Write down the negation of each of the following statements.

- (1) $(\forall x \in \mathbb{N})(\exists y \in \mathbb{N})(y = x + 5)$.
- (2) $(\exists y \in \mathbb{N})(\forall x \in \mathbb{N})(y = x + 5)$.
- (3) $(\forall x \in \mathbb{N})(\forall y \in \mathbb{N})(\exists z \in \mathbb{N})(z < x + y)$.

- (4) $(\exists z \in \mathbb{N})(\forall y \in \mathbb{N})(\forall x \in \mathbb{N})(z < x + y)$.
(5) $(\forall x \in \mathbb{N})(\exists z \in \mathbb{N})(\forall y \in \mathbb{N})(z < x + y)$.

Exercise A.1.7. Determine with proper justification which of the previous statements are true.

Exercise A.1.8. Determine the truth value of each of the following statements.

- (1) If 2021 is even, then 2022 is odd.
(2) For any natural number n , there exists a natural number k such that $k > 5n$.
(3) There exists a natural number k such that $k > 5n$ for any natural number n .
(4) For any prime number p , there exists a prime number q such that $p + q$ is even.
(5) For any prime number p , there exists a prime number q such that $p + q$ is odd.
(6) There exists a real number y such that $y = x^2$ for every real number x .
(7) For every real number x , there exists a real number y such that $y = x^2$.

Exercise A.1.9. Write down *the converse* and *the contrapositive* of each of the following implications.

- (1) $(ab = 0) \Rightarrow [(a = 0) \vee (b = 0)]$.
(2) $[(a \geq b) \wedge (b \geq a)] \Rightarrow (a = b)$.
(3) $(a^2 + b^2 = 0) \Rightarrow [(a = 0) \wedge (b = 0)]$.
(4) $(a > b) \Rightarrow [(\exists c > 0)(a = b + c)]$.

Exercise A.1.10. Write down the negation of the statement

$$(\forall n \in \mathbb{N})(n^2 + n + 41 \text{ is prime})$$

and decide which statement is true.

A.2. Sets

When trying to describe sets, we may think of a collection of certain objects sharing one or several common properties. For instance, we may speak about the set of students at the local, the set of books on Werner's desk, or the set of soccer balls in Sebi's car. Thus, a first naive definition of a set would be as follows:

*A set is a collection of distinct objects possessing some common property or properties*².

For the purpose of this book this naive description of sets completely suffices and, most importantly, it will lead neither to any contradictions nor any inconsistencies. Thus, throughout this book sets are defined as follows:

Definition A.2.1. A **set** is an unordered collection of distinct objects. These objects are called **members** or **elements** of the set.

²This naive approach may lead to problems. One puzzling example could be the collection of all possible sets. For our purposes, we will not be concerned about these types of problems and the former informal definition of sets will suffice. Those who are interested in an axiomatic approach to Set Theory are referred to books as for example [15] or [18].

There are various ways to describe or define sets. One way is by explicitly listing its elements. For example, when writing

$$X = \{\odot, 2, \clubsuit, 7\},$$

we mean that the set X has four elements: \odot , 2, \clubsuit and 7. Note that what matters here are the elements of the set and not their order. In our situation, the set X is the same as $\{2, 7, \odot, \clubsuit\}$ or $\{7, \clubsuit, 2, \odot\}$ or $\{2, \clubsuit, \odot, 7\}$. Another way to describe a set is by stating a certain property or properties exhibited by its elements such as in the following situation:

$$Y = \{n \leq 10 : n \text{ is an even natural number}\}.$$

The same set can be described in several ways. For example, the set Y above can also be represented as $\{2, 4, 6, 8, 10\}$.

Similar to zero and the natural numbers, there is a special set that has no elements.

Definition A.2.2. The set containing no objects is called the **empty set**, and is denoted by \emptyset .

A key fact involving sets is that an element x may either be an element of a given set A or it may not be. For an element x and the set A , we expect full commitment: x is either in A or x is not in A . This is the same dichotomy exhibited by any mathematical statement: it is either true or false, but cannot be both and cannot be neither. If x is an element of A , we write $x \in A$. When x is not an element of A , we write that $x \notin A$.

Example A.2.1. With $X = \{2, \odot, 7, \clubsuit\}$, we have that $\odot \in X$ and $7 \in X$ while $5 \notin X$ and $\odot \notin X$.

Definition A.2.3. Given two sets A and B , we say that A is a **subset** of B , written $A \subseteq B$, whenever any element of A must belong to B :

$$A \subseteq B \Leftrightarrow (\forall x)[(x \in A) \Rightarrow (x \in B)].$$

This definition tells us the meaning of a set A not being a subset of B , which we write as $A \not\subseteq B$:

$$A \not\subseteq B \Leftrightarrow (\exists x)[(x \in A) \text{ and } (x \notin B)].$$

Example A.2.2. If $Z = \{\odot, \clubsuit\}$ and $X = \{2, \odot, 7, \clubsuit\}$, then $Z \subseteq X$, but $X \not\subseteq Z$.

Definition A.2.4. Two sets A and B **coincide** or **are equal**, written as $A = B$, if they contain exactly the same elements:

$$A = B \Leftrightarrow (\forall x)[(x \in A) \Leftrightarrow (x \in B)] \Leftrightarrow [A \subseteq B \text{ and } B \subseteq A].$$

This definition also tells us when two sets are not equal:

$$A \neq B \Leftrightarrow [(A \not\subseteq B) \text{ or } (B \not\subseteq A)],$$

which is the same as

$$(\exists x)[(x \in A) \text{ and } (x \notin B)] \text{ or } [(x \notin A) \text{ and } (x \in B)].$$

There are several operations that one may perform with sets.

Definition A.2.5. The **intersection** $A \cap B$ of the sets A and B is defined by

$$A \cap B := \{x : x \in A \text{ and } x \in B\}.$$

Sets can be described using a Venn diagram which is a graphic representation of the sets as disks.

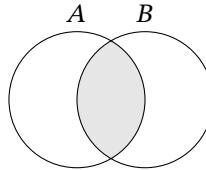


Figure A.2.1. The Venn diagram of the intersection $A \cap B$.

Definition A.2.6. The sets A and B are called **disjoint** if $A \cap B = \emptyset$.

Equivalently, if $x \in A$, then $x \notin B$, or, if $x \in B$, then $x \notin A$.

Example A.2.3. The sets $U = \{3, 4, 5\}$ and $V = \{1, 2\}$ are disjoint.

Note that any set A and \emptyset are disjoint.

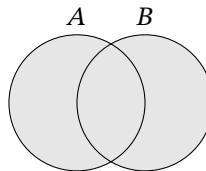


Figure A.2.2. The Venn diagram of the union $A \cup B$.

Definition A.2.7. The **union** $A \cup B$ of the sets A and B is defined as

$$A \cup B := \{x : x \in A \text{ or } x \in B\}.$$

We state below some properties of intersection and union.

Proposition A.2.1. For any sets A and B ,

$$A \cap B \subseteq A \subseteq A \cup B \quad \text{and} \quad A \cap B \subseteq B \subseteq A \cup B.$$

Union and intersection are commutative operations, i.e.,

$$A \cap B = B \cap A \quad \text{and} \quad A \cup B = B \cup A.$$

If $A \subseteq B$, then $A \cap B = A$ and $A \cup B = B$.

The following **distributive law** is true.

Proposition A.2.2. Given three sets A , B , and C , then

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

Proof: To prove the set equality above, we have to verify two inclusions, namely,

$$(A.2.1) \quad A \cap (B \cup C) \subseteq (A \cap B) \cup (A \cap C) \quad \text{and}$$

$$(A.2.2) \quad (A \cap B) \cup (A \cap C) \subseteq A \cap (B \cup C).$$

In order to prove inclusion (A.2.1), let $x \in A \cap (B \cup C)$. Therefore, $x \in A$ and $x \in B \cup C$, meaning x belongs to the set A and at least to one of the two sets B and C . If $x \in B$, then, as $x \in A$, it follows that $x \in A \cap B$, hence $x \in (A \cap B) \cup (A \cap C)$. By the same argument, $x \in C$ implies $x \in A \cap C$, hence also $x \in (A \cap B) \cup (A \cap C)$. This proves (A.2.1).

For the proof of (A.2.2), let y be an element of $(A \cap B) \cup (A \cap C)$. Therefore, $y \in A \cap B$ or $y \in A \cap C$. If $y \in A \cap B$, then $y \in A$ and $y \in B \subseteq B \cup C$ and thus, $y \in A \cap (B \cup C)$. If $y \in A \cap C$, then $y \in A$ and $y \in C \subseteq B \cup C$ and hence, $y \in A \cap (B \cup C)$, finishing our proof. \blacksquare

There is another way to combine two sets, the set difference.

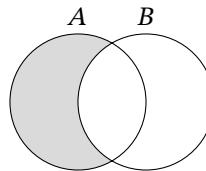


Figure A.2.3. The Venn diagram of the difference $A \setminus B$.

Definition A.2.8. Given sets A and B , the **set difference** between A and B , written as $A \setminus B$, is defined as:

$$A \setminus B = \{x : x \in A \text{ and } x \notin B\}.$$

Example A.2.4. If $X = \{2, \odot, 7, \clubsuit\}$ and $W = \{\diamond, 3, 7\}$, then

$$X \setminus W = \{2, \odot, \clubsuit\} \quad \text{and} \quad W \setminus X = \{\diamond, 3\}.$$

Actually, $A \setminus B$ and $B \setminus A$ are disjoint for any sets A and B .

Sometimes it is useful to describe those elements which belong exactly to one of two given sets A and B .

Definition A.2.9. Given two sets A and B , their **symmetric difference**, written $A \Delta B$, is defined as:

$$A \Delta B = (A \setminus B) \cup (B \setminus A).$$

Example A.2.5. If $X = \{2, \odot, 7, \clubsuit\}$ and $W = \{\diamond, 3, 7\}$, then

$$X \Delta W = \{2, \odot, \clubsuit, \diamond, 3\}.$$

The fact that $A \Delta B = B \Delta A$ justifies the adjective *symmetric*. Another way to express the symmetric difference is as follows.

Proposition A.2.3. *If A and B are sets, then*

$$A \Delta B = (A \cup B) \setminus (A \cap B).$$

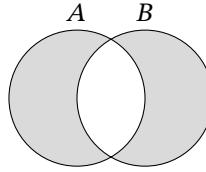


Figure A.2.4. The symmetric difference $A \Delta B$.

Proof: We prove the statement above by double inclusion showing first that

$$A \Delta B \subseteq (A \cup B) \setminus (A \cap B),$$

and

$$(A \cup B) \setminus (A \cap B) \subseteq A \Delta B.$$

In both situations, we will use the definition $A \Delta B = (A \setminus B) \cup (B \setminus A)$. For the first part, let $x \in A \Delta B$. This means that $x \in (A \setminus B) \cup (B \setminus A)$ and therefore, $x \in A \setminus B$ or $x \in B \setminus A$. If $x \in A \setminus B$, we get that $x \in A \subseteq A \cup B$ and $x \notin B$, implying that $x \notin A \cap B$. Thus, $x \in (A \cup B) \setminus (A \cap B)$. If $x \in B \setminus A$, the argument is similar, and we have that $x \in B \subseteq A \cup B$ and $x \notin A$, leading to $x \notin A \cap B$. Hence, $x \in (A \cup B) \setminus (A \cap B)$. This proves that $A \Delta B \subseteq (A \cup B) \setminus (A \cap B)$. For the second part, let $x \in (A \cup B) \setminus (A \cap B)$. Therefore, $x \in A \cup B$ and $x \notin A \cap B$. Hence, x is contained in A or B , but not in both. If $x \in A$ and $x \notin B$, then $x \in A \setminus B \subseteq A \Delta B$. If $x \in B$ and $x \notin A$, then $x \in B \setminus A \subseteq A \Delta B$. In either situation, $x \in A \Delta B$. This shows that $(A \cup B) \setminus (A \cap B) \subseteq A \Delta B$ and finishes our proof. ■

Our next objective is to introduce the complement of a set.

Definition A.2.10. Let S be a set. If A is a subset of S , then the **complement of A** (with respect to S), written A^c , is the set containing all elements in S which do not belong to A :

$$A^c = \{x \in S : x \notin A\}.$$

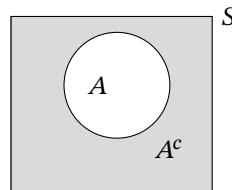


Figure A.2.5. The complement A^c of A with respect to S .

From definition, $A^c = S \setminus A$ so the set S is important when discussing complements and should be specified.

Example A.2.6. If $A = \{1, 2\}$, $S_1 = \{1, 2, 3\}$, and $S_2 = \{1, 2, 3, 4\}$, then the complement of A with respect to S_1 is $\{3\}$ while the complement of A with respect to S_2 is $\{3, 4\}$.

In general, the underlying set S is either explicitly mentioned or is understood from the context.

Proposition A.2.4. *If A and B are subsets of a set S , then*

- (i) $(A^c)^c = A$ and $(A \subseteq B) \Leftrightarrow (B^c \subseteq A^c)$
- (ii) $A \setminus B = A \cap B^c$, $S^c = \emptyset$ and $\emptyset^c = S$,
- (iii) $(A \cap B = \emptyset) \Leftrightarrow (A \subseteq B^c) \Leftrightarrow (B \subseteq A^c)$

We leave the proof as an exercise.

Proposition A.2.5 (De Morgan's Laws). *If A and B are subsets of a set S , then*

$$(A \cup B)^c = A^c \cap B^c \quad \text{and} \quad (A \cap B)^c = A^c \cup B^c.$$

Proof: Let $x \in S$. We have that $x \in (A \cup B)^c$ if and only if $x \in S$ and $x \notin A \cup B$. This is the same as $x \in S$, $x \notin A$, and $x \notin B$ which is equivalent to $x \in S \setminus A$ and $x \in S \setminus B$. Hence, $x \in A^c \cap B^c$. This completes the proof of the first identity. The second identity may either be verified by similar arguments or follows from the first one when applying it to A^c and B^c . Then $(A^c \cup B^c)^c = (A^c)^c \cap (B^c)^c = A \cap B$. Taking the complement of both sides, the second equation follows. ■

Definition A.2.11. Given a set S we denote by $\mathcal{P}(S)$ the **power set** of S , which consists of all possible subsets of S : $\mathcal{P}(S) = \{A : A \subseteq S\}$.

Example A.2.7. If $S = \{a, b, c\}$, then

$$\mathcal{P}(S) = \{\emptyset, \{a\}, \{b\}, \{c\}, \{a, b\}, \{a, c\}, \{b, c\}, \{a, b, c\}\}.$$

Observe that in this example the power set consists of $8 = 2^3$ elements. Proposition 1.7.4 shows that if a set S has n elements, then $\mathcal{P}(S)$ has 2^n elements. This is the justification for the name *power* used in the above definition.

Definition A.2.12. Given two sets A and B , the **Cartesian product** of A and B , written $A \times B$, is defined as the set consisting of all ordered³ pairs (a, b) , where $a \in A$ and $b \in B$:

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

We say that two ordered pairs (a_1, b_1) and (a_2, b_2) are the same which we write as $(a_1, b_1) = (a_2, b_2)$ exactly when $a_1 = a_2$ and $b_1 = b_2$.

Example A.2.8. If $H = \{1, 2, 3\}$ and $V = \{1, 2\}$, then

$$H \times V = \{(1, 1), (1, 2), (2, 1), (2, 2), (3, 1), (3, 2)\}.$$

The readers familiar with calculus have already encountered the Cartesian product of the horizontal x -axis and the vertical y -axis when drawing graphs of functions. Rolling a die we may observe values in the set $A = \{1, \dots, 6\}$. When rolling two dice, then we observe a pair $(d_1, d_2) \in A \times A$, where d_1 is the number of the first die and d_2 is the number of the second one.

³The word *ordered* expresses that $(a, b) \neq (b, a)$ when $a \neq b$. This is different from *unordered* pairs $\{a, b\}$, where $\{a, b\} = \{b, a\}$.

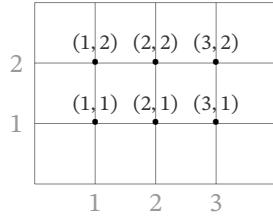


Figure A.2.6. The Cartesian product $\{1, 2, 3\} \times \{1, 2\}$.

In many cases, it is necessary to combine more than two sets, maybe n sets A_1, \dots, A_n , or even infinitely many sets A_1, A_2, \dots . Let us treat here only the first case, i.e., we are given finitely many sets A_1, \dots, A_n . One way to introduce their union and intersection is by an iteration, where each time we combine two sets.

$$A_1 \cup \dots \cup A_n = (A_1 \cup \dots \cup A_{n-1}) \cup A_n \quad \text{and} \quad A_1 \cap \dots \cap A_n = (A_1 \cap \dots \cap A_{n-1}) \cap A_n.$$

One writes

$$\bigcup_{j=1}^n A_j = A_1 \cup \dots \cup A_n \quad \text{and} \quad \bigcap_{j=1}^n A_j = A_1 \cap \dots \cap A_n.$$

A more direct approach to define union and intersection of finitely many sets is as follows:

$$\bigcup_{j=1}^n A_j = \{x : \exists j \leq n, x \in A_j\} \quad \text{and} \quad \bigcap_{j=1}^n A_j = \{x : \forall j \leq n, x \in A_j\}.$$

The basic rules for the union and intersection of more than two sets remain the same as in the case of two sets. The distributive law reads now as follows:

$$A \cap \left(\bigcup_{j=1}^n A_j \right) = \bigcup_{j=1}^n (A \cap A_j).$$

The generalized De Morgan's laws in this setting are now

$$\left(\bigcup_{j=1}^n A_j \right)^c = \bigcap_{j=1}^n A_j^c \quad \text{and} \quad \left(\bigcap_{j=1}^n A_j \right)^c = \bigcup_{j=1}^n A_j^c.$$

The proofs of these formulas use exactly the same arguments as in the case of two sets and are left as an exercise.

The Cartesian product of two sets consists of ordered pairs. In some situations, these ordered pairs may be viewed as 2-dimensional vectors. For a natural number n , the Cartesian product of n sets can be defined similarly as for two sets.

Definition A.2.13. Given n sets A_1, \dots, A_n , the **Cartesian product** $A_1 \times \dots \times A_n$ is defined as:

$$A_1 \times \dots \times A_n = \bigtimes_{j=1}^n A_j = \{x : x = (x_1, \dots, x_n), x_j \in A_j\}.$$

Note that, as before, if $x, x' \in \times_{j=1}^n A_j$ with $x = (x_1, \dots, x_n)$ and $x' = (x'_1, \dots, x'_n)$, then $x = x'$ if and only if $x_1 = x'_1, x_2 = x'_2, \dots, x_n = x'_n$. Moreover, it can be easily seen that the n -fold Cartesian product may be iterated as follows:

$$(A.2.3) \quad \times_{j=1}^n A_j = (\times_{j=1}^{n-1} A_j) \times A_n.$$

To verify this, identify the pair $((x_1, \dots, x_{n-1}), x_n)$ with the vector (x_1, \dots, x_n) .

In the case that $A_1 = \dots = A_n = A$, we will write

$$A^n := \underbrace{A \times \cdots \times A}_{n \text{ times}}.$$

An example of this construction occurs in natural way when tossing a coin n times. If the coin is labeled with H (Head) and T (Tail), then tossing the coin n times, one observes a vector $x = (x_1, \dots, x_n)$ where each x_j is either H or T . A similar Cartesian product $\{0, 1\}^n$ has the same cardinality as the power set $\mathcal{P}(S)$, when S has n elements (see Exercise A.2.9 and Proposition 1.7.4).

Exercise A.2.1. Let $S = \{1, 2, 3, 4\}$. Which of the following assertions are true and which are false?

- (a) $\emptyset \in S$,
- (b) $\emptyset \subseteq S$,
- (c) $\{\emptyset\} \subseteq S$,
- (d) $\emptyset \in \mathcal{P}(S)$,
- (e) $\emptyset \subseteq \mathcal{P}(S)$,
- (f) $\{\emptyset\} \in \mathcal{P}(S)$,
- (g) $\{\emptyset\} \subseteq \mathcal{P}(S)$.

Exercise A.2.2. Let A and B be two sets.

- (1) Prove that $A \cap B = A$ if and only if $A \subseteq B$.
- (2) Prove that $A \cup B = B$ if and only if $A \subseteq B$.

Exercise A.2.3. For any two sets A and B , prove that $A \setminus B = A \setminus (A \cap B)$.

Exercise A.2.4. For any three sets A, B , and C , prove the following assertions:

$$A \setminus (B \cup C) = (A \setminus B) \cap (A \setminus C) \quad \text{and} \quad (A \cap B) \cup C = (A \cup C) \cap (B \cup C).$$

Exercise A.2.5. Prove Proposition A.2.4.

Exercise A.2.6. For any sets $A, B \subseteq S$, prove that

$$(B \setminus A)^c \cap B = A \cap B \text{ and } (A \cup B)^c \cap B = \emptyset.$$

Exercise A.2.7. Let A, B and C be subsets of a set S . Which of the following assertions are correct, which are incorrect?

- | | |
|---|---|
| <ul style="list-style-type: none"> (a) $(A \cup B) \setminus A = B$ (c) $(A \cup B) \cap A = A$ (e) $(A^c \cup B^c)^c = A \cup B$ (g) $A \cap B = \emptyset \Rightarrow A \cap C, B \text{ disjoint}$ (h) $A \cap B = \emptyset \Leftrightarrow A^c \cup B^c = S$ | <ul style="list-style-type: none"> (b) $(A \cup B) \cap C = (A \cap C) \cup (B \cap C)$ (d) $A \setminus B^c = A \cup B$ (f) $A \cup B = (A \setminus (A \cap B)) \cup B$ (h) $(A \cap B) \cup B = B$ (j) $A \subseteq B \Leftrightarrow B^c \subseteq A^c$ |
|---|---|

Exercise A.2.8. Let A, B , and C be three sets.

- (1) Prove that $(A \Delta B) \Delta C = A \Delta (B \Delta C)$.
- (2) Is it true that $(A \Delta B) \Delta C = (A \cup B \cup C) \setminus (A \cap B \cap C)$? Prove or give a counterexample.
- (3) Is it true that $(A \Delta B) \Delta C = (A \setminus (B \cup C)) \cup (B \setminus (A \cup C)) \cup (C \setminus (A \cup B))$? Prove or give a counterexample.

Exercise A.2.9. Write down and count all elements of

$$\{0, 1\}^4 = \{0, 1\} \times \{0, 1\} \times \{0, 1\} \times \{0, 1\}$$

and all the subsets of the set $\{1, 2, 3, 4\}$. Can you explain your findings?

Exercise A.2.10. Let A and B be two subsets of a set S . Which of the following equations are valid? Prove the correct identities, give counterexamples for the false ones.

$$\begin{aligned}(A \times B)^c &= A^c \times B^c, \\ (A \times B)^c &= (A^c \times B^c) \cup (A^c \times S) \cup (S \times B^c), \\ (A \times B)^c &= (A^c \times B) \cup (A \times B^c).\end{aligned}$$

A.3. Functions

Mathematical functions occur everywhere in our life. Our grade in a class is evaluated as a function of our work and effort, the speed of our car is a function of our driving and the traffic conditions, the bank pays us interest as a function of the amount in our account and the temperature at present time in the world is a function of the location. Thus, we have some input such as our work and effort transformed by some fixed rule into an output such as our grade. Informally, a function is a rule which describes the way how input and output are related.

A first, slightly imprecise mathematical definition of a function is as follows:

Definition A.3.1. A **function** f from a set A into a set B , written as $f : A \rightarrow B$, is a rule that assigns to each $x \in A$ precisely one $y \in B$, which is denoted by $f(x)$. The element $f(x)$ is called the **image of x under f** or the **value of f at x** .

$$(f : A \rightarrow B) \Leftrightarrow [(\forall x \in A)(\exists! y \in B)(f(x) = y)]$$

Example A.3.1. Define the function $f : \mathbb{N} \rightarrow \mathbb{N}$ with $f(n) = 2n$ for any $n \in \mathbb{N}$. Similarly, let $g : \mathbb{N} \rightarrow \mathbb{Z}$ with $g(n) = 2n$ for any $n \in \mathbb{N}$.

The key property of a function is that it assigns to each $x \in A$ exactly one $y \in B$. It may be useful to think of a function as a black box or transformation where one input $x \in A$ leads to exactly one output $y \in B$. And if one puts in the same element $x \in A$ the next day, the output will be exactly the same $y \in B$ as the day before. A function is a fixed rule: Neither a change of the rules by time nor by randomness.

Definition A.3.2. Let $f : A \rightarrow B$. The set A is called the **domain** of f and is denoted by $\text{dom}(f)$, while the set B is called the **codomain** of f and is denoted by $\text{codom}(f)$.

Example A.3.2. For f and g in Example A.3.1, $\text{dom}(f) = \text{codom}(f) = \mathbb{N}$ while $\text{dom}(g) = \mathbb{N}$ and $\text{codom}(g) = \mathbb{Z}$.

Definition A.3.3. Two functions f and g **coincide** or **are equal**, written $f = g$, if $\text{dom}(f) = \text{dom}(g)$, $\text{codom}(f) = \text{codom}(g)$ and $f(x) = g(x)$ for any $x \in \text{dom}(f)$.

Note that f and g in Example A.3.1 are not the same since their codomains are different.

Definition A.3.4. Let $f : A \rightarrow B$ be a function. The set

$$\text{range}(f) = \{y \in B : \exists x \in A, y = f(x)\}$$

is called the **range** or the **image** of the function f .

Proposition A.3.1. *The range $\text{range}(f)$ is a subset of the codomain $\text{codom}(f)$ for any function $f : A \rightarrow B$.*

Example A.3.3. For the function f in Example A.3.1, $\text{range}(f)$ is a proper subset of $\text{codom}(f)$ as, for example, $1 \in \text{codom}(f)$, but $1 \notin \text{range}(f)$.

Another important set related to a function is its graph.

Definition A.3.5. Let $f : A \rightarrow B$ be a function. The subset

$$\text{graph}(f) = \{(x, y) : x \in A, y = f(x) \in B\} = \{(x, f(x)) : x \in A\}$$

of $A \times B$ is called the **graph of f** .

In courses such as calculus, when we sketch a certain function on the blackboard or our notebooks, we are actually drawing the graph of the function.

Example A.3.4. The function $h : [-2, 2] \rightarrow \mathbb{R}$ with $h(x) = x^2$ has the following graph plotted in Figure A.3.1.

$$\text{graph}(h) = \{(x, x^2) : x \in [-2, 2]\}.$$

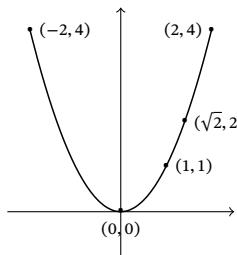


Figure A.3.1. The graph of $h(x) = x^2$ for $-2 \leq x \leq 2$.

Not every subset $F \subseteq A \times B$ is the graph of a suitable function from A to B . Look at the following example.

Example A.3.5. If $A = B = \{1, 2\}$, then the subset $F = \{(1, 1), (1, 2), (2, 2)\}$ of $A \times B$ is not a graph of a function from A to B . Neither is $F = \{(2, 1), (2, 2)\}$.

Why is this so? In the first case the number 1 would be mapped to 1 but also to 2. This contradicts the fact that a function maps each x in the domain to a unique y in

its codomain. Also, the second set is not a graph of a function. It would be so, if the domain consists only of the element 2. But since the domain is $\{1, 2\}$, there is no value in the codomain assigned to 1. And as we defined, a function has to map **every** value in its domain to another value in its codomain.

Subsets F of $A \times B$ being graphs of functions possess a certain special property, namely

$$(A.3.1) \quad (\forall x \in A)(\exists! y \in B)((x, y) \in F).$$

This property is also known as the **vertical line property**, meaning that a vertical line at a point x intersects the graph of a function exactly once if x is in the domain of the function or does not intersect the graph if x is not in the domain.

Now we are in the position to state the more formal and more precise definition of a function $f : A \rightarrow B$. It reads as follows.

Definition A.3.6. A function f from $A \rightarrow B$ is a subset $F \subseteq A \times B$ satisfying (A.3.1). The mapping $f : A \rightarrow B$ behaves in the following way:

$$f(x) := y \Leftrightarrow (x, y) \in F.$$

Of course, when f is given in this way, then $\text{graph}(f) = F$.

Functions can be described in many ways.

Example A.3.6. For small domains, one way to define a function f is by the table of its values. This means writing down the values $f(x)$ for all x in the domain of f . If we play a game where we win \$1 if we roll a die and an even number occurs while we have to pay \$1 for an odd number, then the outcome of this game is expressed by the function $f : \{1, 2, 3, 4, 5, 6\} \rightarrow \{-1, 1\}$ with $f(1) = f(3) = f(5) = -1$ while $f(2) = f(4) = f(6) = +1$. This may be written as a table, in the following way:

| | | | | | | |
|--------|----|---|----|---|----|---|
| x | 1 | 2 | 3 | 4 | 5 | 6 |
| $f(x)$ | -1 | 1 | -1 | 1 | -1 | 1 |

Another possibility to describe f is by its graph

$$\text{graph}(f) = \{(1, -1), (2, 1), (3, -1), (4, 1), (5, -1), (6, 1)\}.$$

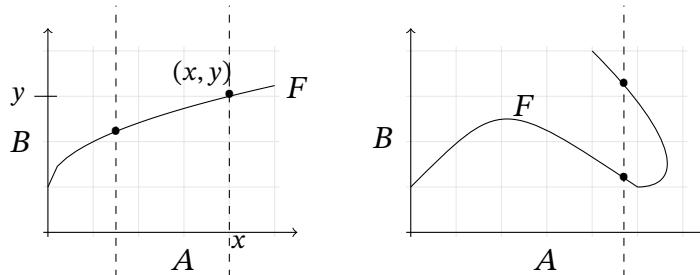


Figure A.3.2. The set $F \subseteq A \times B$ in the left-hand figure possesses the vertical line property, the set in the right-hand one does not. Thus, F on the left defines a function f from A to B by letting $f(x) = y$ if $(x, y) \in F$ (here $f : x \mapsto 1 + \sqrt{x}$).

Some functions are given by an algebraic rule as the ones in Example A.3.1.

Example A.3.7. Let $n \in \mathbb{N}_0$ be a nonnegative integer and a_0, \dots, a_n be some fixed real numbers. Define $p : \mathbb{R} \rightarrow \mathbb{R}$ by

$$p(x) = a_n x^n + a_{n-1} x^{n-1} + \cdots + a_1 x + a_0, \quad x \in \mathbb{R}.$$

This is an example of a polynomial function.

Example A.3.8. A function may be defined recursively as $f : \mathbb{N} \rightarrow \mathbb{N}$, with $f(1) = 1$ and by $f(n+1) = 2f(n)$ if $n \geq 2$.

Another function g defined recursively is $g : \mathbb{N} \rightarrow \mathbb{N}$ where $g(1) = 1$, $g(2) = 2$ and $g(n) = g(n-1) + g(n-2)$ for any $n \geq 3$.

Example A.3.9. In some cases functions are defined implicitly. Logarithms are such examples. Consider the function $h : (0, +\infty) \rightarrow \mathbb{R}$ where $h(x)$ is the unique real number such that $2^{h(x)} = x$. This is the same as $\log_2(x)$, the logarithm taken to base 2.

Definition A.3.7. Let f be a function from a set A into another set B . Given a subset $X \subseteq A$, the **image** of X under f , written as $f(X)$, is defined by

$$f(X) = \{y \in B : \exists x \in X, y = f(x)\} = \{f(x) : x \in X\} \subseteq B.$$

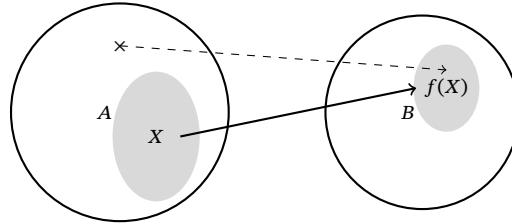


Figure A.3.3. Any element in $f(X)$ is the image of some element in X . In particular, $f(X)$ is a subset of the range of f . Hereby it is possible that also certain elements in $A \setminus X$ are mapped into $f(X)$.

If $f : A \rightarrow B$, then $\text{range}(f) = f(A)$.

Definition A.3.8. For $Y \subseteq B$, the **pre-image** of Y with respect to f , written as $f^{-1}(Y)$, is defined as

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

Example A.3.10. Define a function $f : \mathbb{N} \times \mathbb{N} \rightarrow \mathbb{N}$ by $f(m, n) = m + n$. Assume that $X = \{(m, n) : 1 \leq m, n \leq 4\}$, then $f(X) = \{2, 3, \dots, 8\}$.

On the other hand, if $Y = \{1, 2, 3, 4\}$, then we get

$$f^{-1}(Y) = \{(1, 1), (1, 2), (2, 1), (1, 3), (3, 1), (2, 2)\}.$$

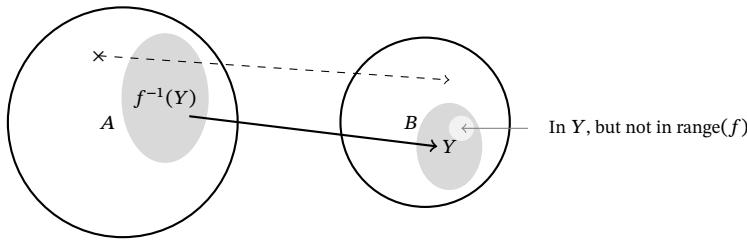


Figure A.3.4. Only elements from $f^{-1}(Y)$ are mapped to Y . All other elements in A have to have their image outside Y . But not every element in Y needs to be the image of some element in A .

Example A.3.11. Let $f : \mathbb{R} \rightarrow \mathbb{R}$ be defined by $f(x) = x^2$, $x \in \mathbb{R}$. Then

$$f([-2, 2]) = [0, 4] \quad \text{and} \quad f^{-1}([-1, 9]) = [-3, 3].$$

Example A.3.12. Let $H : \mathbb{R} \rightarrow \mathbb{R}$ be the so-called Heaviside function⁴ defined by

$$H(x) = \begin{cases} 0 & : x \leq 0 \\ 1 & : x > 0. \end{cases}$$

Then it follows (check) that

$$H([1, 2]) = \{1\}, \quad H([-1, 1]) = \{0, 1\} \quad \text{and} \quad H(\{-1, -2, -3\}) = \{0\}.$$

On the other hand (check too)

$$H^{-1}([2, 3]) = \emptyset, \quad H^{-1}([0.5, 1.5]) = (0, \infty) \quad \text{and} \quad H^{-1}([-1, 1]) = \mathbb{R}.$$

Proposition A.3.2. If $f : A \rightarrow B$ is a function, then the following statements are true.

- (1) $X \subseteq A \Rightarrow X \subseteq f^{-1}(f(X))$
- (2) $Y \subseteq B \Rightarrow f(f^{-1}(Y)) \subseteq Y$
- (3) $X_1, X_2 \subseteq A \Rightarrow f(X_1 \cup X_2) = f(X_1) \cup f(X_2)$
- (4) $X_1, X_2 \subseteq A \Rightarrow f(X_1 \cap X_2) \subseteq f(X_1) \cap f(X_2)$
- (5) $Y_1, Y_2 \subseteq B \Rightarrow f^{-1}(Y_1 \cup Y_2) = f^{-1}(Y_1) \cup f^{-1}(Y_2)$
- (6) $Y_1, Y_2 \subseteq B \Rightarrow f^{-1}(Y_1 \cap Y_2) = f^{-1}(Y_1) \cap f^{-1}(Y_2)$

Proof: To prove (1) recall that $f^{-1}(f(X))$ consists of all elements in A which are mapped into $f(X)$. But elements in X are, again by definition, mapped into $f(X)$, hence $X \subseteq f^{-1}(f(X))$.

Inclusion (2) follows by similar arguments. Recall that elements in $f^{-1}(Y)$ are those mapped into Y . Hence, by definition of the image, $f(f^{-1}(Y)) \subseteq Y$.

To prove (3) let us first choose some $y \in f(X_1 \cup X_2)$. Then $y = f(x)$ for some $x \in X_1 \cup X_2$. If $x \in X_1$, then $y = f(x) \in f(X_1)$. Similarly, if $x \in X_2$, then $y = f(x) \in f(X_2)$. Combining both cases it follows that $y \in f(X_1) \cup f(X_2)$. Hence, we proved

$$(A.3.2) \quad f(X_1 \cup X_2) \subseteq f(X_1) \cup f(X_2).$$

⁴This function is named after Oliver Heaviside (1850–1925), an English self-taught mathematician and physicist.

To verify the reversed inclusion, choose $y \in f(X_1) \cup f(X_2)$. Then either $y \in f(X_1)$ or $y \in f(X_2)$. In the first case, $y = f(x)$ for some $x \in X_1$ while in the latter case $y = f(x)$ for some $x \in X_2$. Consequently, in any case $y = f(x)$ for some $x \in X_1 \cup X_2$, hence $y \in f(X_1 \cup X_2)$, and we proved

$$(A.3.3) \quad f(X_1) \cup f(X_2) \subseteq f(X_1 \cup X_2).$$

Combining (A.3.2) and (A.3.3) proves assertion (3).

The proofs of properties (4), (5) and (6) go along the same lines and are left as exercise for the reader. \blacksquare

Definition A.3.9. Let $f : A \rightarrow B$ and $g : B \rightarrow C$ be two functions. The **composition** $g \circ f$ is the function from A to C , with

$$(g \circ f)(x) := g(f(x)), \quad x \in A.$$

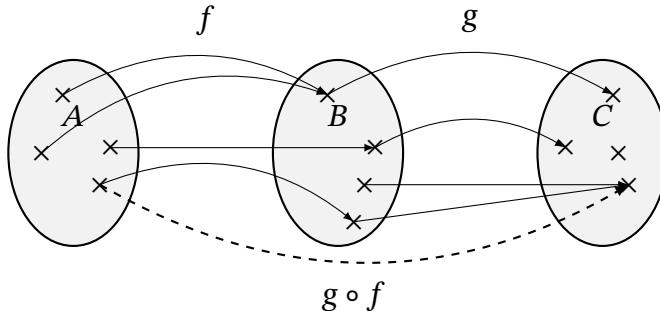


Figure A.3.5. Composition of functions f and g .

Note that $g \circ f$ is well-defined only when the codomain of f equals the domain of g . Hence, one should be careful when composing functions as the order is important. Just because $g \circ f$ is well-defined, it does not mean that $f \circ g$ is well-defined.

Example A.3.13. Let $f : \mathbb{R} \rightarrow [0, +\infty)$, $f(x) = x^2 + 1$ and $g : [0, \infty) \rightarrow [-1, 1]$ with $g(x) = \sin x$. Then $g \circ f$ is well-defined, $g \circ f : \mathbb{R} \rightarrow [-1, 1]$ and

$$(g \circ f)(x) = g(f(x)) = \sin(x^2 + 1),$$

for any $x \in \mathbb{R}$. In this case, $f \circ g$ is not defined because the codomain of g (which is $[-1, 1]$) does not equal the domain of f (which is \mathbb{R}).

If $f : A \rightarrow B$, $g : B \rightarrow C$ and $A = B = C$, then both $f \circ g$ and $g \circ f$ are well-defined, but note that in general $f \circ g \neq g \circ f$.

Example A.3.14. Let $f, g : \mathbb{N} \rightarrow \mathbb{N}$ be defined as follows

$$g(n) = n + 1, n \geq 1, \quad \text{and} \quad f(n) = n - 1, n \geq 2 \text{ while } f(1) = 1.$$

Then $(g \circ f)(n) = n$, $n \geq 2$, and $(g \circ f)(1) = 2$. On the other hand, $(f \circ g)(n) = n$ for all $n \geq 1$. This example shows that in general $f \circ g \neq g \circ f$.

Definition A.3.10. Let A be a set. The **identity map** i_A on A is the function $i_A : A \rightarrow A$ with

$$i_A(x) = x, \quad x \in A.$$

The following result is not too difficult, and we omit its proof.

Proposition A.3.3. If $f : A \rightarrow B$ is a function, then

$$f \circ i_A = f \quad \text{and} \quad i_B \circ f = f.$$

Definition A.3.11. A function $f : A \rightarrow B$ is **injective** or **one-to-one** or an **injection** from A to B if for any $x_1 \neq x_2 \in A$, $f(x_1) \neq f(x_2)$. In other words, whenever $f(x_1) = f(x_2)$ for some $x_1, x_2 \in A$, then necessarily $x_1 = x_2$.

$$(f \text{ injective}) \Leftrightarrow [(\forall x_1, x_2 \in A)[(f(x_1) = f(x_2)) \Rightarrow (x_1 = x_2)]]$$

Definition A.3.12. A function f from A to B is **surjective** (equivalently, f is called a **surjection**) provided that for each $y \in B$ there is at least one $x \in A$ with $f(x) = y$. In other words, each element $y \in B$ appears as the image of some $x \in A$. Another way to express that a function is surjective is that it satisfies $\text{range}(f) = B$.

$$(f \text{ surjective}) \Leftrightarrow [(\forall y \in B)(\exists x \in A)(f(x) = y)]$$

Definition A.3.13. Finally, a function is said to be **bijection** or called a **bijection**, if it is at the same time injective and surjective. That is, f from A to B is a bijection if and only if for each $y \in B$ there is exactly one $x \in A$ for which $f(x) = y$.

$$\begin{aligned} (f \text{ bijective}) &\Leftrightarrow (f \text{ injective and surjective}) \\ &\Leftrightarrow [(\forall y \in B)(\exists! x \in A)(f(x) = y)]. \end{aligned}$$

Remark A.3.1. Another way to characterize injective, surjective or bijective functions $f : A \rightarrow B$ is as follows: Given an arbitrary $y \in B$, then the equation $f(x) = y$

$$\begin{aligned} \text{has at most one solution} &\Leftrightarrow \text{if } f \text{ is injective} \\ \text{is always solvable} &\Leftrightarrow \text{if } f \text{ is surjective} \\ \text{is always uniquely solvable} &\Leftrightarrow \text{if } f \text{ is bijective.} \end{aligned}$$

There exists a tight relation between injective functions from A to B and bijections from A to a certain subset of B . Recall the definition of the range of a function as given in Definition A.3.4. The proof is straightforward, therefore we leave it as an easy but instructive exercise.

Proposition A.3.4. A function $f : A \rightarrow B$ is injective if and only if it is a bijection from A to $\text{range}(f) \subseteq B$.

$$(f : A \rightarrow B \text{ is injective}) \Leftrightarrow (f : A \rightarrow \text{range}(f) \text{ is bijective}).$$

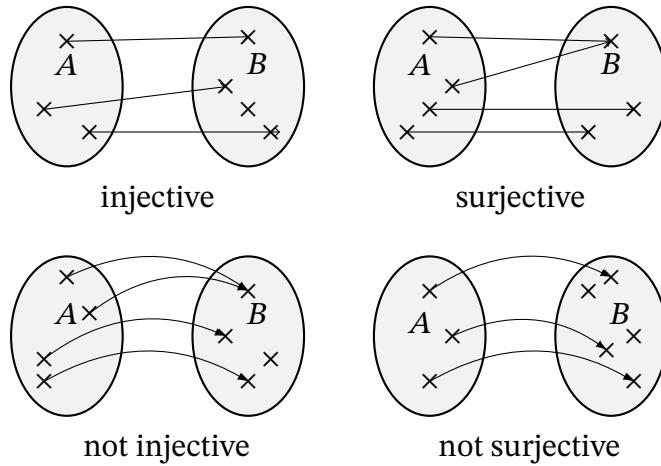


Figure A.3.6. Injective and surjective functions.

The following characterization of injective, surjective, and bijective functions is important.

Proposition A.3.5. *Let f be a function from a set A into a set B . Then the following are valid*

$$\begin{aligned} f \text{ is injective} &\Leftrightarrow [(\exists g : B \rightarrow A)(\forall x \in A)(g(f(x)) = x)] \\ f \text{ is surjective} &\Leftrightarrow [(\exists g : B \rightarrow A)(\forall y \in B)(f(g(y)) = y)] \\ f \text{ is bijective} &\Leftrightarrow (\exists! g : B \rightarrow A)[(\forall x \in A)(g(f(x)) = x) \text{ and} \\ &\quad (\forall y \in B)(f(g(y)) = y)]). \end{aligned}$$

Before proving Proposition A.3.5, let us rewrite the characterizations of injective, surjective and bijective functions. Recall that the identity function was introduced in Definition A.3.10.

$$\begin{aligned} f \text{ is injective} &\Leftrightarrow (\exists g : B \rightarrow A)(g \circ f = i_A) \\ f \text{ is surjective} &\Leftrightarrow (\exists g : B \rightarrow A)(f \circ g = i_B) \\ f \text{ is bijective} &\Leftrightarrow (\exists! g : B \rightarrow A)[(g \circ f = i_A) \text{ and } (f \circ g = i_B)]. \end{aligned}$$

Proof: Suppose first that there is a function g with $g(f(x)) = x$. Then the equation $f(x_1) = f(x_2)$ for some $x_1, x_2 \in A$ implies $x_1 = g(f(x_1)) = g(f(x_2)) = x_2$, hence f is injective.

Assume now that f is an injection. Let $f(A) := \{f(x) : x \in A\}$ be the range of f in B . Fix some $x_0 \in A$ and set

$$g(y) := \begin{cases} x & : y \in f(A), f(x) = y \\ x_0 & : y \notin f(A). \end{cases}$$

Note that the injectivity of f implies that g is well-defined. Of course, then $g(f(x)) = x$ for all $x \in A$.

To prove the second assertion we first assume the existence of a function g from B into A with $f(g(y)) = y$ for all $y \in B$. For a given $y \in B$, let $x := g(y)$. It follows that $f(x) = y$. Hence, f is a surjection.

Conversely, if f is surjective, for each $y \in B$ we define a subset $A_y \subseteq A$ by

$$(A.3.4) \quad A_y := \{x \in A : f(x) = y\}.$$

Since f is surjective, we always have $A_y \neq \emptyset$, and, moreover, $\bigcup_{y \in B} A_y = A$.

Summing up, we have a collection $\mathcal{A} = \{A_y : y \in B\}$ of nonempty sets whose union is the set A . In general, i.e., if f is not injective, the sets A_y will contain many elements. Our aim is now to single out exactly one element from each set A_y . And this has to be done for all sets A_y simultaneously. This is easy if there are only finitely many sets A_y , but turns out to be a difficult problem in the case of infinite sets B , i.e., in the case of infinitely many sets A_y . To overcome this problem we need the following axiom of set theory⁵⁶.

Axiom A.3.1 (Axiom of Choice). *Let \mathcal{A} be any collection of nonempty sets. Then there is a function $\varphi : \mathcal{A} \rightarrow \bigcup_{A \in \mathcal{A}} A$ such that for all $A \in \mathcal{A}$ it follows that $\varphi(A) \in A$. In other words, the element $\varphi(A)$ is the one chosen from the set A . Therefore, φ is called a **choice function**.*

Equivalently, for any collection \mathcal{A} of nonempty sets we may single out in each $A \in \mathcal{A}$ exactly one element which we denote by $\varphi(A)$.

Now we are in position to finish the proof of the second property. For each $y \in B$ let A_y be defined as above. Due to the axiom of choice, for each $y \in B$ we may choose some element $x \in A_y$ which we denote by $g(y)$. In the formulation of the axiom we have

$$g(y) = \varphi(A_y), \quad y \in B.$$

Then g is a well-defined mapping from B into A which, by construction, satisfies $g(y) \in A_y$, $y \in B$, hence $f(g(y)) = y$.

The proof of the third assertion is easier. If such a function g exists, then by the two previous results f is as well injective as surjective, hence bijective.

Conversely, if f is a bijection, the function g may be defined by

$$g(y) = x \quad \text{provided that} \quad f(x) = y, \quad y \in B.$$

⁵⁶The axiom of choice was originally formulated in 1904 by Ernst Zermelo. In 1938, Kurt Gödel (1906–1978) showed that the axiom of choice is consistent with the common axioms of set theory, and in 1963, the American mathematician Paul J. Cohen (1934–2007) proved that its negation is also consistent with them. Thus, neither the axiom of choice nor its negation can be derived from other properties of sets. The axiom of choice has far-reaching consequences in several fields of mathematics. For example, it implies the existence of a linear basis in each vector space or of nonmeasurable subsets in the real line. We refer to [17] for further reading about this exciting topic.

⁶Kurt Gödel (1906–1978) was born in Brünn, Austria-Hungary (now Brno, Czech Republic), and was educated and worked at University of Vienna, Austria. In 1939, he fled Europe to US and worked for the rest of his life at the Institute for Advanced Studies at Princeton.

It remains to show the uniqueness of the function g satisfying the third assertion. Thus, let us assume the existence of another function h from B into A for which $h \circ f = i_A$ and $f \circ h = i_B$. Then this implies

$$h = i_A \circ h = g \circ f \circ h = g \circ i_B = g.$$

Thus, g is uniquely determined and this completes the proof. ■

Definition A.3.14. Let f be a function from a set A into a set B . A function $g : B \rightarrow A$ is said to be

- (1) **left-inverse** to f if $g \circ f = i_A$,
- (2) **right-inverse** to f if $f \circ g = i_B$ and
- (3) **inverse** to f if $g \circ f = i_A$ and $f \circ g = i_B$.

Remark A.3.2. Proposition A.3.5 asserts that f has a left-inverse if and only if f is injective, it is surjective if and only if a right-inverse function exists. And, finally, f is bijective if and only if it possesses an inverse function.

$$\begin{aligned} f \text{ is injective} &\Leftrightarrow f \text{ has a left-inverse } g \\ f \text{ is surjective} &\Leftrightarrow f \text{ has a right-inverse } g \\ f \text{ is bijective} &\Leftrightarrow f \text{ has an inverse } g. \end{aligned}$$

Example A.3.15. Suppose $f : \mathbb{N}_0 \rightarrow \mathbb{N}_0$ is defined by

$$f(n) = \left\lfloor \frac{n}{10} \right\rfloor, \quad n = 0, 1, 2, \dots$$

Then f is a surjection but no injection. For each $m \in \mathbb{N}_0$ let us define the sets A_m as in (A.3.4) by

$$A_m = \{n \in \mathbb{N}_0 : f(n) = m\}.$$

For example,

$$A_0 = \{0, 1, \dots, 9\} \quad \text{and} \quad A_1 = \{10, \dots, 19\}.$$

Then we have to choose (here we do not need the axiom of choice) one element $n = g(m)$ with $n \in A_m$. For example, let us choose the smallest element $10m$ in A_m . Hence, a suitable right-inverse function from \mathbb{N}_0 to \mathbb{N}_0 is given by $g(m) = 10m$, $m = 0, 1, \dots$. Another right-inverse function would be $g(m) = 10m + 1$. The reader is encouraged to find more right-inverse functions of f .

While, in general, a function f may have several left- or right-inverse functions, the inverse function, whenever it exists, is unique. Therefore, the following definition makes sense.

Definition A.3.15. Let f be a bijection from A to B . Its inverse function from B to A is denoted by f^{-1} .

$$(f^{-1} : B \rightarrow A \text{ is inverse to } f : A \rightarrow B) \Leftrightarrow (f^{-1} \circ f = i_A \text{ and } f \circ f^{-1} = i_B).$$

The following properties of the inverse function are crucial.

Proposition A.3.6.

- (1) For any bijection f its inverse is bijective as well, and, moreover, $(f^{-1})^{-1} = f$.
- (2) Let f be a bijection from A to B and g a bijection from B to C . Then $g \circ f$ is a bijection from A to C and $(g \circ f)^{-1} = f^{-1} \circ g^{-1}$.

Proof: The first assertion follows directly by

$$f^{-1} \circ f = i_A \quad \text{and} \quad f \circ f^{-1} = i_B .$$

By Proposition A.3.5, part 3, the inverse f^{-1} is bijective and, moreover, f is the inverse of f^{-1} . Recall that g from A to B is inverse to f^{-1} if and only if

$$f^{-1} \circ g = i_A \quad \text{and} \quad g \circ f^{-1} = i_B .$$

The second part follows by

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

as well as

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

■

For later applications, the following result will be useful.

Proposition A.3.7. *Let $f : A \rightarrow B$ be some function.*

- (1) *Let g be a left-inverse function of (the injection) f . Then g is surjective.*
- (2) *If g is right-inverse to (the surjection) f , then g is injective.*

Proof: The proof is a direct consequence of Proposition A.3.5. Indeed, if g is left-invariant to f , then $g \circ f = i_A$. But this says that f is right-inverse to $g : B \rightarrow A$. Hence, by Proposition A.3.5 the function g is surjective. The second part follows by exactly the same reasoning. If g is right-inverse to f , then f is left-inverse to g . Hence, by Proposition A.3.5 it is injective. ■

Corollary A.3.8. *There exists an injection from a set A into a set B if and only if there is a surjection from B to A .*

Proof: This is a direct consequence of Proposition A.3.7. ■

Example A.3.16.

- (1) Define f from $A = \mathbb{N} \rightarrow \mathbb{N}$ by $f(n) = 2n$ for $n \in \mathbb{N}$. The function f is injective, but not surjective. A left-inverse function g of f is for example given by

$$g(n) := \begin{cases} n/2 & : n \text{ even} \\ 1 & : n \text{ odd} \end{cases}$$

This example shows that, in general, there may be many left-inverse functions g for an injective f . For example, here we may change g easily to \tilde{g} with $\tilde{g}(n) = n/2$ for even n and $\tilde{g}(n) = 2$ if n is odd.

- (2) Let A and B as before and define $f(n) = \lfloor \frac{n+1}{2} \rfloor$, where $\lfloor \cdot \rfloor$ denotes the floor function, i.e., it maps x to the largest integer smaller than x . Then f is surjective, but because of $f(1) = 1 = f(2)$ it is not injective. As right-inverse function g from B to A we may take $g(n) = 2n - 1$.
- (3) Let as before $A = \mathbb{N}$, but now B is $\{1, 4, 9, 16, 25, 36, \dots\}$. Then f with $f(n) = n^2$ is injective and surjective, hence a bijection. The inverse function of f is given by $f^{-1}(m) = \sqrt{m}$, $m \in B$.

Definition A.3.16. Let A be a nonempty set. A bijective function $f : A \rightarrow A$ is said to be a **permutation** on A . Then

$$S_A := \{f : f \text{ bijection from } A \text{ to } A\}$$

is called the group of bijections (permutations) on A .

Why is S_A called a group? Define on S_A a binary operation

$$(f, g) \mapsto f \circ g, \quad (f, g) \in S_A \times S_A.$$

Then this operation possesses the following properties:

- (1) This operation is associative, i.e., for all f, g and h in S_A it follows that

$$(f \circ g) \circ h = f \circ (g \circ h).$$

- (2) There is a unit or neutral element e in S_A satisfying

$$e \circ f = f \circ e = f, \quad f \in S_A.$$

Take $e = i_A$ with identity map i_A on A .

- (3) And finally each element in S_A possesses an inverse element. That is, for each $f \in S_A$ there is a (unique) $g \in S_A$ such that

$$f \circ g = g \circ f = e.$$

To see this choose as g the inverse bijection f^{-1} .

As already mentioned in Remark 2.4.1, a group is a set with a binary operation satisfying the previous three properties. Examples of other important groups are \mathbb{R} with the binary operation

$$(x, y) \mapsto x + y, \quad (x, y) \in \mathbb{R} \times \mathbb{R},$$

or the set $\mathbb{Z}_p^* = \{1, \dots, p-1\}$ for some prime $p \geq 2$ where the binary operation is

$$(a, b) \mapsto a \cdot b \pmod{p}.$$

Other groups are $(\mathbb{Z}, +)$, $(\mathbb{R} \setminus \{0\}, \cdot)$, $(\mathbb{C}, +)$ and many more. All these groups possess an additional property which in general is not satisfied in the case of (S_A, \circ) . Let us make this precise.

Definition A.3.17. A group G with binary operation $*$ is called **commutative** or **abelian**⁷ provided that

$$x * y = y * x, \quad \forall x, y \in G.$$

⁷In honor of the Norwegian mathematician Niels Henrik Abel (1802–1829).

Clearly, addition and multiplication are always commutative operations, but as soon as $|A| > 2$, the composition of bijections on A is not so. Compare Exercise A.3.19 to see that (S_A, \circ) is not abelian in the case $A = \{1, 2, 3\}$. Of course, this implies that (S_A, \circ) is nonabelian as soon as A consists of three or more elements.

If A is finite, say $A = \{a_1, \dots, a_n\}$ then any permutation of A is generated by a permutation of $[n] = \{1, \dots, n\}$. Consequently, in this case it suffices to investigate bijections from $[n]$ to $[n]$. We write

$$S_n := S_{[n]} = \{\pi : \pi \text{ is a bijection from } [n] \text{ to } [n]\}.$$

The group S_n is called the **symmetric group of permutations** of $\{1, \dots, n\}$. Its elements are called permutations of order n and they are usually denoted by Greek letters: π, μ, σ . Recall that we already proved in Theorem 1.7.2, that

$$|S_n| = n! = 1 \cdot 2 \cdots (n-1) \cdot n.$$

Remark A.3.3. A permutation $\pi \in S_n$ reorders the numbers from 1 to n . Instead of the order $1, 2, \dots, n$, the new order is now given by $\pi(1), \pi(2), \dots, \pi(n)$. For example, if $n = 3$ and $\pi \in S_3$ is given by $\pi(1) = 2, \pi(2) = 3$ and $\pi(3) = 1$, the generated order of the three numbers is now 2, 3, 1.

One way to describe permutations is by its table of values. That is, given $\pi \in S_n$, one writes

$$\pi = \begin{pmatrix} 1 & \cdots & n \\ \pi(1) & \cdots & \pi(n) \end{pmatrix}.$$

For example, if a permutation $\pi \in S_4$ is given by

$$\pi = \begin{pmatrix} 1 & 2 & 3 & 4 \\ 3 & 1 & 4 & 2 \end{pmatrix},$$

this means that $\pi(1) = 3, \pi(2) = 1, \pi(3) = 4$ and $\pi(4) = 2$.

Note that the identity map i_n (identity permutation) is in this setting equals

$$i_n = \begin{pmatrix} 1 & \cdots & n \\ 1 & \cdots & n \end{pmatrix}$$

Given two permutations π and μ in S_n , their composition (product) $\pi \circ \mu$ can be visualized as follows

$$\begin{pmatrix} 1 & \cdots & \mu(k) & \cdots & n \\ \pi(1) & \cdots & \pi(\mu(k)) & \cdots & \pi(n) \end{pmatrix} \circ \begin{pmatrix} 1 & \cdots & k & \cdots & n \\ \mu(1) & \cdots & \mu(k) & \cdots & \mu(n) \end{pmatrix}.$$

For example, we have

$$\begin{pmatrix} 1 & 2 & 3 & 4 \\ 3 & 1 & 4 & 2 \end{pmatrix} \circ \begin{pmatrix} 1 & 2 & 3 & 4 \\ 4 & 3 & 2 & 1 \end{pmatrix} = \begin{pmatrix} 1 & 2 & 3 & 4 \\ 2 & 4 & 1 & 3 \end{pmatrix}.$$

How to get in this notation the inverse of a permutation $\pi \in S_n$? Because of the way two permutations are composed, the inverse π^{-1} of $\pi \in S_n$ may be represented as

$$\pi^{-1} = \begin{pmatrix} \pi(1) & \cdots & \pi(n) \\ 1 & \cdots & n \end{pmatrix}.$$

If we now reorder the upper line in its natural way, we get a representation of π^{-1} in its canonical form. For example, if

$$\pi = \begin{pmatrix} 1 & 2 & 3 & 4 \\ 3 & 1 & 4 & 2 \end{pmatrix},$$

then

$$\pi^{-1} = \begin{pmatrix} 3 & 1 & 4 & 2 \\ 1 & 2 & 3 & 4 \end{pmatrix} = \begin{pmatrix} 1 & 2 & 3 & 4 \\ 2 & 4 & 1 & 3 \end{pmatrix}.$$

A special type of permutations are so-called **transpositions**. They interchange two numbers and leave all other unchanged.

Definition A.3.18. A permutation $\pi \in S_n$ is said to be a **transposition** if there are two numbers $1 \leq k < \ell \leq n$ such that $\pi(k) = \ell$, $\pi(\ell) = k$ and $\pi(j) = j$ for all j different from k and ℓ .

For example, $\pi = \begin{pmatrix} 1 & 2 & 3 & 4 \\ 3 & 2 & 1 & 4 \end{pmatrix}$ is a transposition because it only interchanges 1 and 3 while the two other numbers 2 and 4 remain fixed. Note that any transposition π is idempotent, i.e., it satisfies $\pi \circ \pi = i_n$. The next result shows that any permutation can be obtained by composing transpositions. The intuition behind its proof is related to ordering a group of n people by height, where we start from one end and iteratively, place a person in their place by at most one permutation.

Proposition A.3.9. Any permutation $\pi \in S_n$ can be written as a product of transpositions. Moreover, although this representation is not unique, the parity of the number of transpositions is determined by π .

Proof: If π is the identity permutation, then π is a product of zero transpositions. Otherwise, let j be the smallest number such that $\pi(j) \neq j$. Denote by $\ell \neq j$ the number such that $\pi(\ell) = j$. Let (j, ℓ) be the transposition interchanging the numbers j and ℓ . Then the permutation $\pi' = \pi \circ (j, \ell)$ has the property that $\pi'(j) = \pi(\ell) = j$, $\pi'(\ell) = \pi(j)$ and $\pi'(a) = \pi(a)$ for any $a \neq j, \ell$. The number of fixed points of π' is greater than the corresponding one of π . By repeating this process with π' , we must eventually reach the identity permutation. When we do, we can reverse it and write π as a product of transpositions. ■

Remark A.3.4. A permutation π is said to be **even** if it can be written as an even product of transpositions. Otherwise, it is said to be an **odd** permutation of order n .

Another useful way of describing permutations is the canonical cyclic notation.

Definition A.3.19. A sequence (i_1, \dots, i_m) of m different numbers in $\{1, \dots, n\}$ is called a **cycle** of length m of a permutation $\pi \in S_n$ if

$$\pi(i_1) = i_2, \pi(i_2) = i_3, \dots, \pi(i_m) = i_1.$$

Example A.3.17. The permutation

$$\pi = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 \\ 3 & 2 & 4 & 5 & 1 \end{pmatrix}$$

has the cycles $(1, 3, 4, 5)$ and (2) .

In this way any permutation is completely described by its cycles where it is common not to mention cycles of length 1. That is, if a number $k \leq n$ does not occur in any of the cycles of π , then this tells us that $\pi(k) = k$.

Example A.3.18. Suppose $\pi \in S_7$ has the cycles $(1, 4, 7)$ and $(2, 6, 5)$. Then the corresponding permutation is

$$\pi = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 \\ 4 & 6 & 3 & 7 & 2 & 5 & 1 \end{pmatrix}.$$

In Definition 1.2.1 we introduced $\sum_{k=1}^n a_k$ and $\prod_{k=1}^n a_k$ as abbreviations for $a_1 + \dots + a_n$ and $a_1 \cdot \dots \cdot a_n$, respectively. After having investigated permutations, we are now able to state some important and useful properties of these notations. The proof of the subsequent formulas is straightforward, and therefore we omit it.

Proposition A.3.10. Let a_1, \dots, a_n be arbitrary numbers. The following statements are true.

- (1) If π is a permutation in S_n , then we may **change the order of summation and multiplication** in the following way.

$$\sum_{k=1}^n a_{\pi(k)} = \sum_{k=1}^n a_k \quad \text{and} \quad \prod_{k=1}^n a_{\pi(k)} = \prod_{k=1}^n a_k.$$

- (2) A special case of the preceding property is the useful method of **backward summation or multiplication**.

$$\sum_{k=1}^n a_{n-k+1} = \sum_{k=1}^n a_k \quad \text{and} \quad \prod_{k=1}^n a_{n-k+1} = \prod_{k=1}^n a_k.$$

- (3) A tightly related technique for the evaluation of finite sums and products is to **shift the index**⁸. For all $m \in \mathbb{Z}$ one may shift the index of the sum and of the product in the following way.

$$\sum_{k=m+1}^{n+m} a_{k-m} = \sum_{k=1}^n a_k \quad \text{and} \quad \prod_{k=m+1}^{n+m} a_{k-m} = \prod_{k=1}^n a_k.$$

So, for example, this implies

$$\sum_{k=1}^n a_k = \sum_{k=0}^{n-1} a_{k+1} = \sum_{k=3}^{n+2} a_{k-2}.$$

Example A.3.19. Say we want to evaluate $\sum_{k=m+1}^n q^k$ for $m < n$ and $q \neq 1$. One possibility to do so is as follows: Shifting the index and applying the summation formula for a geometric progression (Proposition 1.3.2) leads to

$$\sum_{k=m+1}^n q^k = q^{m+1} \sum_{k=m+1}^n q^{k-m-1} = q^{m+1} \sum_{k=0}^{n-m-1} q^k = q^{m+1} \cdot \frac{1 - q^{n-m}}{1 - q} = \frac{q^{m+1} - q^{n+1}}{1 - q}.$$

⁸Although this technique is obvious and easy to handle, we observed quite often that many students have serious problems to understand it. That is the reason why we discuss this technique here in more detail.

An alternative, maybe shorter, approach is to use

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

Example A.3.20. Because of the usefulness of index shifts let us give another example where this technique turns out to be very helpful.

$$\sum_{k=1}^n k \binom{n}{k} = n \sum_{k=1}^n \frac{(n-1)!}{(k-1)!(n-k)!} = n \cdot \sum_{k=0}^{n-1} \frac{(n-1)!}{k!(n-1-k)!} = n \cdot \sum_{k=0}^{n-1} \binom{n-1}{k} = n \cdot 2^{n-1}.$$

The last step of the calculation rests upon an application of Theorem 1.7.6 with $n-1$ and with $a=b=1$.

Exercise A.3.1. Suppose f and g are two linear functions from \mathbb{R} to \mathbb{R} . That is, for some real numbers a, b, c and d we have

$$f(x) = a + bx \quad \text{and} \quad g(x) = c + dx, \quad x \in \mathbb{R}.$$

Which assumptions about the coefficients a, b, c , and d imply

$$f \circ g = g \circ f ?$$

Exercise A.3.2. For any function $f : A \rightarrow A$ one may define its n th iteration by

$$f^1(x) = f(x) \quad \text{and} \quad f^n(x) = f(f^{n-1}(x)), \quad x \in A.$$

Another way to describe the iteration of order n is

$$f^n = \underbrace{f \circ \cdots \circ f}_{n \text{ times}}.$$

Describe f^n and g^n for f and g given by

$$f(x) = \frac{x}{1+x} \quad \text{and} \quad g(x) = \sqrt{x}, \quad x > 0.$$

Exercise A.3.3. Prove (4), (5) and (6) in Proposition A.3.2

Exercise A.3.4. Find examples which show that in general the inclusions in properties (1), (2) and (4) of Proposition A.3.2 cannot be replaced by equality signs.

Exercise A.3.5. Verify the following: Let $f : A \rightarrow B$ be any function. Then for all subsets $X \subseteq A$ and $Y \subseteq B$ one has

$$f^{-1}(f(X)) = \{a \in A : \exists x \in X \ f(a) = f(x)\} \quad \text{and} \quad f(f^{-1}(Y)) = Y \cap f(X).$$

Why does this imply $f^{-1}(f(X)) = X$ if f is injective and $f(f^{-1}(Y)) = Y$ for surjective functions f ?

Exercise A.3.6. Define f from \mathbb{N} to \mathbb{Z} by

$$f(n) = \begin{cases} 0 & : n \text{ even} \\ 1 & : n \text{ odd} \end{cases}$$

Describe $f^{-1}(Y)$ for all $Y \subseteq \mathbb{Z}$.

Exercise A.3.7. Determine

$$f([0, 5]) \quad \text{and} \quad f^{-1}([0, \infty))$$

where $f : \mathbb{R} \rightarrow \mathbb{R}$ denotes the floor function. That is,

$$f(x) = \lfloor x \rfloor, \quad x \in \mathbb{R}.$$

Exercise A.3.8. Let f be a function from a set A to a set B and let g be a function from B to a set C . Prove the following assertions:

- (1) If f and g are injective, then so is $g \circ f$.
- (2) If f and g are surjective, then so is $g \circ f$.

Exercise A.3.9. Define $f : \mathbb{N} \times \mathbb{N} \rightarrow \mathbb{N}$ by

$$f(m, n) = \min\{m, n\}.$$

Why is f surjective but not injective? Find a right-inverse function $g : \mathbb{N} \rightarrow \mathbb{N} \times \mathbb{N}$ of f .

Exercise A.3.10. Let $f : \mathbb{N} \rightarrow \mathbb{N}$ be given by $f(n) = 3n^2 - 2$. Show that f is injective. Give a left-inverse function $g : \mathbb{N} \rightarrow \mathbb{N}$ of f .

Exercise A.3.11. Suppose $f_1 : A_1 \rightarrow B_1$ and $f_2 : A_2 \rightarrow B_2$ are both bijections. Prove that then $g : A_1 \times A_2 \rightarrow B_1 \times B_2$ with

$$g(a_1, a_2) := (f_1(a_1), f_2(a_2)), \quad (a_1, a_2) \in A_1 \times A_2,$$

is bijective as well. How can the inverse function of g expressed by the inverse functions of f_1 and f_2 ?

Exercise A.3.12. Define $f : (0, \infty) \rightarrow (0, 1)$ by

$$f(x) = \frac{x}{1+x}, \quad 0 < x < \infty.$$

Show that f is bijective and determine its inverse function.

Exercise A.3.13. Suppose $f : A \rightarrow B$ is a bijection. Show that the graph of its inverse function is a subset of $B \times A$ which may be represented as

$$\text{graph}(f^{-1}) = \{(f(x), x) : x \in A\}.$$

In other words, if $R : A \times B \rightarrow B \times A$ defined by

$$R : (x, y) \mapsto (y, x), \quad (x, y) \in A \times B,$$

reflects any point $(x, y) \in A \times B$ to (y, x) , then it follows that

$$R(\text{graph}(f)) = \text{graph}(f^{-1}).$$

In the case $A = B$, the function R is a reflection at the diagonal $\{(x, x) : x \in A\}$

Exercise A.3.14. Define $f : [-1, 1] \rightarrow [-1, 1]$ by

$$f(x) = x|x|, \quad -1 \leq x \leq 1.$$

- (1) Sketch the graph of f .
- (2) Is f bijective?
- (3) If the answer to (2) is affirmative, determine the inverse function of f .

Exercise A.3.15. As is well-known, the functions $x \mapsto \sin x$ and $x \mapsto \cos x$ from \mathbb{R} to $[-1, 1]$ are surjective but not injective. Find in both cases suitable right-inverse functions.

Exercise A.3.16. Suppose $f : \mathbb{N} \rightarrow \mathbb{N}$ is defined by $f(n) = n + 2$. Then f is injective. Find at least two different functions from \mathbb{N} to \mathbb{N} which are left-inverse to f . Is f also surjective? Justify your answer.

Exercise A.3.17. (\star) As mentioned before, in general it is impossible without the axiom of choice to pick out exactly one element of each set from a collection of sets. Even if $\mathcal{A} = \{A_1, A_2, \dots\}$ with $|A_j| = 2$, without the axiom of choice, in general there is no way to select one element of each of the A_j 's at the same time.

There exist some examples where one may define a choice function without using the axiom of choice. Show that this is so in each of the following cases:

- (1) The collection of sets is finite. That is, $\mathcal{A} = \{A_1, \dots, A_n\}$ for some nonempty sets A_j .
- (2) The sets in \mathcal{A} are all finite intervals of real numbers.
- (3) The sets in \mathcal{A} are all nonempty subsets of \mathbb{N} .

Exercise A.3.18. Evaluate for any $n \geq 1$ the sum

$$\sum_{k=1}^n k^2 \binom{n}{k}.$$

Hint: Shift the index of summation in a suitable way.

Exercise A.3.19. Find all permutations in S_3 . Evaluate $\pi \circ \mu$ for all $\pi, \mu \in S_3$ and show that, in general, we do not have $\pi \circ \mu = \mu \circ \pi$. Determine π^{-1} for all $\pi \in S_3$.

A.4. Cardinality of Sets

In Section 1.7 we investigated the cardinality of finite sets. The set A was said to be finite if either $A = \emptyset$, then its cardinality is $|A| = 0$, or there exists a natural number $n \geq 1$ such that each element in A has a unique label between 1 and n . In other words, the set A may be written as

$$A = \{a_1, \dots, a_n\}.$$

In this case we define its cardinality as $|A| = n$. Thus, $|A| = n$ if and only if there is a bijective function from $[n] = \{1, \dots, n\}$ to A which assigns to each $k \leq n$ the (unique) element $a_k \in A$. That is, the number $k \leq n$ is mapped to the element in A with label k . If we use the notation introduced in Section 1.7, then this may be formulated as follows:

$$|A| = n \Leftrightarrow (\exists f : [n] \rightarrow A)(f \text{ bijective}).$$

Note that we get for free from the definition that $|A| = n$ is also equivalent with the existence of a bijective function g from A to $[n]$. Take as g the inverse of the bijection f from $[n]$ to A .

Informally, we use this definition all the time when counting finite sets (fingers, steps, stairs and so on) and we essentially create a bijection between the set of first n

natural numbers (for some n) and whatever set interests us. We have 5 fingers as we implicitly create a bijection between $[5]$ and the set of fingers F such as for example $f : [5] \rightarrow F$, $f(1)=\text{pinkie}$, $f(2)=\text{ring finger}$, $f(3)=\text{middle finger}$, $f(4)=\text{pointer}$ and $f(5)=\text{thumb}$. Or we do it the other way round: We take the inverse of f and assign the pinkie to 1, the ring finger to 2, and so on. For example, the set $A = \{\emptyset, 1, \{\emptyset\}, \{2, 3\}\}$ is definitely a weird looking set and its cardinality is 4 as the function $f : [4] \rightarrow A$, $f(1) = \emptyset$, $f(2) = 1$, $f(3) = \{\emptyset\}$, $f(4) = \{2, 3\}$ is a bijective function.

Since the composition of bijective functions is bijective as well, the following obvious result is true:

Proposition A.4.1. *If A and B are two sets such that there exists a bijection from A to B , then A is finite if and only if B is so. Moreover, it follows $|A| = |B|$.*

For example, if we are able to assign to each of our fingers in an one-to-one way one apple of a set of apples, then we can deduce from this that there are exactly as many apples as fingers. And since we know, that we have 5 fingers at one hand, it follows that there are also 5 apples.

It seems to be obvious that subsets of finite sets are finite as well. But let us give a rigorous proof of this.

Proposition A.4.2. *Let A be a finite set. If $B \subseteq A$, then B is also finite with $|B| \leq |A|$.*

Proof: Assume $|A| = n$ and let $f : [n] \rightarrow A$ be a bijection. The pre-image $f^{-1}(B)$ consists of numbers in $\{1, \dots, n\}$. Say $f^{-1}(B) = \{i_1, \dots, i_m\}$ where we order these numbers such that $1 \leq i_1 < \dots < i_m \leq n$. Of course, $m \leq n$, and we can define a bijection g from $[m]$ to $f^{-1}(B)$ by $g(k) := i_k$, $1 \leq k \leq m$. By the construction $f^{-1} \circ g$ is a bijection from $[m]$ to B . Hence, B is finite with $|B| = m \leq n = |A|$. ■

The following result is a bit more general.

Proposition A.4.3. *Suppose that B is a finite set. Then for any set A the following are equivalent:*

- (1) *The set A is finite with $|A| \leq |B|$*
- (2) *There exists an injection $f : A \rightarrow B$.*
- (3) *There exists a surjection $g : B \rightarrow A$.*

Proof: We first prove that properties (2) and (3) are equivalent. This is an immediate consequence of Corollary A.3.8 which asserts that there exists an injection from a set A to a set B if and only if there is a surjection from B to A .

Next let us assume that (1) is true, i.e. A is finite with $|A| \leq |B|$. Say $|A| = m$ and $|B| = n$ with $m \leq n$. As we mentioned above, there exists a bijection $g : A \rightarrow [m]$ and, moreover, by definition we also have a bijection $h : [n] \rightarrow B$. Let $J : [m] \rightarrow [n]$ be

given by

$$J(k) = k, \quad 1 \leq k \leq m.$$

Then, as can be easily seen,

$$f := h \circ J \circ g$$

is an injective function from A to B . This proves (2).

Suppose now that there is an injective function $f : A \rightarrow B$. Let

$$B_0 := \text{range}(f) = \{f(a) : a \in A\}.$$

Of course, f is a bijection from A to B_0 , hence it follows that $|A| = |B_0|$. But $B_0 \subseteq B$, which by Proposition A.4.2 implies $|B_0| \leq |B|$. Putting these two estimates together we finally obtain $|A| \leq |B|$ as claimed in (1). This completes the proof. ■

Our next aim is to extend the previous considerations to **infinite** sets. Before let us precise what an infinite set is.

Definition A.4.1. A set A is said to be **infinite** if is not finite. That is, $A \neq \emptyset$ and there is no $n \geq 1$ such that there exist a bijection between $[n]$ and A . In other words, for any a_1, \dots, a_n in A there exists an $a \in A$ such that $a \notin \{a_1, \dots, a_n\}$.

What does it mean that two infinite sets possess the same cardinality? For example, one may ask whether \mathbb{Q} and \mathbb{R} are of the same size. Before we answer those and similar questions, let us first fix the notation.

Definition A.4.2. Let A and B be two (finite or infinite) sets. We say that they possess the same cardinality provided that there exists a bijection f mapping A onto B . We write

$$|A| = |B| \Leftrightarrow (\exists f \text{ bijection})(f : A \rightarrow B).$$

Similarly, the size of the set A is equal or smaller than the size of B provided that

$$|A| \leq |B| \Leftrightarrow (\exists f \text{ injection})(f : A \rightarrow B).$$

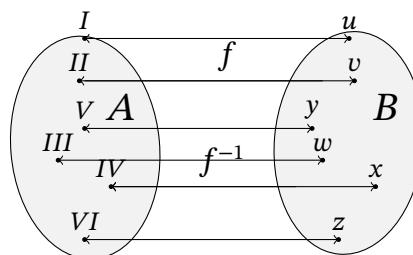


Figure A.4.1. If $A = \{I, \dots, VI\}$ and $B = \{u, \dots, z\}$, then $|A| = |B|$.

Remark A.4.1. In view of Corollary A.3.8 it follows that

$$|A| \leq |B| \Leftrightarrow (\exists g \text{ surjection})(g : B \rightarrow A).$$

Example A.4.1. (1) Let $2\mathbb{N}$ be the set of even integers. If we define $f : \mathbb{N} \rightarrow 2\mathbb{N}$ by

$$f(k) = 2k, \quad k \in \mathbb{N},$$

then f is bijective from \mathbb{N} to $2\mathbb{N}$, hence

$$|\mathbb{N}| = |2\mathbb{N}|.$$

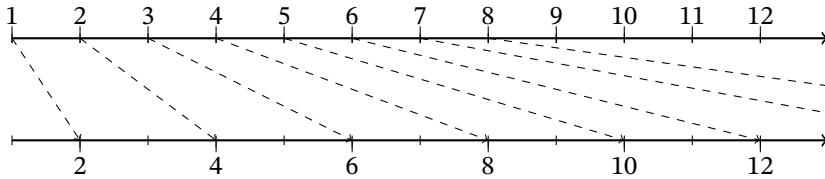


Figure A.4.2. $|\mathbb{N}| = |2\mathbb{N}|$.

(2) In the same way one may show that all the sets

$$\{2, 3, 4, \dots\} \quad \{1, 10, 100, 10^4, \dots\} \quad \text{and} \quad \left\{1, \frac{1}{2}, \frac{1}{3}, \dots\right\}$$

possess the same cardinality as \mathbb{N} .

These examples show that there appear quite new phenomenons in the setting of infinite sets. Note that for finite sets B one has always $|A| < |B|$ whenever $A \subseteq B$ with $A \neq B$. Here $2\mathbb{N}$ or $\{2, 3, 4, \dots\}$ are strict subsets of \mathbb{N} , but nevertheless they are of the same size as \mathbb{N} . The example that $\{2, 3, 4, \dots\}$ is of the same size as \mathbb{N} is also known as Hilbert's Hotel⁹.

The previous example suggests the following definition.

Definition A.4.3. A set A is called **countably infinite** provided there exists a bijection from \mathbb{N} to A . It is said to be **countable** whenever it is either finite or countably infinite.

An (infinite) noncountable set A is said to be **uncountable**. That is, A is neither finite nor does there exist a bijection between A and \mathbb{N} .

Remark A.4.2. Suppose A is countably infinite and let $f : \mathbb{N} \rightarrow A$ be a bijection. Set $a_1 = f(1), a_2 = f(2)$ and so on, it follows that

$$A = \{f(n) : n \in \mathbb{N}\} = \{a_n : n \geq 1\} = \{a_1, a_2, \dots\}.$$

That is why countable sets are also called **denumerable**, i.e., we can enumerate the elements in A in the same way as we can do with either finite sets or with the natural numbers.

⁹A guest enters a hotel and asks for a free room, but the hotel is booked up completely; no room is available. Fortunately, the hotel has infinitely many rooms, so the concierge asks the guest in room 1 to move to room 2, the guest in room 2 has to move to room 3, the one in room 3 to room 4 and so on. At the end, nobody had to leave the hotel, but nevertheless, room 1 is now free and the visitor has got an accommodation.

One may ask whether there exist nonfinite sets with cardinality strictly less than that of \mathbb{N} . The answer is negative as the following proposition shows.

Proposition A.4.4. *A subset of a countably infinite set is either finite or countably infinite as well.*

Proof: Of course, it suffices to show that any subset $A \subseteq \mathbb{N}$ is either finite or countably infinite. More precisely, we have to verify that any nonfinite subset of \mathbb{N} is countably infinite. Set $A_1 = A$. By the well-ordering property of \mathbb{N} there exists a minimal element $a_1 \in A_1$. Let us look at

$$A_2 := A \setminus \{a_1\}.$$

Since A is assumed to be nonfinite, we conclude that $A_2 \neq \emptyset$. Another application of the well-ordering principle yields the existence of a minimal element $a_2 \in A_2$. By the construction of A_2 we obtain

$$a_1 < a_2.$$

Then by the same reasoning as before the set $A_3 := A \setminus \{a_1, a_2\}$ is nonempty, and there exists a minimal element a_3 of A_3 .

Proceeding further in this way we get subsets $A = A_1 \supseteq A_2 \supseteq A_3 \supseteq A_4 \supseteq \dots$ and elements $a_1 < a_2 < \dots$ in A such that a_k is minimal in A_k .

Define now $f : \mathbb{N} \rightarrow A$ by $f(k) := a_k$. Because of $a_1 < a_2 < \dots$ the function f is injective. But is it also surjective? Take any $a \in A$ and assume there is **no** element a_k with $a_k = a$. Since $a \in A = A_1$, by the choice of a_1 and because of $a \neq a_1$ we conclude that $a > a_1$ as well as $a \in A_2$. Recall that $A_2 = A_1 \setminus \{a_1\}$. By the same reason $a > a_2$ and $a \in A_3$. Proceeding further we get for all $k \geq 1$ that $a \in A_k$ and therefore $a > a_k$. But $a_1 \geq 1$, hence $a_2 \geq 2$ and so on, we obtain $a > a_k \geq k$ for all $k \geq 1$. The theorem of Eudoxos (cf. Theorem 4.5.3) asserts that the set of natural numbers is not bounded above. That is, there is no real number x with $n \leq x$ for all $n \in \mathbb{N}$. Consequently, an $a \in \mathbb{N}$ with $a > a_k \geq k$ for all $k \geq 1$ cannot exist. Therefore, we cannot have $a \neq a_k$ for all $k \geq 1$. That is, any $a \in A$ is in the range of f . Thus, f is also surjective, hence also bijective as claimed. This completes the proof. ■

Remark A.4.3. Another way to look at the previous construction is as follows: We start with the smallest element a_1 of A . Then we choose the second smallest a_2 and so on. In this way we ensure that there are no elements in A which lie between a_k and a_{k+1} . Less obvious is the fact that any $a \in A$ is the k th smallest for a suitable $k \geq 1$. This follows from the observation that there are no $a \in A$ with $a > a_k$ for all $k \geq 1$. So, finally we get $A = \{a_1, a_2, \dots\}$ where a_k is the k th smallest element in A .

Corollary A.4.5. *A set A is countable if and only if there exists a surjective function from \mathbb{N} to A .*

Proof: Recall that Corollary A.3.8 asserts that there exists an injection from a set A to a set B if and only if there is a surjection from B to A . Hence, the corollary is an immediate consequence of Proposition A.4.4 and the fact that $|A| \leq |\mathbb{N}|$ if and only if there is an injection from A to \mathbb{N} . ■

The following permanence properties for countable sets are important.

Proposition A.4.6.

(1) Suppose there is a surjection from a set B to a set A . If B is countable, then so is A .

$$\exists g : B \rightarrow A \text{ surjective} \Rightarrow [B \text{ countable} \Rightarrow A \text{ countable}]$$

(2) The union of finitely many countable sets is countable too.

$$A_1, \dots, A_n \text{ countable} \Rightarrow A_1 \cup \dots \cup A_n \text{ countable}$$

(3) The Cartesian product of finitely many countable sets is countable as well.

$$A_1, \dots, A_n \text{ countable} \Rightarrow A_1 \times \dots \times A_n \text{ countable}$$

Proof: The first assertion easily follows by Corollary A.4.5. Indeed, since B is countable, there exists a surjection f from \mathbb{N} to B . By assumption there exists another surjection g from B to A . Since the composition of surjective functions is also surjective, the function $g \circ f : \mathbb{N} \rightarrow A$ is surjective as well. Another application of Corollary A.4.5 shows that A is countable.

We prove the second assertion only for two sets which we denote for simplicity by A and B . The general case easily follows from this by induction. So suppose we are given two countable sets A and B and we want to show that $A \cup B$ is countable too. By assumption there exist two surjections f and g from \mathbb{N} to A and B , respectively. Define a function $h : \mathbb{N} \rightarrow A \cup B$ as follows

$$h(n) := \begin{cases} f\left(\frac{n}{2}\right) & : n \text{ even} \\ g\left(\frac{n+1}{2}\right) & : n \text{ odd.} \end{cases}$$

Of course, h is surjective with values in $A \cup B$, hence the union is countable as well. This proves the second assertion.

The proof of the third property is a bit more involved. Again we restrict ourselves to two countable sets A and B and our goal is to show that $A \times B$ is countable as well.

Let us first treat a special case, namely that $A = B = \mathbb{N}$. That is, we are going to prove the following:

Proposition A.4.7. *The set $\mathbb{N} \times \mathbb{N}$ of pairs of integers is countable.*

Proof: (Cantor's first diagonal argument) Let us reorder $\mathbb{N} \times \mathbb{N}$ as follows:

$$(A.4.1) \quad \mathbb{N} \times \mathbb{N} = \bigcup_{k=1}^{\infty} B_k \quad \text{where} \quad B_k := \{(i, j) \in \mathbb{N} \times \mathbb{N} : i + j = k + 1\}.$$

The cardinality of the sets B_k equals k . Let $N_0 = 0$ and

$$N_k = 1 + 2 + \dots + k = \frac{k(k+1)}{2}, \quad k = 1, 2, \dots$$

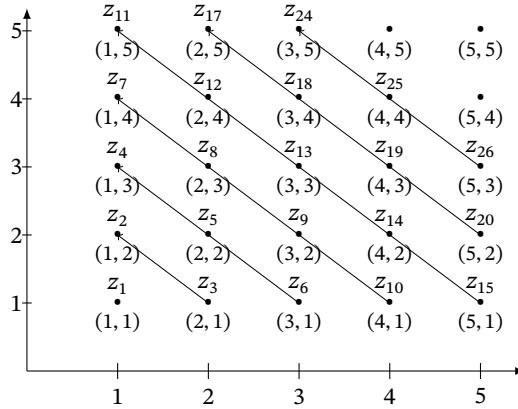
Then we may enumerate the elements in B_k as follows:

$$B_k = \{z_{N_{k-1}+1}, \dots, z_{N_k}\}, \quad k = 1, 2, \dots$$

For example, $N_1 = 1$, $N_2 = 3$, $N_3 = 6$, $N_4 = 10$ and $N_5 = 15$, hence

$$B_1 = \{z_1\}, \quad B_2 = \{z_2, z_3\}, \quad B_3 = \{z_4, \dots, z_6\}, \quad B_4 = \{z_7, \dots, z_{10}\} \text{ and} \\ B_5 = \{z_{11}, \dots, z_{15}\},$$

where, for example, we may choose $z_1 = (1, 1)$, $z_2 = (1, 2)$, $z_3 = (2, 1)$, $z_4 = (1, 3)$, and so on.



Define now $\varphi : \mathbb{N} \rightarrow \mathbb{N} \times \mathbb{N}$ by $\varphi(k) = z_k$ for all $k \geq 1$. The function φ is surjective because of (A.4.1). Moreover, since the B_k 's are pairwise disjoint, φ is also injective. That is, $\varphi : \mathbb{N} \rightarrow \mathbb{N} \times \mathbb{N}$ is a bijection. So, by definition $\mathbb{N} \times \mathbb{N}$ is countably infinite. ■

Now let us come back to the general case. Since A and B are assumed to be countable, there are two surjections $f : \mathbb{N} \rightarrow A$ and $g : \mathbb{N} \rightarrow B$. Let $h : \mathbb{N} \times \mathbb{N} \rightarrow A \times B$ be defined by

$$h(k, \ell) := (f(k), g(\ell)), \quad (k, \ell) \in \mathbb{N} \times \mathbb{N}.$$

We claim that h is surjective. To verify this, we take an arbitrary pair $(x, y) \in A \times B$. Since f and g are surjective, there are $k \in \mathbb{N}$ and $\ell \in \mathbb{N}$ with $f(k) = x$ and $g(\ell) = y$. Of course, this implies

$$h(k, \ell) = (f(k), g(\ell)) = (x, y).$$

So, h is a surjection from $\mathbb{N} \times \mathbb{N}$ to $A \times B$. But as we proved in Proposition A.4.7, the set $\mathbb{N} \times \mathbb{N}$ is countable. Now the first property in the current proposition applies and yields that $A \times B$ is countable as well. This completes the proof. ■

Remark A.4.4. The second property of Proposition A.4.6 may be generalized as follows:

Let A_1, A_2, \dots be countable sets. Then their union is countable too.

$$(A_1, A_2, \dots \text{ countable}) \Rightarrow (A_1 \cup A_2 \cup \dots \text{ countable}).$$

The idea of the proof is as follows: Enumerate the elements in each of the A_k 's. Then enumerate the elements in the union in the following way: Start with the first element in A_1 , then take the second element in A_1 , after that the first element in A_2 . In the next step choose the third element in A_1 , then the second one in A_2 , then the first one in A_3 . After that go back to A_1 till the first element of A_4 and so on.

From Proposition A.4.6 we obtain the following important result.

Theorem A.4.8. *The following sets are all countably infinite:*

- (1) *The set \mathbb{Z} of integers.*
- (2) *The set \mathbb{Q} of rational numbers.*
- (3) *The set \mathbb{Z}^n of vectors of length n with entries in \mathbb{Z} .*
- (4) *The set \mathbb{Q}^n of vectors of length n with entries in \mathbb{Q} .*
- (5) *The set \mathcal{A} of algebraic numbers. Recall that these are solutions of equations*

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

for certain $a_j \in \mathbb{Z}$ and some $n \geq 1$.

Proof:

1. The set \mathbb{Z} is union of the three countable sets \mathbb{N} , $-\mathbb{N}$ and $\{0\}$. Consequently, because of Proposition A.4.6 the set \mathbb{Z} is countable as well. Another more direct approach is to enumerate the elements in \mathbb{Z} in the following way:

$$\mathbb{Z} = \{0, 1, -1, 2, -2, 3, \dots\}.$$

2. In a first step we show that

$$\mathbb{Q}_+ = \{q \in \mathbb{Q} : q > 0\}$$

is countable. To this end we define a function f from $\mathbb{N} \times \mathbb{N}$ to \mathbb{Q}_+ by

$$f(m, n) := \frac{m}{n}.$$

By the definition of rational numbers, f is surjective. Moreover, as we saw above, the set $\mathbb{N} \times \mathbb{N}$ is countable, hence \mathbb{Q}_+ is so as well. Of course, this implies that also

$$\mathbb{Q}_- := \{q \in \mathbb{Q} : q < 0\}$$

is countable. Hence, in view of Proposition A.4.6 the assertion follows by

$$\mathbb{Q} = \mathbb{Q}_+ \cup \mathbb{Q}_- \cup \{0\}.$$

Properties (3) and (4) are a direct consequence of properties (1) and (2) combined with Proposition A.4.6. Recall that

$$\mathbb{Z}^n = \underbrace{\mathbb{Z} \times \cdots \times \mathbb{Z}}_{n \text{ times}} \quad \text{and} \quad \mathbb{Q}^n = \underbrace{\mathbb{Q} \times \cdots \times \mathbb{Q}}_{n \text{ times}}.$$

5. Fix a number $n \geq 0$ as well as $n + 1$ integers a_0, a_1, \dots, a_n . By the fundamental theorem of algebra there are at most n real numbers $x \in \mathbb{R}$ satisfying

$$a_n x^n + a_{n-1} x^{n-1} + \cdots + a_1 x + a_0 = 0.$$

Thus, if

$$A(a_0, \dots, a_n) := \{x \in \mathbb{R} : a_n x^n + a_{n-1} x^{n-1} + \cdots + a_1 x + a_0 = 0\},$$

then it follows that

$$|A(a_0, \dots, a_n)| \leq n.$$

Setting

$$A_n := \bigcup_{(a_0, \dots, a_n) \in \mathbb{Z}^{n+1}} A(a_0, \dots, a_n),$$

the set A_n consists of those real numbers which are zeroes of a polynomial of degree less than or equal to n and with integer coefficients. Recall that \mathbb{Z}^{n+1} is countable, hence A_n is the countable union of finite sets. And as we mentioned in Remark A.4.4, this union is again countable. Summing up, for each $n \geq 0$, the sets A_n are countable. Since

$$\mathcal{A} = \bigcup_{n=1}^{\infty} A_n,$$

another application of Remark A.4.4 implies that the set \mathcal{A} of algebraic numbers is countable. This completes the proof. \blacksquare

We turn now to the question to determine the size of \mathbb{R} . In a first step we show that \mathbb{R} and the open interval $(0, 1)$ are of the same size.

Proposition A.4.9. *The set \mathbb{R} of real numbers and the open interval satisfy*

$$|\mathbb{R}| = |(0, 1)|.$$

Proof: There are many ways to verify this. We present here the direct approach by constructing a bijection φ from $(0, 1)$ to \mathbb{R} . To this end consider the tangent function $x \mapsto \tan x$ which is a bijection from $(-\frac{\pi}{2}, \frac{\pi}{2})$ to \mathbb{R} . Thus, the function

$$\varphi(x) := \tan\left(\pi x - \frac{\pi}{2}\right), \quad 0 < x < 1,$$

is a bijection from $(0, 1)$ onto \mathbb{R} . By the definition of the cardinality of sets this completes the proof. \blacksquare

Corollary A.4.10. *For any two numbers $a < b$ it follows that*

$$|(a, b)| = |\mathbb{R}|.$$

Proof: This easily follows by the fact that the function ψ with

$$\psi(x) := \frac{x-a}{b-a}, \quad a < x < b,$$

is a bijection from (a, b) onto $(0, 1)$. By Proposition A.4.9 this leads immediately to $|(a, b)| = |(0, 1)| = |\mathbb{R}|$. \blacksquare

For all those who are not convinced that there are as many numbers in a finite interval as there are in \mathbb{R} , let us construct another, maybe more concrete, bijection between a finite interval and \mathbb{R} . This bijection is the so-called stereographic projection defined as follows: Identify the interval $(-\pi, \pi)$ with a circle \mathcal{S} of radius 1 where we cut out the North Pole N corresponding to π or $-\pi$, respectively. Then we draw a straight line \overline{NP} from the North Pole N to a given point $P \neq N$ on the circle. This line cuts the real line at exactly one point $P' \in \mathbb{R}$ (see Figure A.4.3). The mapping $P \mapsto P'$ is then the desired bijection from \mathcal{S} to \mathbb{R} , respectively from the finite interval $(-\pi, \pi)$ to \mathbb{R} .

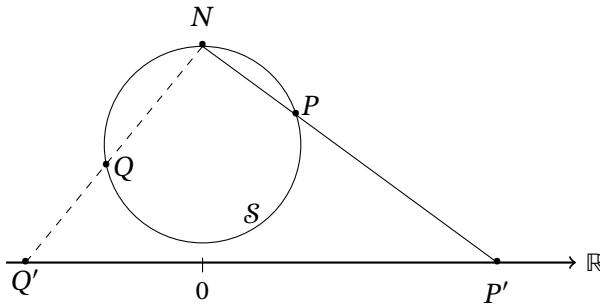


Figure A.4.3. The stereographic projection as bijection from $(-\pi, \pi)$ to \mathbb{R} .

One can generalize this construction to the stereographic projection from a sphere with the North Pole removed into the plane $\mathbb{R}^2 = \mathbb{R} \times \mathbb{R}$. The construction of this projection is done almost in the same way as we did above for the circle. But now any straight line from N to a point $P \neq N$ on the sphere cuts the plane at a unique point $P' \in \mathbb{R}^2$.

We are going to prove now an important result about the size of \mathbb{R} . Since $\mathbb{N} \subseteq \mathbb{R}$ it follows $|\mathbb{N}| \leq |\mathbb{R}|$. The next theorem shows that the inequality is strict.

Theorem A.4.11. *The set \mathbb{R} of real numbers is **not** countable. In other words,*

$$|\mathbb{N}| < |\mathbb{R}|.$$

Proof: In a first step we observe that it suffices to prove that the open interval $(0, 1)$ is not countable. Indeed, if \mathbb{R} would be countable, then by Proposition A.4.9 this would be the same for $(0, 1)$.

So, let us assume the contrary, that is, the elements in the interval $(0, 1)$ could be enumerated as follows:

$$(A.4.2) \quad (0, 1) = \{x^1, x^2, x^3, \dots\}.$$

As we have shown in Theorem 4.7.3, each real number x^k in $(0, 1)$ admits a fractional expansion with respect to the base $b = 3$. Thus, there are numbers $x_j^k \in \{0, 1, 2\}$ such that

$$x^k =_3 0.x_1^k x_2^k \dots, \quad k = 1, 2, \dots$$

If the expansion of x^k is finite, i.e., if x^k is 3-rational, we fill it up with zeroes in order to obtain an infinite sequence $(x_j^k)_{j \geq 1}$.

Next we apply an argument which is nowadays called **Cantor's (second) diagonal argument**. We change the expansion of each x^k at position k in a clever way. If the k th digit of the expansion of x^k equals 0, then we change it to 1 while if $x_k^k = 1$ we replace it by 0. Finally, if the k th digit equals 2, the changed digit equals 1.

That is, for each $k \geq 1$ we define a number $y_k \in \{0, 1\}$ as follows:

$$y_k = \begin{cases} 1 & : \text{ if } x_k^k = 0 \\ 0 & : \text{ if } x_k^k = 1 \\ 1 & : \text{ if } x_k^k = 2. \end{cases}$$

Let us illustrate the construction of y in a table.

| | | | | | | | |
|-------|---|---|---|---|---|---|-----|
| x^1 | 2 | 2 | 0 | 0 | 2 | 1 | ... |
| x^2 | 2 | 1 | 2 | 0 | 2 | 0 | ... |
| x^3 | 0 | 0 | 0 | 2 | 1 | 0 | ... |
| x^4 | 2 | 1 | 1 | 2 | 0 | 2 | ... |
| x^5 | 0 | 2 | 0 | 0 | 1 | 2 | ... |
| x^6 | 2 | 0 | 1 | 0 | 2 | 0 | ... |
| . | . | . | . | . | . | . | ... |
| y | 1 | 0 | 1 | 1 | 0 | 1 | ... |

The choice of the y_k 's implies that for all $k \geq 1$ we have $y_k \neq x_k^k$. Moreover, and this was the reason we have chosen $b = 3$, not $b = 2$, the sequence $(y_k)_{k \geq 1}$ is admissible, that is, there is no integer k_0 such that $y_k = 2$ for $k \geq k_0$. Theorem 4.7.3 implies that there exists $y \in (0, 1)$ with

$$y =_3 0.y_1y_2y_3\cdots.$$

So we got a number $y \in (0, 1)$ which cannot coincide with any of the x^k 's. Why? Because if there were some $k \geq 1$ with $y = x^k$, then by the uniqueness of the expansion the k th coordinate of x^k should coincide with y_k . But this is not so by the choice of y . Consequently, we found a $y \in (0, 1)$ with

$$y \notin \{x^1, x^2, x^3, \dots\}.$$

This contradicts (A.4.2), hence the interval $(0, 1)$ is not countable and \mathbb{R} is not so too. ■

The previous theorem has the following consequences.

Theorem A.4.12. *The following sets are all uncountable, i.e., neither finite nor countable infinite.*

- (1) *Each nonempty interval in \mathbb{R} , open or closed, finite, or infinite.*
- (2) *The set of irrational numbers in \mathbb{R} .*
- (3) *The set of transcendental numbers. Recall that these are the nonalgebraic numbers.*
- (4) *The set*

$$\{0, 1\}^{\mathbb{N}} := \{(x_1, x_2, \dots) : x_j = 0 \text{ or } x_j = 1\}$$

of infinite sequences with entries 0 or 1.

- (5) *The power sets $\mathcal{P}(\mathbb{N})$ and $\mathcal{P}(\mathbb{Z})$.*

Proof:

1. As we saw above each finite open interval has the same cardinality as \mathbb{R} . This tells us that any open interval cannot be countable. Because open and closed intervals differ only by at most two points, this is so also for closed intervals.

2. Let $\mathbb{I} := \mathbb{R} \setminus \mathbb{Q}$ denote the set of irrational numbers. If \mathbb{I} were countable, then so would

$$\mathbb{R} = \mathbb{I} \cup \mathbb{Q}.$$

But, as we saw \mathbb{R} is uncountable, hence \mathbb{I} has to be so as well.

3. This follows by the same arguments because we know that the set \mathcal{A} of algebraic numbers is countable.

4. Assume that $\{0, 1\}^{\mathbb{N}}$ would be countable. Then we define a function φ from $\{0, 1\}^{\mathbb{N}}$ to $(0, 1)$ by

$$\varphi : (x_1, x_2, \dots) \mapsto x =_2 0.x_1x_2 \dots .$$

In view of Theorem 4.7.3, the function φ is surjective. Hence, if $\{0, 1\}^{\mathbb{N}}$ would be countable, then by Proposition A.4.6 the interval $(0, 1)$ would be countable too. But this contradicts the fact that $(0, 1)$ is uncountable.

5. Given a subset $A \subseteq \mathbb{N}$ we assign to it an infinite sequence x_A with entries 0 and 1, that is an element $x_A \in \{0, 1\}^{\mathbb{N}}$, as follows.

$$A \Leftrightarrow x_A = (x_1, x_2, \dots) \text{ where } x_k = 1 \text{ if } k \in A \text{ and } x_k = 0 \text{ if } k \notin A.$$

So for example, the set of even integers is mapped to the sequence $(0, 1, 0, 1, 0, 1, \dots)$.

The function φ is a bijection, hence

$$(A.4.3) \quad |\{0, 1\}^{\mathbb{N}}| = |\mathcal{P}(\mathbb{N})|,$$

which shows that $\mathcal{P}(\mathbb{N})$ cannot be countable. The proof for $\mathcal{P}(\mathbb{Z})$ follows by the same arguments, so we omit it. \blacksquare

Our final goal is to prove that the Cantor set \mathcal{C} is uncountable. This is a quite surprising result. Why? Recall that we presented in Example 4.7.3 the Cantor set as an infinite intersection of certain sets C_n , where each of the sets C_n is the union of 2^n intervals of length $1/3^n$. Consequently, for each $n \geq 1$ the overall length of \mathcal{C} is less than $(2/3)^n$. As n becomes big, the sequence $(2/3)^n$ tends to zero, saying that \mathcal{C} is contained in sets which are arbitrarily small. This suggests that \mathcal{C} is a tiny set. But indeed, in contrast to that, it turns out that it is even uncountable, i.e., much bigger than \mathbb{N} .

Let us state and prove now the announced result. We shall give two different proofs. One is based on a Cantor-type diagonal technique while the second proof is more direct by constructing a surjection from \mathcal{C} to $[0, 1]$.

Proposition A.4.13. *The Cantor set \mathcal{C} is uncountable.*

Proof: During the proof we use the notation introduced in Example 4.7.3.

Let us assume the contrary, i.e., that there are numbers c_1, c_2, \dots in $[0, 1]$ such that

$$(A.4.4) \quad \mathcal{C} = \{c_1, c_2, \dots\}.$$

In a first step we use that

$$\mathcal{C} \subseteq [0, \frac{1}{3}] \cup [\frac{2}{3}, 1].$$

Consequently, c_1 , the first number of the enumeration of \mathcal{C} , belongs either to $[0, \frac{1}{3}]$ or to $[\frac{2}{3}, 1]$. Note that these two sets are disjoint and that \mathcal{C} is a subset of their union.

Let I_1 be the interval which does **not** contain c_1 . That is, either $I_1 = [0, \frac{1}{3}]$ or $I_1 = [\frac{2}{3}, 1]$ in dependence on the value of c_1 .

In the next step, we split I_1 into three intervals of equal size, take off the middle one, and obtain two disjoint closed intervals of length $1/9$. Now, c_2 , the second element in the enumeration, does belong to at most one of these intervals. Of course, it could also happen, that c_2 is neither in the left nor in the right part of I_1 . Anyway, there exist one interval I_2 contained in I_1 of length $1/9$ which does **not** contain c_2 .

Proceeding further in this way we get intervals I_1, I_2, \dots with

$$I_1 \supseteq I_2 \supseteq \dots \quad \text{such that} \quad |I_n| = \frac{1}{3^n} \quad \text{and with} \quad c_n \notin I_n.$$

Moreover, as observed in Example 4.7.3, the endpoints $a_n < b_n$ of each $I_n = [a_n, b_n]$ are always elements in \mathcal{C} .

An application of the nested interval theorem, Proposition 5.4.7, implies the existence of an element $x \in [0, 1]$ such that

$$\bigcap_{n=1}^{\infty} I_n = \{x\}.$$

We claim now¹⁰ that $x \in \mathcal{C}$. So let us assume on the contrary that $x \notin \mathcal{C}$. By the construction of \mathcal{C} this implies $x \in \bigcup_{n=1}^{\infty} D_n$. Recall that the sets D_n were those which we cut out at step n and that $\mathcal{C} = [0, 1] \setminus \left(\bigcup_{n=1}^{\infty} D_n \right)$. Consequently, there exists $m \geq 1$ such that $x \in D_m$. But D_m is the disjoint union of 2^m open intervals, hence x has to be in (exactly) one of them. As said, these intervals are open, therefore there exists some $\varepsilon > 0$ so that a whole ε -neighborhood $(x - \varepsilon, x + \varepsilon)$ lies in D_m . In other words, there is an ε -neighborhood of x without any element of \mathcal{C} . But $x \in I_n = [a_n, b_n]$ for all $n \geq 1$, hence

$$|x - a_n| \leq |a_n - b_n| \leq \frac{1}{3^n}.$$

Thus, if $n \geq 1$ satisfies $1/3^n < \varepsilon$ it follows that $|x - a_n| < \varepsilon$ or, equivalently, that $a_n \in (x - \varepsilon, x + \varepsilon)$, which by $a_n \in \mathcal{C}$ contradicts $(x - \varepsilon, x + \varepsilon) \subseteq D_m$. This contradiction proves $x \in \mathcal{C}$ as asserted.

So we see that there is $x \in \mathcal{C}$ with $x \in I_n$ for all $n \geq 1$. Since by the choice of the I_n we have $c_n \notin I_n$ for all $n \geq 1$, this implies $x \neq c_n$ for all $n \geq 1$. So we found an element $x \in \mathcal{C}$ in which does not belong to $\{c_1, c_2, \dots\}$. This contradicts (A.4.4), hence \mathcal{C} cannot be countable. This completes the proof. ■

Let us give another more direct proof of Proposition A.4.13.

Proof: We claim that there exists a surjection ϕ from \mathcal{C} to $[0, 1]$. If such a surjection exists, then \mathcal{C} cannot be countably infinite because otherwise by Proposition A.4.6 this would imply that $[0, 1]$ is countable too, contradicting Theorem A.4.12.

So let us construct the surjection $\phi : \mathcal{C} \rightarrow [0, 1]$. If $x \in \mathcal{C}$ possesses the infinite ternary expansion

$$x =_3 x_1 x_2 \dots \quad \text{with} \quad x_j \in \{0, 2\},$$

then $\phi(x)$ is defined as follows:

$$(A.4.5) \quad \phi(x) = y \Leftrightarrow y =_2 0.y_1 y_2 \dots \quad \text{with} \quad y_j = \frac{x_j}{2}.$$

¹⁰This easily follows from the fact that \mathcal{C} is a closed set. But since we did not prove this (although this easily follows by Proposition 5.8.1), some extra considerations are necessary.

Note that all x_j s are either 0 or 2, hence the y_j s are all either 0 or 1. So for example, because of $2/3 =_3 0.2 =_3 0.200\dots$ it follows $\phi(2/3) =_2 0.1 = 1/2$. Or $x =_3 0.\overline{2022}$ is mapped to $y =_2 0.1011$.

Now, if $x \in \mathcal{C}$ has a finite ternary expansion

$$x =_3 0.x_1 \dots x_{n-1} 1 \quad \text{with } x_j \in \{0, 2\},$$

then we let

$$(A.4.6) \quad \phi(x) = y \quad \text{where} \quad y =_2 0.y_1 \dots y_{n-1} 1 \quad \text{with} \quad y_j = \frac{x_j}{2}.$$

So, for example $x =_3 0.1 =_3 1/3$ is mapped to $y =_2 0.1 =_2 1/2$. Or $x =_3 0.221 =_3 \frac{25}{27}$ is mapped to $y =_2 0.111 =_2 \frac{7}{8}$.

Finally, we let $\phi(1) = 1$. Hence, $\phi(x)$ is well-defined for all $x \in \mathcal{C}$. Moreover, an application of Theorem 4.7.4 proves that ϕ is in fact surjective. For example, if

$$y =_2 0.y_1 \dots y_{n-1} \underbrace{1}_{n} 00 \dots,$$

with certain $y_j \in \{0, 1\}$, then $\phi(x) = y$ where

$$x =_3 0.x_1 \dots x_{n-1} \underbrace{1}_{n} 00 \dots \quad \text{with} \quad x_j = 2y_j.$$

The case of $y \in [0, 1)$ with infinite binary expansion is even easier to handle and therefore left as an exercise. ■

Remark A.4.5. It is worthwhile to mention that the surjection ϕ is **not** injective. For example, if $x =_3 0.21 =_3 7/9$ and $x' =_3 0.22 =_3 8/9$, then

$$\phi(x) =_2 0.11 =_2 \frac{3}{4} \quad \text{as well as} \quad \phi(x') =_2 0.11 =_2 \frac{3}{4}.$$

Remark A.4.6. The second proof gives even a slightly stronger result than Proposition A.4.13. It shows that $|\mathbb{R}| \leq |\mathcal{C}|$. On the other hand, since $\mathcal{C} \subseteq \mathbb{R}$ it follows that $|\mathcal{C}| \leq |\mathbb{R}|$. Theorem A.4.14 then yields the remarkable fact that even

$$|\mathcal{C}| = |\mathbb{R}|.$$

This tells us that there are as many elements in \mathcal{C} as there are in \mathbb{R} .

Epilogue:

(1) **Is the comparison of the size of sets anti-symmetric?**

Recall that by definition the size of A is less than or equal that of a set B if

$$|A| \leq |B| \Leftrightarrow \exists \text{ injective } f : A \rightarrow B.$$

This relation is of course reflexive. Moreover, since the composition of injective mappings is injective as well, it is also transitive. That is,

$$|A| \leq |B| \quad \text{and} \quad |B| \leq |C| \Rightarrow |A| \leq |C|.$$

But is this relation also anti-symmetric? Equivalently, does $|A| \leq |B|$ and $|B| \leq |A|$ imply $|A| = |B|$? According to the definition this problem can be formulated as follows:

Problem: Let A and B be sets such that there are an injection from A to B and another injection from B to A . Does this imply the existence of a bijection from A to B or, equivalently, of a bijection from B to A ?

This is answered in the affirmative by the following important theorem¹¹. We refer to [15] for a proof and for more details.

Theorem A.4.14 (Cantor–Bernstein–Schröder Theorem). *If two sets A and B satisfy $|A| \leq |B|$ as well as $|B| \leq |A|$, then it follows that $|A| = |B|$.*

(2) Does there exist a largest set?

A (natural) question is about the existence of a largest set B_{\max} , that is a set which fulfills

$$(A.4.7) \quad |A| \leq |B_{\max}| \quad \text{for all sets } A.$$

The answer to this question is in the negative. This follows from the next result:

Theorem A.4.15 (Cantor's Theorem, 1891). *For any set A ,*

$$|A| < |\mathcal{P}(A)|.$$

Proof: In a first step we show $|A| \leq |\mathcal{P}(A)|$. So we define $f : A \rightarrow \mathcal{P}(A)$ by

$$f(a) := \{a\}, \quad a \in A.$$

That is, the function f maps an element $a \in A$ to the singleton containing the element $a \in A$. Of course, f is injective from A to $\mathcal{P}(A)$, hence $|A| \leq |\mathcal{P}(A)|$.

Next we show that A and $\mathcal{P}(A)$ do not possess the same cardinality. To this end we will prove that there is no surjective function from A to $\mathcal{P}(A)$. Of course, if we are able to verify this, then $|A| < |\mathcal{P}(A)|$.

Thus take any function f from A to $\mathcal{P}(A)$. That is, for any $a \in A$ the image $f(a)$ is a subset of A . Then for each $a \in A$ two cases are possible:

$$a \in f(a) \quad \text{or} \quad a \notin f(a).$$

We introduce now a subset $B \subseteq A$ as follows:

$$B := \{a \in A : a \notin f(a)\}.$$

Note that $B \in \mathcal{P}(A)$. Our aim is to show that B is not in the range of f . If we are able to verify this, then f is not surjective, and we are done. Assume the contrary, i.e., suppose there is a $b \in A$ such that $f(b) = B$.

We ask now whether $b \in B$ or $b \notin B$. If $b \in B$, then by the definition of B it follows that $b \notin f(b)$, but $f(b) = B$, hence $b \in B$ implies $b \notin B$. On the other side, if $b \notin B$, then $b \notin f(B)$, hence, by the definition of B it follows that $b \in B$. Summing up, $b \in B$ implies $b \notin B$ and vice versa. This cannot happen. The conclusion is, our assumption

¹¹The result was formulated by Georg Cantor in 1887. In 1897 Felix Bernstein found a correct proof. Before that, in 1896, Ernst Schröder had given a proof of the theorem, but it was observed that his proof was erroneous. After the publication of Bernstein's proof it turned out that already before Bernstein the mathematician Richard Dedekind had found a correct proof in 1887. Nevertheless, surprisingly his name does not show up in the denomination of the result.

about the existence of $b \in A$ with $f(b) = B$ was wrong. Consequently, there exists at least one element in $\mathcal{P}(A)$, namely B , which is not in the range of f , hence f cannot be surjective. So we see that there are no surjective functions from A to $\mathcal{P}(A)$, hence A and $\mathcal{P}(A)$ never possess the same cardinality, which completes the proof. ■

Remark A.4.7. For better understanding let us explain the construction of the set B at an easy example. Suppose $A = \{0, 1\}$. Then $\mathcal{P}(A)$ contains the four elements

$$\emptyset, \quad \{0\}, \quad \{1\} \quad \text{and} \quad A = \{0, 1\}.$$

Say, for example, the function f from A to $\mathcal{P}(A)$ is defined by

$$f(0) = A \quad \text{and} \quad f(1) = \emptyset.$$

Then

$$B := \{a \in A : a \notin f(a)\} = \{1\}.$$

And, indeed, $B = \{1\}$ is not in the range of f . If we change f by letting $f(1) = \{1\}$, then $B = \emptyset$, and again B is not in the range of the modified function f .

Example A.4.2. If we consider

$$\mathcal{S} := \{S : S \subseteq \mathcal{P}(\mathbb{N})\},$$

then Theorem A.4.15 implies that

$$|\mathbb{R}| = |\mathcal{P}(\mathbb{N})| < |\mathcal{S}|.$$

Note that \mathcal{S} consists of all (finite or infinite) collections of subsets of \mathbb{N} . For example, the set $\{\{1\}, \{1, 2\}, \{1, 2, 3\}, \dots\}$ is an element of \mathcal{S} . Another element is $\{\mathbb{N}, 2\mathbb{N}, 3\mathbb{N}, \dots\}$, but as we saw there are many more.

Remark A.4.8. Theorem A.4.15 is valid for all sets, infinite and finite ones. In the latter case we even know more. Recall that we proved in Proposition 1.7.4 that in the case of finite sets

$$|\mathcal{P}(A)| = 2^{|A|}.$$

Because of $n < 2^n$ for all $n \geq 0$ we rediscover Theorem A.4.15 for those sets.

Remark A.4.9. By (A.4.3) we know that

$$|\mathcal{P}(\mathbb{N})| = |\{0, 1\}^{\mathbb{N}}|.$$

Thus, Theorem A.4.15 implies

$$|\mathbb{N}| < |\mathcal{P}(\mathbb{N})| = |\{0, 1\}^{\mathbb{N}}|.$$

This is an alternative proof of the fact that the set $\{0, 1\}^{\mathbb{N}}$ of infinite $\{0, 1\}$ -sequences is not countable.

Corollary A.4.16. *There is no set B_{\max} with $|A| \leq |B_{\max}|$ for all sets A .*

Proof: If such a set B_{\max} were to exist, then estimate (A.4.7) should also be valid for $A = \mathcal{P}(B_{\max})$. But this contradicts Theorem A.4.15. ■

(3) Are the sizes of two sets always comparable?

Suppose we are given two arbitrary sets A and B . Is it always possible to say that either A is bigger than B or that vice versa B is bigger than A ? In other words, does for any pair A and B of sets always follow either

$$|A| < |B|, \quad |B| < |A| \quad \text{or} \quad |A| = |B| ?$$

The question is answered by the following theorem.

Theorem A.4.17 (Zermelo's¹² Comparability Theorem, 1905). *Assuming the axiom of choice (cf. Axiom A.3.1) the size of two arbitrary sets is always comparable. That is, for all sets A and B one has either*

$$|A| < |B| \quad \text{or} \quad |B| < |A| \quad \text{or} \quad |A| = |B| .$$

Remark A.4.10. Theorem A.4.17 is the consequence of a more general result due to Ernst Zermelo, the so-called well-ordering theorem. It asserts that each set may be well-ordered. This result is known to be equivalent to the axiom of choice.

(4) Are there uncountable sets $A \subset \mathbb{R}$ with $|A| < |\mathbb{R}|$?

In 1878 Georg Cantor asked whether there exist sets with a cardinality which is strictly between that of \mathbb{N} and \mathbb{R} . In other words, do there exist **uncountable** subsets $A \subseteq \mathbb{R}$ with $|A| < |\mathbb{R}|$? Later on, in 1900 David Hilbert presented this question in his famous talk about the most important 23 open problems in Mathematics. There are two alternatives for the answer to this question:

$$\text{CH : } (\forall \mathbb{N} \subseteq A \subseteq \mathbb{R}) (|A| = |\mathbb{N}| \text{ or } |A| = |\mathbb{R}|)$$

$$\neg\text{CH : } (\exists \mathbb{N} \subseteq A \subseteq \mathbb{R}) (|\mathbb{N}| < |A| < |\mathbb{R}|)$$

Here CH means that the **Continuum Hypothesis** is satisfied while property $\neg\text{CH}$ says the contrary, namely that there exists at least one set with size strictly between that of \mathbb{N} and \mathbb{R} .

In 1940, Kurt Gödel proved that one may add property CH to the other axioms of Set Theory without getting a contradiction. Around 1960, Paul J. Cohen proved the same results in the case of adding $\neg\text{CH}$ to the axioms of Set Theory. Consequently, there exist two types of basic Mathematics: One where the continuum hypotheses is valid and another one where this is not so. This may not be very satisfying, yet shows that on the base of the current axioms of Set Theory we cannot decide whether the Continuum Hypothesis or its contrary are satisfied. For further information about this interesting topic we refer to [4].

Exercise A.4.1. If A and B are sets with $A \subseteq B$. Argue why this implies $|A| \leq |B|$.

Exercise A.4.2. Let $S \subseteq \mathbb{R}$ be defined by

$$S := \{2^n + 3^m : n, m \in \mathbb{N}\}.$$

Why is S countably infinite?

¹²Ernst Zermelo (1871–1953) was a German mathematician.

Exercise A.4.3. Construct a bijection between \mathbb{N} and $\mathbb{N}^3 = \{(k, \ell, m) : k, \ell, m \in \mathbb{N}\}$.

Exercise A.4.4. Show that the collection of all **finite** subsets of \mathbb{N} is countable. Recall that the collection $\mathcal{P}(\mathbb{N})$ of **all** subsets of \mathbb{N} is uncountable.

Exercise A.4.5. Let $a < b$ and $c < d$ be four real numbers. Construct bijective functions between the following pairs of intervals:

- (1) $[a, b]$ and $[c, d]$,
- (2) $[a, b)$ and $[c, d)$,
- (3) $[a, b]$ and $[a, b)$,
- (4) $[a, b)$ and (a, b) .

Exercise A.4.6. Complete the proof of the result mentioned in Remark A.4.4. Recall that we claimed that for given countable sets A_1, A_2, \dots also their infinite union $A_1 \cup A_2 \cup \dots$ is countable.

Exercise A.4.7. Given a base $b \geq 2$, show that the set of b -rational numbers in $[0, 1]$ is countable.

Exercise A.4.8. Let \mathbb{Z}_0^∞ be the set of all infinite sequences of integers which are constant from a certain point. That is, a sequence x_1, x_2, \dots belongs to \mathbb{Z}_0^∞ if all $x_j \in \mathbb{Z}$ and, moreover, there exists a $k \geq 1$ (depending on the sequence) such that $x_k = x_{k+1} = \dots$. Argue why \mathbb{Z}_0^∞ is countable infinite. How about \mathbb{Z}^∞ , the set of all infinite sequences of integers?

Exercise A.4.9. Let S be the set of all infinite sequences x_1, x_2, \dots with $x_j = 0$ or $x_j = 1$ such that for any $k \in \mathbb{N}$ there is some $j > k$ such that $x_j = 0$. In other words, we exclude sequences for which there is an integer k such that $x_j = 1$ for all $j > k$. Construct a bijection between S and $[0, 1)$.

Exercise A.4.10. Show that the set of all infinite points in the Cantor set \mathcal{C} is uncountable. Recall that an element $x \in \mathcal{C}$ is an infinite point if its ternary expansion is infinite or, equivalently, if it is not an endpoint in \mathcal{C} . Why do in contrast to that exist only countably many endpoints in \mathcal{C} ?

Exercise A.4.11. Let $\phi : \mathcal{C} \rightarrow [0, 1]$ be defined by (A.4.5) and (A.4.6), respectively.

- (1) Complete the proof of the fact that ϕ is a surjection.
- (2) Show that there are at most two elements $x \neq y$ in \mathcal{C} with $\phi(x) = \phi(y)$. Characterize those pairs (x, y) .

A.5. Relations

We talk quite often about relations between certain objects or animate beings. For instance, we say that a person x is related to another person y if x is a cousin of y . Or x is related to y if x and y attended the same high school. In mathematics relations occur in many circumstances. In the previous sections we investigated already some

quite important ones. For example, we said that

- an integer n is in relation with another integer m if $n \leq m$ or that
- an integer $a \neq 0$ is in relation with $b \in \mathbb{Z}$ if a divides b or that
- a set A is in relation with a set B provided that $A \subseteq B$ or that
- two integers a and b are in relation if $a \equiv b \pmod{n}$ or that
- two finite sets A and B are in relation whenever $|A| = |B|$.

Already these few examples show that relations between different objects play an important role in mathematics. But what is a relation between two objects? For example, how can we express that a natural number n is in relation with some positive integer m provided that $n \mid m$? This may be described by the subset $R \subseteq \mathbb{N} \times \mathbb{N}$ with

$$R := \{(n, m) : n, m \in \mathbb{N} \text{ and } n \mid m\}.$$

The pair $(2, 6)$ belongs to R while the pair $(5, 7)$ does not.

In other words, a relation is nothing other than a set of pairs where the entries are those pairs which are related by some common property (in our case that property is $n \mid m$). The precise definition of a (mathematical) relation is as follows.

Definition A.5.1. Let A and B be two nonempty sets. A subset $R \subseteq A \times B$ is said to establish a (binary¹³) **relation** between certain elements of the set A and of the set B . An element $x \in A$ is in relation with a certain $y \in B$ whenever $(x, y) \in R$. Otherwise, i.e., if $(x, y) \notin R$, then x is not related to y . In the case $A = B$, the subset $R \subseteq A \times A$ is called a relation on A .

$$\begin{aligned}(R \text{ is a relation between elements of } A \text{ and of } B) &\Leftrightarrow [R \subseteq A \times B] \\(x \in A \text{ is in relation } R \text{ with } y \in B) &\Leftrightarrow [(x, y) \in R].\end{aligned}$$

Definition A.5.2. Given a relation $R \subseteq A \times B$, the set $\text{dom}(R)$ defined by

$$\text{dom}(R) = \{x \in A : \exists y \in B, (x, y) \in R\}$$

is called the **domain** of the relation R while $\text{range}(R) \subseteq B$ defined by

$$\text{range}(R) = \{y \in B : \exists x \in A, (x, y) \in R\}$$

is said to be the range of R .

$$\begin{aligned}\text{dom}(R) &= \{x \in A : \exists y \in B, (x, y) \in R\} \\ \text{range}(R) &= \{y \in B : \exists x \in A, (x, y) \in R\}.\end{aligned}$$

If A and B are finite or countably infinite, relations can be easily interpreted by a table. Writing the elements of A at the left margin of the table and that of B on the top, a mark in the box (x, y) says that x is in relation with y while an empty box signalizes that these two elements are not in relation.

¹³There also exist relations between more than two objects. For example, let three real numbers x, y , and z be related if $x \cdot y = z$. But here and later on we will restrict ourselves to relations between two objects. That is, all relations we treat later on will always be binary.

Example A.5.1. Say $A = \{a, b, c, d\}$ while $B = \{e, f, g, h\}$. Then a relation R can, for instance, be described by the table

| | e | f | g | h |
|---|---|---|---|---|
| a | • | | • | |
| b | • | | | |
| c | | | • | |
| d | | • | • | |

For example, this says that $a \in A$ is in relation with e and g in B , but not with f and h . While the domain of R is the whole set A , the element $h \in B$ is not contained in the range of R . That is, $\text{range}(B) = \{e, f, g\}$.

Remark A.5.1. In the case of an at most countably infinite set $A = \{a_1, a_2, \dots\}$ there exists another way to describe relations R on A . Draw an arrow from a_i to a_j whenever $(a_i, a_j) \in R$. For example, if $A = \{a, b, c\}$ and

$$(A.5.1) \quad R = \{(a, a), (a, b), (b, a), (c, a), (c, b), (c, c)\},$$

the corresponding figure is given by the below drawing.

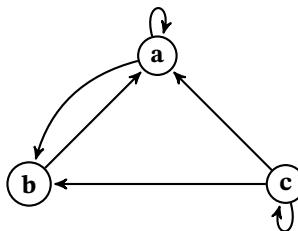


Figure A.5.1. $R = \{(a, a), (a, b), (b, a), (c, a), (c, b), (c, c)\}$.

Remark A.5.2. For any two sets A and B there always exist two trivial relations. The first one appears when $R = \emptyset$. Then none of the $x \in A$ is in relation with some $y \in B$. The second trivial relation occurs if $R = A \times B$. Then all $x \in A$ are related with all $y \in B$. Of course, both relations are not of very special interest.

Remark A.5.3. There is a tight connection between functions f from a set A into B and relations R on $A \times B$. Indeed, let

$$R = \text{graph}(f) = \{(x, y) : y = f(x)\} \subseteq A \times B.$$

Then R is a relation with an additional feature, the so-called *vertical line property* as introduced in (A.3.1):

$$(A.5.2) \quad (\forall x \in A)(\exists! y \in B)((x, y) \in R).$$

In particular, it follows that $A = \text{dom}(f) = \text{dom}(R)$ and $\text{range}(R) = \text{range}(f)$.

Conversely, if $R \subseteq A \times B$ is a relation with property (A.5.2), then it generates a function $f : A \rightarrow B$ by setting

$$f(x) = y \quad \text{provided that } (x, y) \in R.$$

Moreover, then $\text{graph}(f) = R$.

$$\{\text{Functions from } A \text{ to } B\} \Leftrightarrow \{\text{Relations on } A \times B \text{ with property (A.5.2)}\}$$

The definition of a relation as subset of $A \times B$ is very general. They become of greater interest if they possess some additional properties. Let us state the most important of those properties.

Definition A.5.3. Let $R \subseteq A \times A$ be a relation on a set A

- (1) The relation R is **reflexive** if for all $x \in A$ it follows that $(x, x) \in R$.

$$(\forall x \in A)[(x, x) \in R]$$

- (2) The relation R is called **symmetric** provided that $(x, y) \in R$ implies $(y, x) \in R$.

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

- (3) R is **anti-symmetric**: $(x, y) \in R$ and $(y, x) \in R$ yield $x = y$.

$$(\forall x, y \in A)[(x, y) \in R \text{ and } (y, x) \in R \Rightarrow x = y]$$

- (4) The relation R is **transitive** if $(x, y) \in R$ and $(y, z) \in R$ imply $(x, z) \in R$.

$$(\forall x, y, z \in A)[(x, y) \in R \text{ and } (y, z) \in R \Rightarrow (x, z) \in R]$$

Example A.5.2. As first example we investigate the relation R on $A = \{a, b, c, d\}$ given by

| | a | b | c | d |
|---|---|---|---|---|
| a | • | | • | |
| b | | • | | • |
| c | • | | • | |
| d | | • | | • |

- (1) Each element in A is in relation with itself. Hence, R is reflexive.
- (2) Whenever $x \in A$ is in relation with some $y \in A$, then also the converse is true. Thus, R is also symmetric.
- (3) The relation R is **not** anti-symmetric. Indeed, $(a, c) \in R$ and $(c, a) \in R$, but $a \neq c$.
- (4) It is easy to check that R is also transitive. Investigate all possible ways to compose two pairs of elements.

Let us investigate some more examples of relations.

Example A.5.3. Let A be the set of inhabitants in some country. Say a person x is in relation with another person y if both were born at the same day of the year, but not necessarily in the same year. The generated relation $R \subseteq A \times A$ possesses the following properties:

- (1) For each $x \in A$ it follows $(x, x) \in R$, hence R is reflexive.
- (2) If $(x, y) \in R$, then also $(y, x) \in R$. That is, R is symmetric.
- (3) Suppose $(x, y) \in R$ and $(y, z) \in R$ for some $z \in A$, then also $(x, z) \in R$. Thus, R is also transitive.

Example A.5.4. A relation R on \mathbb{N} may be defined as follows: $(n, m) \in R$ provided that $n \mid m$ (n divides m). In Proposition 1.6.1 we already have proved that this relation is symmetric, anti-symmetric and transitive.

Example A.5.5. Let us say that two natural numbers m and n are related if they are coprime, i.e., if $\gcd(n, m) = 1$. This relation is **not** reflexive. Indeed, for all $n \in \mathbb{N}$ we have $\gcd(n, n) = n$, thus only 1 is in relation with itself. Because of $\gcd(n, m) = \gcd(m, n)$ the relation is symmetric. But it is neither transitive nor anti-symmetric. For example, 3 and 7 as well as 7 and 12 are coprime, but 3 and 12 are not so.

Let us discuss a few more examples.

- Say a student x is in relation with a student y provided that both of them achieved the same grade in a certain class. Then this relation is reflexive, symmetric and transitive, but not anti-symmetric.
- A person x is in relation with a person y whenever x is married with y . Of course, this relation is not reflexive, nor anti-symmetric nor transitive. But it is symmetric.
- Two natural numbers n and m are in relation if $n \leq m$. As mentioned in (1.1.2), this relation is reflexive, anti-symmetric and transitive but, of course, it is not symmetric.

Definition A.5.4. Let R be a binary relation on a nonempty set S . It is said to be a **partial order** on S if it is reflexive, anti-symmetric and transitive.

$$(R \text{ partial order on } S) \Leftrightarrow (R \text{ is reflexive, anti-symmetric and transitive})$$

Remark A.5.4. Let R on S is a partial order on S , then it is common to write $x \leq y$ whenever $(x, y) \in R$.

$$x \leq y \Leftrightarrow (x, y) \in R$$

Moreover,

$$x < y \Leftrightarrow x \leq y \text{ and } x \neq y$$

Then the properties of R go over to

- (1) $(\forall x \in S)(x \leq x)$
- (2) $(\forall x, y \in S)[(x \leq y \text{ and } y \leq x) \Rightarrow (x = y)]$
- (3) $(\forall x, y, z \in S)[(x \leq y \text{ and } y \leq z) \Rightarrow (x \leq z)]$

Definition A.5.5. The following notation is used often:

$$S \text{ is a } \mathbf{poset} \text{ (partial ordered set)} \Leftrightarrow \exists \text{ a partial order } \leq \text{ on } S$$

Remark A.5.5. Let S be a poset and $A \subseteq S$ be a nonempty subset. Then restricting the order on S to A , the set A becomes a poset as well. A typical example is the order on \mathbb{Z} as restriction of the given order on \mathbb{Q} or on \mathbb{R} , respectively.

Remark A.5.6. Given a partial order on S , this does not necessarily say that any two elements in S are comparable. More precisely, for any $x, y \in S$ there exist three possibilities: $x \leq y$ or $y \leq x$ or neither of both, that is, x and y are not comparable.

Example A.5.6. (1) Let $S = \mathbb{N} \times \mathbb{N} = \{(n, m) : n, m \in \mathbb{N}\}$ and define

$$(n_1, m_1) \leq (n_2, m_2) \text{ provided that } n_1 \leq n_2 \text{ and } m_1 \leq m_2.$$

Then the relation \leq is a partial order on $\mathbb{N} \times \mathbb{N}$. With respect to this order it follows that $(1, 1) \leq (1, 3)$ or $(2, 2) \leq (3, 4)$, but we have neither $(1, 2) \leq (2, 1)$ nor $(2, 1) \leq (1, 2)$.

(2) Consider the so-called lexicographical order¹⁴ on $\mathbb{N} \times \mathbb{N}$. It is defined by

$$(n_1, m_1) \leq (n_2, m_2) \text{ if either } n_1 < n_2 \text{ or if } m_1 \leq m_2 \text{ in the case } n_1 = n_2.$$

In this way $(1, 7) \leq (2, 3)$ or $(1, 6) \leq (1, 8)$.

(3) Let X be a set with power set $\mathcal{P}(X)$. A partial order on $\mathcal{P}(X)$ is then defined by

$$A \leq B : \Leftrightarrow A \subseteq B.$$

Of course, in general there exist subsets A and B of X where neither $A \subseteq B$ nor $B \subseteq A$.

(4) For each pair (n, m) of natural numbers define $n \leq m$ if n divides m , i.e., if $n \mid m$. As shown in Proposition 1.6.1 this is indeed an order. But again, there exist noncomparable integers in this order. For example, we have neither $5 \leq 9$ nor $9 \leq 5$.

Remark A.5.7. Sometimes it is useful to illustrate a partial order by a diagram. Elements lying on the same branch are comparable, and $x < y$ if y is on the same branch as x , but above x . On the contrary, elements lying on different branches are not comparable. Of special interest is when the diagram has a root, i.e., there is some $x_0 \in S$ (the root) with $x_0 \leq x$ for all $x \in S$. In this case the diagram becomes a tree with root x_0 .

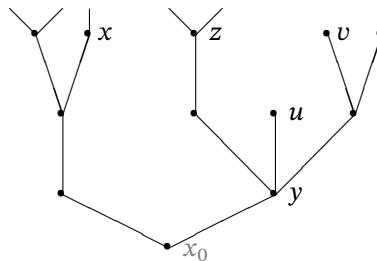


Figure A.5.2. The diagram (tree) of a partial order with root x_0 .

In Figure A.5.2 we have $y < z$, $y < u$ and $y < v$, but x is not comparable with v , y and z . The elements v and z are also not in relation.

Definition A.5.6. Let S be a poset with partial order \leq . An element $M \in S$ is said to be a **maximal element** of S provided there is no $x \in S$ such that $M < x$.

Similarly, $m \in S$ is a **minimal element** whenever there is no $x \in S$ being strictly smaller than m .

$$(M \text{ is a maximal element of } S) \Leftrightarrow [M \in S \text{ and } (\nexists x \in S)(M < x)]$$

and

$$(m \text{ is a minimal element of } S) \Leftrightarrow [m \in S \text{ and } (\nexists x \in S)(x < m)].$$

¹⁴This is the way how words are alphabetically ordered in a dictionary.

Remark A.5.8. It may happen that a poset has no maximal or no minimal elements, but there may also exist many such elements. For example, if we look at the order described by Figure A.5.2, then u and v are maximal elements while the only minimal element is x_0 .

Another trivial example is as follows: Given an arbitrary nonempty set S we define a partial order on S by

$$x \leq y \Leftrightarrow x = y, \quad x, y \in S.$$

Then any element in S is at the same time a maximal and a minimal element.

In the case of posets one has to distinguish carefully between maximal elements and maxima and in the same way between minimal elements and minima. Recall that an element $M \in S$ is said to be a **maximum** of the poset S provided that $x \leq M$ for all $x \in S$. In the same way $x \in S$ is a minimum of S whenever $m \leq x$ for all $x \in S$.

$$\begin{aligned} (M \text{ is a maximum of } S) &\Leftrightarrow [M \in S \text{ and } (\forall x \in S)(x \leq M)] \\ (m \text{ is a minimum of } S) &\Leftrightarrow [m \in S \text{ and } (\forall x \in S)(m \leq x)]. \end{aligned}$$

Remark A.5.9. Of course, a maximum of the poset is also a maximal element in S . But as the example in Remark A.5.7 shows, there are easy examples of posets without maximum but with several maximal elements.

Proposition A.5.1. *Whenever a poset S has a maximum (minimum), then this is unique.*

Proof: Suppose there are two maxima $M_1, M_2 \in S$, i.e., we have $M_1, M_2 \in S$ and $x \leq M_1$ as well as $x \leq M_2$ for all $x \in S$. In particular, this implies $M_2 \leq M_1$ as well as $M_1 \leq M_2$. But the relation \leq is anti-symmetric, hence $M_1 = M_2$. By similar reasoning, a minimum of a poset S is unique, provided it exists. ■

Consequently, if a maximum or a minimum of a poset S exist, we may denote them by $\max(S)$ or $\min(S)$, respectively.

$$\begin{aligned} [M = \max(S)] &\Leftrightarrow [(M \in S) \text{ and } (\forall x \in S)(x \leq M)] \\ [m = \min(S)] &\Leftrightarrow [(m \in S) \text{ and } (\forall x \in S)(m \leq x)]. \end{aligned}$$

Example A.5.7.

(1) If the partial order on a set $\mathcal{P}(X)$ is defined by $A \leq B$ if $A \subseteq B$, the empty set \emptyset is a minimum while the set X is a maximum with respect to the given order on $\mathcal{P}(X)$.

(2) Consider the real numbers with its natural order. If $A = \{\frac{n}{n+1} \mid n \geq 1\}$, then there is no element $M \in A$ satisfying $\frac{n}{n+1} \leq M$, for all $n \geq 1$. Of course, $x \leq 1$ for all $x \in A$, but 1 is not in A , hence 1 is not a maximum of A . But there are also no other maximal elements as can be seen easily.

(3) If the partial order on \mathbb{N} is defined by $n \leq m$ if $n \mid m$, then 1 is a minimum, but there are no maximal elements. Indeed, for any $n \in \mathbb{N}$ it follows $n < 2n$, so n is not maximal.

(4) The integers \mathbb{Z} with its natural order have neither a minimal nor a maximal element.

Definition A.5.7. A relation $R \subseteq S \times S$ is an **order** (sometimes also called total order) if it is a partial order and, furthermore, for all $x, y \in S$ either

$$(x, y) \in R \text{ or } (y, x) \in R \Leftrightarrow x \leq y \text{ or } y \leq x.$$

$$[\text{A partial order } \leq \text{ on } S \text{ is an order}] \Leftrightarrow (\forall x, y \in S)(x \leq y \text{ or } y \leq x).$$

Remark A.5.10. In the visualization of orders by trees, an order is a tree consisting only of one branch. But note, that this does not necessarily imply that this tree has a root as the example $S = \mathbb{Z}$ endowed with its natural order shows.

Another important class of relations is that of equivalence relations.

Definition A.5.8. Let S be some nonempty set. A relation R on S is said to be an **equivalence relation** on S if it is reflexive, symmetric and transitive.

Remark A.5.11. Let R be an equivalence relation on S . Then it is common to write $x \sim y$ whenever $(x, y) \in R$.

$$x \sim y \Leftrightarrow (x, y) \in R$$

Then the properties of R go over to

- (1) $(\forall x \in S)(x \sim x)$
- (2) $(\forall x, y \in S)[(x \sim y) \Leftrightarrow (y \sim x)]$
- (3) $(\forall x, y, z \in S)[(x \sim y \text{ and } y \sim z) \Rightarrow (x \sim z)]$

Example A.5.8. In Section 2.4, we investigated a very important example of an equivalence relation. Fix a modulus of congruence $n \geq 1$. Then, as introduced in Definition 2.4.1, two integers a and b are said to be congruent modulo n which we write as $a \equiv b \pmod{n}$ if n divides $a - b$. In formulas this says

$$a \equiv b \pmod{n} \Leftrightarrow n \mid a - b.$$

Defining $a \sim b \Leftrightarrow a \equiv b \pmod{n}$, then Proposition 2.4.3 asserts that “ \sim ” is an equivalence relation on \mathbb{Z} .

Example A.5.9. In Section 3.1 we investigated another equivalence relation, this time given on $\mathbb{Z} \times \mathbb{Z}^*$. Recall that $\mathbb{Z}^* = \mathbb{Z} \setminus \{0\}$. The relation was defined in (3.1.2) as follows: If (m, n) and (p, q) are in $\mathbb{Z} \times \mathbb{Z}^*$, then these two pairs of integers are in relation if

$$(A.5.3) \quad (m, n) \sim (p, q) \Leftrightarrow mq = np.$$

Indeed, this is an equivalence relation as proved in Proposition 3.1.1.

So, for example,

$$(-3, -6) \sim (2, 4) \text{ but } (1, 2) \not\sim (3, 4).$$

Let us give some more examples of equivalence relations.

Example A.5.10.

(1) Suppose that two persons are in relation if they were born at the same day of the year. Then this generates an equivalence relation on the set of all inhabitants in some country.

(2) Let f be an arbitrary function from a set A into a set B . Given two elements $x_1, x_2 \in A$, then x_1 is in relation with x_2 if $f(x_1) = f(x_2)$. Check (it is easy) that this is an equivalence relation.

(3) If two finite sets A and B are in relation whenever $|A| = |B|$, i.e., if they contain exactly the same number of elements, then this is an equivalence relation on all finite sets.

Each equivalence relation splits in natural way the underlying set into disjoint subsets. To clarify this statement, let us introduce the notion of equivalence classes.

Definition A.5.9. Let \sim be an equivalence relation on a set S . Given $x \in S$ set

$$\hat{x} := \{y \in S : x \sim y\}.$$

The set $\hat{x} \subseteq S$ is called the **equivalence class** generated by $x \in S$. Note that \hat{x} is a **subset** of S , not an element of S .

Example A.5.11. Let us describe the equivalence classes of the relation treated in Example A.5.8. If $n \geq 1$ is the modulus of congruence, then there are exactly n different equivalence classes, namely

$$\hat{0}, \hat{1}, \dots, \widehat{n-1}.$$

Another way to formulate this is as follows: Given an integer $0 \leq r < n$, then

$$\hat{r} = \{a \in \mathbb{Z} : a \text{ has remainder } r \text{ when divided by } n\}.$$

For example, if $n = 2$, then $\hat{0}$ consists of all even integers while $\hat{1}$ is the set of all odd numbers.

If $n = 5$, then

$$\hat{0} = \{\dots, -10, -5, 0, 5, 10, \dots\} \quad \text{while} \quad \hat{3} = \{\dots, -12, -7, -2, 3, 8, 13, \dots\}$$

It is important to mention, that, for example,

$$\hat{0} = \hat{5} = \hat{35} = \widehat{-10}.$$

Example A.5.12. How do the equivalence classes look like in the case of the relation investigated in Example A.5.9? We answered this question in Section 3.1. Given a pair $(m, n) \in \mathbb{Z} \times \mathbb{Z}^*$ one writes

$$\frac{m}{n} \quad \text{instead of} \quad (m, n).$$

Then, for example, the equivalence class generated by $1/2$ consists of all fractions m/n for which

$$(m, n) \sim (1, 2) \Leftrightarrow n = 2m.$$

So we got

$$\widehat{\left(\frac{1}{2}\right)} = \left\{ \frac{1}{2}, \frac{2}{4}, \frac{-1}{-2}, \frac{-2}{-4}, \dots \right\} = \left\{ \frac{k}{2k} : k \in \mathbb{Z}^* \right\}.$$

After describing the equivalence class generated by $1/2$, let us treat now the general case. To do so recall Definition 3.2.1. According to that a fraction is said to be irreducible provided that nominator and denominator are coprime.

The next proposition is a consequence of Proposition 2.3.3. We leave its proof as an exercise.

Proposition A.5.2. *Let m_1/n_1 and m_2/n_2 be two irreducible fractions with positive denominators n_1, n_2 . Then these two fractions are equivalent with respect to relation (A.5.3) if and only if $m_1 = m_2$ and $n_1 = n_2$.*

Moreover, if m/n is irreducible, then any other fraction p/q is equivalent to m/n if and only if there is $k \in \mathbb{Z}^*$ such that

$$p = k m \quad \text{and} \quad q = k n \quad \text{or, equivalently, if} \quad \frac{p}{q} = \frac{k m}{k n}.$$

As a consequence we get the following: If m/n is irreducible, then

$$\widehat{\left(\frac{m}{n}\right)} = \left\{ \frac{k m}{k n} : k \in \mathbb{Z}^* \right\}.$$

Example A.5.13. In the first case of Example A.5.10 there exist (disregard leap years) at most 365 different equivalence classes. For example, in class 1 are those persons born at January 1, in the second one those born at January 2, and so on. Of course, the number of different equivalence classes may be strictly less than 365. This happens if there are days of the year at which nobody of the regarded persons was born.

Let us state now the main properties of equivalence classes.

Proposition A.5.3. *Let \sim be an equivalence relation on a set S with generated equivalence classes \hat{x} , $x \in S$. The following statements are true.*

- (1) *For all $x \in S$ we have $x \in \hat{x}$.*
- (2) *It follows that*

$$S = \bigcup_{x \in S} \hat{x}.$$

- (3) *Given $x, y \in S$, then either $\hat{x} = \hat{y}$ or $\hat{x} \cap \hat{y} = \emptyset$. The former happens if and only if $x \sim y$. Consequently, the generated equivalence classes are disjoint if and only if $x \not\sim y$.*

$$(\hat{x} = \hat{y}) \Leftrightarrow (x \sim y) \quad \text{and} \quad (\hat{x} \cap \hat{y} = \emptyset) \Leftrightarrow (x \not\sim y)$$

Proof: Since the relation \sim is reflexive it follows that $x \sim x$, hence $x \in \hat{x}$. Of course, this implies property (2).

To prove the third part assume first $x \sim y$. Take an arbitrary element $z \in \hat{x}$. Then $x \sim z$, hence by the transitivity of the relation we also get $y \sim z$. Thus, $z \in \hat{y}$. Interchanging the roles of x and y implies $\hat{x} = \hat{y}$ whenever $x \sim y$. Suppose now $x \not\sim y$. If there is $z \in S$ with $z \in \hat{x} \cap \hat{y}$, then by definition $z \sim x$ as well as $z \sim y$. Again we conclude that this implies $x \sim y$ contradicting our assumption. This completes the proof. ■

Remark A.5.12. Another way to express the statement of Proposition A.5.3 is as follows. The set S splits into disjoint subsets. Each of these subsets consists of elements which are pairwise equivalent. In particular, any element of S belongs to one and only one equivalence class.

Example A.5.14. Define an equivalence relation on \mathbb{Z} by $x \sim y$ if $x^2 = y^2$. Then, for example, $-1 \sim 1$. Given an integer $n \in \mathbb{N}$, the generated equivalence class \hat{n} is

$$\hat{n} = \{-n, n\} \quad \text{while} \quad \hat{0} = \{0\}.$$

Thus $\mathbb{Z} = \hat{0} \cup \hat{1} \cup \hat{2} \cup \dots$. But observe that also $\mathbb{Z} = \hat{0} \cup \widehat{(-1)} \cup \widehat{(-2)} \cup \dots$ because of $\hat{n} = \widehat{(-n)}$.

Example A.5.15. Let two natural numbers n and m be equivalent if we have $[n/10] = [m/10]$, then $\hat{1} = \{1, \dots, 9\}$, $\hat{10} = \{10, \dots, 19\}$ and so on. Of course, also $\hat{1} = \hat{9}$ or $\hat{10} = \hat{28}$.

Remark A.5.13. Maybe it is a bit confusing that the same set \hat{x} may have different names. Instead of \hat{x} we can denote it also by \hat{y} for any $y \in S$ with $x \sim y$. Nevertheless, it is always the same set. This is similar to the following: Suppose we name a city by one of its inhabitants. So, for example, assume the city is called Maria's city. But we could call it also Frank's city if Frank is another inhabitant. The city remains the same, only its name has changed.

Finally, we want to introduce a common notation for the collection of all equivalence classes.

Definition A.5.10. Let \sim be an equivalence relation on a set S . Then

$$S/\sim := \{\hat{x} : x \in S\}$$

is the **quotient set** of S with respect to the equivalence relation \sim .

Example A.5.16. Fix $n \geq 1$ and consider the equivalence relation on \mathbb{Z} defined by

$$a \sim b \quad \text{if} \quad a \equiv b \pmod{n}.$$

In this case it is common to write $\mathbb{Z}/n\mathbb{Z}$ instead of \mathbb{Z}/\sim . That is

$$\mathbb{Z}/\sim = \mathbb{Z}/n\mathbb{Z} = \{\hat{0}, \dots, \widehat{n-1}\}.$$

Example A.5.17. Let the equivalence relation on $\mathbb{Z} \times \mathbb{Z}^*$ be defined by (A.5.3). Then by Proposition A.5.2 the quotient set may be described as follows:

$$\mathbb{Z} \times \mathbb{Z}^*/\sim = \left\{ \left(\widehat{\frac{m}{n}} \right) : \frac{m}{n} \text{ irreducible, } n > 0 \right\}.$$

Observe that the only equivalence classes are those generated by irreducible pairs m/n with $n > 0$.

Remark A.5.14. Summing up, when we deal with fractions m/n , then in fact we treat the equivalence class $\widehat{m/n}$. So, the equation $1/2 = 2/4$ as taught in school, means in fact that

$$\widehat{\left(\frac{1}{2}\right)} = \widehat{\left(\frac{2}{4}\right)}.$$

Thereby it is important that, as mentioned in Section 3.1, neither the algebraic operations nor the order depend on the chosen elements in the equivalence classes. So, for example,

$$\frac{1}{2} + \frac{1}{4} = \frac{3}{6} + \frac{3}{12} \quad \text{or} \quad \frac{1}{2} \cdot \frac{1}{4} = \frac{3}{6} \cdot \frac{3}{12}$$

or

$$\frac{1}{4} < \frac{1}{2} \quad \text{is equivalent to saying that} \quad \frac{2}{8} < \frac{5}{10}.$$

Exercise A.5.1. Suppose a relation R on $A = \{a, b, c, d\}$ is given by the table

| | a | b | c | d |
|---|---|---|---|---|
| a | • | | | |
| b | • | • | | • |
| c | • | • | • | |
| d | | | | • |

Is this relation reflexive, symmetric, anti-symmetric, or transitive?

Exercise A.5.2. Let $R \subseteq A \times B$ be a relation. Then we may define the dual relation \tilde{R} on $B \times A$ as follows:

$$(A.5.4) \quad (y, x) \in \tilde{R} \Leftrightarrow (x, y) \in R.$$

- (1) How do the dual relations look like in the case of the relation in Example A.5.1 and for the one given in (A.5.1)?
- (2) Suppose the relation R is generated by a function $f : A \rightarrow B$. Under what conditions is the dual relation \tilde{R} generated by a function g from B to A ? How are f and g related whenever such a function g exists?

Exercise A.5.3. Let R be a relation on a set A . Define the dual relation \tilde{R} on A by (A.5.4). Which properties does \tilde{R} possess in the case of a reflexive or symmetric or anti-symmetric or transitive relation R ?

Exercise A.5.4. Describe the relation (A.5.1) in form of a table.

Exercise A.5.5. Prove the following: Suppose that a poset S has a maximum $\max(S)$. Then $\max(S)$ is the only maximal element in S .

Exercise A.5.6. Let \leq be a (total) order on S . Prove that there exists at most one maximal element. Moreover, this happens if and only if $\max(S)$ exists.

Exercise A.5.7. Is the (partial) lexicographical order in (2) of Example A.5.6 a (total) order? Justify your answer.

Exercise A.5.8. Let S consist of those natural numbers less than 60 which divide 60. That is,

$$S = \{n \in \mathbb{N} : n < 60, n \mid 60\}.$$

Define a relation on S by

$$a \leq b \Leftrightarrow a \mid b.$$

- (1) Show that this is indeed a partial order on S .
- (2) Why is \leq not a (total) order?
- (3) Determine maximal and minimal elements in S . Do $\max(S)$ and/or $\min(S)$ exist?

Exercise A.5.9. A relation on $S = \{a, b, c, d, e, f\}$ is defined by the following table:

| | a | b | c | d | e | f |
|---|---|---|---|---|---|---|
| a | • | • | • | • | • | • |
| b | | • | | | | • |
| c | | | • | • | • | |
| d | | | | • | | |
| e | | | | | • | |
| f | | | | | | • |

Show that it is an order relation. Do there exist minimal and/or maximal elements in S ? Draw a tree related to this partial order.

Exercise A.5.10. Let X be a finite nonempty set. Define a partial order on

$$S = \{A : A \subseteq X : A \neq \emptyset, A \neq X\} = \mathcal{P}(X) \setminus \{\emptyset, X\}$$

by

$$A \leq B \Leftrightarrow A \subseteq B.$$

Describe maximal and minimal elements in S .

Exercise A.5.11. A subset K of a poset (S, \leq) is called **chain** provided that for all $x, y \in K$ one has either $x \leq y$ or $y < x$. Give three different examples of infinite chains in (\mathbb{N}, \leq) where $n \leq m$ if $n \mid m$ and give two different examples of infinite chains in $(\mathcal{P}(\mathbb{N}), \subseteq)$.

Exercise A.5.12. Let \leq be a partial order on a set S and let $R = \{(x, y) : x \leq y\}$ be the generating relation. Define the dual relation \tilde{R} as in (A.5.4). Is \tilde{R} also an order relation? If this is so, how does the generated order \leq look like? Here we set

$$x \leq y \Leftrightarrow (x, y) \in \tilde{R}.$$

For example, if “ \leq ” on $\mathcal{P}(M)$ is defined by

$$A \leq B \Leftrightarrow A \subseteq B,$$

describe $A \leq B$ for subsets A, B of M .

Exercise A.5.13. Define a relation R on \mathbb{N} by $(n, m) \in R$ if $n \geq m$. Which properties does this relation possess? Is it an order relation? If yes, do there exist maximal and/or minimal elements in \mathbb{N} ? How is this problem related to Exercise A.5.12

Exercise A.5.14. Suppose a relation R on $\{a, b, c\}$ is given by

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

Show that this is an equivalence relation and determine the equivalence classes.

Exercise A.5.15. Let us say that two real numbers are equivalent if $\sin x = \sin y$. That is,

$$x \sim y \Leftrightarrow \sin x = \sin y, \quad x, y \in \mathbb{R}.$$

Describe the equivalence classes generated by this relation. Why one can identify \mathbb{R}/\sim on one hand with $[0, 2\pi)$ and on the other hand with a circle of radius 1 in the plane?

Exercise A.5.16. Recall that two triangles in the plane are **congruent** if their corresponding sides are equal in length. Show that this is an equivalence relation on the set of all triangles in the plane. Describe the equivalence classes generated by this relation.

Exercise A.5.17. Two triangles in the plane are said to be **similar** if the corresponding angles coincide. Show that the similarity of triangles is an equivalence relation on the set of all triangles. Describe the generated equivalence classes.

Exercise A.5.18. We define a relation “ \sim ” on the set of nonzero real numbers as follows:

$$(A.5.5) \quad x \sim y \Leftrightarrow \frac{x}{y} = 2^k \text{ for some } k \in \mathbb{Z}.$$

Show that this is an equivalence relation on $\mathbb{R} \setminus \{0\}$ and determine the equivalence classes.

What happens if we define the relation by

$$x \sim y \Leftrightarrow \frac{x}{y} = a^k \text{ for some } k \in \mathbb{Z},$$

where $a \in \mathbb{R}$ is some fixed positive number? In other words, do we get also an equivalence relation if we replace the number 2 in (A.5.5) by an arbitrary (fixed) number $a > 0$?

Exercise A.5.19. Is the relation in Example A.5.2 an equivalence relation? If yes, then find all equivalence classes.

Exercise A.5.20. The following relation is given on $S = \{a, b, c, d, e, f\}$:

| | a | b | c | d | e | f |
|---|---|---|---|---|---|---|
| a | • | | • | | | |
| b | | • | | • | | • |
| c | • | | • | | | |
| d | | • | | • | | • |
| e | | | | | • | |
| f | | • | | • | | • |

Show that it is an equivalence relation and determine the equivalence classes.

A.6. Proofs

In each mathematical statement that we wish to prove, we have a hypothesis which consists of the statements that we assume to be true and a conclusion which is the statement that we wish to prove. For some problems such as *$\sqrt{2}$ is irrational* or *there are infinitely many primes*, only the conclusion is stated while the hypothesis is omitted. One can restate *$\sqrt{2}$ is irrational* as *if x is the positive number such that $x^2 = 2$, then $x \neq a/b$ for any nonzero integers a and b* .

A proof is like a journey from our hypothesis to our conclusion, where at each step we have to obey the traffic rules of mathematics. Sometimes, we remember a proof by recalling each tree that we passed by on our way and other times when we have a better understanding, we just know our way by knowing the big crossroads or junctions on our trip.

There are several types of proofs one may encounter or use in mathematics. We briefly describe some of them here.

A.6.1. Direct proof. As the name suggested, this method involves using the hypothesis in a direct way to lead us to the conclusion.

Example A.6.1. If a is an even integer, then a^2 is an even integer.

Proof: Our hypothesis is that a is even. This means that $a = 2k$ for some integer k . Since our hypothesis involves a^2 , it makes sense to square the both sides of the previous equation and get that $a^2 = 4k^2$. To finish the proof, we note that $4k^2 = 2(2k^2)$ is an even integer. Hence, a^2 is an even integer. ■

Example A.6.2. If $3 \mid a$ and $3 \mid b$, then $3 \mid a + b$.

Proof: Our hypothesis is that a and b are divisible by 3. Therefore, $a = 3a_1$ and $b = 3b_1$, for some integers a_1 and b_1 . Since our hypothesis deals with $a + b$, we add the previous two equations and get that $a + b = 3a_1 + 3b_1 = 3(a_1 + b_1)$. Because $a_1 + b_1$ is an integer, this means that $a + b$ is divisible by 3. ■

In some situations, we can use a direct proof method by starting with the hypothesis.

Example A.6.3. Let $f : A \rightarrow B$ and $g : B \rightarrow C$ be two functions. If $g \circ f$ is injective, then f is injective.

Proof: We are asked to prove that f is injective. This means that we have to show that for any $a_1, a_2 \in A$, if $f(a_1) = f(a_2)$, then $a_1 = a_2$. Let $a_1, a_2 \in A$ such that $f(a_1) = f(a_2)$. Our only hypothesis is that $g \circ f : A \rightarrow C$ is injective. It makes sense to get g involved in the action and from $f(a_1) = f(a_2)$, we deduce that $(g \circ f)(a_1) = g(f(a_1)) = g(f(a_2)) = (g \circ f)(a_2)$. Because $g \circ f : A \rightarrow C$ is injective, we get that $a_1 = a_2$. Hence, f is injective. ■

Example A.6.4. Let $f : A \rightarrow B$ and $g : B \rightarrow C$ be two functions. If $g \circ f : A \rightarrow C$ is surjective, then g is surjective.

Proof: We have to prove that for any $c \in C$, there exists $b \in B$ such that $g(b) = c$. So we start with an arbitrary $c \in C$. Our hypothesis is that $g \circ f : A \rightarrow C$ is surjective. That implies that there exists $a \in A$ such that $c = (g \circ f)(a)$. The right-hand side can be rewritten as $g(f(a))$ so $c = g(f(a))$. We denote $f(a) \in B$ by b and therefore, $c = g(f(a)) = g(b)$. This means that g is surjective. ■

A.6.2. Proof by cases. Perhaps a good way to remember this method is to think of Yogi Berra's quote:

When you come to a fork in a road, take it!

Assume we are trying to prove that a statement q is true. It may happen that we need additional information (such a statement or predicate p) to our hypothesis to be able to prove q and may be helpful to split the proof into cases: if p is true, then q is true, and if p is false, then q is true. The logical foundation of this method is that q is equivalent with $(p \Rightarrow q) \wedge ((\neg p) \Rightarrow q)$ (see Exercise A.1.3).

Example A.6.5. Let $f : \mathbb{R} \rightarrow \mathbb{R}$ be a function defined as follows:

$$f(x) = \begin{cases} 2x + 1, & \text{if } x \leq 2, \\ 3x - 1, & \text{if } x > 2. \end{cases}$$

Prove that f is injective and surjective.

Proof: To prove that f is injective, we have to show that for any $x_1, x_2 \in \mathbb{R}$, if $f(x_1) = f(x_2)$, then $x_1 = x_2$.

Let $x_1, x_2 \in \mathbb{R}$ such that $f(x_1) = f(x_2)$. The values of $f(x_1)$ and $f(x_2)$ depend on the locations of x_1 and x_2 with respect to 2. This leads to the following cases.

Case 1: $x_1 \leq 2$ and $x_2 \leq 2$.

In this case, $f(x_1) = 2x_1 + 1$ and $f(x_2) = 2x_2 + 1$. From $f(x_1) = f(x_2)$, we get that $2x_1 + 1 = 2x_2 + 1$. Subtracting 1 from both sides and then dividing by 2, we deduce that $x_1 = x_2$. This finishes the proof in this case.

Case 2: $x_1 \leq 2$ and $x_2 > 2$.

In this case, $f(x_1) = 2x_1 + 1$ and $f(x_2) = 3x_2 - 1$. Because $x_1 \leq 2$, we get that $f(x_1) = 2x_1 + 1 \leq 5$. Because $x_2 > 2$, we deduce that $f(x_2) = 3x_2 - 1 > 5$. Thus, $f(x_1) \leq 5 < f(x_2)$. This contradicts $f(x_1) = f(x_2)$ and it means that this case cannot happen.

Case 3: $x_1 > 2$ and $x_2 \leq 2$.

The analysis of this case is similar to Case 2 above, and we leave to the interested reader to complete.

Case 4: $x_1 > 2$ and $x_2 > 2$.

We have that $f(x_1) = 3x_1 - 1$ and $f(x_2) = 3x_2 - 1$. Adding one to both sides and then dividing by 3, we obtain that $x_1 = x_2$. This finishes the proof of this final case and our proof that f is injective.

To show that f is surjective, we have to prove that for any $y \in \mathbb{R}$, there exists $x \in \mathbb{R}$ such that $f(x) = y$. The previous arguments pointed to the following case analysis.

Case A: $y \leq 5$.

In this case, let $x = \frac{y-1}{2}$. Because $y \leq 5$, $x \leq 2$. Therefore, $f(x) = 2x + 1 = y$.

Case B: $y > 5$.

In this situation, let $x = \frac{y+1}{3}$. Because $y > 5$, $x > 2$. Hence, $f(x) = 3x - 1 = y$. This finishes our proof that f is surjective. ■

Example A.6.6. A real number x satisfies

$$(A.6.1) \quad |x - 2| < |x + 2|$$

if and only if it is positive.

Proof: There are several ways to verify this. One possibility is to investigate the following three different cases separately:

$$(a) \quad -\infty < x < -2, \quad (b) \quad -2 \leq x \leq 2 \quad \text{and} \quad (c) \quad 2 < x < \infty.$$

Case (a): Here we have

$$|x - 2| < |x + 2| \Leftrightarrow 2 - x < -2 - x,$$

which is never satisfied.

Case (b): In this case it follows that

$$|x - 2| < |x + 2| \Leftrightarrow 2 - x < x + 2 \Leftrightarrow x > 0.$$

Case (c): If $x > 2$, then we observe that

$$|x - 2| < |x + 2| \Leftrightarrow x - 2 < x + 2,$$

which is always true.

Summing up, it follows that (A.6.1) is satisfied exactly if $x > 0$.

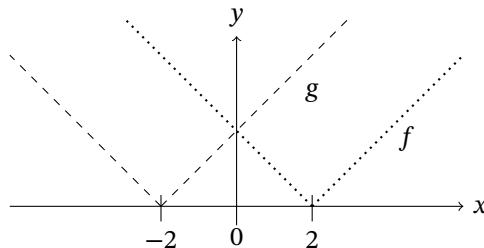


Figure A.6.1. The graphs of $f(x) = |x - 2|$ and $g(x) = |x + 2|$. Then $f(x) < g(x)$ if and only if $x > 0$.

Another way to get an intuition and to prove this result is to consider the following functions:

$$f(x) := |x - 2| \quad \text{and} \quad g(x) := |x + 2|, \quad x \in \mathbb{R}.$$

Draw the graphs of f and g as in Figure A.6.1 and check where the graph of g is above the graph of f . ■

A.6.3. Proof by contradiction. This method of proof is applicable when there is no clear path on how to approach the problem or how to use the hypothesis directly. Informally, we assume that the statement we are trying to prove is false and deduce a contradiction from this assumption.

For any mathematical statement p , either p is true or $\neg p$ is true, but not both. If we wish to prove that a proposition p is true, but we do not know how to start a direct proof, we may consider using a proof by contradiction which would start by assuming that $\neg p$ is true. One should state this explicitly by writing *We use proof by contradiction*.

Assume that $\neg p$ is true.. The proof would then proceed by working towards a false statement or contradiction. Once that is found, the proof may be concluded by writing *We have found a contradiction. Therefore, our assumption that $\neg p$ is true, must have been incorrect. Hence, p must be true.*

Example A.6.7. The number $\sqrt{3}$ is irrational¹⁵.

Proof: Assume that $\sqrt{3}$ is not irrational, hence rational. This means that $\sqrt{3} = a/b$, for some natural numbers a and b . From $\sqrt{3} = a/b$, a natural step is to square both sides and get that $3 = (a/b)^2 = a^2/b^2$. Therefore, $a^2 = 3b^2$.

From here, there are different ways to get a contradiction. In the factorization of a^2 into primes, 3 will have an even exponent (0 if 3 does not divide a and positive, otherwise). In the factorization of $3b^2$ into primes, 3 has an odd exponent (1 if 3 does not divide b and at least 3, otherwise). Considering that every natural number has a unique factorization as a product of primes, this gives a contradiction as a number cannot be even and odd at the same time. Hence, $\sqrt{3}$ is irrational.

A similar way to get a contradiction goes as follows. We may assume that $\gcd(a, b) = 1$. Indeed, if $d = \gcd(a, b)$, then taking $a' = a/d \in \mathbb{N}$ and $b' = b/d \in \mathbb{N}$, we have that $\sqrt{3} = a/b = a'/b'$ and $\gcd(a', b') = 1$ by the previous example. Hence, $a^2 = 3b^2$ and $\gcd(a, b) = 1$. Thus, $3 \mid a^2$. Therefore, $3 \mid a$ meaning that $a = 3k$ for some natural number k . From $a^2 = 3b^2$, we deduce that $(3k)^2 = 3b^2$ which gives that $3k^2 = b^2$. Thus, $3 \mid b^2$ implying that $3 \mid b$. Hence, $3 \mid a$ and $3 \mid b$, contradiction with $\gcd(a, b) = 1$. Therefore, $\sqrt{3}$ must be irrational. ■

Example A.6.8. There are infinitely many prime numbers.

Proof: We use proof by contradiction and assume that there are finitely many primes. We can list them as p_1, \dots, p_k for some natural number k . Consider the number $p_1 \cdots p_k + 1$. Because $p_1, \dots, p_k \geq 2$, $p_1 \cdots p_k + 1$ is strictly greater than any of the numbers p_1, \dots, p_k . There exists a prime number that divides $p_1 \cdots p_k + 1$ and since $\{p_1, \dots, p_k\}$ is the list of all primes, such a prime number must come from this list. Assume that $p_j \mid p_1 \cdots p_k + 1$ for some $j \in \{1, \dots, k\}$. Because $p_j \mid p_1 \cdots p_k$, we deduce that $p_j \mid p_1 \cdots p_k + 1 - p_1 \cdots p_k = 1$. This is impossible as p_j is a prime. Therefore, our initial assumption that there are finitely many primes must have been false. Hence, there are infinitely many primes. ■

Similarly, if we wish to prove that $p \Rightarrow q$ is true, and we do not see a way of proceeding directly from the hypothesis p to reach the conclusion q , we can start by assuming that the implication $p \Rightarrow q$ is actually false, and work to find a contradiction. This contradiction means that our assumption was incorrect and therefore, $p \Rightarrow q$ must be true. The logical foundation of this method is that the implication $p \Rightarrow q$ is equivalent with $(\neg p) \vee q$. This can be proved using truth tables (see Exercise A.1.2). Therefore, the negation $\neg(p \Rightarrow q)$ is equivalent with $\neg(p \vee (\neg q))$ which from De Morgan's laws, is the same as $p \wedge (\neg q)$. Hence, whenever we start a proof by contradiction by assuming that $\neg(p \Rightarrow q)$ is true, we begin by assuming that $p \wedge (\neg q)$ is true.

¹⁵We give other proofs of this statement in Section 3.2.

Example A.6.9. Let a and b be two natural numbers. If $d = \gcd(a, b)$, then prove that $\gcd(a/d, b/d) = 1$.

Proof: Our hypothesis is $d = \gcd(a, b)$ and our conclusion is $\gcd(a/d, b/d) = 1$. We start our proof by assuming the hypothesis and by assuming that the conclusion is false, namely suppose that $d = \gcd(a, b)$ and $\gcd(a/d, b/d) \neq 1$. This implies that there is a natural number $k \geq 2$ such that $k \mid a/d$ and $k \mid b/d$. Therefore, $dk \mid a$ and $dk \mid b$. Because $k > 1$, $dk > d$. Hence, dk is a common divisor of a and b that is strictly greater than $d = \gcd(a, b)$. This contradicts the definition of $d = \gcd(a, b)$ as the largest common divisor of a and b . Therefore, our assumption that $\gcd(a/d, b/d) \neq 1$ must have been false. Hence, $\gcd(a/d, b/d) = 1$. ■

A.6.4. Proof by contrapositive. This method of proof is used when there is no clear way on how to use the hypothesis. Its logical foundation is that $p \Rightarrow q$ is equivalent to $(\neg q) \Rightarrow (\neg p)$.

Some of our examples in this subsection can also be proved using proof by contradiction.

Example A.6.10. Let a be a natural number. If a^2 is odd, then a is odd.

Proof: Our hypothesis is that a^2 is odd. That means that $a^2 = 2k + 1$ for some natural number k . Trying to get a from the previous equation leaves us with the equation $a = \sqrt{2k + 1}$. It seems difficult to get our proof wagon out of this ditch. An alternative to the direct proof attempt above starts by considering the contrapositive statement:

If a is not odd, then a^2 is not odd,

which is the same as

If a is even, then a^2 is even.

This statement seems more approachable. We start with a being even which means that $a = 2\ell$ for some natural number ℓ . Since our conclusion involves a^2 , it makes sense to square the both sides of the previous equation and get that $a^2 = (2\ell)^2 = 4\ell^2$. Since $4\ell^2 = 2(2\ell^2)$ and $2\ell^2 \in \mathbb{N}$, the number $a^2 = 4\ell^2 = 2(2\ell^2)$ must be even. This finishes our proof. ■

Example A.6.11. If $2^n - 1$ is a prime, then n is a prime.

Proof: Our hypothesis is $2^n - 1$ being a prime. We leave to the reader to ponder what possible useful information can be deduced from this fact. The contrapositive of the statement above is the following:

If n is not a prime, then $2^n - 1$ is not a prime.

Our hypothesis now is that n is not a prime. This means $n = 1$ or $n \geq 2$ and $n = ab$, for some natural number a and b with $1 < a, b < n$. If $n = 1$, then $2^n - 1 = 1$ which is not a prime. If $n \geq 2$ and $n = ab$ with $1 < a, b < n$, then

$$2^n - 1 = 2^{ab} - 1 = (2^b)^a - 1 = (2^b - 1)[(2^b)^{a-1} + \cdots + (2^b)^1 + 1].$$

Because $a > 1$, $2^b - 1 < 2^{ab} - 1$. Also, since $b > 1$, $2^b - 1 > 1$. Hence, $2^n - 1$ has a divisor that is not equal to 1 nor to itself. Therefore, $2^n - 1$ is not prime. ■

In many situations, one needs to combine various proof methods. The following example starts with proof by contrapositive and then uses proof by contradiction.

Example A.6.12. Let $n \geq 2$ be a natural number. If $(n - 1)! \equiv -1 \pmod{n}$, then n is prime.

Proof: Our hypothesis is equivalent to $n \mid (n - 1)! + 1$. Our conclusion is that n is prime. The contrapositive of the statement above is:

If $n \geq 2$ is not a prime, then n does not divide $(n - 1)! + 1$.

The hypothesis for the contrapositive statement is that $n \geq 2$ is not a prime. Thus, $n \geq 4$ and there exists a prime number k such that $k \mid n$ and $2 \leq k \leq n - 1$.

We will prove that n does not divide $(n - 1)! + 1$ using proof by contradiction. Assume that n divides $(n - 1)! + 1$. Because $k \mid n$, $k \mid (n - 1)! + 1$. On the other hand, $2 \leq k \leq n - 1$ implies that k is a factor in the product $(n - 1)! = 1 \cdot 2 \cdot \dots \cdot (n - 1)$ and $k \mid (n - 1)!$. Therefore, $k \mid (n - 1)! + 1 - (n - 1)! = 1$ meaning that $k = 1$. This is a contradiction with our assumption that k is prime. ■

A.6.5. Proof by induction. The basis of this method is the **principle of induction** (see Axiom 1.2.1 or property (4) of Axiom A.7.1).

Axiom A.6.1. *The natural numbers \mathbb{N} possess the following property:*

$$(\forall A \subseteq \mathbb{N})[(1 \in A) \text{ and } (n \in A \Rightarrow n + 1 \in A)] \Rightarrow A = \mathbb{N}$$

We have met this method already in Section 1.2 where the most general result is below (see Proposition 1.2.3).

Proposition A.6.1 (Principle of Mathematical Induction). *Let $n_0 \in \mathbb{N}_0$. For each natural number $n \geq n_0$ let $P(n)$ be a proposition that depends on n . Suppose that*

- (1) *The proposition $P(n_0)$ is true.*
- (2) *If $P(n)$ is true for some $n \geq n_0$, then $P(n + 1)$ is true.*

Then $P(n)$ is true for all $n \geq n_0$.

Proof: Let $A = \{n \in \mathbb{N} : P(n) \text{ is true}\}$. By assumption, $n_0 \in A$, and, moreover, if $n \in A$, then $n + 1 \in A$. Denote by B the set of natural numbers that are strictly less than n_0 . If $n_0 = 1$, $B = \emptyset$. The principle of induction lets us conclude that $A \cup B = \mathbb{N}$, meaning that $P(n)$ is true for all $n \geq n_0$. ■

The first item above “ $P(n_0)$ true” is usually called **the base case** of the induction proof. The second item “ $P(n)$ implies $P(n + 1)$ ” or $P(n) \Rightarrow P(n + 1)$ is often referred to as **the induction step** and in its proof, $P(n)$ is called **the induction hypothesis**.

Example A.6.13. Prove that

$$(A.6.2) \quad \frac{1}{1 \cdot 2} + \frac{1}{2 \cdot 3} + \dots + \frac{1}{n(n+1)} = \frac{n}{n+1}.$$

Proof: For $n \in \mathbb{N}$, let $P(n)$ be the statement in the above equation. Thus, $P(1)$ means that $\frac{1}{1 \cdot 2} = \frac{1}{1+1}$, $P(2)$ is that $\frac{1}{1 \cdot 2} + \frac{1}{2 \cdot 3} = \frac{2}{2+1}$ and so on.

For this problem $n_0 = 1$. The base case $P(1)$ is true as it can be observed from above as $\frac{1}{1 \cdot 2} = \frac{1}{2}$. For the induction step, let $n \geq 1$ be a natural number such that $P(n)$ is true. This means that

$$(A.6.3) \quad \frac{1}{1 \cdot 2} + \frac{1}{2 \cdot 3} + \dots + \frac{1}{n(n+1)} = \frac{n}{n+1}.$$

We wish to prove that $P(n+1)$ is true which is the same as

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

We start with the statement that we know it is true, namely $P(n)$ above, and look for any connections between it and the result that we want to prove which is $P(n+1)$. The left-hand sides of these equations looks quite similar, and it makes sense to add $\frac{1}{(n+1)(n+2)}$ to both sides of the equation (A.6.3). Therefore,

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

Now we just have to manipulate the right-hand side above:

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

Thus, we have proved that the statement $P(n+1)$ is true. This finishes our proof. ■

Remark A.6.1. There exists an easier and more direct way to prove (A.6.2). Namely, use

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

The right-hand side is an example of a telescoping sum¹⁶. Hence, we get

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

Next we give an example where the telescoping sum trick may not be applicable, but the induction method can be used.

Example A.6.14. Show that for all $n \geq 1$ it follows that

$$(A.6.4) \quad 1^3 + 2^3 + \dots + n^3 = (1 + 2 + \dots + n)^2.$$

Before proving (A.6.5), let us rewrite it. Applying Proposition 1.2.2, it follows that (A.6.4) is equivalent to

$$(A.6.5) \quad 1^3 + 2^3 + \dots + n^3 = \left(\frac{n(n+1)}{2} \right)^2 = \frac{n^2(n+1)^2}{4}.$$

We will prove (A.6.5) by induction.

¹⁶A telescoping sum is one which may be written as

$$\sum_{k=1}^n [a_k - a_{k+1}] = a_1 - a_{n+1}.$$

The base case with $n = 1$ is true as $1^3 = 1^2$. For the induction step suppose that (A.6.5) is true for some fixed $n \geq 1$. Our objective is to verify its validity for $n + 1$. That is, we have to show that

$$(A.6.6) \quad 1^3 + 2^3 + \dots + n^3 + (n+1)^3 = \frac{(n+1)^2(n+2)^2}{4}.$$

To this end we use that (A.6.5) is true for n . Hence, adding $(n+1)^3$ to both side of (A.6.5), we deduce that

$$1^3 + 2^3 + \dots + n^3 + (n+1)^3 = \frac{n^2(n+1)^2}{4} + (n+1)^3.$$

Now we have to do some calculations and simplify the right-hand side above as follows:

$$\begin{aligned} \frac{n^2(n+1)^2}{4} + (n+1)^3 &= \frac{n^2(n+1)^2 + 4(n+1)^3}{4} \\ &= \frac{(n+1)^2(n^2 + 4n + 4)}{4} = \frac{(n+1)^2(n+2)^2}{4}. \end{aligned}$$

Hence, the equation (A.6.6) is true. This shows that (A.6.4) is valid for all $n \geq 1$.

The method of induction can be used to prove other types of mathematical statements such as inequalities or divisibility results.

Example A.6.15. Show that for any natural number n ,

$$1 + \frac{1}{\sqrt{2}} + \dots + \frac{1}{\sqrt{n}} > 2\sqrt{n+1} - 2.$$

Proof: We use induction on n . The base case corresponds to $n = 1$ and means that $1 > 2\sqrt{1+1} - 2$ which is the same as $3 > 2\sqrt{2}$. This inequality is true because $9 > 8$.

For the induction step, let $n \geq 1$ be a natural number and assume that the inequality above is true for n , meaning that

$$(A.6.7) \quad 1 + \frac{1}{\sqrt{2}} + \dots + \frac{1}{\sqrt{n}} > 2\sqrt{n+1} - 2.$$

Our goal is to prove that (A.6.7) is true for $n + 1$, namely that

$$(A.6.8) \quad 1 + \frac{1}{\sqrt{2}} + \dots + \frac{1}{\sqrt{n}} + \frac{1}{\sqrt{n+1}} > 2\sqrt{n+2} - 2.$$

Noting the close resemblance of the left-hand sides of these inequalities, we add $\frac{1}{\sqrt{n+1}}$ to both sides of the inequality (A.6.7) and get that

$$1 + \frac{1}{\sqrt{2}} + \dots + \frac{1}{\sqrt{n}} + \frac{1}{\sqrt{n+1}} > 2\sqrt{n+1} - 2 + \frac{1}{\sqrt{n+1}}.$$

The inequality (A.6.8) would follow by transitivity if we could prove that

$$2\sqrt{n+1} - 2 + \frac{1}{\sqrt{n+1}} > 2\sqrt{n+2} - 2.$$

This inequality is the same as

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

This inequality is true as $\sqrt{n+2} > \sqrt{n+1}$. Hence, inequality (A.6.8) is true and our proof is finished. ■

Example A.6.16. Prove that for any natural number n , 8 divides $5^{n+1} - 6 \cdot 3^n + 1$.

Proof: Denote by $P(n)$ the statement above. We want to show that $P(n)$ is true for any $n \in \mathbb{N}$ using induction on n . The base case is $n = 1$ and means that $8 \mid 5^{1+1} - 6 \cdot 3 + 1 = 25 - 18 + 1 = 8$ which is true.

For the induction step, let $n \in \mathbb{N}$ be such that $8 \mid 5^{n+1} - 6 \cdot 3^n + 1$. We want to prove that $8 \mid 5^{n+2} - 6 \cdot 3^{n+1} + 1$.

One approach is to try to somehow relate the two numbers $5^{n+1} - 6 \cdot 3^n + 1$ (which we know is divisible by 8) and $5^{n+2} - 6 \cdot 3^{n+1} + 1$ (which we want to prove is divisible by 8). Perhaps the simplest way to go about this is to subtract one number from the other and see if the result is divisible by 8

$$\begin{aligned}(5^{n+2} - 6 \cdot 3^{n+1} + 1) - (5^{n+1} - 6 \cdot 3^n + 1) &= (5^{n+2} - 5^{n+1}) - 6(3^{n+1} - 3^n) \\&= 5^{n+1}(5 - 1) - 6 \cdot 3^n(3 - 1) \\&= 4 \cdot 5^{n+1} - 12 \cdot 3^n \\&= 4(5^{n+1} - 3^{n+1}).\end{aligned}$$

Because 5 and 3 are odd, 5^{n+1} and 3^{n+1} are both odd and therefore, 2 divides $5^{n+1} - 3^{n+1}$. Hence, 8 divides $4(5^{n+1} - 3^{n+1})$. Since

$$(5^{n+2} - 6 \cdot 3^{n+1} + 1) = (5^{n+1} - 6 \cdot 3^n + 1) + 4(5^{n+1} - 3^{n+1}),$$

we observe that both terms on the right-hand side are divisible by 8. Consequently, we deduce that $8 \mid 5^{n+2} - 6 \cdot 3^{n+1} + 1$ which finishes our proof. ■

Exercise A.6.1. Prove that an integer is odd if and only if its cube is odd.

Exercise A.6.2. Let $x > 0$ such that $x^2 \neq 2$. If $y = \frac{1}{2}(x + \frac{2}{x})$, then prove that $y^2 > 2$.

Exercise A.6.3. Let n be a natural number. Prove that

$$\left\lfloor \frac{n}{2} \right\rfloor^2 + \left\lfloor \frac{n+1}{2} \right\rfloor^2 = \left\lfloor \frac{n^2+1}{2} \right\rfloor.$$

Exercise A.6.4. Let x and y be two real numbers. Suppose we know that $x = y$. Then one may argue as follows:

$$x = y \Rightarrow x^2 = xy \Rightarrow x^2 - y^2 = xy - y^2 \Rightarrow (x-y)(x+y) = y(x-y).$$

Next we divide both sides of the last equation by $x - y$ and obtain

$$x + y = y \Rightarrow 2y = y \Rightarrow 2 = 1.$$

What is wrong with this proof¹⁷?

Exercise A.6.5. Suppose we want to show that $1 = 3$. To this end we argue as follows:

$$1 = 3 \Rightarrow 1 - 2 = 3 - 2 \Rightarrow -1 = 1 \Rightarrow (-1)^2 = 1^2 \Rightarrow 1 = 1.$$

As the statement $1 = 1$ is surely true, $1 = 3$ has to be valid as well. What error did we make?

¹⁷Taken from [32], Problem 2 in Chapter I.

Exercise A.6.6. Show that for any natural number n ,

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

Exercise A.6.7. Show that for any natural number n ,

$$1 + \frac{1}{\sqrt{2}} + \dots + \frac{1}{\sqrt{n}} < 2\sqrt{n}.$$

Exercise A.6.8. For which integers $n \geq 1$ do we have that

$$3^n > 2^{n+4} ?$$

Use induction to prove your answer.

Exercise A.6.9. Let x, y, z , and w be complex numbers such that $x + y = z + w$ and $x^2 + y^2 = z^2 + w^2$. Prove that $x^n + y^n = z^n + w^n$ for any natural number n .

Exercise A.6.10. Let a, b, c be positive real numbers.

- (1) Prove that $\frac{a+b}{c} + \frac{b+c}{a} + \frac{c+a}{b} \geq 6$ and that equality happens if and only if $a = b = c$.
- (2) Prove that $\frac{a}{b+c} + \frac{b}{c+a} + \frac{c}{a+b} \geq \frac{3}{2}$ with equality if and only if $a = b = c$.

A.7. Peano's Axioms and the Construction of Integers

In Sections 1.1 and 2.1 we introduced natural numbers and integers more or less heuristically. The sets \mathbb{N} and \mathbb{Z} were defined as collections of certain objects, called numbers, which can be added and multiplied, where these operations fulfill certain rules, and where their elements can be ordered in a way compatible with addition and multiplication. This approach suffices in most cases, yet does not meet the demands for building a rigorous mathematical theory upon them. Therefore, we decided to include a short section about the precise introduction of \mathbb{N} and \mathbb{Z} for those who want to know more. Since some proofs are quite technical, we will not give all details, restrict ourselves to the main ingredients of the constructions. In this setting, natural numbers are introduced by Peano's axioms¹⁸ while integers are described by equivalence classes of pairs of natural numbers.

A.7.1. Natural Numbers by Peano's Axioms. Let us start with stating the axioms which define the set \mathbb{N} of natural numbers.

Axiom A.7.1 (Peano's Axioms). *The set \mathbb{N} is given by the following properties:*

- (1) *There exists a distinguished element in \mathbb{N} which we denote by “1”.*
- (2) *There exists an injective mapping $\sigma : \mathbb{N} \rightarrow \mathbb{N}$.*
- (3) *There is no $n \in \mathbb{N}$ with $\sigma(n) = 1$. In particular, σ is not surjective.*
- (4) *The following induction principle is valid: If a subset $S \subseteq \mathbb{N}$ possesses the two properties $1 \in S$ and whenever $n \in S$, then also $\sigma(n) \in S$, then necessarily $S = \mathbb{N}$.*

¹⁸Giuseppe Peano (1858–1931) was an Italian mathematician who formulated these axioms in 1889.

Remark A.7.1. Which role does the function σ play? The Greek letter “ σ ” stands for the Latin “*s*” as **successor**. Thus, having always in mind that $\sigma(n)$ denotes the successor of n (later on denoted by $n + 1$), the properties stated in the axioms become quite natural and may be rephrased as follows:

- (1) 1 is a natural number.
- (2) Each natural number n has a unique successor $\sigma(n)$.
- (3) 1 is not a successor of any $n \in \mathbb{N}$.
- (4) If a set $S \subseteq \mathbb{N}$ has the properties $1 \in S$ and if with every $n \in S$ also its successor belongs to S , then this implies $S = \mathbb{N}$.

The next lemma shows that each $m \in \mathbb{N}$ different of 1 is successor of a unique element in \mathbb{N} .

Lemma A.7.1. *For each $m \in \mathbb{N}$, $m \neq 1$, there exists a unique $n \in \mathbb{N}$ with $\sigma(n) = m$.*

Proof: Define a subset $S \subseteq \mathbb{N}$ as follows:

$$S := \{m \in \mathbb{N} : m = 1 \text{ or } m = \sigma(n) \text{ for some } n \in \mathbb{N}\}.$$

Of course, $1 \in S$ by definition. Suppose now $m \in S$. We want to show that then $\sigma(m) \in S$ as well. But this is true by trivial reason. Note that S contains 1 and elements in the range of σ . Thus, for any $m \in S$ it follows $\sigma(m) \in S$ by the definition of S . The induction principle implies now $S = \mathbb{N}$. Consequently, any $m \neq 1$ admits a representation $m = \sigma(n)$ for a suitable $n \in \mathbb{N}$. Moreover, since σ is supposed to be injective, there is one and only one n with $\sigma(n) = m$. ■

Remark A.7.2. Another way to formulate Lemma A.7.1 is as follows: The range of σ is $\mathbb{N} \setminus \{1\}$, and since σ is assumed to be injective, it is even a bijection between \mathbb{N} and $\mathbb{N} \setminus \{1\}$. The number $n \in \mathbb{N}$ with $\sigma(n) = m$ is the **predecessor** of m (keep in mind that this is $m - 1$ in the common notation). Equivalently, it holds $n = \sigma^{-1}(m)$. Thus, Lemma A.7.1 asserts that each element different of 1 has a unique predecessor.

How to define the sum of two elements in \mathbb{N} . The approach is very natural. First one defines $n + 1$ by setting

$$(A.7.1) \quad n + 1 := \sigma(n).$$

If one wants to define $n + m$, this can be achieved by applying m times σ to n . That is

$$n + m = \underbrace{(\sigma \circ \cdots \circ \sigma)(n)}_{m \text{ times}}.$$

An easier way two formulate this is recursively as follows:

$$(A.7.2) \quad n + \sigma(m) := \sigma(n + m), \quad n, m \in \mathbb{N}.$$

By Lemma A.7.1 this is well-defined because every number different of 1 may be written as $\sigma(m)$ for a suitable $m \in \mathbb{N}$. Maybe the procedure becomes more understandable if one replaces $\sigma(m)$ by $m + 1$ and gets

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

So, if one knows the value of $n + m$, this defines the sum of $n + (m + 1)$. Consequently, starting with n one gets $n + 1$ as successor of n , then $n + 1 + 1$ as successor of $n + 1$ and so on. And applying this procedure m times one finally arrives at $n + m$.

After knowing how two elements in \mathbb{N} are added, one may introduce an order on \mathbb{N} in the following way:

$$(A.7.3) \quad (n < m) \Leftrightarrow (\exists k \in \mathbb{N} \text{ such that } m = k + n).$$

In particular, this implies

$$1 < 1 + 1 < 1 + 1 + 1 < \dots.$$

How about the multiplication of two elements in \mathbb{N} ? The easiest case is multiplication by 1. Here we set

$$(\forall n \in \mathbb{N})(n \cdot 1 := n).$$

The definition of $n \cdot m$ for arbitrary $n, m \in \mathbb{N}$ is again done recursively. Suppose we already know the value of $n \cdot m$. How do we obtain $n \cdot (m + 1) = n \cdot \sigma(m)$? Of course, it should be $n \cdot m + n$. Therefore, the multiplication may be defined recursively as follows:

$$n \cdot \sigma(m) := n \cdot m + n, \quad n, m \in \mathbb{N}.$$

So, for example, by this rule

$$n \cdot (1+1) = n \cdot \sigma(1) = n \cdot 1 + n = n + n \quad \text{and} \quad n \cdot \sigma(1+1) = n \cdot (1+1) + n = n = n + n + n.$$

After introducing addition, multiplication and an ordering on \mathbb{N} , several issues have to be shown.

- (1) Addition and multiplication on \mathbb{N} are associative operations.
- (2) These operations are also commutative.
- (3) The cancellation laws are valid for addition and multiplication. That is, given n, m and ℓ in \mathbb{N} , then

$$n + m = n + \ell \Rightarrow m = \ell \quad \text{and} \quad n \cdot m = n \cdot \ell \Rightarrow m = \ell.$$

- (4) Addition and multiplication satisfy the distributive law. Thus, given n, m and ℓ in \mathbb{N} , then

$$n \cdot (m + \ell) = n \cdot m + n \cdot \ell.$$

- (5) The binary relation " \leq " is an order on \mathbb{N} . Here $n \leq m$ if either $n < m$ or $n = m$.
- (6) The order " \leq " is compatible with addition and multiplication. In other words, given n, m and ℓ in \mathbb{N} , then

$$(m \leq \ell) \Rightarrow (n + m \leq n + \ell \quad \text{and} \quad n \cdot m \leq n \cdot \ell).$$

- (7) \mathbb{N} is well-ordered. That says, that every nonempty subset of \mathbb{N} possesses a least element with respect to the order defined in (A.7.3). As we saw in Proposition 1.4.3, this property is equivalent to the validity of induction principle.

It is quite clear that the verification of all these properties is very technical and cumbersome. We avoid presenting it here. But in order to get an impression how the proofs are done, let us verify two of the properties stated above.

Proposition A.7.2. *The addition on \mathbb{N} defined by (A.7.1) and by (A.7.2) is associative.*

Proof: Our goal is to show

$$(n + m) + \ell = n + (m + \ell)$$

for all $n, m, \ell \in \mathbb{N}$. To this end fix $n, m \in \mathbb{N}$ and define a subset $S \subseteq \mathbb{N}$ by

$$S := \{\ell \in \mathbb{N} : (n + m) + \ell = n + (m + \ell)\}.$$

If we are able to show $S = \mathbb{N}$, then we are done. Note that n and m are fixed, but arbitrary elements in \mathbb{N} . To verify $S = \mathbb{N}$ we are going to use the induction principle. Hence, we have to verify two facts:

- (1) $1 \in S$.
- (2) If $\ell \in S$, then also $\sigma(\ell) \in S$.

The first assertion follows by using three times the definition of addition in the following way:

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

Consequently, $1 \in S$.

To verify property (2) choose an $\ell \in S$. That is, we already know that

$$(A.7.4) \quad (n + m) + \ell = n + (m + \ell).$$

Our aim is to show that this remains valid with ℓ replaced by $\sigma(\ell)$. Using three times the definition of addition as well as (A.7.4) the assertion follows by

$$\begin{aligned} n + (m + \sigma(\ell)) &= n + \sigma(m + \ell) = \sigma(n + (m + \ell)) = \sigma((n + m) + \ell) \\ &= (n + m) + \sigma(\ell). \end{aligned}$$

Thus, $\ell \in S$ leads to $\sigma(\ell) \in S$, and by (4) in Peano's axioms it follows that $S = \mathbb{N}$. As already mentioned above, this completes the proof. \blacksquare

Proposition A.7.3. *The addition on \mathbb{N} defined by (A.7.1) and (A.7.2) is commutative.*

Proof: Define $S \subseteq \mathbb{N}$ by

$$(A.7.5) \quad S := \{m \in \mathbb{N} : \forall n \in \mathbb{N}, n + m = m + n\}.$$

The strategy for the proof is as before. We want to show that $1 \in S$ and that $m \in S$ implies $\sigma(m) \in S$. This lets us conclude that $S = \mathbb{N}$ and completes the proof.

Step 1: Our aim is to show that $1 \in S$. To this end set

$$S_1 := \{n \in \mathbb{N} : 1 + n = n + 1\}.$$

By $1 + 1 = 1 + 1$ it follows $1 \in S_1$. Now suppose $n \in S_1$, i.e., we have

$$1 + n = n + 1.$$

This and the definition of addition imply

$$1 + \sigma(n) = \sigma(1 + n) = \sigma(n + 1) = \sigma(\sigma(n)) = \sigma(n) + 1.$$

Hence, $n \in S_1$ yields $\sigma(n) \in S_1$, which proves $S_1 = \mathbb{N}$. Consequently, we know that all $n \in \mathbb{N}$ satisfy

$$1 + n = n + 1.$$

Step 2: Let us come back to the set S defined in (A.7.5). By step 1 we know that $1 \in S$. So, suppose now $m \in S$ for some fixed $m \in \mathbb{N}$. Then this $m \in \mathbb{N}$ satisfies

$$(A.7.6) \quad m + n = m + n \quad \text{for all } n \in \mathbb{N}.$$

Again we want to show that this is so also for $\sigma(m)$. Using twice $1 \in S$, twice the associativity of the addition, twice the definition of addition together with (A.7.6) one obtains

$$\begin{aligned} n + \sigma(m) &= n + (m + 1) = (n + m) + 1 = (m + n) + 1 = 1 + (m + n) \\ &= (1 + m) + n = (m + 1) + n = \sigma(m) + n. \end{aligned}$$

This being true for all $n \in \mathbb{N}$ implies $\sigma(m) \in S$ provided that $m \in S$. The induction principle lets us conclude that $S = \mathbb{N}$ which completes the proof by the definition of the set S in (A.7.5). ■

Two basic questions about Peano's axioms arise immediately:

- (1) Does there exist a mathematical model for \mathbb{N} ? In other words, may we construct a set \mathbb{N} with distinguished element 1 and with a mapping $\sigma : \mathbb{N} \rightarrow \mathbb{N}$ satisfying the properties stated in Peano's axioms?
- (2) What can be said about the uniqueness of \mathbb{N} and σ satisfying the axioms? In other words, do there exist different models for the natural numbers?

An affirmative answer to the first question gave John von Neumann (1903–1957) in [22]. He constructed a model of \mathbb{N}_0 based on the so-called Zermelo-Fraenkel axioms of set theory. Let us give an idea¹⁹ how this construction looks like.

A collection \mathcal{S} of sets is said to be a **successor set** provided that $\emptyset \in \mathcal{S}$ and, moreover, whenever $A \in \mathcal{S}$, then also $A \cup \{A\} \in \mathcal{S}$. For example, since $\emptyset \in \mathcal{S}$, this is also so for $\emptyset \cup \{\emptyset\} = \{\emptyset\}$ and for $\{\emptyset, \{\emptyset\}\}$. An axiom (called the axiom of infinity) of set theory ensures the existence of infinite successor sets.

It is easy to see that the intersection of successor sets is a successor set as well. Hence, there exists a smallest successor set \mathcal{S}_0 , taken as model of \mathbb{N}_0 . The distinguished element is \emptyset , denoted by 0, and the mapping σ on \mathcal{S}_0 is defined by

$$\sigma(A) = A \cup \{A\}, \quad A \in \mathcal{S}_0.$$

In this setting it follows that

$$\begin{aligned} 0 &= \emptyset \\ 1 &= \sigma(0) = \emptyset \cup \{\emptyset\} = \{\emptyset\} \\ 2 &= \sigma(1) = 1 \cup \{1\} = \{\emptyset, \{\emptyset\}\} \\ 3 &= \sigma(2) = 2 \cup \{2\} = \{\emptyset, \{\emptyset\}, \{\emptyset, \{\emptyset\}\}\}, \end{aligned}$$

and so on. The model for \mathbb{N} is then $\mathcal{S}_0 \setminus \{\emptyset\}$. In this way each number $n \in \mathbb{N}$ is associated with a set in \mathcal{S}_0 which contains exactly n elements. It is a nice exercise to determine the sets in \mathcal{S}_0 which correspond to the numbers 4 and 5.

¹⁹More interesting details about this topic can be found in the interesting recent note Ueberberg, Johannes: Successor Sets and the Axiom of Peano. Bonn: MATH-Garden. Version 1.0.3. url: <https://www.math-garden.com/unit/nst-successor-sets> (2022).

Note that σ maps $\mathcal{S}_0 \setminus \{\emptyset\}$ to $\mathcal{S}_0 \setminus \{\emptyset\}$. It remains to show that $1 = \{\emptyset\}$ and the restriction σ to $\mathcal{S}_0 \setminus \{\emptyset\}$ fulfill the four properties stated in Axiom A.7.1.

While the first three properties are an immediate consequence of the construction of σ , the induction principle follows by the choice of \mathcal{S}_0 as minimal successor set. Indeed, whenever a collection \mathcal{S} of sets in \mathcal{S}_0 satisfies $\emptyset \in \mathcal{S}$ and $(A \in \mathcal{S}) \Rightarrow (\sigma(A) \in \mathcal{S})$, then it is a successor set, hence cannot be smaller than the minimal one \mathcal{S}_0 which implies $\mathcal{S} = \mathcal{S}_0$. Of course, then the induction principle is also valid if one starts the induction at 1 and not at 0.

Let us formulate the answer to the second question as a proposition.

Proposition A.7.4. *Suppose that there are two sets \mathbb{N}_1 and \mathbb{N}_2 together with distinguished elements 1_1 and 1_2 and mappings σ_1 and σ_2 , both satisfying Peano's axioms. Then there is a bijection $f : \mathbb{N}_1 \rightarrow \mathbb{N}_2$ such that*

$$\begin{aligned} f(1_1) &= 1_2 \quad \text{and} \\ f(\sigma_1(n)) &= \sigma_2(f(n)), \quad n \in \mathbb{N}_1. \end{aligned}$$

Thus, up to a bijection, there is one and only one model $(\mathbb{N}, \sigma, 1)$ for which Peano's axioms are valid.

A.7.2. Integers as Equivalence Classes. Our next objective is to introduce integers in a mathematical exact way. After we described the construction of \mathbb{N} and after we investigated general equivalence relations, now we are in the position to precise the construction of negative integers as indicated in Section 2.1. Say we know \mathbb{N} and our aim is to add some new numbers, the negative ones. How can we do this? For example, if we want to define the number -1 , one possibility is to introduce it as the object (number) which, when added to 2, gives 1. Thus, we could characterize -1 by the pair $(1, 2)$. But in the same way we can describe -1 by the pair $(3, 4)$ because adding -1 to 4 leads to 3. And so on. Therefore, we have a whole collection of pairs of natural numbers all characterizing -1 . The basic idea is to put all these pairs of natural numbers together into a single equivalence class. In the same way we can do this construction for $-2, -3$ and so on. But thereby we also have to rediscover \mathbb{N} in this construction. Because 1 added to 2 gives 3, this is easily done by assigning the number 1 to the pair $(3, 2)$.

The corresponding relation \sim on $\mathbb{N} \times \mathbb{N}$ is given by

$$(A.7.7) \quad (n_1, m_1) \sim (n_2, m_2) \Leftrightarrow n_1 + m_2 = n_2 + m_1.$$

In this way we get the equivalence of the pairs $(2, 1) \sim (4, 3)$ or $(3, 6) \sim (5, 8)$.

Remark A.7.3. One may ask why we do not replace (A.7.7) by

$$(n_1, m_1) \sim (n_2, m_2) \Leftrightarrow n_1 - m_1 = n_2 - m_2.$$

The answer is very easy: Of course, we can do so if $m_1 < n_1$, hence also $m_2 < n_2$. But as soon as $m_1 \geq n_1$ the difference $n_1 - m_1$ does not make sense within the framework of natural numbers. But for better understanding of the following calculations we may always think of (n, m) as $n - m$.

Proposition A.7.5. *The relation \sim defined by (A.7.7) is an equivalence relation on $\mathbb{N} \times \mathbb{N}$.*

Proof: Of course, the relation \sim is reflexive and symmetric. Therefore, it remains to show that it is transitive. Choose three pairs (n_1, m_1) , (n_2, m_2) and (n_3, m_3) of natural numbers such that $(n_1, m_1) \sim (n_2, m_2)$ and $(n_2, m_2) \sim (n_3, m_3)$. By definition, this says that

$$n_1 + m_2 = n_2 + m_1 \quad \text{as well as} \quad n_2 + m_3 = n_3 + m_2.$$

Adding m_3 to the first equation as well as m_1 to the second leads to

$$n_1 + m_2 + m_3 = n_2 + m_1 + m_3 \quad \text{as well as} \quad n_2 + m_1 + m_3 = n_3 + m_1 + m_2.$$

The right-hand side of the left equation coincides with the left-hand side of the right one. Hence, we get

$$n_1 + m_2 + m_3 = n_3 + m_1 + m_2,$$

which implies $n_1 + m_3 = n_3 + m_1$ or, equivalently, $(n_1, m_1) \sim (n_3, m_3)$. This shows the transitivity of \sim and completes the proof. \blacksquare

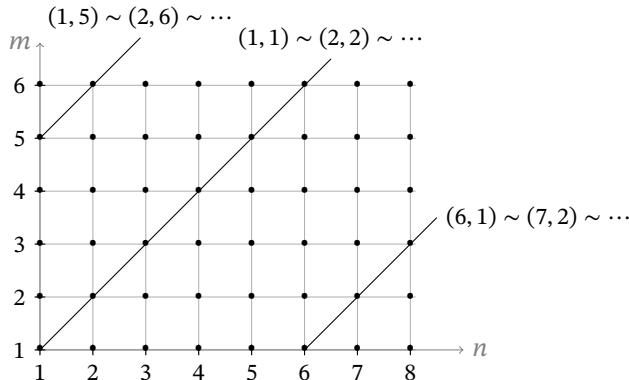


Figure A.7.1. The lattice points equivalent to $(1, 5)$, to $(1, 1)$ and to $(6, 1)$.

Definition A.7.1. Given two pairs (n_1, m_1) and (n_2, m_2) of integers, their sum is defined by

$$(n_1, m_1) + (n_2, m_2) := (n_1 + n_2, m_1 + m_2).$$

Proposition A.7.6. *The addition $+$ on $\mathbb{N} \times \mathbb{N}$ is compatible with the equivalence relation. That is, for all pairs (n_1, m_1) , (n'_1, m'_1) , (n_2, m_2) and (n'_2, m'_2) the following holds:*

$$\begin{aligned} (n_1, m_1) &\sim (n'_1, m'_1), \quad (n_2, m_2) \sim (n'_2, m'_2) \\ \Rightarrow \quad (n_1, m_1) + (n_2, m_2) &\sim (n'_1, m'_1) + (n'_2, m'_2). \end{aligned}$$

Proof: The proof is straightforward and therefore left as a nice exercise for the reader. See Figure A.7.2 for a graphic illustration of this result. \blacksquare

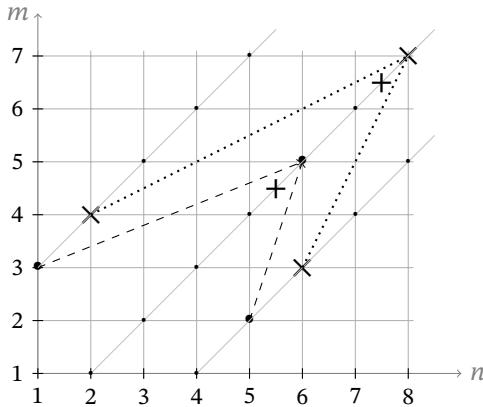


Figure A.7.2. $(5, 2) + (1, 3) \sim (6, 3) + (2, 4)$.

We saw in Proposition A.5.3 that every equivalence relation splits the underlying set into disjoint subsets, the so-called equivalence classes. The collection of all these equivalence classes forms the quotient set, also called quotient space. In view of Proposition A.7.5 this approach leads in this case to

$$\mathbb{N} \times \mathbb{N} / \sim = \{\widehat{(n, m)} : n, m \in \mathbb{N}\}.$$

Recall that the equivalence class $\widehat{(n, m)}$ consists of all pairs (k, ℓ) of natural numbers for which

$$(n, m) \sim (k, \ell) \iff n + \ell = m + k.$$

One may visualize $\mathbb{N} \times \mathbb{N} / \sim$ as a set of certain straight lines in $\mathbb{N} \times \mathbb{N}$ as drawn in Figure A.7.1.

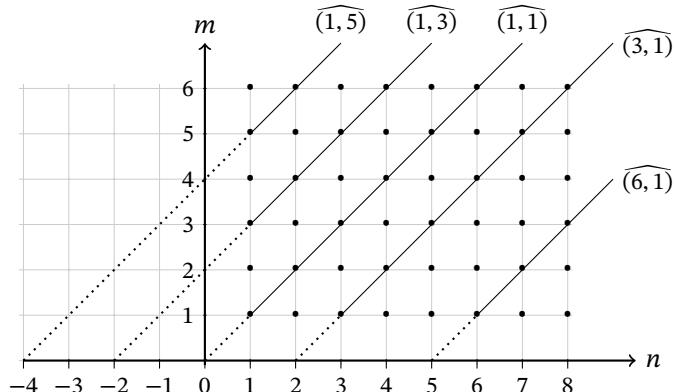


Figure A.7.3. $\widehat{(1, 5)} = -4$, $\widehat{(1, 3)} = -2$, $\widehat{(1, 1)} = 0$, $\widehat{(3, 1)} = 2$, $\widehat{(6, 1)} = 5$.

In view of Proposition A.7.6 we may now define an addition on the quotient set $\mathbb{N} \times \mathbb{N}/\sim$, the set of equivalence classes on $\mathbb{N} \times \mathbb{N}$, by

$$\widehat{(n_1, m_1)} + \widehat{(n_2, m_2)} := \widehat{(n_3, m_3)} \quad \text{where } (n_3, m_3) = (n_1, m_1) + (n_2, m_2).$$

Our next aim is to identify the quotient set $\mathbb{N} \times \mathbb{N}/\sim$ with the entirety of integers. In a first step we introduce

$$0 := \widehat{(1, 1)} = \{(n, n) : n \in \mathbb{N}\}.$$

Furthermore, we identify $n \in \mathbb{N}$ with

$$\widehat{(n+1, 1)} = \{(n+1, 1), (n+2, 2), \dots\}$$

and since

$$\widehat{(n+1, 1)} + \widehat{(1, n+1)} = \widehat{(1, 1)} = 0$$

we define $-n$ by

$$-n := \widehat{(1, n+1)} = \{(1, n+1), (2, n+2), \dots\}$$

So, at the end we get \mathbb{Z} , the set of all integers in the following way

$$\mathbb{Z} = \mathbb{N} \times \mathbb{N}/\sim = \{\widehat{(n, m)} : n, m \in \mathbb{N}\}.$$

In other words, we identify \mathbb{N} with those pairs (n, m) where $n > m$, while the new negative integers occur for $n < m$. The number 0 corresponds to pairs (n, n) . Compare Figure A.7.3.

The addition on \mathbb{Z} is that induced by the one in $\mathbb{N} \times \mathbb{N}/\sim$. So, for example,

$$3 + (-4) = \widehat{(4, 1)} + \widehat{(1, 5)} = \widehat{(5, 6)} = \widehat{(1, 2)} = -1.$$

We may proceed and introduce an order on \mathbb{Z} by

$$(n_1, m_1) < (n_2, m_2) \Leftrightarrow n_1 + m_1 < n_2 + m_2.$$

Again one has to show (which is easy) that the order is independent of the choice of an element in a given equivalence class. In other words, we have to prove that

$$\begin{aligned} (A.7.8) \quad (n_1, m_1) &\sim (n'_1, m'_1), (n_2, m_2) \sim (n'_2, m'_2) \\ &\Rightarrow [(n_1, m_1) < (n_2, m_2) \Leftrightarrow (n'_1, m'_1) < (n'_2, m'_2)]. \end{aligned}$$

So in this way the “old” order on \mathbb{N} is retained and the “new” order on \mathbb{Z} gives

$$\dots < -3 < -2 < -1 < 0 < 1 < 2 < \dots.$$

Finally, we may also extend multiplication to \mathbb{Z} by defining the product of two pairs of integers by

$$(A.7.9) \quad (n_1, m_1) \cdot (n_2, m_2) = (n_1 n_2 + m_1 m_2, n_1 m_2 + n_2 m_1).$$

This multiplication of pairs of natural numbers extends the one on \mathbb{N} because of

$$(n_1 - m_1)(n_2 - m_2) = [n_1 n_2 + m_1 m_2] - [n_1 m_2 + n_2 m_1]$$

in the case $n_1 > m_1$ and $n_2 > m_2$.

Again, one has to verify that multiplication is independent of the chosen element in the equivalence class. Here the main point is not the proof of this fact (these are elementary algebraic calculations), but to understand that something has to be proved

because otherwise multiplication would not have been defined uniquely. For example, we have to make sure that

$$(3, 4) \cdot (8, 5) \sim (5, 6) \cdot (10, 7) \sim (1, 2) \cdot (4, 1).$$

Otherwise, the multiplication of integers would depend on the special element which we have chosen from the equivalence classes -1 and 3 , respectively.

It is a bit cumbersome but not difficult to show that addition and multiplication on \mathbb{Z} satisfy all properties (associativity, commutativity, distributive law, etc.) We omit these proofs because they are very technical and do not lead to new insight into the structure of integers.

Example A.7.1. We want to multiply 4 and -3 in the setting of pairs of natural numbers. Since 4 corresponds to $(5, 1)$, and -3 is represented by $(1, 4)$, formula (A.7.9) leads to

$$(5, 1) \cdot (1, 4) = (5 \cdot 1 + 1 \cdot 4, 5 \cdot 4 + 1 \cdot 1) = (9, 21).$$

Since $(9, 21)$ corresponds to -12 , this is a different way to evaluate $4(-3) = -12$. If we replace $(5, 1)$ by $(7, 3)$ and/or $(1, 4)$ by $(6, 9)$, then we end up with a different pair of numbers but which is equivalent to $(9, 21)$, hence also gives -12 .

Exercise A.7.1. Let \mathbb{N} and σ possess the properties as stated in Peano's axioms. For each $n \geq 1$ let

$$\sigma^n := \underbrace{\sigma \circ \cdots \circ \sigma}_{n \text{ times}}.$$

Show that each σ^n is injective and determine its range.

Is it true that then

$$\mathbb{N} = \{1, \sigma(1), \sigma^2(1), \dots\}?$$

Exercise A.7.2. Let $f : \mathbb{N}_1 \rightarrow \mathbb{N}_2$ be a bijection possessing the two properties stated in Proposition A.7.4. Show that then its inverse function f^{-1} satisfies

$$\begin{aligned} f^{-1}(1_2) &= 1_1 \quad \text{and} \\ f^{-1}(\sigma_2(m)) &= \sigma_1(f^{-1}(m)), \quad m \in \mathbb{N}_2. \end{aligned}$$

Exercise A.7.3. Prove Proposition A.7.6.

Exercise A.7.4. Prove the implication in (A.7.8).

Exercise A.7.5. Show that the multiplication (A.7.9) of pairs is compatible with the given equivalence relation (A.7.7) on $\mathbb{N} \times \mathbb{N}$.

Exercise A.7.6. Let $S \subseteq \mathbb{Z} \times \mathbb{Z}$ consist of all pairs (a, b) of integers satisfying $2a = 3b$. So, for example the pairs $(3, 2)$ or $(-6, -4)$ and so on, they all belong to the set S . We define now a relation on $\mathbb{Z} \times \mathbb{Z}$ as follows:

$$(a, b) \sim (c, d) \iff (a, b) - (c, d) = (a - c, b - d) \in S.$$

Show that this is an equivalence relation on $\mathbb{Z} \times \mathbb{Z}$. What are the equivalence classes?

Exercise A.7.7. Given a pair $(n, m) \in \mathbb{N} \times \mathbb{N}$ we set $-(n, m) := (m, n)$. This is justified because of $(n, m) + (m, n) \sim (1, 1)$ where we consider $\widehat{(1, 1)}$ as zero. Show that then for all pairs (n_1, m_1) and (n_2, m_2) it follows that

$$-(n_1, m_1) \cdot (n_2, m_2) = (n_1, m_1) \cdot [-(n_2, m_2)] = -[(n_1, m_1) \cdot (n_2, m_2)].$$

Moreover,

$$[-(n_1, m_1)] \cdot [- (n_2, m_2)] = (n_1, m_1) \cdot (n_2, m_2).$$

Hereby, the multiplication of pairs of natural numbers is defined by (A.7.9).

Exercise A.7.8. Say a pair (x_1, y_1) of real numbers is related to another pair (x_2, y_2) of those numbers provided that

$$x_1 + y_1 = x_2 + y_2.$$

Show that this is an equivalence relation on $\mathbb{R} \times \mathbb{R}$. Describe the equivalence classes.

Exercise A.7.9. Define a binary relation on pairs in $\mathbb{N} \times \mathbb{N}$ as follows: (m_1, n_1) is in relation with (m_2, n_2) if $m_1 n_1 = m_2 n_2$. Determine if this relation is an equivalence relation and if so, describe the equivalence classes.

Exercise A.7.10. Consider the following binary relation defined on triples of real numbers as follows: $(x_1, x_2, x_3) \sim (y_1, y_2, y_3)$ if

$$\sum_{k=1}^3 x_k = \sum_{k=1}^3 y_k, \quad \sum_{1 \leq j < k \leq 3} x_j x_k = \sum_{1 \leq j < k \leq 3} y_j y_k, \text{ and } x_1 x_2 x_3 = y_1 y_2 y_3.$$

Determine whether this relation is reflexive, symmetric or transitive.

A.8. More Exercises

Exercise A.8.1. Suppose a permutation $\pi \in S_n$ has the cycle (i_1, \dots, i_m) . Which cycle does the inverse permutation π^{-1} possess?

Exercise A.8.2. Find all cycles of the permutations

$$\begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 \\ 2 & 7 & 4 & 6 & 1 & 5 & 3 \end{pmatrix} \quad \text{and} \quad \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 \\ 7 & 1 & 4 & 5 & 3 & 2 & 6 \end{pmatrix}.$$

Exercise A.8.3. Suppose $\pi \in S_8$ has the cycles $(1, 4, 6)$ and $(2, 8, 3)$. Represent π in form of a table.

Exercise A.8.4. Write

$$\pi = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 \\ 3 & 2 & 4 & 5 & 1 \end{pmatrix}$$

as product of transpositions. Is π an even or an odd permutation?

Exercise A.8.5. Why is the composition of two even or of two odd permutations always even? Similarly, the composition of an even and an odd permutation is always odd. Why?

A.8.1. Exercises about Dedekind Cuts. The following assertions and problems deal with a special class of subsets of \mathbb{Q} , the class of so-called Dedekind cuts of rational numbers. An example of such a cut already occurred in Example 4.4.1. The aim of the subsequent exercises is to give those, who want to know more, some additional background about the construction of real numbers. This is done via a special technique, which was first introduced by the French mathematician Joseph Bertrand (1822–1900), then later on further developed by the German mathematician Richard Dedekind (1831–1916), and therefore bears his name.

It is important to mention, that all subsequent operations, as for example addition, subtraction or multiplication, are those of rational numbers, in the way as they were introduced in (3.1.3). Similarly, the order on \mathbb{Q} is the one defined in (3.1.5). Thus, everything happens in the setting of rational numbers. Real numbers did not show up yet!

Definition A.8.1. A subset $A \subseteq \mathbb{Q}$ is said to be a **Dedekind cut** (on \mathbb{Q}) provided it satisfies the following three properties:

- (1) The set A is neither empty nor the whole set \mathbb{Q} .
- (2) The set A is downward closed. That is, if $x \in A$ and $y \in \mathbb{Q}$ satisfies $y < x$, then also $y \in A$.
- (3) The set A does not contain a greatest element. In other words, if $x \in A$, then there exists a $y \in A$ such that $y > x$.

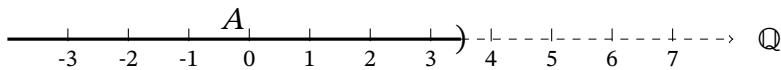


Figure A.8.1. One may visualize a Dedekind cut by a ray, unbounded at its left, bounded and open at its right-hand side.

Let us denote the set of all Dedekind cuts over \mathbb{Q} by

$$\mathcal{D} := \{A \subseteq \mathbb{Q} : A \text{ is a Dedekind cut}\}.$$

For each Dedekind cut A its complementary set $A^c = \mathbb{Q} \setminus A$ plays an important role. This is due to the fact that the pair (A, A^c) divides the set \mathbb{Q} of rational numbers into two parts (a lower part A and an upper part A^c). This explains why elements in \mathcal{D} are named *cut*²⁰.

Exercise A.8.6. Verify the following assertions.

- For any rational number q the set $A_q \subset \mathbb{Q}$ defined by

$$A_q := \{x \in \mathbb{Q} : x < q\}$$

is a Dedekind cut.

- Given two rational numbers q_1 and q_2 , then $A_{q_1} = A_{q_2}$ happens if and only if $q_1 = q_2$. In other words, the mapping $q \mapsto A_q$ is an injection from \mathbb{Q} to \mathcal{D} .

²⁰Quite often Dedekind cuts are defined as pairs (A, B) where $B = A^c$ consists of the upper bounds of A . But since B is completely determined by A , it suffices to deal with A only. On the other hand, the notation (A, B) explains better that after a cut one obtains two pieces, not only the left-hand one.

- The set

$$A := \{x \in \mathbb{Q}_+ : x^2 < 2\} \cup \{x \in \mathbb{Q} : x \leq 0\}$$

is a Dedekind cut with $A \neq A_q$ for all $q \in \mathbb{Q}$.

Hint: Apply the technique used in the proof of Theorem 4.6.3 together with the fact that there is no $q \in \mathbb{Q}$ for which $q^2 = 2$.

Definition A.8.2. Dedekind cuts represented as A_q for a certain $q \in \mathbb{Q}$ are said to be **rational**. All other cuts are called **irrational**²¹.

Examples of rational cuts are

$$A_0 = \{x \in \mathbb{Q} : x < 0\} = \mathbb{Q}_- \quad \text{and} \quad A_1 = \{x \in \mathbb{Q} : x < 1\}.$$

On the contrary, $A := \{x \in \mathbb{Q}_+ : x^2 < 2\} \cup \{x \in \mathbb{Q} : x \leq 0\}$ is an irrational cut.

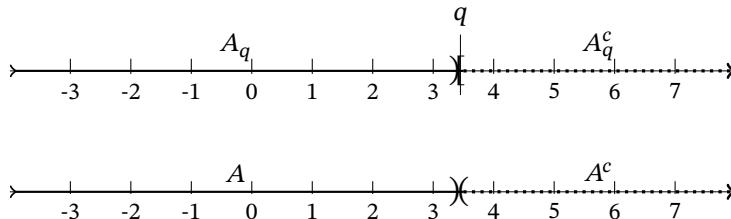


Figure A.8.2. Rational and irrational Dedekind cuts.

Exercise A.8.7. Prove the following properties of the complements of cuts.

- If $A \in \mathcal{D}$, then it follows that $A^c \neq \emptyset$ and $A^c \neq \mathbb{Q}$.
- A^c is upwards closed. That is, if $x \in A^c$ and the rational number y satisfies $y > x$, then this implies $y \in A^c$ as well.
- The complement A^c of a cut A consists of all upper bounds of A . This says that

$$A^c = \{y \in \mathbb{Q} : \forall x \in A, x \leq y\}.$$

- Describe the complement of rational cuts.
- A Dedekind cut A is rational if and only if A^c has a minimal element (in \mathbb{Q}).

Definition A.8.3. Suppose A and B are two Dedekind cuts. Then we define two binary relations on \mathcal{D} by

$$A \leq B \Leftrightarrow A \subseteq B \quad \text{and} \quad A < B \Leftrightarrow A \leq B, A \neq B.$$

The cut A is said to be **positive** if $A > A_0$ where $A_0 = A_q$ with $q = 0$. The cut is **negative** in the case $A < A_0$.

²¹The way to think about Dedekind cuts is that you are cutting the line of rational numbers by an infinitely thin knife. Then two cases may occur: Either the knife hits a rational number or the knife meets a hole (compare Figure A.8.2).

Exercise A.8.8. Prove that the introduced relations own the following properties.

- “ \leq ” fulfills the properties of an order as stated in Definition A.5.7.
- For any $q_1, q_2 \in \mathbb{Q}$ the relation $A_{q_1} < A_{q_2}$ is equivalent to $q_1 < q_2$. Consequently, the mapping $q \mapsto A_q$ is not only injective, but it is also compatible with the ordering on \mathbb{Q} and on \mathcal{D} , respectively.
- The cut A is positive if and only if $0 \in A$. Is $A \in \mathcal{D}$ negative if and only if $0 \in A^c$? Prove this or give a counterexample.

Definition A.8.4. Let A and B be two Dedekind cuts. Define their sum by

$$A + B := \{x + y : x \in A, y \in B\}.$$

Exercise A.8.9. Verify the following properties of the summation on \mathcal{D} .

- For all $A, B \in \mathcal{D}$ the sum $A + B$ is a Dedekind cut as well.
- The summation of cuts is an associative operation. That is

$$(\forall A, B, C \in \mathcal{D})(A + (B + C) = (A + B) + C).$$

- The summation of cuts is commutative:

$$(\forall A, B \in \mathcal{D})(A + B = B + A).$$

- Let as before $A_0 = \{x \in \mathbb{Q} : x < 0\}$. Then it follows that

$$(\forall A \in \mathcal{D})(A + A_0 = A_0 + A = A).$$

- For any cut A there exists a unique cut B such that $A + B = A_0$. Do not forget, that B has to be a cut!
- Denote B by $-A$. That is, $-A$ is the unique cut with $A + (-A) = A_0$. How does $-A$ look like in the case of rational cuts A and of irrational ones.
- A Dedekind cut A is negative if and only if $-A$ is positive.
- For all $q, q_1, q_2 \in \mathbb{Q}$ it follows that

$$A_{q_1+q_2} = A_{q_1} + A_{q_2} \quad \text{and} \quad -A_q = A_{-q}.$$

Conclusion: The pair $(\mathcal{D}, +)$ is a commutative group with neutral element A_0 and inverse $-A$ for each $A \in \mathcal{D}$. Moreover, the mapping

$$q \mapsto A_q, \quad q \in \mathbb{Q},$$

is a group homomorphism from $(\mathbb{Q}, +)$ to $(\mathcal{D}, +)$ (cf. Exercise A.8.15 for the definition of a group homomorphism).

Exercise A.8.10.

- Let A_1, \dots, A_n be all Dedekind cuts. Show that then

$$\bigcup_{j=1}^n A_j \quad \text{as well as} \quad \bigcap_{j=1}^n A_j$$

are also Dedekind cuts. How can these two cuts be described as cut A_q with suitable $q \in \mathbb{Q}$, provided that for all $j \leq n$ one has $A_j = A_{q_j}$ with rational numbers q_1, \dots, q_n ?

- (\star) Let \mathcal{A} be a set of Dedekind cuts. Suppose, furthermore, that there is a $B \in \mathcal{D}$ such that for all $A \in \mathcal{A}$ one has $A \leq B$ (i.e., \mathcal{A} is bounded above). Prove that then

$$A^\infty := \bigcup_{A \in \mathcal{A}} A = \{x \in \mathbb{Q} : \exists A \in \mathcal{A}, x \in A\}$$

is also a Dedekind cut. Furthermore,

$$(\forall A \in \mathcal{A})(A \leq A^\infty) \quad \text{and} \quad (\forall A \in \mathcal{A}, A \leq C) \Rightarrow (A^\infty \leq C)$$

Remark A.8.1. The latter property says the following: The cut A^∞ is not only an upper bound of the set \mathcal{A} , it is even the least one. Thus, we conclude that every set of Dedekind cuts, which is bounded above, possesses a least upper bound.

One important feature is missing: The multiplication of Dedekind cuts in a way such that $A_{q_1} \times A_{q_2} = A_{q_1 \cdot q_2}$. This can be done, but is quite complicated. Let us shortly sketch why this is not an easy task. A first idea to define $A \times B$ would be to use the same approach as in the case of addition, namely to define the product of A and B by

$$A \times B := \{x \cdot y : x \in A, y \in B\}.$$

But this does not work. Why? In this setting $A_1 \times A_1$ would contain $(-2)(-2) = 4$ which should not be in the product $A_1 \times A_1$. Recall that we expect and want that $A_1 \times A_1 = A_{1 \cdot 1} = A_1$.

To avoid such undesirable phenomenons the product has to be defined in several steps. In a first one we define the product of two **positive** cuts as follows:

$$A \times B := \{x \cdot y : x \in A, y \in B, \text{ where } x, y \text{ are not both negative}\}.$$

If A is negative and B is positive, then $-A$ is positive, so the following definition makes sense:

$$A \times B := -[(-A) \times B].$$

If both cuts are negative, set

$$A \times B := [(-A) \times (-B)].$$

Finally for all $A \in \mathcal{D}$ let

$$A \times A_0 = A_0 \times A = A_0.$$

It is obviously clear that it is technical difficult and very cumbersome to verify that the product defined in this way is associative, commutative, that $A \times A_1 = A$, that each $A \neq A_0$ has an inverse and, moreover, that the distributive law is valid. Finally, it has to be shown that $A_{q_1} \times A_{q_2} = A_{q_1 \cdot q_2}$ for all $q_1, q_2 \in \mathbb{Q}$. So we even do not formulate the verification of all these properties as exercise. Those who are interested in the proofs are referred to [29] or to Chapter 28 in [32]. A nicely written introduction to Dedekind cuts can be found in the classical book [16] by G. H. Hardy (1877–1947)²².

All this can be done. And at the end one gets a model for the real numbers, namely \mathcal{D} with two binary operations, $+$ and \times defined above, as well as with an order which is compatible with the algebraic operations. Furthermore, and this was the actual target, due to Exercise A.8.10 the field \mathcal{D} is order complete, i.e., every set bounded above has

²²G. H. Hardy, was one of the greatest English mathematicians of his time and did significant work in number theory.

a least upper bound. And, most important, identifying \mathbb{Q} with the set $\{A_q : q \in \mathbb{Q}\}$ of rational cuts, this model extends naturally the field of rational numbers. Hereby, the new ingredients are irrational Dedekind cuts, i.e., those which cannot be represented as A_q for some $q \in \mathbb{Q}$, and which correspond to irrational numbers, in the way as the cut in Example 4.4.1 corresponds to $\sqrt{2}$.

A.8.2. Algebraic Exercises. For additional information and helpful hints about the subsequent exercises we recommend the recent book [30].

Definition A.8.5. Let G be a nonempty set. A function $h : G \times G$ into G is said to be a **binary operation** on G . Typical examples are

$$h : \mathbb{N} \times \mathbb{N} \rightarrow \mathbb{N} \quad \text{with} \quad h(n, m) = n + m \quad \text{or} \quad h(n, m) = n \cdot m.$$

Another important binary operation on S_n , the set of permutations of order n , is

$$(A.8.1) \quad (\pi, \mu) \mapsto \pi \circ \mu, \quad \pi, \mu \in S_n.$$

Or we may consider the binary operations on \mathbb{R} given by

$$(x, y) \mapsto \min\{x, y\} \quad \text{or} \quad (x, y) \mapsto \max\{x, y\}.$$

Let $h : G \times G \rightarrow G$ be a binary operation. It is common to write²³

$$x * y \quad \text{instead of} \quad h(x, y), \quad x, y \in G.$$

In other words, a binary operation assigns to each pair (x, y) of elements in G a third one, denoted by $x * y$.

$$(x, y) \mapsto x * y.$$

One usually writes $(G, *)$ to describe at the same time the underlying set as well as the binary operation. Typical examples are $(\mathbb{Z}, +)$ or $(\mathbb{R}, +)$ or (\mathbb{Q}_+, \cdot) .

Exercise A.8.11. Answer the following questions and give the relevant examples.

- (1) Find at least 3 examples of binary operations, all different from the ones already mentioned.
- (2) A binary operation $*$ on a set G is said to be **associative** provided that

$$(x * y) * z = x * (y * z), \quad x, y, z \in G.$$

Give two examples of associative binary operations and at least one example of a nonassociative one.

- (3) A binary operation $*$ is said to be **commutative** provided that

$$x * y = y * x, \quad x, y \in G.$$

Give examples of two commutative binary operations and at least one noncommutative one.

- (4) An element $e \in G$ is called **unit** or **neutral element** if

$$x * e = e * x = x, \quad x \in G.$$

- Show that any unit is unique, provided there exists one.

²³For example, we are used to writing $3 + 4 = 7$ instead of $+(3, 4) = 7$. Indeed, addition is nothing else than an operation which maps every pair of numbers by a fixed rule to another number (called their sum).

- Does there exist a unit for the binary operation on S_n defined by (A.8.1)?
- Do there exist units for the binary operations on \mathbb{N} defined by

$$(n, m) \mapsto \min\{n, m\} \quad \text{or} \quad (n, m) \mapsto \max\{n, m\}, \quad n, m \in \mathbb{N}?$$

- (5) Suppose there exists a unit e on G with respect to the binary operation $*$. An element $a \in G$ is said to be inverse to x whenever

$$x * a = a * x = e.$$

- Show that any $x \in G$ has at most one inverse element.
- Give two examples of binary operations with unit such that any element has an inverse one. Find another binary operation with unit where there exist elements **without** an inverse element.
- Determine the unit and the inverse elements in \mathbb{Z} with binary operation $x * y = x + y$. Do the same for the binary operation on \mathbb{R}_+ given by $x * y = x \cdot y$.
- If we denote the inverse of an element $x \in G$ by x^{-1} , show the following: If x and y possess both inverse elements, then this also so for $x * y$. Moreover, one has

$$(x * y)^{-1} = y^{-1} * x^{-1}.$$

Definition A.8.6. A set G with a binary operation $*$ is called **group** whenever $*$ is associative, there exists a unit and each $x \in G$ has an inverse x^{-1} . It is said to be a **commutative group** (or abelian group) provided that the binary operation $*$ on G is commutative.

Exercise A.8.12. Answer the following questions and give relevant examples.

- (1) Why are $(\mathbb{N}, +)$ and (\mathbb{Z}, \cdot) **no** groups?
- (2) Find two examples of commutative groups and one of a noncommutative one.
- (3) Given $n \geq 2$, one defines on

$$\mathbb{Z}_n^* = \mathbb{Z}_n \setminus \{0\} = \{1, \dots, n - 1\}$$

a binary operation $*$ by

$$a * b = a \cdot b \pmod{n}, \quad a, b \in \mathbb{Z}_n^*.$$

Characterize those $n \geq 2$ for which $(\mathbb{Z}_n^*, *)$ is a group?

Exercise A.8.13. Let the set $G = \{e, a, b, c\}$ be endowed with the binary operation defined by the following table:

| ★ | e | a | b | c |
|---|---|---|---|---|
| e | e | a | b | c |
| a | a | e | c | b |
| b | b | c | e | a |
| c | c | b | a | e |

Show that $(G, *)$ is a group. Is it commutative? Which element is the unit and what are the inverse elements of e, a, b , and c ?

The group G with the given binary operation is usually denoted by \mathbb{K}_4 and called Klein's²⁴ 4-group.

Prove that \mathbb{K}_4 is isomorphic to the group $\mathbb{Z}_2 \times \mathbb{Z}_2$ introduced in Exercise 2.6.6. That is, construct a bijection between $\mathbb{Z}_2 \times \mathbb{Z}_2$ and \mathbb{K}_4 which preserves the binary operations on $\mathbb{Z}_2 \times \mathbb{Z}_2$ and \mathbb{K}_4 , respectively.

Exercise A.8.14. Let $(G_1, *_1)$ and $(G_2, *_2)$ be two groups. Define on $G := G_1 \times G_2$ a binary operation $*$ by

$$(x_1, x_2) * (y_1, y_2) = (x_1 *_1 y_1, x_2 *_2 y_2), \quad (x_1, y_1), (x_2, y_2) \in G = G_1 \times G_2.$$

- (1) Show that $(G, *)$ is also a group. Moreover, if both groups are commutative, then so is $(G, *)$.
- (2) Suppose $G_1 = G_2 = \{0, 1\}$ with binary operation

$$0 * 0 = 0, \quad 0 * 1 = 1 * 0 = 1 \text{ and } 1 * 1 = 0.$$

Argue why G_1 , hence also G_2 , are groups. Describe the generated binary operation on $G = G_1 \times G_2 = \{0, 1\} \times \{0, 1\}$.

- (3) Describe G and $*$ in the case $(G_1, *_1) = (G_2, *_2) = (\mathbb{R}, +)$.

Exercise A.8.15. Let $(G_1, *_1)$ and $(G_2, *_2)$ be two groups. A function ϕ from G_1 to G_2 is said to be a (group) **homomorphism** if

$$\phi(x_1 *_1 y_1) = \phi(x_1) *_2 \phi(y_1), \quad x_1, y_1 \in G_1.$$

Typical examples are ϕ from $(\mathbb{Z}, +)$ to (\mathbb{R}_+, \cdot) with

$$\phi(k) = 10^k \quad \text{or} \quad \phi(k) = 2^k, \quad k \in \mathbb{Z},$$

and ϕ from (\mathbb{R}_+, \cdot) to $(\mathbb{R}, +)$ with

$$\phi(x) = \log x, \quad x \in \mathbb{R}_+.$$

- (1) Give further examples of homomorphisms, e.g., from $(\mathbb{Z}, +)$ to $(\mathbb{Z}, +)$ or from (\mathbb{R}_+, \cdot) to (\mathbb{R}_+, \cdot) .
- (2) Check whether ϕ from $(\mathbb{Z}, +)$ to $(\mathbb{Z}_n, +)$ with

$$\phi(n) = r \quad \text{if} \quad n \equiv r \pmod{n}$$

is a homomorphism.

- (3) Recall (cf. Example A.5.16) that

$$\mathbb{Z}/n\mathbb{Z} = \{\hat{0}, \dots, \widehat{n-1}\}$$

denotes the set of all equivalence classes modulo n . Moreover, addition (and in the same way also multiplication) of these classes are defined by

$$\hat{a} + \hat{b} = \widehat{a+b}, \quad a, b \in \mathbb{Z}.$$

Define φ from \mathbb{Z}_n to $\mathbb{Z}/n\mathbb{Z}$ by

$$\varphi(r) := \hat{r}, \quad r \in \mathbb{Z}_n.$$

Show that φ is a bijective homomorphism. Describe its inverse φ^{-1} .

²⁴Felix Klein (1849–1925) was a German mathematician.

Exercise A.8.16. Let ϕ be a homomorphism from a group $(G_1, *_1)$ to a group $(G_2, *_2)$. Denote by $e_1 \in G_1$ and $e_2 \in G_2$ the corresponding units. Show the following:

- (1) $\phi(e_1) = e_2$.
- (2) For all $x \in G_1$, $\phi(x^{-1}) = \phi(x)^{-1}$. Here, on the left-hand side the inverse element is taken in G_1 while on the right-hand side it is the inverse in G_2 .
- (3) The homomorphism ϕ is injective if and only if only $\phi(x) = e_2$ implies that $x = e_1$.
- (4) Suppose that ϕ is moreover bijective. Prove that then its inverse ϕ^{-1} is a homomorphism from $(G_2, *_2)$ to $(G_1, *_1)$.

A bijective homomorphism is usually called a (group) **isomorphism**. The existence of an isomorphism from a group G_1 to a group G_2 tells us, that these two groups may be considered as identically.

Definition A.8.7. Suppose a set R has two binary operations, denoted by “+” and by “·”. It is common to use the same signs as for the (usual) addition and (usual) multiplication of numbers, although they may be quite different operations. That is, + and · are mappings from $R \times R$ to R with

$$+ : (x, y) \mapsto x + y \quad \text{and} \quad \cdot : (x, y) \mapsto x \cdot y,$$

and nothing more. Then $(R, +, \cdot)$ is said to be a **ring** if

- (i) $(R, +)$ is a commutative group and
- (ii) The following distributive laws are valid:

$$(\forall x, y, z \in R)[(x \cdot (y + z) = x \cdot y + x \cdot z) \text{ and } (x + y) \cdot z = x \cdot z + y \cdot z].$$

A ring is said to be **commutative** provided that

$$(\forall x, y \in R)(x \cdot y = y \cdot x).$$

Exercise A.8.17. Prove the following.

- (1) If $(R, +, \cdot)$ is a ring, then $x \cdot 0 = 0 \cdot x = 0$, for any $x \in R$. Here $0 \in R$ denotes the neutral element with respect to +, that is $x + 0 = 0 + x = x$ for all $x \in R$.
- (2) If $-x$ denotes the additive inverse of $x \in R$, that is $x + (-x) = (-x) + x = 0$, then

$$(\forall x, y \in R)((-x) \cdot y = x \cdot (-y) = -(x \cdot y))$$

and

$$(\forall x, y \in R)((-x) \cdot (-y) = x \cdot y).$$

Exercise A.8.18. Check whether the following sets with given binary operations are rings:

- (1) The set \mathbb{N} of natural numbers with ordinary addition and multiplication.
- (2) The set \mathbb{Z} of integers with ordinary addition and multiplication.
- (3) The set $2\mathbb{Z}$ of even integers with ordinary addition and multiplication.
- (4) The set $\text{Pol}(\mathbb{R})$ of all real polynomials with

$$(p + q)(x) = p(x) + q(x) \quad \text{and} \quad (p \cdot q)(x) = p(x) \cdot q(x), \quad x \in \mathbb{R}.$$

- (5) The set $\text{Pol}(\mathbb{C})$ of all complex polynomials with analogue addition and multiplication taken in \mathbb{C} .
- (6) The set $\mathbb{Z}_n = \{0, \dots, n - 1\}$ with addition and multiplication modulo $n \geq 2$.
- (7) The set C_0 of real sequences tending to zero. That is,

$$C_0 := \{(x_k)_{k \geq 1} : x_k \in \mathbb{R}, \lim_{k \rightarrow \infty} x_k = 0\}$$

with binary operations

$$(x_k)_{k \geq 1} + (y_k)_{k \geq 1} = (x_k + y_k)_{k \geq 1} \quad \text{and} \quad (x_k)_{k \geq 1} \cdot (y_k)_{k \geq 1} = (x_k \cdot y_k)_{k \geq 1}.$$

- (8) Define the set $\mathbb{Z}[i]$ of so-called **Gaussian integers** by

$$\mathbb{Z}[i] := \{a + bi : a, b \in \mathbb{Z}\}$$

Endow this set with addition and multiplication as defined for complex numbers.

- (9) The set $\mathcal{B}(A)$ of bounded functions from a set A with values in \mathbb{R} . Hereby, the operations are defined by

$$(f + g)(a) := f(a) + g(a) \quad \text{and} \quad (f \cdot g)(a) := f(a)g(a), \quad a \in A.$$

Note that a function f from A to \mathbb{R} is bounded provided that $|f(a)| \leq c$ for some $c > 0$ and all $a \in A$.

Exercise A.8.19. Let $(R, +, \cdot)$ be a ring. An element $e \in R$ is said to be a (multiplicative) **unit** provided that $x \cdot e = e \cdot x = x$, for any $x \in R$. Which of the rings in Example A.8.18 possesses a (multiplicative) unit?

Exercise A.8.20. Two nonzero elements x, y in a ring R are called (nontrivial) **divisors of zero** if $x \cdot y = y \cdot x = 0$.

- (1) Show that the ring \mathbb{Z}_n with addition and multiplication modulo $n \geq 2$ has (nontrivial) divisors of zero if and only if n is a composite natural number. In other words, \mathbb{Z}_n does not possess nontrivial divisors of zero if and only if n is prime.
- (2) Do there exist (nontrivial) divisors of zero in \mathbb{Z} , in $\text{Pol}(\mathbb{R})$, in $\mathcal{B}(A)$ or in C_0 ?

Exercise A.8.21. A set \mathbb{F} with two binary operations $+$ and \cdot is said to be a **field** if it is a ring and, moreover, $(\mathbb{F} \setminus \{0\}, \cdot)$ is a commutative group. Here 0 denotes the unit in $(\mathbb{F}, +)$.

- (1) Give two examples of fields.
- (2) Of course, every field is a commutative ring with unit. Find two rings which are not fields.
- (3) Show that in any field there do not exist nontrivial divisors of zero.
- (4) For which $n \in \mathbb{N}$ is $(\mathbb{Z}_n, +, \cdot)$ a field? Here addition and multiplication are defined as addition and multiplication modulo n .

Bibliography

- [1] Petr Beckmann, *A history of π (pi)*, 2nd ed., The Golem Press, Boulder, Colo., 1971. MR0449960
- [2] Ray C. Bose, Sharadchandra S. Shrikhande, and Ernest T. Parker, *Further results on the construction of mutually orthogonal Latin squares and the falsity of Euler's conjecture*, Canadian J. Math. **12** (1960), 189–203, DOI 10.4153/CJM-1960-016-5. MR122729
- [3] Neil Calkin and Herbert S. Wilf, *Recounting the rationals*, Amer. Math. Monthly **107** (2000), no. 4, 360–363, DOI 10.2307/2589182. MR1763062
- [4] Paul J. Cohen, *Set theory and the continuum hypothesis*, W. A. Benjamin, Inc., New York-Amsterdam, 1966. MR0232676
- [5] Henry Cohn, *A short proof of the simple continued fraction expansion of e*, Amer. Math. Monthly **113** (2006), no. 1, 57–62, DOI 10.2307/27641837. MR2202921
- [6] John H. Conway and Richard K. Guy, *The book of numbers*, Copernicus, New York, 1996, DOI 10.1007/978-1-4612-4072-3. MR1411676
- [7] Richard Courant and Herbert Robbins, *What is mathematics? An elementary approach to ideas and methods*, Oxford University Press, New York, 1979. MR552669
- [8] William Dunham, *Journey through genius: The great theorems of mathematics*, Penguin Books, New York, 1991. MR1147417
- [9] William Dunham, *Euler: the master of us all*, The Dolciani Mathematical Expositions, vol. 22, Mathematical Association of America, Washington, DC, 1999. MR1669154
- [10] Jordan Ellenberg, *How not to be wrong: The power of mathematical thinking*, Penguin Press, New York, 2014. MR3236985
- [11] Ronald L. Graham, Donald E. Knuth, and Oren Patashnik, *Concrete mathematics: A foundation for computer science*, 2nd ed., Addison-Wesley Publishing Company, Reading, MA, 1994. MR1397498
- [12] Ben Green and Terence Tao, *The primes contain arbitrarily long arithmetic progressions*, Ann. of Math. (2) **167** (2008), no. 2, 481–547, DOI 10.4007/annals.2008.167.481. MR2415379
- [13] Scott B. Guthery, *A motif of mathematics: History and application of the mediant and the Farey sequence*, Docent Press, Boston, MA, 2011. MR2895290
- [14] Paul R. Halmos, E. E. Moise, and George Piranian, *The Problem of Learning to Teach*, Amer. Math. Monthly **82** (1975), no. 5, 466–476, DOI 10.2307/2319737. MR1537724
- [15] Paul R. Halmos, *Naive set theory*, The University Series in Undergraduate Mathematics, D. Van Nostrand Co., Princeton, N.J.-Toronto-London-New York, 1960. MR0114756
- [16] Godfrey H. Hardy, *A course of pure mathematics*, reprint of the tenth (1952) edition with a foreword by T. W. Körner, Centenary edition, Cambridge University Press, Cambridge, 2008, DOI 10.1017/CBO9780511989469. MR2400109
- [17] Thomas J. Jech, *The axiom of choice*, Studies in Logic and the Foundations of Mathematics, Vol. 75, North-Holland Publishing Co., Amsterdam-London; American Elsevier Publishing Co., Inc., New York, 1973. MR0396271
- [18] Thomas J. Jech, *Set theory*, The third millennium edition, revised and expanded, Springer Monographs in Mathematics, Springer-Verlag, Berlin, 2003. MR1940513
- [19] Morris Kline, *Mathematical thought from ancient to modern times*, Oxford University Press, New York, 1972. MR0472307

-
- [20] James Maynard, *Small gaps between primes*, Ann. of Math. (2) **181** (2015), no. 1, 383–413, DOI 10.4007/annals.2015.181.1.7. MR3272929
 - [21] Steven J. Miller and David Montague, *Picturing irrationality*, Math. Mag. **85** (2012), no. 2, 110–114, DOI 10.4169/math.mag.85.2.110. MR2910300
 - [22] John von Neumann, *Über die Definition durch transfinite Induktion und verwandte Fragen der allgemeinen Mengenlehre* (German), Math. Ann. **99** (1928), no. 1, 373–391, DOI 10.1007/BF01459102. MR1512455
 - [23] Ivan Niven, *A simple proof that π is irrational*, Bull. Amer. Math. Soc. **53** (1947), 509, DOI 10.1090/S0002-9904-1947-08821-2. MR21013
 - [24] Ivan Niven, *Irrational numbers*, The Carus Mathematical Monographs, No. 11, Mathematical Association of America; distributed by John Wiley & Sons, Inc., New York, N.Y., 1956. MR0080123
 - [25] Ivan Niven, Herbert S. Zuckerman, and Hugh L. Montgomery, *An introduction to the theory of numbers*, 5th ed., John Wiley & Sons, Inc., New York, 1991. MR1083765
 - [26] Marc Noy, *A Short Solution of a Problem in Combinatorial Geometry*, Math. Mag. **69** (1996), no. 1, 52–53. MR1573137
 - [27] Oskar Perron, *Die Lehre von den Kettenbrüchen* (German), 2nd ed, Chelsea Publishing Co., New York, N. Y., 1950. MR0037384
 - [28] George Polya, *How to solve it: A new aspect of mathematical method*, with a foreword by John H. Conway; reprint of the second (2004) edition [MR2183670], Princeton Science Library, Princeton University Press, Princeton, NJ, 2014. MR3289212
 - [29] Walter Rudin, *Principles of mathematical analysis*, 2nd ed., McGraw-Hill Book Co., New York, 1964. MR0166310
 - [30] Joseph H. Silverman, *Abstract algebra—an integrated approach*, Pure and Applied Undergraduate Texts, vol. 55, American Mathematical Society, Providence, RI, [2022] ©2022. MR4423367
 - [31] Joanne E. Snow, *Views on the real numbers and the continuum*, Rev. Mod. Log. **9** (2001/03), no. 1-2, 95–113. MR2040859
 - [32] Michael Spivak, *Calculus*, Houston, TX: Publish or Perish, 2008 (English).
 - [33] Yitang Zhang, *Bounded gaps between primes*, Ann. of Math. (2) **179** (2014), no. 3, 1121–1174, DOI 10.4007/annals.2014.179.3.7. MR3171761

Index

- \mathbb{C} , 369
 \mathbb{N} , 3
 \mathbb{N}_0 , 4
 Π , 13, 452
 \mathbb{Q} , 160
 \mathbb{Q}_+ , 163
 \mathbb{R} , 256
 \mathbb{R}_+ , 247
 Σ , 13, 452
 \mathbb{Z} , 82, 500
 \mathbb{Z}^* , 161
 \mathbb{Z}_n , 108
 \mathbb{Z}_p^* , 113, 137
 \emptyset , 431
 \exists , 425
 $\exists!$, 428
 \forall , 425
 \mathcal{D} , 506
 \mathcal{C} , 276
 z_p , 369
- Abel, Niels Henrik, 391, 449
Abel–Ruffini theorem, 391
abelian group, 450, **511**
 $a \equiv b \pmod{n}$, 105
absolute value
 of a complex number, 375
 of a real number, 250
 of an integer, 82
absolutely summable sequence
 complex, 406
 real, 335
accumulation point
- of a sequence, 321
achievable number, 96
acyclic graph, 23
addition
 of complex numbers, 369
 of natural numbers, 4
 of rational numbers, 162
addition principle, 66
adjacent points
 of a graph, 21
Adleman, Leonard, 152
admissible sequence of integers, 187, 272
Al-Khwarizmi, Muhammad bin Musa, 391
algebraic integer, 288
algebraic number, 285
altitude, 400
AM–GM inequality, 18
AM–GM–HM inequality, 20
ancestor
 of a vertex, 224
Archimedean property, 166, **259**
Archimedes of Syracuse, 238, 259
Argand, Jean-Robert, 368, 389
argument
 of a complex number, 381
arithmetic mean, 8, 18
arithmetic progression, 29
Artin, Emil, 149, 186
associative operation, 4, 107, 108, 244, 371,
 510
average
 of n numbers, 5
axiom of choice, **446**, 471

- backward summation, 452
 base case, 11
 Basel problem, 18
 Bernoulli's inequality, 28
 Bernoulli, Jacob, 28
 Bernstein, Felix, 469
 Bertrand, Joseph, 506
 Bézout's lemma, 90
 Bézout, Étienne, 90
b-fraction, 179
 as infinite series, 347
 finite, 178
 of rational numbers, 179
 of real numbers, 272
 periodic, 183
 bijection, 444
 binary code, 116
 binary fractions, 179
 binary operation, 510
 binary relation, 473
 binary tree
 infinite complete, 225
 binomial coefficient, 69
 binomial formula, 71, 373
 binomial theorem, 71
 bipartite graph, 26
 Bolzano, Bernard, 297, 328
 Bolzano–Weierstrass theorem, 321
 Bombelli, Rafael, 193, 368
 Borel, Émile, 322
 bounded above
 sequence, 293
 set, 255
 bounded below
 sequence, 293
 set, 255
 bounded sequence, 293
 bounded set, 255
 boundedness criterion
 for nonnegative sequences, 334
 Bourbaki, Nicolas, 82
 box principle, 6
 Brahmagupta, 81
b-rational number, 179
 Brocot, Achille, 225
 Brun, Viggo, 334
 Bunyakovsky, Viktor, 290
 Caesar cipher, 122, 126
 Calkin, Neil, 235
 Cantor dust, 287
 Cantor set, 276, 466
 Cantor's diagonal argument
 first one, 460
 second one, 464
 Cantor's theorem, 469
 Cantor, Georg, 241, 469, 471
 Cantor–Bernstein–Schröder theorem, 469
 Cardano's formula, 367
 Cardano, Gerolamo, 367
 cardinality
 of a finite set, 65, 455
 of arbitrary sets, 457
 Cartesian product
 of finitely many sets, 436
 of two sets, 435
 Cataldi, Pietro, 193
 Cauchy criterion
 for infinite sums, 332, 406
 Cauchy product
 of two infinite series, 410
 Cauchy sequence
 of complex numbers, 404
 of real numbers, 326
 Cauchy, Augustin-Louis, 206, 290, 328, 389
 ceiling function, 87
 centroid, 375
 Cesáro, Ernesto, 363
 chain
 in a poset, 484
 characteristic equation
 of a linear recurrence sequence, 41, 393
 Chebyshev, Pafnuty Lvovich, 289
 child
 of a vertex, 224
 Chinese remainder theorem, 131
 choice function, 446
 Chuquet, Nicholas, 207
 closed set, 365
 closed walk
 in a graph, 23
 cluster point
 of a sequence, 321
 cluster set
 of a sequence, 324
 code, 115
 codeword, 115
 codom(f), 438
 codomain
 of a function, 438
 Cohen, Paul J., 446, 471
 commutative group, 450, 511

- commutative operation, 2, 4, 107, 244, 432, **510**
comparison test
 for infinite series, 336
complement, 434
complete bipartite graph, 27
complete graph, 26
complex number, 369
complex plane, 369
composite number, 53
composition
 of functions, 443
 of permutations, 450
conclusion
 of a statement, 423
cone of positive elements, 247
congruent numbers, 105
congruent triangles, 485
conjugate complex number, 375
connected graph, 23
constructible polygon, 140
continued fraction
 finite, 194
 infinite, 204, **353**
continuum hypothesis, 471
contradiction, 424
contrapositive
 of an implication, 424
convergent
 of a finite continued fraction, 197
 of an infinite continued fraction, 353
convergent infinite sum, 330
convergent sequence
 of complex numbers, 401
 of real numbers, 299
converse
 of an implication, 423
coprime
 integers, 93
 natural numbers, 57
Costas array, 149
Costas, John P., 149
countably infinite set, 458
Cousin, Pierre, 323
cycle
 in a graph, 23
 of a permutation, 451
cycle graph, 27
cyclic notation
 of a permutation, 451
cyclic number, 186
de Moivre's theorem, 383
De Morgan's laws, 424, 435
De Morgan, Augustus, 424
decimal fraction
 finite, 174
 infinite, 176
Dedekind cut, 506
 irrational, 507
 rational, 507
Dedekind, Richard, 241, 469, 506
degree
 of a vertex, 22
denominator
 of a fraction, 160
Descartes, René, 368, 391
descendant
 of a vertex, 224
Diophantine equation, 94
 linear, 95
Diophantus of Alexandria, 94, 103
Dirichlet, Peter Gustav Lejeune, 114, 210, 323
discriminant
 of a quadratic polynomial, 391
disjoint sets, 432
distance between two words, 116
distributive law
 for modular operations, 108
 for sets, 432
 in \mathbb{C} , 372
 in \mathbb{N} , 4
 in \mathbb{R} , 244
divisibility, 46, 52, 85
division
 of rational numbers, 162
division with remainder, 46, 86
divisor, 85
divisors of zero, 514
 $\text{dom}(f)$, 438
domain
 of a function, 438
 of a relation, 473
dyadic rational number, 179
edge
 of a graph, 21
Egyptian fraction, 229
Einstein, Albert, xiii
element
 of a set, 430
empty set, 431
endpoints

- of an edge, 21
- ϵ -neighborhood
 - of a complex number, 401
 - of a real number, 298
- equivalence class, 480
- equivalence relation, 106, **479**
- Eratosthenes of Cyrene, 53
- Erdős, Paul, 232
- Erdős-Strauss conjecture, 232
- Euclid, 2
- Euclid's Elements, 62, 166, 259
- Euclid's lemma, 56
 - for integers, 94
- Euclidean algorithm, 59, 89
- Euclidean division, 46, 86
- Eudoxos, 239, 459
- Eudoxos of Cnidos, 259
- Euler number e , 317, **338**, 341, 360
- Euler's ϕ -function, 141
- Euler's corollary, 138
- Euler's formula, 411
- Euler's theorem, 145
- Euler's totient, 141
- Euler, Leonhard, 18, 63, 193, 238, 334, 351, 368, 389
- Euler-Mascheroni constant, 286, 345
- even number, 9
- even permutation, 451
- existence of infinite sums, 330
- exponential function
 - complex, 411
 - real, 270
- face, 24
- Farey sequence, 208
- Farey, John, 206
- Fermat number, **63**, 79, 139
- Fermat's last theorem, 63
- Fermat's little theorem, 136
- Fermat, Pierre de, 63, 102, 136
- Ferrari, Ludovico, 391
- Fibonacci number, 77, 353
- Fibonacci sequence, **39**, 202, 292, 312
- Fibonacci, Leonardo, 39, 230
- field, 113, 162, **244**, 514
- finite expansion, 178
- floor function, 48, 87
- Ford Sr., Lester Randolph, 220
- fraction in lowest terms, 168, 205
- function
 - between sets, 438
 - bijective, 444
 - injective, 444
 - one-to-one, 444
 - surjective, 444
- fundamental theorem
 - of algebra, 389
 - of arithmetic, 55
- Gauss, Carl Friedrich, 12, 103, 141, 297, 368, 389
- Gauss-Wantzel theorem, 141
- Gaussian integers, 514
- gcd, 57
- Gelfond, Alexander Osipovich, 286
- Gelfond-Schneider theorem, 286
- geometric mean, 8, 18
- geometric progression, 30
- geometric series, 331
- Germain, Sophie, 103
- Gilbert, Edgar, 149
- Gödel, Kurt, 446, 471
- Goldbach's conjecture, **64**, 427
- Goldbach, Christian, 63
- golden ratio, **40**, 60, 173, 202, 205
- Golomb ruler, 158
- Golomb, Solomon, 158
- graph, 21
- graph of a function, 439
- greatest common divisor, 57, 89
 - of n numbers, 100
- greatest lower bound
 - of a set in \mathbb{R} , 261
- Green, Ben, 64
- group, 108, 113, **449**, 511
- Halmos, Paul, xiii
- handshaking lemma, 22
- Hardy, Godfrey Harold, 509
- harmonic mean, 20
- harmonic series, 333
- Haros, Charles, 206
- Heaviside, Oliver, 442
- height
 - of a rooted tree, 224
- Heine, Eduard, 241, 322
- Heine-Borel covering theorem, 323
- Hermite, Charles, 286, 340, 341
- Heron of Alexandria, 317
- Heron sequence, 313, 317
- hexadecimal fractions, 179
- Hilbert's hotel, 458
- Hilbert, David, 286, 340, 341, 458, 471
- homomorphism

- between groups, 512
- Hui, Yang, 70
- Hurwitz, Adolf, 211
- Huygens, Christiaan, 193
- identity map
 - of a set, 444
- image
 - of a function, 439
 - of a set, 441
- imaginary part
 - of a complex number, 371
- implication
 - of a statement, 423
- incidence matrix, 22
- incident vertex, 21
- indeterminate form
 - of limits, 306
- induction
 - mathematical, 11
- induction hypothesis, 11
- induction step, 11
- inductive subset of \mathbb{R} , 283
- infimum
 - of a sequence, 295
 - of a set in \mathbb{R} , 261
 - of an unbounded set, 262
- infinite point
 - of the Cantor set, 279
- infinite series, 329
- infinite sum, 330
- injection, 444
- integer numbers, 82
- integer part
 - of a division, 86
- integers mod n , 105
- integral test
 - for infinite series, 341
- intersection
 - of two sets, 432
- intervals
 - of real numbers, 250
- inverse element, 511
- inverse function, 447
- irrational number, 168
- irreducible fraction, 168, 205
- ISBN, 114
- isomorphism
 - between groups, 513
- Jones, William, 238
- Kepler, Johannes, 312
- Khayyam, Omar, 70
- Klein's 4-group, 512
- Klein, Felix, 512
- Koch snowflake, 362
- Koch, Helge von, 362
- k -subset, 69
- Lagrange, Joseph Louis, 147, 193
- Lamé, Gabriel, 60
- Lambert, Johann Heinrich, 193, 240
- Latin square, 116
- Latin squares
 - mutually orthogonal, 120
- lattice point, 103, 173, 212
- lcm, 65
- Le Blanc, Auguste Antoine, 103
- leaf
 - in a tree, 23
- least common multiple, 65
- least element principle
 - for subsets of \mathbb{N} , 34, 108
- least upper bound
 - of a set in \mathbb{R} , 257
- Lebesgue, Henri, 323
- left-child
 - of a vertex, 224
- left-inverse function, 447
- Legendre symbol, 151
- Legendre, Adrien-Marie, 151
- Leibniz's theorem, 343
- Leibniz, Gottfried Wilhelm, 343
- length
 - of a cycle, 23
 - of a face, 24
 - of a path, 23
 - of a walk, 23
- level
 - of a vertex, 224
- Liber Abaci, 39, 230
- limit
 - of a complex sequence, 401
 - of a real sequence, 298
- Lindemann, Ferdinand von, 286
- linear recurrence
 - of second order, 40, 393
- Liouville, Joseph, 286
- long prime, 186
- loop
 - of a graph, 21
- lower bound
 - of a set, 255

- lower limit
 - of a sequence, 318
- Lucas numbers, 42
- Lucas, François Édouard Anatole, 42
- main branch
 - of complex logarithm, 414
- maximal element
 - of a poset, 478
- maximum
 - of n numbers, 254
 - of a poset, 478
- mediant
 - of two fractions, 206
- Mengoli, Pietro, 18
- Mersenne prime, 62, 63
- Mersenne, Marin, 62
- method of exhaustion, 239
- minimal element
 - of a poset, 478
- minimum
 - of n numbers, 254
 - of a poset, 478
- Mirzakhani, Maryam, 367
- modular equation, 124
- modulo n residue, 105
- modulus
 - of a complex number, 375
- modulus of congruence, 105
- Moivre, Abraham de, 368, 383
- monotone convergence theorem, 311
- Montessori, Maria, xiii
- multiple, 85
- multiple edge
 - of a graph, 21
- multiplication
 - of complex numbers, 369
 - of natural numbers, 4
 - of rational numbers, 162
- multiplicative inverse
 - modulo n , 110
- multiplicative order, 148
- $N(n)$, 120
- natural numbers, 3, 495
- $\binom{n}{k}$, 69
- n choose k, 69
- negation
 - of a statement, 422
- negative Dedekind cut, 507
- negative integer, 3
- negative rational number, 164
- negative real number, 247
- neighborhood
 - of a point in \mathbb{R} , 298
- nested interval theorem, 325
- Newton, Isaac, 343
- n -factorial, 28, 67
- Noy, Marc, 76
- numerator
 - of a fraction, 160
- odd number, 9
- odd permutation, 451
- open set, 324, 364
- order
 - in \mathbb{N} , 4
 - in \mathbb{Q} , 164
 - in \mathbb{R} , 247
 - in \mathbb{Z} , 83
- order completeness
 - of \mathbb{R} , 255
- order of $a \in \mathbb{Z}_p$ modulo p , 148
- order relation, 476
- ordered field, 163
- ordered pair
 - of two elements, 435
- $\text{ord}_p(a)$, 148
- Oresme, Nicole, 333
- Orlin, Ben, 1
- orthocenter, 400
- Ostrowski, Alexander, 389
- $p \wedge q$, p , and q , 422
- parent
 - of a vertex, 224
- partial order, 52
 - on a set, 476
- partial sum, 330, 405
- partite sets
 - of a graph, 26
- Pascal's triangle, 70
- Pascal, Blaise, 70
- path
 - in a graph, 23
- peak point, 320
- Peano's axioms, 495
- Peano, Giuseppe, 495
- perfect number, 62
- periodic fractional representation, 184
- permutation
 - of order n , 450
 - on a set A , 449
- permutation matrix, 149

- Perron, Oskar, 193
Pick's theorem, 212
Pick, Georg Alexander, 212
pigeonhole principle, 6, 184
Pingala, Acharya, 3, 39, 70
planar graph, 24
polar representation
 of a complex number, 379
Polignac, Alphonse de, 64
Pólya, George, xiii, 59
 $p \vee q$, p or q , 423
poset, 476
positive Dedekind cut, 507
positive integer, 3
positive rational number, 163
power
 of an integer, 82
power function, 270
power of positive numbers
 w.r.t. rational numbers, 268
 w.r.t. real numbers, 269
power set, 435
pre-image
 of a set, 441
predecessor
 of a natural number, 496
predicate, 425
prime
 full reptend, 186
 full-period, 186
prime number, 53
primitive root modulo p , 148
primitive root of unity, 395
primitive triangle, 213
principle of inclusion and exclusion, 73
principle of induction, **10**, 16, 491
product
 of two Dedekind cuts, 509
proof
 by cases, 426, 487
 by contradiction, 211, **425**, 488, 490
 by contrapositive, 490
 by induction, 491
 direct, 486
Ptolemy, Claudius, 399
Pythagoras, 8
Pythagorean theorem, 2
Pythagorean triples, 99
quadratic reciprocity theorem, 152
quadratic residue modulo p , 151
quotient, 46, 86
quotient set
 w.r.t. an equivalence relation, 482
range
 of a function, 439
 of a relation, 473
ratio test
 for infinite series, 337
rational Cauchy sequence, 328
rational numbers, 160
real numbers, 256
real part
 of a complex number, 371
reduced fraction, 168, 205
relation, 473
 anti-symmetric, 475
 reflexive, 475
 symmetric, 475
 transitive, 475
relatively prime
 integers, 93
 natural numbers, 57
remainder, 46, 86
reordering finite sums, 452
representation in base b
 of a natural number, 47
reverse triangle inequality
 for complex numbers, 378
 for real numbers, 252
Rhind Papyrus, 29, 78, 229
right-child
 of a vertex, 224
right-inverse function, 447
ring, 113, 513
ring of integers modulo n , 105
Rivest, Ron, 152
Roman numerals, 45
root
 of a tree, 223
root test
 for infinite series, 352
roots
 of complex numbers, 386
 of positive real numbers, 267
roots of unity, 386
RSA encryption algorithm, 152
Ruffini, Paolo, 391
sandwich theorem, 304
Schinzel, Andrzej, 232
Schneider, Theodor, 286
Schönflies, Arthur Moritz, 323

- Schröder, Ernst, 469
 Schwarz, Karl Hermann Amandus, 290
 sequence
 absolutely summable, 335
 alternating, 293
 bounded, 293
 convergent, 298
 decreasing, 292
 divergent, 299
 increasing, 292
 monotone, 293
 nondecreasing, 292
 nonincreasing, 292
 of complex numbers, 400
 of real numbers, 291
 properly divergent, 302
 summable, 330
 unbounded, 293
 series representation
 of real numbers, 347
 set, 430
 countable, 458
 denumerable, 458
 finite, 65, 455
 infinite, 457
 uncountable, 458
 Shamir, Ron, 152
 shift of the index, 452
 Sidon set, 158
 Sidon, Simon, 158
 Sierpiński carpet, 281
 Sierpiński, Wacław, 232, 281
 sieve of Eratosthenes, 53
 similar triangles, 485
 simple graph, 21
 snail and rubber band, 344
 Sophie Germain primes, 103
 stereographic projection, 463
 Stern, Moritz, 225
 Stern–Brocot tree, 225
 Steven, Simon, 391
 Stirling's formula, 308
 Stirling, James, 308
 Straus, Ernst, 232
 strong induction, 36
 subsequence, 319
 subset, 431
 subtraction
 of rational numbers, 162
 of real numbers, 244
 successor
 of a natural number, 496
 successor set, 499
 sum
 of two Dedekind cuts, 508
 summable sequence, 330
 Sun-tzu, 130
 supremum
 of a sequence, 295
 of a set in \mathbb{R} , 258
 of an unbounded set, 262
 supremum is attained, 258
 surjection, 444
 Sylvester, James Joseph, 96, 141
 symmetric difference
 of sets, 433
 Tao, Terry, 64
 Tarry, Gaston, 121
 tautology, 424
 Tennenbaum, Stanley, 170
 Thales of Miletus, x, 5
 theorem of Eudoxos, 259
 totient function, 141
 transcendental number, 285
 transposition, 451
 tree, 23, 223
 binary, 224
 rooted, 223
 triangle inequality
 for complex numbers, 377
 for real numbers, 252
 triangular numbers, 12
 twin prime conjecture, 64
 twin primes, 64
 unbounded face, 24
 union of sets, 432
 unit
 of a binary operation, 510
 of a ring, 514
 unit fraction, 229
 upper bound
 of a set, 255
 upper limit
 of a sequence, 318
 Vandermonde's identity, 77
 Vandermonde, Alexandre-Théophile, 77
 vanishing condition
 for summable sequences, 333
 Venn diagram, 432
 vertex

- of a graph, 21
- vertical line property, 440
- Viète, François, 391
- von Neumann, John, 499
- walk
 - in a graph, 23
- Wallis, John, 193
- Wantzel, Pierre, 141
- Waring, Edward, 147
- Weierstrass, Karl, 241, 297
- Weil, André, 229
- well-ordering property
 - of natural numbers, **34**, 108
- Wessel, Caspar, 368
- Wiles, Andrew, 102
- Wilf, Herbert, 235
- Wilson's theorem, 147
- Wilson, John, 147
- Zeno of Elea, 32
- Zeno's paradox, 32
- Zermelo's comparability theorem, 471
- Zermelo, Ernst, 446, 471
- zeroes
 - of a complex polynomial, 389
 - of a complex quadratic equation, 391
 - of a real quadratic equation, 393

Selected Published Titles in This Series

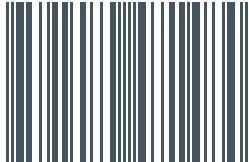
- 58 **Sebastian M. Cioabă and Werner Linde**, A Bridge to Advanced Mathematics, 2023
- 57 **Meighan I. Dillon**, Linear Algebra, 2023
- 55 **Joseph H. Silverman**, Abstract Algebra, 2022
- 54 **Rustum Choksi**, Partial Differential Equations, 2022
- 53 **Louis-Pierre Arguin**, A First Course in Stochastic Calculus, 2022
- 52 **Michael E. Taylor**, Introduction to Differential Equations, Second Edition, 2022
- 51 **James R. King**, Geometry Transformed, 2021
- 50 **James P. Keener**, Biology in Time and Space, 2021
- 49 **Carl G. Wagner**, A First Course in Enumerative Combinatorics, 2020
- 48 **Róbert Freud and Edit Gyarmati**, Number Theory, 2020
- 47 **Michael E. Taylor**, Introduction to Analysis in One Variable, 2020
- 46 **Michael E. Taylor**, Introduction to Analysis in Several Variables, 2020
- 45 **Michael E. Taylor**, Linear Algebra, 2020
- 44 **Alejandro Uribe A. and Daniel A. Visscher**, Explorations in Analysis, Topology, and Dynamics, 2020
- 43 **Allan Bickle**, Fundamentals of Graph Theory, 2020
- 42 **Steven H. Weintraub**, Linear Algebra for the Young Mathematician, 2019
- 41 **William J. Terrell**, A Passage to Modern Analysis, 2019
- 40 **Heiko Knospe**, A Course in Cryptography, 2019
- 39 **Andrew D. Hwang**, Sets, Groups, and Mappings, 2019
- 38 **Mark Bridger**, Real Analysis, 2019
- 37 **Mike Mesterton-Gibbons**, An Introduction to Game-Theoretic Modelling, Third Edition, 2019
- 36 **Cesar E. Silva**, Invitation to Real Analysis, 2019
- 35 **Álvaro Lozano-Robledo**, Number Theory and Geometry, 2019
- 34 **C. Herbert Clemens**, Two-Dimensional Geometries, 2019
- 33 **Brad G. Osgood**, Lectures on the Fourier Transform and Its Applications, 2019
- 32 **John M. Erdman**, A Problems Based Course in Advanced Calculus, 2018
- 31 **Benjamin Hutz**, An Experimental Introduction to Number Theory, 2018
- 30 **Steven J. Miller**, Mathematics of Optimization: How to do Things Faster, 2017
- 29 **Tom L. Lindstrøm**, Spaces, 2017
- 28 **Randall Pruim**, Foundations and Applications of Statistics: An Introduction Using R, Second Edition, 2018
- 27 **Shahriar Shahriari**, Algebra in Action, 2017
- 26 **Tamara J. Lakins**, The Tools of Mathematical Reasoning, 2016
- 25 **Hossein Hosseini Giv**, Mathematical Analysis and Its Inherent Nature, 2016
- 24 **Helene Shapiro**, Linear Algebra and Matrices, 2015
- 23 **Sergei Ovchinnikov**, Number Systems, 2015
- 22 **Hugh L. Montgomery**, Early Fourier Analysis, 2014
- 21 **John M. Lee**, Axiomatic Geometry, 2013
- 20 **Paul J. Sally, Jr.**, Fundamentals of Mathematical Analysis, 2013
- 19 **R. Clark Robinson**, An Introduction to Dynamical Systems: Continuous and Discrete, Second Edition, 2012
- 18 **Joseph L. Taylor**, Foundations of Analysis, 2012
- 17 **Peter Duren**, Invitation to Classical Analysis, 2012

For a complete list of titles in this series, visit the
AMS Bookstore at www.ams.org/bookstore/amstextseries/.

Most *introduction to proofs* textbooks focus on the structure of rigorous mathematical language and only use mathematical topics incidentally as illustrations and exercises. In contrast, this book gives students practice in proof writing while simultaneously providing a rigorous introduction to number systems and their properties. Understanding the properties of these systems is necessary throughout higher mathematics. The book is an ideal introduction to mathematical reasoning and proof techniques, building on familiar content to ensure comprehension of more advanced topics in abstract algebra and real analysis with over 700 exercises as well as many examples throughout. Readers will learn and practice writing proofs related to new abstract concepts while learning new mathematical content. The first task is analogous to practicing soccer while the second is akin to playing soccer in a real match. The authors believe that all students should practice and play mathematics.

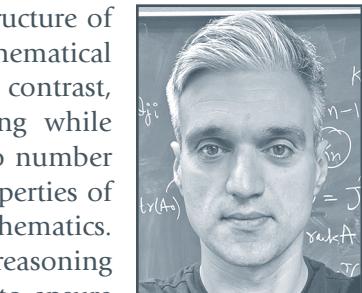
The book is written for students who already have some familiarity with formal proof writing but would like to have some extra preparation before taking higher mathematics courses like abstract algebra and real analysis.

ISBN 978-1-4704-7148-4

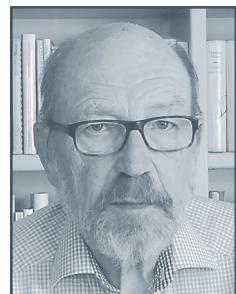


9 781470 471484

AMSTEXT/58



Courtesy of Sebastian M. Cioba.



Courtesy of Werner Linde.



For additional information
and updates on this book, visit
www.ams.org/bookpages/amstext-58



www.ams.org



This series was founded by the highly respected
mathematician and educator, Paul J. Sally, Jr.