Undergraduate Texts in Mathematics
Readings in Mathematics

UTM

Gabor Toth

Elements of Mathematics

A Problem-Centered Approach to History and Foundations

Springer

Undergraduate Texts in Mathematics

Undergraduate Texts in Mathematics

Readings in Mathematics

Undergraduate Texts in Mathematics are generally aimed at third- and fourth-year undergraduate mathematics students at North American universities. These texts strive to provide students and teachers with new perspectives and novel approaches. The books include motivation that guides the reader to an appreciation of interrelations among different aspects of the subject. They feature examples that illustrate key concepts as well as exercises that strengthen understanding.

For further volumes:
http://www.springer.com/series/666
and
http://www.springer.com/series/4672

Gabor Toth

Elements of Mathematics

A Problem-Centered Approach to History and Foundations

Springer

Gabor Toth
Department of Mathematics
Rutgers University-Camden
Camden, NJ, USA

ISSN 0172-6056 ISSN 2197-5604 (electronic)
Undergraduate Texts in Mathematics
ISBN 978-3-030-75050-3 ISBN 978-3-030-75051-0 (eBook)
https://doi.org/10.1007/978-3-030-75051-0

Mathematics Subject Classification: 26AXX, 11BXX, 51F05, 12DXX, 01AXX

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

To my children
Evelyn, Isabel, Gerald,
Gregory, Gabriel, and Gerda.

Preface

"If you're teaching a class, you can think about elementary things that you know very well. These things are kind of fun and delightful. It doesn't do any harm to think them over again. Is there a better way to present them? Are there any new problems associated with them? Are there any new thoughts you can make about them?"
Richard P. Feynman (1918–1988)

Why This Book?

This textbook aims for a rigorous, precise, and transparent presentation of mathematics before the advent of calculus. In developing naïve and axiomatic theories alike, and with geometry and algebra hand in hand, the text takes a new and fresh look at many a mathematical concept, never losing sight of the importance of intuition, and the ultimate quest for mathematical rigor.

Every experienced instructor knows that curious students always ask many questions. This book is written for them, the inquisitive and demanding readers who are seeking real challenge. Questions should always be encouraged and welcomed; as Francis Bacon (1561–1616) put it, *"Who questions much, shall learn much, and retain much."* In this book we answer many: What are the foundations of mathematics? Why did the Sumerians and the Babylonians chose sexagesimal arithmetic? What is a real number? What is the meaning of irrational powers? What is metric geometry? Why is the Pythagorean Theorem important in Archimedes' approximation of π? How much did the ancient Greeks know about conics? Why do we have different approaches to exponentiation?

One of the primary goals of this book is to offer an honest and in-depth text for the readers. Its appeal rests in the clarity of the gradually and carefully built up material and the transparency of the explanations; the emphasis on interconnections among seemingly unrelated topics (in algebra, geometry, number theory, etc.); correct and

unglossed answers to many fundamental questions that the student may ask; and intriguing historical notes based on most recent scholarship.
Throughout the entire book we insist on elementary approach, and leisurely pace, taking many side tours when opportunities arise. The text is sprinkled with a variety of thought-provoking examples, often inspired by problems posed in mathematical contests around the world.
There are over 150 challenging exercises at the end of the sections. A solutions manual can be found in the author's website:
https://math.camden.rutgers.edu/faculty/gabor-toth/

Audience

This book is intended to serve: (1) talented high school students in training for regional, national, and international mathematical contests; (2) college seniors with a certain level of mathematical maturity to better prepare them to graduate school; and (3) leaders of mathematical circles who wish to enrich and deepen their student's knowledge and understanding of mathematics beyond the standard textbooks.

(1) Various parts of this book have been used by the author in his mathematics contest-training course for high school students in the Princeton Campus of the Art of Problem Solving Academy. A contest preparation course for these students should cover only parts of Sections 1.3, 2.1–2.3, 3.1–3.4 and 6.2–6.7, 7.4–7.5, and should focus on problem solving strategies without much theoretical material or proofs. Within the main text in these sections, there are a total of 123 worked out and challenging examples, and, in addition, these sections end with 71 additional exercises. These should provide enough material for a one-semester course.
(2) The latter part of the book can also be adopted for a senior capstone course in mathematics for advanced undergraduate students. In this capacity, the author used various parts of the text in the last 30 years as material for the capstone unit Mathematics Seminar at Rutgers University–Camden for graduating seniors. A typical college course should essentially cover Chapters 10–11 along with some preliminary material in Chapters 5 and 8, and with a nice balance between the theoretical material and various specific applications expounded in the exercises. Although mathematics seniors are expected to master basic precalculus concepts and understand how to work with limits, the instructor will need to spend time on recalling some preliminary material contained in Sections 5.2–5.5, and especially Sections 8.2–8.4 and 5.9 as preparatory to trigonometry in Chapter 11. The technically demanding Sections 10.2 and 11.7 could be bypassed and included only for exceptionally strong classes. The exercises in Chapters 10–11 are written for college seniors.

(3) The material for mathematical circles can be used for individual lectures highlighting topics of exceptional beauty. Assuming weekly sessions in an ordinary 14-week college semester, the lectures may cover Sections 2.4, 5.9, 9.5–9.6, 10.2, 11.7–11.8, and a specific session on the famous problem #6 in the International Mathematical Olympiad in 1988 with two solutions (along with background material in Section 8.4) in Examples 6.6.8 and 8.4.1.

The Historical Context

"The history of mathematics is one of the large windows through which the philosophic eye looks into the past ages and traces the line of intellectual development."
Florian Cajori (1859–1930)

It is fashionable to scatter historical notes throughout a book to place the material in historical context and to enlighten the text. To the surprise of the author, most of these books swarm with historical inaccuracies, fashionable but unverified anecdotes, and hearsay. For example, analyzing the writings of Cicero, Plutarch, and others, scholars nowadays have serious doubts whether Pythagoras of Samos ever did any mathematics, let alone discovered the theorem that is often named after him. Note, in contrast, that the biographer Diogenes Laërtius (3rd century CE), quoting Apollodorus, explicitly attributes the Pythagorean Theorem to him, but his credibility is disputed as he notoriously relied on information that he failed to examine critically. Moreover, in many books it is usually and erroneously stated that René Descartes (1596–1650) invented Cartesian coordinates and analytic geometry. The origins of the use of coordinate systems can actually be traced back to antiquities, to Archimedes of Syracuse and Apollonius of Perga, and the invention of modern analytic geometry is due to Pierre de Fermat (posthumously published). Books often attribute the Pascal triangle to Blaise Pascal, but there is abundant evidence that it was known by the Indian mathematician Pingala well over 2000 years ago in the Vedic period (and independently by Al-Karajī and Omar Khayyám in Persia and Jia Xian and Yang Hui in China several centuries before Pascal). Moreover, references (by Nilakantha Somayaji in his *Tantrasanghara*) to the lost works of the Indian mathematician Mādhava, the founder of the Kerala School of Astronomy and Mathematics, point to the fact that he could expand certain transcendental functions into power series, predating James Gregory, Brook Taylor, and Colin Maclaurin for more than two centuries. Last but not least, in books the role of Sir Isaac Newton and Leonhard Euler are often confused about the discovery of the properties of the natural exponential function; it is a little known fact that Newton considered (and explicitly stated that) calculus is an algebraic counterpart of arithmetic that deals with infinite decimals.

One of the special features of our book is that it is a myth breaker; it sets the historical records straight and gives precise references.

In Closing: Gelfand's Teaching Legacy

"From my long experience with young students all over the world, I know that they are curious and inquisitive and I believe if they have some clear material presented in a simple form, they will prefer this to all artificial means of attracting their attention - much as one buys books for their content and not for their dazzling jacket designs that engage only for the moment. The most important thing a student can get from the study of mathematics is the attainment of a higher intellectual level."
Israel M. Gelfand (1913–2009)

The four booklets of I.M. Gelfand and his collaborators, *Algebra*, *The Method of Coordinates*, *Functions and Graphs*, and *Trigonometry* (Birkhäuser, 2001, 2003, Dover 2002, 2011), are beautiful expositions on precalculus concepts. Gelfand's fifth and final book *Geometry* (Birkhäuser, 2020) in this sequence covers the classical geometries. These were conceived in the early 1960s to satisfy the need for improved mathematics education in high schools and colleges. Gelfand's brilliant exposition served as a benchmark throughout this book. In addition to his elegant writing style, many of his ideas play fundamental and influential roles here. For example, the author adopted his point of view on placing pivotal role on the AM-GM inequality in many extremal problems, and also using continuity of the exponential functions over the rationals to establish real exponentiation. The unfortunate drawback of Gelfand's booklets is that, even when put together, they cannot be adopted as a (continuous) text for an undergraduate college course. They are separate "gems" in mathematics, and can be viewed individually. His fifth book is well suited for a geometric course, but the content is separate from our present book.

Camden, NJ, USA Gabor Toth

Acknowledgment

Finally, it is the author's pleasure to record his thanks to those who gave their generous assistance in writing this book. Catherine Meehan, a former student and colleague, spent many hours in developing the illustrations. Several students of the author's mathematical contest training class at the Art of Problem Solving Academy at Princeton, always in fierce competition and sometimes in the middle of heated arguments, but always having the greatest fun in doing mathematics, (unknowingly) contributed to the many contest level examples throughout the book. Two of them, Bhaumik Mehta and Sri Sai Nandan Indla, suggested several improvements in the book. Several reviewers underwent the painstaking work of going through the manuscript and provided many critical remarks for improvements. Last but not least, the smooth and speedy publication of this book is due to a large extent to the always enthusiastic and unwavering support and advice of Loretta Bartolini, editor of Springer, New York.

Contents

Chapter 0
Preliminaries: Sets, Relations, Maps

"A set is a gathering together into a whole of definite, distinct objects of our perception or of our thought - which are called elements of the set."
Georg Cantor (1845–1918)

In this chapter we give an account on the foundations of mathematics: naïve and axiomatic set theory. We introduce here several concepts that will play principal roles later: The Least Upper Bound Property for ordered sets, relations, maps, infinite sequences, the principle of inclusion-exclusion, cardinality, and classes vs. sets. The reader familiar with these basic concepts may skip this chapter altogether as the primary goal here is to "set the stage" by introducing some fairly standard notations and recalling a few well-known facts. This chapter ends with a short optional[1] introduction to the Zermelo–Fraenkel axiom system. This is not intended to be a thorough exposition in axiomatic set theory; only to provide a glimpse into how set theory can be put onto a rigorous foundation.[2]

In general, **naïve theory** in mathematics is a term referring to a mathematical theory that employs natural language to describe its objects of study. Many terms in a naïve theory are not defined with mathematical rigor, and thus the theory is prone to "excesses," possibly leading to inconsistencies.

A naïve theory is not necessarily inconsistent, however. A naïve theory may be recast into an **axiomatic theory**[3] in which some loosely defined concepts turn into **undefined terms** or **primitives** whose existence and basic properties are postulated by **axioms**. Axioms are statements or assertions without any justification.

[1]Sections marked with asterisk contain some more challenging (and therefore optional) material than the main text.

[2]For a classical text on set theory including recent major advances, see Jech, T. *Set Theory*, 3rd ed. Springer, New York, 2002. Note that, for readers wishing to go deeper in some topics, additional recommended material is listed in the "Further Reading" at the end of the book.

[3]For contrast, a naïve theory is also called a **non-axiomatic** theory.

G. Toth, *Elements of Mathematics*, Undergraduate Texts in Mathematics,
https://doi.org/10.1007/978-3-030-75051-0_0

In axiomatic theory, every subsequent assertion about primitives, called **theorems**,[4] must be proved as rigorous and logical consequences of the axioms and previously proved theorems. Other loosely defined notions of a naïve theory turn into **formal definitions**. These establish new object names for complex combinations of primitives and previously formally defined terms.

0.1 Sets

In **naïve set theory** the concept of a **set** is **undefined**. A set is "described" as a collection of "definite, distinct objects."[5] (See the epitaph of this chapter.) Sets are usually denoted by uppercase letters of the English alphabet.[6]

Naïve set theory **postulates** a fundamental relation between an object and a set. If this relation exists between an object x and a set X, then we say that x is an **element**, or a **member**, of the set X, or that x **belongs to** the set X, and write $x \in X$. Thus, the objects that belong to a set are called the elements, or members, of the set.

Whenever feasible, a generic element of a set will be denoted by the corresponding lowercase letter. Thus, as above, x is an element of a set X, and a is an element of a set A, and so on.

The negation of the relation $x \in X$, x is **not an element** of X (or x does not belong to X, etc.), is denoted by $x \notin X$.

History
The German word "Menge," translated as "set," or "aggregate," in English, appeared first in *The Paradoxes of the Infinite* (German *Paradoxien des Unendlichen*) of the Bohemian mathematician Bernard Bolzano (1781 – 1848). As many of his works, this was published posthumously in 1851 by František Přihonský, Bolzano's student and friend.
The special mathematical symbol $\in$ was introduced by Giuseppe Peano (1858 – 1932) in 1889 as the first letter of the Greek word $\epsilon\sigma\tau\iota$ for "is." Typographically, it is a derivation, not the same as the Greek epsilon ϵ or its variant ε.

Specific sets that play fundamental roles in mathematics are denoted by special letters or symbols. The sets of all **natural numbers**, **integers**, **rational numbers**, and **real numbers** are denoted, respectively, by $\mathbb{N}$ (from the word "natural," or the German "natürlich"), $\mathbb{Z}$ (from the German "Zahl," number), $\mathbb{Q}$ (from the Italian "quoziente" by Peano in 1895), and $\mathbb{R}$ (from the word "real"), respectively. In this chapter we first discuss these number sets naïvely, and in the next chapter axiomatically.

History
Modern set theory was initiated in the 1870s by Georg Cantor (1845 – 1918) and Richard Dedekind (1831 – 1916). Cantor was aware of some of the inconsistencies and paradoxes of his naïve set

[4] Or propositions, lemmas, etc.

[5] The words "collection," "family," "ensemble," "system" are only synonyms of the word "set;" therefore none of them serve as precise definitions.

[6] Not the Latin alphabet in which there are no separate letters for J, U or V.

theory but did not believe that they were serious. Due to these inconsistencies and paradoxes, a need for axiomatization of naïve set theory became more and more apparent. The first axiomatic system was put forward by the German mathematician Ernst Zermelo (1871 – 1953) in 1908. Subsequently, the German-Israeli mathematician Abraham Adolf Fraenkel (1891 – 1965) and the Norwegian Thoralf Albert Skolem (1887 – 1963) initiated some revisions of the Zermelo axioms, and added a new axiom. This new revised system became known as the Zermelo–Fraenkel set theory, ZF for short. We give a short account of the Zermelo–Fraenkel set theory in Section 0.5.

A set X is a **subset** of a set Y if every element that belongs to X also belongs to Y. The "inclusion" symbol $\subset$ is used to designate that a set is a subset of another. In other words, $X \subset Y$ means: $z \in X \Rightarrow z \in Y$.

Clearly, the inclusion as a (binary) relation[7] is **reflexive** in the sense that $X \subset X$ for any set X; that is, any set is a subset of itself. The inclusion is also **transitive** in the sense that $X \subset Y$ and $Y \subset Z$ imply $X \subset Z$.

As a specific example, we have $\mathbb{N} \subset \mathbb{Z} \subset \mathbb{Q} \subset \mathbb{R}$ for the number sets above in increasing generality.

We define two sets X and Y to be **equal** if they have the same elements. Therefore, a set is uniquely determined by its elements. Using the inclusion relation, $X = Y$ means that $X \subset Y$ and $Y \subset X$; in other words, $z \in X \Leftrightarrow z \in Y$. Thus, the inclusion as a relation is **antisymmetric**: $X \subset Y$ and $Y \subset X$ imply $X = Y$.

Remark As noted above, a subset may be equal to the set itself. Some authors use $X \subseteqq Y$ instead of $X \subset Y$, and specify $X \subsetneqq Y$ if X is a **proper** subset of Y, where proper means $X \neq Y$. This notation is somewhat cumbersome for our purposes; in cases of ambiguity we will explicitly indicate if a respective subset is proper.

In naïve set theory sets can be described **extensionally**, by "listing" their elements in braces (or curly brackets), or **intensionally**, that is, specifying a set of attributes for the elements.

The sets of natural numbers and integers (in decimal, base ten, representation) can be described extensionally as

$$\mathbb{N} = \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10, \ldots\},$$

and

$$\mathbb{Z} = \{0, \pm 1, \pm 2, \pm 3, \ldots\}.$$

These definitions are naïve because of the use of the ambiguous ellipsis ... which meant to indicate the continuation of the list in an "obvious way."

Continuing, the set of **rational** numbers is described as

$$\mathbb{Q} = \left\{ \frac{a}{b} \,\middle|\, a, b \in \mathbb{Z} \text{ and } b \neq 0 \right\}.$$

[7] Relations will be discussed in detail in Section 0.2.

Remark Some mathematicians count the integer 0 as a natural number, in view of providing a slightly more convenient setting for the Peano Axioms for $\mathbb{N}$ to be discussed in Section 1.1. We will occasionally adopt this and define

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

Given a set X, a **predicate** on X is a Boolean-valued function on X; that is, P is a statement concerning the elements of X which may be true or false depending on the elements of X. We write a predicate on X as a map[8] $P : X \to \{\text{true, false}\}$ with $P(x)$, $x \in X$, referred to as the (true-false) statement on the element x, the placeholder of the predicate P.

Given a predicate P on X, the set of elements $x \in X$ such that $P(x)$ is true is described intensionally as

$$\{x \in X \mid P(x)\}.$$

The predicate P on X is usually a Boolean expression, a logical statement which is either true or false on the elements of X. The predicate P may also spell out the ambient set X in which case X is omitted.

History
In his first proposed axiomatic set theory in 1908 Zermelo called the predicate P on X defining the set $\{x \in X \mid P(x)\}$ the "definite property" of the elements of X. The operational meaning of this concept is ambiguous. As noted above, Fraenkel and Skolem (independently) put forward a replacement of this term by introducing the concept of a **well-formed formula**. This will be discussed in detail in Section 0.5.

Example 0.1.1 For the set of integers $\mathbb{Z}$, let $P(x) = (x > 0)$, $x \in \mathbb{Z}$. Then we have

$$\mathbb{N} = \{n \in \mathbb{Z} \mid n > 0\}.$$

Similarly, for $P_0(x) = (x \geq 0)$, $x \in \mathbb{Z}$, we obtain

$$\mathbb{N}_0 = \{n \in \mathbb{Z} \mid n \geq 0\}.$$

A set that contains no elements is called the **empty set**, and it is denoted by $\emptyset$. Thus, the empty set $\emptyset$ is a set such that, for all x, we have $x \notin \emptyset$. The empty set is the subset of any set: $\emptyset \subset X$ for any set X.

History
In some axiomatic treatments, the existence of the empty set is postulated. In other treatments the existence (and uniqueness) of the empty set follows from other axioms. (See Section 0.5 again.)

Given a set X, the **power set** of X, denoted by $\mathcal{P}(X)$, is the set of all subsets of X. It is described as

[8] For maps, see Section 0.3.

$$\mathcal{P}(X) = \{Z \mid Z \subset X\}.$$

Equivalently: $Z \in \mathcal{P}(X) \Leftrightarrow Z \subset X$.

Example 0.1.2 We have $\mathcal{P}(\emptyset) = \{\emptyset\}$.[9]

The operations of **union** and **intersection** on two sets X and Y are defined, respectively, as

$$X \cup Y = \{z \mid z \in X \textbf{ or } z \in Y\} \quad \text{and} \quad X \cap Y = \{z \mid z \in X \textbf{ and } z \in Y\}.$$

They satisfy the following identities (with obvious proofs):

$X \cup X = X \cap X = X$	(idempotence)
$X \cup Y = Y \cup X$	(commutativity of the union)
$X \cup (Y \cup Z) = (X \cup Y) \cup Z$	(associativity of the union)
$X \cap Y = Y \cap X$	(commutativity of the intersection)
$X \cap (Y \cap Z) = (X \cap Y) \cap Z$	(associativity of the intersection)
$X \cap (Y \cup Z) = (X \cap Y) \cup (X \cap Z)$	(distributivity of union over intersection)
$X \cup (Y \cap Z) = (X \cup Y) \cap (X \cup Z)$	(distributivity of intersection over union)

Example 0.1.3 Let $\triangle[A, B, C]$ be a (non-degenerate) triangle[10] in the plane with (non-collinear) vertices A, B, C. Let S_A, S_B, S_C be the sides of the triangle opposite to the respective vertices A, B, C. Then, we have $S_A \cap S_B = \{C\}$, $S_B \cap S_C = \{A\}$, $S_C \cap S_A = \{B\}$, and $S_A \cap S_B \cap S_C = \emptyset$.

Two sets X and Y are called **disjoint** if $X \cap Y = \emptyset$.

The empty set is the **additive identity** with respect to the union, and it plays the role of the "zero" for the intersection; that is, for any set X, we have

$$X \cup \emptyset = X \quad \text{and} \quad X \cap \emptyset = \emptyset.$$

The (set-theoretic) **difference** of two sets X and Y is the set

$$X \setminus Y = \{z \mid z \in X \text{ and } z \notin Y\}.$$

The operation of difference satisfies the following properties:

[9] Note that $\emptyset$ is different from $\{\emptyset\}$. The former is the set with no elements; the latter is non-empty; it is the set whose only element is $\emptyset$.

[10] Real analytic plane geometry will be studied axiomatically (Birkhoff metric geometry) in Chapter 5.

$$X \setminus X = \emptyset \text{ and } X \setminus (X \setminus Y) = X \cap Y,$$

for any sets X and Y.

Example 0.1.4 For any two sets X and Y, the following are equivalent:

$$X \subset Y \iff X \cap Y = X \iff X \cup Y = Y \iff X \setminus Y = \emptyset.$$

Let U be a fixed set which we declare to be **universal** in the sense that, in a specific study, all sets considered are subsets of U. Equivalently, we restrict our study to elements of $\mathcal{P}(U)$. We define the **complement** of a set $X \in \mathcal{P}(U)$ as the difference $X^c = U \setminus X \in \mathcal{P}(U)$ (with respect to U). Clearly, $(X^c)^c = X$, $X \in \mathcal{P}(U)$, and $X \subset Y$ implies $Y^c \subset X^c$, $X, Y \in \mathcal{P}(U)$. In addition, the complement satisfies **De Morgan's identities** with respect to union and intersection:

$$(X \cup Y)^c = X^c \cap Y^c \text{ and } (X \cap Y)^c = X^c \cup Y^c, \quad X, Y \in \mathcal{P}(U).$$

History
De Morgan's identities can be traced back to Archimedes of Syracuse (c. 287 – 212 BCE), and can also be found in the works of the English Franciscan friar William of Ockham (c. 1287 – 1347), and the French philosopher Jean Buridan (c. 1300 – c. 1358/61). Augustus De Morgan (1806 – 1871) formulated these laws in terms of propositional (zeroth order) logic as valid rules of inference.

The operations of union and intersection can be extended to arbitrary collections of sets. Let $\mathcal{X}$ be a set of sets. Then we define the union and intersection of $\mathcal{X}$ by

$$\bigcup \mathcal{X} = \{x \mid x \in X \text{ for some } X \in \mathcal{X}\}$$

$$\bigcap \mathcal{X} = \{x \mid x \in X \text{ for all } X \in \mathcal{X}\}.$$

Clearly, we have $\bigcup \mathcal{P}(X) = X$ and $\bigcap \mathcal{P}(X) = \emptyset$ for any set X.

The set of sets $\mathcal{X}$ can be given as a labelled family $\mathcal{X} = \{X_a \mid a \in A\}$, where A is a so-called **index set**. In this case we write

$$\bigcup \mathcal{X} = \bigcup \{X_a \mid a \in A\} = \bigcup_{a \in A} X_a$$

$$\bigcap \mathcal{X} = \bigcap \{X_a \mid a \in A\} = \bigcap_{a \in A} X_a.$$

Returning to the complement (with respect to a universal set U), if $X_a \in \mathcal{P}(U)$ for all $a \in A$, then we have **De Morgan's identities**

$$\left(\bigcup_{a \in A} X_a\right)^c = \bigcap_{a \in A} X_a^c \text{ and } \left(\bigcap_{a \in A} X_a\right)^c = \bigcup_{a \in A} X_a^c.$$

The **Cartesian product** of two sets X and Y is defined as

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

As the notation indicates, (x, y) is the **ordered pair** of x (first) and y (second), as opposed to the **unordered pair** $\{x, y\} = \{y, x\}$ as a set. (In particular, in the Cartesian product $X \times X$, the elements (x, y) and (y, x) are different unless $x = y$.) In axiomatic set theory the existence of unordered and ordered pairs is guaranteed by an axiom; see Section 0.5.

History
"Cartesius" is "Renatus Cartesius," the Latinized name of René Descartes. Based on an appendix *La géométrie* of his famous work *Discours de la méthode* (published in 1637), it is usually and erroneously believed that he invented the coordinate system on the plane $\mathbb{R}^2 = \mathbb{R} \times \mathbb{R}$ as well as analytic geometry. The origins of the use of coordinate systems can actually be traced back to antiquities, to Archimedes and Apollonius of Perga (c. 262 – c. 190 BCE). Modern analytic geometry was inaugurated by Pierre de Fermat (1601 – 1655) in his *Introduction to Plane and Solid Loci*, a work written in 1629 but not published in Fermat's lifetime.

Example 0.1.5 For $X = \{a, b, c, d, e, f, g, h\}$ and $Y = \{1, 2, 3, 4, 5, 6, 7, 8\}$, the Cartesian product $X \times Y$ consists of 64 ordered pairs

$$(a, 1), (a, 2), \ldots, (a, 8),$$
$$(b, 1), (b, 2), \ldots, (b, 8),$$
$$\ldots$$
$$(h, 1), (h, 2), \ldots, (h, 8).$$

This set is used to describe the possible positions (squares) on a chessboard.

Example 0.1.6 Let $X = \{2, 3, 4, 5, 6, 7, 8, 9, 10, J, Q, K, A\}$ and $Y = \{\clubsuit, \diamondsuit, \heartsuit, \spadesuit\}$. The Cartesian product $X \times Y$ is consists of the $13 \times 4 = 52$ standard playing cards; X is the set of 13 ranks, and Y is the set of 4 suits:

$$(2, \clubsuit), (3, \clubsuit), \ldots, (9, \clubsuit), (10, \clubsuit), (J, \clubsuit), (Q, \clubsuit), (K, \clubsuit), (A, \clubsuit),$$
$$(2, \diamondsuit), (3, \diamondsuit), \ldots, (9, \diamondsuit), (10, \diamondsuit), (J, \diamondsuit), (Q, \diamondsuit), (K, \diamondsuit), (A, \diamondsuit),$$
$$(2, \heartsuit), (3, \heartsuit), \ldots, (9, \heartsuit), (10, \heartsuit), (J, \heartsuit), (Q, \heartsuit), (K, \heartsuit), (A, \heartsuit),$$
$$(2, \spadesuit), (3, \spadesuit), \ldots, (9, \spadesuit), (10, \spadesuit), (J, \spadesuit), (Q, \spadesuit), (K, \spadesuit), (A, \spadesuit).$$

Example 0.1.7 For any set X, the Cartesian product $X \times \emptyset$ is the empty set. Thus, in general, the equality $X \times Z = Y \times Z$ does not imply $X = Y$ unless Z is non-empty.

The Cartesian product satisfies the following properties:

$$X \times (Y \cup Z) = (X \times Y) \cup (X \times Z) \text{ and } X \times (Y \cap Z) = (X \times Y) \cap (X \times Z).$$

In an ordered pair $(x, y) \in X \times Y$, x is called the **first element** and y the **second element**. In a Cartesian product $X \times Y$, a **coordinate system** can be defined in the usual way. The choice of an element $(x_0, y_0) \in X \times Y$ specifies the **origin**, and the subsets

$$X \times \{y_0\} = \{(x, y_0) \mid x \in X\} \quad \text{and} \quad \{x_0\} \times Y = \{(x_0, y) \mid y \in Y\}$$

serve as the first and second **coordinate axes**. With respect to this coordinate system, an element $(x, y) \in X \times Y$ has first coordinate (x, y_0) and second coordinate (x_0, y).

The operation of Cartesian product can be naturally extended to finitely many sets $X_1, X_2, \ldots X_n$, $n \in \mathbb{N}$. By definition, the elements of the Cartesian product $X_1 \times X_2 \times \cdots \times X_n$ are ordered n-tuples $(x_1, x_2, \ldots, x_n)$ such that $x_1 \in X_1, x_2 \in X_2, \ldots, x_n \in X_n$.[11]

Exercises

0.1.1. Find a set A such that $A \not\subset \mathcal{P}(A)$.
0.1.2. Give an example of three sets A, B, and C such that $A \in B$, $B \subset C$ but $A \not\subset C$.

0.2 Relations

Let X and Y be (non-empty) sets. A (binary) **relation** R **from** X **to** Y is a subset of the Cartesian product $X \times Y$, that is $R \subset X \times Y$.[12] If $(x, y) \in R$, then we say that x **is** R**-related to** y, and write xRy. If $(x, y) \notin R$, then we say that x **is not** R**-related to** y, and we write $x \not R y$. If $X = Y$, then we say that R is a **relation on** X.

Still in naïve set theory, in this section we assemble a few facts about relations on a given set X. In the next section we will discuss the most prominent class of relations from a set X to a set Y, called maps or functions.

Relations with special properties play paramount roles in mathematics. Let X be a set and $R \subset X \times X$ be a relation on X. The specific properties that R may have (and used throughout this book) are given in the following list:

Reflexivity: For any $x \in X$, we have xRx;
Symmetry: For $x, y \in X$, xRy implies yRx;
Transitivity: For $x, y, z \in X$, xRy and yRz imply xRz;

[11] Axiomatically, this definition requires Peano's Axiom of Induction; see Section 1.1.

[12] Let $X_1, \ldots X_n$, $n \in \mathbb{N}$, be sets. An n**ary relation** R is a subset $R \subset X_1 \times \cdots \times X_n$. We will not need this concept.

Trichotomy: For any $x, y \in X$, exactly one of the following is true: xRy, $x = y$, yRx;
Antisymmetry: For $x, y \in X$, xRy and yRx imply $x = y$;
Totality: For any $x, y \in X$, either xRy or yRx.

Example 0.2.1 As noted in Section 0.1, the inclusion relation on the set of all subsets of a fixed set[13] is reflexive, antisymmetric and transitive.

An **equivalence relation** on a non-empty set X is a reflexive, symmetric and transitive relation on X. An equivalence relation on X is usually denoted by $\sim$.

Let $\sim$ be an equivalence relation on X. For $x \in X$, we define the **equivalence class** of x as $[x]_\sim = \{y \in X \mid x \sim y\}$. We say that x is a **representative** of the equivalence class $[x]_\sim$. By the properties of the equivalence relation, for $x, y \in X$, we have $[x]_\sim \cap [y]_\sim \neq \emptyset$ if and only if $x \sim y$ if and only if $[x]_\sim = [y]_\sim$. Indeed, if $z \in [x]_\sim \cap [y]_\sim$, then $x \sim z$ and $y \sim z$ so that, by symmetry, $(x \sim)z \sim y$, and, by transitivity, $x \sim y$. If $x \sim y$, $x, y \in X$, then $z \in [x]_\sim$ implies $x \sim z$, so that, by symmetry, $y \sim x(\sim z)$, and, by transitivity, $y \sim z$. This means that $z \in [y]_\sim$, and we obtain $[x]_\sim \subset [y]_\sim$. Reversing the roles of x and y (by symmetry), we arrive at $[x]_\sim = [y]_\sim$.

It follows that the equivalence classes partition the set X into mutually disjoint subsets. The set of equivalence classes is denoted by $X/\sim$, and it is called the **quotient** of X by the equivalence relation $\sim$.

Example 0.2.2 In plane geometry the relation being "parallel" (equal or disjoint) on the set of all lines is an equivalence relation. An equivalence class is called a **pencil of parallel lines**. In **projective plane geometry** a **projective point** is a point of the plane $\mathbb{R}^2$, or a pencil of parallel lines; the latter is also called an **ideal point**. A **projective line** is either a line in the plane $\mathbb{R}^2$ plus the ideal point (the pencil of parallel lines) that the line participates in, or the **ideal line** consisting of all ideal points. Incidence is defined by set membership. Clearly, in projective geometry any two distinct projective points are incident to a unique projective line; and every two distinct projective lines are incident to a unique projective point. Therefore, in projective plane geometry, (projective) points and lines play dual roles. Note finally that in projective plane geometry there are no parallel (projective) lines.

Example 0.2.3 The relations "similarity" and "congruence" on the set of all triangles in the plane are equivalence relations.

Example 0.2.4 On the set of integers $\mathbb{Z}$ having the same "parity" (even-odd) is an equivalence relation $\sim$. Here $a, b \in \mathbb{Z}$ have the same parity if and only if $a - b$ is even. There are two equivalence classes: the set of all even integers, $[0]_\sim$, and the set of all odd integers $[1]_\sim$.

A **strict total order** on a non-empty set X is a transitive, irreflexive,[14] and trichotomous binary relation on X. A strict total order on X is usually denoted by $<$

[13] Or on the **class** of all sets; see Section 0.5.

[14] To rule out "=" as a relation. Note also that the prefix ir- is a variant of the Latin negative prefix in- by assimilation for words that begin with "r" such as ir-rational, ir-reducible, ir-regular, etc.

(or $>$). A **total order** on a non-empty set X is a transitive, antisymmetric and total binary relation on X. A total order on X is usually denoted by $\leq$ (or $\geq$).

Let $<$ be a strict total order on X. We **define** a binary relation $\leq$ on X as follows: For $x, y \in X$, let $x \leq y$ if $x = y$ or $x < y$. It follows that $\leq$ is a total order on X.

Conversely, let $\leq$ be a total order on X. We **define** a binary relation $<$ on X as follows: For $x, y \in X$, let $x < y$ if $x \neq y$ and $x \leq y$. It follows that $<$ is a strict total order on X.

Thus, a strict total order and a total order mutually determine each other. In what follows, we will use these terms alternatively. If X has a (strict) total order, then we say that X is a **totally ordered set**.

Example 0.2.5 On the sets of natural numbers $\mathbb{N}$, integers $\mathbb{Z}$, rational numbers $\mathbb{Q}$ and real numbers $\mathbb{R}$, the usual strict order $<$ and order[15] $\leq$ are strict total order and total order relations, respectively.

Let X be a totally ordered set and $A \subset X$ a non-empty subset. An **upper bound** of A is an element $z \in X$ such that, for any $a \in A$, we have $a \leq z$. We say that A is **bounded above** if A has an upper bound. If A is bounded above, a **least upper bound** or **supremum** of A, denoted by $\sup A$, is an upper bound of A such that, for any upper bound z of A, we have $\sup A \leq z$. Clearly, the supremum may or may not exist. If the supremum exists, then it is unique (trichotomy), but it may not be **attained in** A; that is, $\sup A \in A$ may not hold.

Example 0.2.6 The set

$$A = \{1 - 1/n \mid n \in \mathbb{N}\} = \{0, 1/2, 2/3, 3/4, \ldots\}$$

of rational numbers has $\sup A = 1 \notin A$. Indeed, 1 is clearly an upper bound for A. Assume that $a/b \in \mathbb{Q}$, $a, b \in \mathbb{N}$, is an upper bound of A less than 1, that is, we have $0 < a < b$. This means that $1 - 1/n < a/b$ for all $n \in \mathbb{N}$. Rearranging, we obtain $n(b - a) < b$ for all $n \in \mathbb{N}$. That this is impossible (since $b - a > 0$) is intuitively obvious, and, rigorously, it is the consequence of the Archimedean Property of the natural numbers discussed at the end of Section 1.1.

In a similar vein, a **lower bound** of $A \subset X$ is an element $y \in X$ such that, for any $a \in A$, we have $a \geq y$. We say that A is **bounded below** if A has a lower bound. If A is bounded below, a **greatest lower bound** or **infimum** of A, denoted by $\inf A$, is a lower bound of A such that, for any lower bound y of A, we have $\inf A \geq y$. As before, the infimum may or may not exist. If $\inf A$ exists, then it is unique, but it may not be **attained in** A, that is, $\inf A \in A$ may not hold. Finally, we say that a non-empty set $A \subset X$ is **bounded** if it is bounded above and below.

Remark Let X be a totally ordered set. If a subset $A \subset X$ is defined by a predicate P on X, $A = \{x \in X \mid P(x)\}$, as in Section 0.1, then we will write the supremum and infimum of A as

[15] These order relations will be defined axiomatically in the forthcoming sections.

$$\sup A = \sup\{x \in X \mid P(x)\} = \sup_{\mathcal{P}(x)} x \quad \text{and} \quad \inf A = \inf\{x \in X \mid P(x)\} = \inf_{\mathcal{P}(x)} x.$$

Proposition 0.2.1 *Let X be a totally ordered set. Then the following are equivalent:*

I. For any non-empty subset $A \subset X$ which is bounded above, sup A *exists in X.*
II. For any non-empty subset $A \subset X$ which is bounded below, inf A *exists in X.*

Proof We will show that I implies II; the converse is analogous. Assume I holds. Let $A \subset X$ be non-empty and bounded below.

Let $B \subset X$ be the set of all lower bounds of A. By assumption, B is non-empty, and, by definition, it is bounded above, since any element in A is an upper bound of B. Thus, by I, sup B exists in X. We claim that sup B is the greatest lower bound of A.

First, if $b < \sup B$ for some $b \in X$, then b is not an upper bound of B so that b cannot be an element of A. Hence, for all $a \in A$, we have $a \geq \sup B$. This means that sup B is a lower bound of A.

Second, if $b \in X$ is a lower bound of A, then $b \in B$, and consequently, we have $b \leq \sup B$. Thus, sup B is the greatest lower bound of A. The proposition follows.

A totally ordered set X is said to have the **Least Upper Bound Property** if I (or II) of Proposition 0.2.1 holds. As we will see in Section 2.1, with respect to their natural orders, the set of rational numbers $\mathbb{Q}$ does not have the Least Upper Bound Property, while the set of real numbers $\mathbb{R}$ does.

Finally, as a much more restrictive property, a totally ordered set X is said to be **well-ordered** if, for **any** non-empty subset $A \subset X$, the infimum inf A exists and **belongs to** A.

As we will see in Section 1.1, the set of natural numbers $\mathbb{N}$ is well-ordered with respect to its natural total order. On the other hand, $\mathbb{Z}$, $\mathbb{Q}$ and $\mathbb{R}$ are not well-ordered with respect to their natural total orders.

Remark The **Well-Ordering Theorem** or **Zermelo's Theorem** states that every set can be well-ordered. This is, in fact, equivalent to the **Axiom of Choice**: Given a set A, for every collection of non-empty sets $\{X_a \mid a \in A\}$, there exists a set $\{x_a \mid a \in A\}$ such that $x_a \in X_a$, for every $a \in A$.

Exercise

0.2.1. Let A be a set of at least two elements. Show that the inclusion relation $\subset$ is not a total order on $\mathcal{P}(A)$.

0.3 Maps and Real Functions

A prominent class of relations is comprised by maps. We first introduce the relevant auxiliary concepts.

Let X and Y be sets, and consider a (binary) relation $R \subset X \times Y$ from X to Y. We define the **domain** of R as

$$X_R = \{x \in X \mid xRy \text{ for some } y \in Y\}.$$

Clearly, we have $R \subset X_R \times Y \subset X \times Y$; in fact, X_R is the smallest subset of X such that $R \subset X_R \times Y$. Oftentimes, it is convenient to **restrict** a relation R to its domain, and replace X with X_R.

The **range** of R is defined by

$$Y_R = \{y \in Y \mid xRy \text{ for some } x \in X\}.$$

The relation is called **surjective** if $Y_R = Y$. We have $R \subset X \times Y_R \subset X \times Y$; in fact, Y_R is the smallest subset of Y such that $R \subset X \times Y_R$. As before, we can replace Y by Y_R, and with this R becomes surjective.

A relation $R \subset X \times Y$ satisfies the **vertical intersection property** if R intersects every subset $\{x\} \times Y$, $x \in X$, at most once (exactly once if $X_R = X$). A relation $R \subset X \times Y$ is called **functional** if $X_R = X$ and R satisfies the vertical intersection property.

Functionality of R can be reformulated by saying that, for any $x \in X$, there is a **unique** $y \in Y$ such that xRy. To express the unique dependency of $y \in Y$ on $x \in X$ with xRy, we write $x \mapsto y$. This way R becomes a **map**[16] between the sets X and Y, that is, a specific relation that relates to each element $x \in X$ a unique element $y \in Y$. The map $x \mapsto y$, $x \in X$, $y \in Y$, associated with a functional relation R is symbolically denoted by $f : X \to Y$, with the element $y \in Y$ R-related to $x \in X$ written as $y = f(x)$.

At times it will be convenient to relax the condition $X = X_R$ in functionality, and define a map $f : X \to Y$ with domain $D_f \subset X$. Here D_f is the domain X_R of the relation R corresponding to f. In this more general case a map $f : X \to Y$ is called **total** if $D_f = X$; otherwise we have a **partial map** whose domain D_f is a proper subset of X. As for relations, oftentimes it is convenient to restrict a map $f : X \to Y$ to its domain and thereby obtain a total map. From now on, unless stated otherwise, we will tacitly assume that our maps are total.

The **range of the map** $f : X \to Y$, the range Y_R of the corresponding relation R, is denoted by

$$f(X) = \{y \in Y \mid y = f(x) \text{ for some } x \in X\} = \{f(x) \mid x \in X\} \subset Y.$$

The element $x \in X$ is unspecified and unconstrained within X, hence it is considered as an **independent variable** in X, also customarily called the **domain variable**. On the other hand, the **range variable** $y \in Y$ depends on x through f; hence it is called the **dependent variable**. This dependence is made explicit by the

[16] Some authors use the term **function** instead of **map**. Following widespread practice, we reserve the former only for maps whose range is a subset of the set of real numbers $\mathbb{R}$.

traditional notation $y = f(x)$, read "y is equal to f of x." We also say that $y = f(x)$ is the **value** of f at x.

For a map, the notation $y = f(x)$ with the dependence on x explicitly indicated has the clear advantage of being more specific than the symbolic $f : X \to Y$. On the other hand, the traditional notation $y = f(x)$ often does not indicate the relevant domain, and hence it either needs to be specified or determined.[17]

The functional relation R can be recovered from the respective map $f : X \to Y$ since

$$R = \{(x, y) \in X \times Y \mid y = f(x)\} = \{(x, f(x)) \mid x \in X\}.$$

In this context R is called the **graph** of the map $f : X \to Y$, and it is denoted by $G_f = R$.

An important (binary) operation on maps is **composition**. Given maps $f : X \to Y$ and $g : Y \to Z$, the composition[18] $g \circ f : X \to Z$ is defined by $(g \circ f)(x) = g(f(x))$, $x \in X$.

The identity map $\mathrm{id}_X : X \to X$ given by $\mathrm{id}_X(x) = x$, $x \in X$, is a right-identity under composition; that is, we have $f \circ \mathrm{id}_X = f$ for any map $f : X \to Y$. Similarly, the identity map $\mathrm{id}_Y : Y \to Y$ is a left-identity under composition; that is, we have $\mathrm{id}_Y \circ f = f$ for any map $f : X \to Y$.

We call a map $f : X \to Y$ **surjective** (or onto) if the corresponding relation R is surjective. Thus f is surjective if and only if $f(X) = Y$.

A map $f : X \to Y$ is called **injective** (or one-to-one) if $f(x) = f(x')$, $x, x' \in X$, implies $x = x'$. A map is injective if and only if the corresponding functional relation R satisfies the **horizontal intersection property**: The graph $G_f = R$ intersects every subset $X \times \{y\}$, $y \in Y$, at most once (exactly once if f is surjective).

Finally, a map is called **bijective** (or a **bijection**) if it is injective and surjective. A bijective map $f : X \to Y$ is also called a **one-to-one correspondence** between X and Y.

Given a map $f : X \to Y$, an **inverse** of f is a map $g : Y \to X$ such that $g \circ f = \mathrm{id}_X$ and $f \circ g = \mathrm{id}_Y$ hold.

An inverse of $f : X \to Y$ exists if an only if f is bijective. Indeed, if an inverse $g : Y \to X$ exists, then, for $x, x' \in X$, $f(x) = f(x')$ implies $x = g(f(x)) = g(f(x')) = x'$, so that f must be injective. Moreover, if $y \in Y$, then $x = g(y) \in X$ satisfies $f(x) = f(g(y)) = y$, so that f must be surjective. Thus, if the inverse of f exists, then f must be bijective.

Conversely, let $f : X \to Y$ be bijective. We define the map $g : Y \to X$ as follows. For $y \in Y$ let $g(y) \in X$ be an element x such that $f(x) = y$. Since f is surjective, $x = g(y)$ exists. Since f is injective, x is unique. Thus, $g : Y \to X$ is

[17] Some authors use the combined notation $X \ni x \mapsto y = f(x) \in Y$. We will not need this.

[18] There is a more general concept of composition of relations (called relative multiplication). Given sets X, Y, Z and relations $R \subset X \times Y$ and $S \subset Y \times Z$, the composition $S \circ R \subset X \times Z$, a relation from X to Z, is defined as follows: $(x, z) \in S \circ R$, $x \in X$, $z \in Z$, if there exists $y \in Y$ such that $(x, y) \in R$ and $(y, z) \in S$. We will not need this concept.

well-defined. With this, for $x \in X$ and $y \in Y$, we have $f(x) = y$ if and only if $g(y) = x$, and we obtain $g \circ f = \mathrm{id}_X$ and $f \circ g = \mathrm{id}_Y$.

The inverse of a map $f : X \to Y$ (if it exists) is uniquely determined by f. Clearly, if $g : Y \to X$ is the inverse of $f : X \to Y$, then f is also the inverse of g. Henceforth, we denote the inverse of a map $f : X \to Y$ by $f^{-1} : Y \to X$. With this we have $(f^{-1})^{-1} = f$.

If $f : X \to Y$ is a bijective map, then the graphs $G_f \subset X \times Y$ and $G_{f^{-1}} \subset Y \times X$ can be obtained from each other by the map $X \times Y \to Y \times X$ that swaps the coordinates,[19] that is, $(x, y) \mapsto (y, x)$, $x \in X$, $y \in Y$.

Example 0.3.1 Let X be a non-empty set and $Y = X \times \{0, 1\}$. Define the maps $f : X \to Y$, by $f(x) = (x, 0)$, $x \in X$, and $g : Y \to X$ by $g(x, y) = x$, $x \in X$, $y = 0, 1$. Then we have $g \circ f = \mathrm{id}_X$ but $f \circ g \neq \mathrm{id}_Y$.

On the other hand, if X and Y are finite sets with the same number of elements, and $f : X \to Y$ and $g : Y \to X$ such that $g \circ f = \mathrm{id}_X$, then we have $f \circ g = \mathrm{id}_Y$. Indeed, as above, $g \circ f = \mathrm{id}_X$ implies that $f : X \to Y$ is injective. Since X and Y have the same number of elements, $f : X \to Y$ must be surjective.[20] Hence, f is a bijection, and its inverse $f^{-1} : Y \to X$ exists. With this, we have $f \circ g = f \circ (g \circ f) \circ f^{-1} = f \circ f^{-1} = \mathrm{id}_Y$.

A map $f : X \to \mathbb{R}$ whose range is a set of real numbers is called a **real(-valued) function**. If the domain is also a set of real numbers, then f is called a **single-variable real function**. It is usually given by an equation $y = f(x)$, where $f(x)$ is a (real-valued) expression depending on the real indeterminate $x \in X \subset \mathbb{R}$. If the domain of a real function $f : X \to \mathbb{R}$ is a subset of the plane $\mathbb{R}^2$, the 3-space $\mathbb{R}^3$, etc., then f is called a **multivariate real function**. It is usually given by equations $z = f(x, y)$, $w = f(x, y, z)$, etc., where all the variables are real, and $f(x, y)$, $f(x, y, z)$, etc. are multivariate expressions in (x, y), (x, y, z), etc. in X.

For simplicity (and brevity) we will call all these **real functions**.

If a real function is given by equations $y = f(x)$, $z = f(x, y)$, etc., then the **domain of definition** of f is the domain of the expressions $f(x)$, $f(x, y)$, etc., that is, the largest set of real numbers $x \in \mathbb{R}$, points $(x, y) \in \mathbb{R}^2$ in the plane, etc. for which the expression f is defined.

History

One may contemplate that Hipparchus of Nicaea (c. 190 – c. 120 BCE), the first compiler of a trigonometric table, already had an implicit notion of what about eighteen centuries later in 1692 Gottfried Wilhelm Leibniz (1646 – 1716) called a "function." Credit should be given to Leibniz not only because in his works the concept of function appears explicitly but also because he used this term in many geometric settings.

In many examples maps and their variables are "named" using (uppercase and lowercase) letters from the English alphabet (f, g, F, G, r, v, t, etc.). This is convenient not only for referencing purposes but also in instances when the map or its variable(s) carry specific (usually geometric or physical) meaning. For example,

[19] This is the key property to define the inverse of a relation $R \subset X \times Y$ as $R^{-1} \subset Y \times X$ where $yR^{-1}x$, $x \in X$, $y \in Y$, if xRy. Once again, we will not need this.

[20] Albeit intuitively obvious, this will be shown rigorously as an easy application of Peano's Principle of Induction in Section 1.3; see Example 1.3.2.

$\mathcal{L}$ may denote the arc length of a circle depending on the domain variable $r > 0$, the radius of the circle, and v usually stands for the velocity of a point-mass moving along a line depending on the domain variable t, the time.

Once the concept of relation, and hence the concept of map, are defined, the intensional definition of a set $\{x \mid P(x)\}$ using a predicate P on X (Section 0.1) can be replaced by the concept of **indicator function** on X. This is the subject of the next example.

Example 0.3.2 Let X be a set. Consider a function $\chi : X \to \{0, 1\}$ with range the two-element set $\{0, 1\}$. Clearly, χ determines and is determined by the subset $A \subset X$ consisting of those elements $x \in X$ for which $\chi(x) = 1$. To indicate this, we set $\mathbf{1}_A = \chi$. We call $\mathbf{1}_A$ the **indicator function** (or characteristic function) of the subset $A \subset X$. To put this into a somewhat wider scope, we see that the map associating to a subset $A \subset X$ its indicator function $\mathbf{1}_A : X \to \{0, 1\}$ establishes a one-to-one correspondence between the power set $\mathcal{P}(X)$ and the set of all functions $\chi : X \to \{0, 1\}$.

This correspondence behaves well under intersection and union of subsets of X: If $A, B \subset X$, then we have[21]

$$\mathbf{1}_{A\cap B} = \min(\mathbf{1}_A, \mathbf{1}_B) = \mathbf{1}_A \cdot \mathbf{1}_B \text{ and } \mathbf{1}_{A\cup B} = \max(\mathbf{1}_A, \mathbf{1}_B) = \mathbf{1}_A + \mathbf{1}_B - \mathbf{1}_A \cdot \mathbf{1}_B.$$

Moreover, for $A \subset X$, we have $\mathbf{1}_{X\setminus A} = 1 - \mathbf{1}_A$.

If X is a finite set consisting of $n \in \mathbb{N}$ elements, then the power set $\mathcal{P}(X)$ consists of 2^n elements. Indeed, this is because the number of functions $\chi : X \to \{0, 1\}$ is 2^n since, for each $x \in X$, the value $\chi(x)$ has two choices, 0 or 1.

Amongst the infinite sets, the indicator function $1_{\mathbb{Q}} : \mathbb{R} \to \mathbb{R}$ of the set of rational numbers within $\mathbb{R}$ plays a prominent role. It is called the **Dirichlet function**:

$$1_{\mathbb{Q}}(x) = \begin{cases} 1 & \text{if } x \in \mathbb{Q} \\ 0 & \text{if } x \in \mathbb{R} \setminus \mathbb{Q}. \end{cases}$$

Example 0.3.3 Given a set X, a map $P : \{1, 2\} \to X$ (with domain the two-element set $\{1, 2\}$) is defined by specifying its two values $P(1)$ and $P(2)$. These are elements of X, and the order how they are listed is determined by the domain variable: $P(1)$ comes first, and $P(2)$ is the second. We thus have the ordered pair $(P(1), P(2))$ in the Cartesian product $X \times X$. Conversely, an ordered pair $(x_1, x_2) \in X \times X$ uniquely determines a map $P : \{1, 2\} \to X$ by setting $P(1) = x_1$ and $P(2) = x_2$.

In summary, to give a map $P : \{1, 2\} \to X$ amounts to specifying a point $(P(1), P(2))$ in the Cartesian product $X \times X$. In particular, for $X = \mathbb{R}$, a function $P : \{1, 2\} \to \mathbb{R}$ can be viewed as a point in the plane $\mathbb{R}^2$, the point being $(P(1), P(2))$.

[21] The notation here indicates that arithmetic operations in $\mathbb{R}$, such as addition, multiplication, etc., naturally carry over to the corresponding operations on real(-valued) functions; so that we can add, multiply, etc. real(-valued) functions.

Example 0.3.4 The previous example can be generalized to demonstrate that an **infinite sequence** of points in a set X can be interpreted as a map. Indeed, let $a : \mathbb{N} \to X$ be a map with domain $\mathbb{N}$, the set of natural numbers, and range X. We list the values of this map in the form of an (ordered) **infinite sequence** $(a_n)_{n\in\mathbb{N}} = (a(1), a(2), a(3), \ldots)$ of points in X. This sequence uniquely determines the map $a : \mathbb{N} \to X$. Conversely, if an infinite sequence $(a_1, a_2, a_3, \ldots)$ of points in X is given, then $a : \mathbb{N} \to X$ can be constructed by setting $a(n) = a_n$, $n \in \mathbb{N}$.

For the next example, note that the cube function $f : \mathbb{R} \to \mathbb{R}$, $f(x) = x^3$, $x \in \mathbb{R}$, is a strictly increasing and surjective real function, and thereby has an inverse, the cube root function given by $f^{-1}(x) = \sqrt[3]{x}$, $x \in \mathbb{R}$, which is also strictly increasing and surjective.[22] Although the strictly monotonic (and surjective) functions provide a large family of functions with inverses, there are many examples of invertible non-monotonic functions. The next example is an extreme case of this.

Example 0.3.5 Define the function $g : \mathbb{R} \to \mathbb{R}$ by

$$g(x) = \begin{cases} x^3 & \text{if } x \in \mathbb{Q} \\ -x^3 & \text{if } x \in \mathbb{R} \setminus \mathbb{Q}. \end{cases}$$

We claim that g has an inverse.

To show injectivity, let $x, x' \in \mathbb{R}$, and assume $g(x) = g(x')$. If $x, x' \in \mathbb{Q}$, then $x^3 = x'^3$ implies $x = x'$. Similarly, if $x, x' \in \mathbb{R} \setminus \mathbb{Q}$, then $-x^3 = -x'^3$ implies $x = x'$. Finally, if $x \in \mathbb{Q}$ and $x' \in \mathbb{R} \setminus \mathbb{Q}$, then $x^3 = -x'^3 = (-x')^3$ implies $x = -x'$. This cannot happen. Thus g is injective.

To show surjectivity, let $y \in \mathbb{R}$. Then $\sqrt[3]{y} \in \mathbb{R}$. If $\sqrt[3]{y} \in \mathbb{Q}$, then $g(\sqrt[3]{y}) = (\sqrt[3]{y})^3 = y$. If $\sqrt[3]{y} \in \mathbb{R} \setminus \mathbb{Q}$, then $g(-\sqrt[3]{y}) = -(-\sqrt[3]{y})^3 = y$. Surjectivity follows.

We conclude that g is bijective and therefore has an inverse.

Exercises

0.3.1. Let A and B be sets. Use the Axiom of Choice to show that there exists an injective map $f : A \to B$ if and only if there exists a surjective map $g : B \to A$.

0.3.2. Let A be a set. Show that an equivalence relation on A which is functional must be the identity id_A as a function.

0.3.3. Let A and B be finite sets. If B has 56 more subsets than A, then how many elements are in A and B?

[22] A function $f : X \to \mathbb{R}$ with $X \subset \mathbb{R}$, is **increasing** if, for $x, x' \in X$, $x < x'$ implies $f(x) \leq f(x')$. Replacing the last inequality sign with strict inequality we obtain the notion of **strictly increasing** function. The function f is **(strictly) decreasing** if its negative $-f$ is **(strictly) increasing**. Finally, f is called **(strictly) monotonic** if it is (strictly) increasing or decreasing. Note also that we treat here the cube root naïvely; it will be treated rigorously in Section 3.2.

0.4 Cardinality

We "identify" two sets via one-to-one correspondence. More precisely, we say that the set X has the same **cardinality** as the set Y if there is a one-to-one correspondence (bijection) $f : X \to Y$. The relation of having the same cardinality is an equivalence relation amongst sets.[23] Indeed, X has the same cardinality as itself via the identity map $\mathrm{id}_X : X \to X$. If X has the same cardinality as Y via $f : X \to Y$, then Y also has the same cardinality as X via the inverse $f^{-1} : Y \to X$. Finally, if X has the same cardinality as Y via $f : X \to Y$, and Y has the same cardinality as Z via $g : Y \to Z$, then X has the same cardinality as Z via the composition $g \circ f : X \to Z$.

We write $|X| = |Y|$ if X and Y have the same cardinality.

Remark A well-known example is the spectacle of a cavalry passing by. In a large crowd it may be hard to count exactly how many horses or horsemen are there, but there is a clear one-to-one correspondence between the set of horses and the set of horsemen; to each horseman there corresponds the respective horse. Clearly, the one-to-one correspondence no longer holds if there is horseman walking alone, or in an unlikely scenario of a stray horse.

A simple example for one-to-one correspondence between infinite sets, and thereby having the same cardinality, is furnished by writing the natural numbers as Hindu-Arabic and Roman numerals. Recall the set of Roman numerals

$$\{\mathrm{I}, \mathrm{II}, \mathrm{III}, \mathrm{IV}, \mathrm{V}, \mathrm{VI}, \mathrm{VII}, \mathrm{VIII}, \mathrm{IX}, \mathrm{X}, \mathrm{XI}, \ldots\},$$

where I=1, V=5, X=10, L=50, C=100, D=500, M=1000. Note also that, to avoid four identical Roman numerals to pile up (up to 4000), a subtractive notation is used; for example, IX= $10 - 1 = 9$ (instead of VIIII), XC= $100 - 10 = 90$ (instead of LXXXX), etc.) For example, the natural number 48 corresponds to XLVIII and 2021, our Gregorian calendar year, corresponds to MMXXI.

History
Leonardo Pisano Bigollo (c. 1175 – c. 1250), an Italian mathematician, is credited for advocating the Hindu-Arabic numeral system (notably the use of $0, 1, 2, \ldots, 9$ as digits and place value) in medieval Europe (as opposed to the clumsy Roman numeral system). During his extensive travels around the Mediterranean coast, meetings with many merchants, and learning about their systems of doing arithmetic, he realized the many advantages of the Hindu-Arabic numeral system. In 1202, he completed his book *Liber Abaci* (Book of Abacus or Book of Calculation) which popularized the Hindu-Arabic numerals in Europe. Leonardo Pisano Bigollo is known to us by the name "Fibonacci" (an abbreviated version of filius Bonacci, son of Bonacci), the latter name concocted in 1838 by the Franco-Italian historian Guillaume Libri.

If X and Y are **finite** sets (as in the example of the cavalry above), then $|X| = |Y|$ if and only if they have the same number of elements.[24]

[23] More precisely, on the class of all sets; see Section 0.5.

[24] The proof of this a simple application of Peano's Principle of Induction; see Section 1.3.

Example 0.4.1 Let X and Y be finite sets; $|X| = m$ and $|Y| = n$, $m, n \in \mathbb{N}$. How many maps $X \to Y$ are there? How many injective maps $X \to Y$ are there?[25]

Clearly, the number of maps $X \to Y$ is n^m since each element in X can be mapped to any element of Y (a choice of n).[26]

To count the number of injective maps $X \to Y$ we select a (first) element in X. This element can be mapped to any element in Y, a choice of n. Once this is done, we select a (second) element. Due to injectivity, this can be mapped to another element, a choice of $n-1$. Thus, so far, the number of choices made is $n(n-1)$. Continuing this way, the number of injective maps is $n(n-1)(n-2)\cdots(n-m+1)$. In particular, we must have $m \leq n$.

Example 0.4.2 Let X be a finite set of $n \in \mathbb{N}$ elements. A **permutation** of X is a bijective map $X \to X$. Determine the number of permutations of X.

As noted in Example 0.3.1, an injective map $f : X \to X$ must be surjective, therefore it must be a permutation. By the second part of the previous example, we see that the number of injective maps $X \to X$ is $n(n-1)(n-2)\cdots 2 \cdot 1$.

Based on this, we define the **factorial** of a natural number $n \in \mathbb{N}$, denoted by $n!$, as the product of all natural numbers less than equal to n. We conclude that there are $n!$ permutations of a set X of n elements.

Remark The sequence of factorials increases very rapidly. Here are the first few:

$1! = 1,$	$8! = 40,320,$	$15! = 1,307,674,368,000,$
$2! = 2,$	$9! = 362,880,$	$16! = 20,922,789,888,000,$
$3! = 6,$	$10! = 3,628,800,$	$17! = 355,687,428,096,000,$
$4! = 24,$	$11! = 39,916,800,$	$18! = 6,402,373,705,728,000,$
$5! = 120,$	$12! = 479,001,600,$	$19! = 121,645,100,408,832,000,$
$6! = 720,$	$13! = 6,227,020,800,$	$20! = 2,432,902,008,176,640,000,$
$7! = 5,040,$	$14! = 87,178,291,200,$	$21! = 51,090,942,171,709,440,000.$

Example 0.4.3 Let $n \in \mathbb{N}$, and write

$$\frac{(n!)!}{n} = M \cdot N!, \quad M, N \in \mathbb{N},$$

where N is as large as possible. Find $M + N$.[27]

We have

$$\frac{(n!)!}{n} = \frac{n! \cdot (n!-1)!}{n} = (n-1)! \cdot (n!-1)!$$

[25] In Section 6.3 (Example 6.3.7) we will determine the number of surjective maps $X \to Y$, $|X| = m$ and $|Y| = n$, $m, n \in \mathbb{N}$.

[26] For this reason, the set of all maps $X \to Y$ is usually denoted by Y^X. Note the special case $\mathcal{P}(X) = 2^X$ as discussed in Section 0.3.

[27] A special case ($n = 3! = 6$) was a problem in the American Invitational Mathematics Examination, 2003.

This gives $M = (n-1)!$ and $N = n! - 1$ with sum $M + N = n! + (n-1)! - 1$.

In the next example we return to the indicator function (Example 0.3.2).

Example 0.4.4 (Principle of Inclusion-Exclusion) We saw that the indicator function satisfies the equality

$$\mathbf{1}_{A\cup B} = \mathbf{1}_A + \mathbf{1}_B - \mathbf{1}_A \cdot \mathbf{1}_B, \quad A, B \subset X.$$

We generalize this to a (finite) collection of subsets $\{A_i \mid i = 1, \dots, n\}$, $n \in \mathbb{N}$, of a given set X. We claim[28]

$$\mathbf{1}_{\bigcup_{i=1}^n A_i} = \sum_{\emptyset \neq J \subset \{1,\dots,n\}} (-1)^{|J|+1} \mathbf{1}_{\bigcap_{j\in J} A_j}.$$

To show this, we consider the product $\prod_{i=1}^n \left(1 - \mathbf{1}_{A_i}\right)$. This is a function on X with values 0, 1. The ith factor of this product vanishes (precisely) on A_i, so that the entire product vanishes (precisely) on the union $\bigcup_{i=1}^n A_i$. We obtain that this product is the indicator function

$$\prod_{i=1}^n \left(1 - \mathbf{1}_{A_i}\right) = \mathbf{1}_{X\setminus\bigcup_{i=1}^n A_i} = 1 - \mathbf{1}_{\bigcup_{i=1}^n A_i}.$$

On the other hand, expanding the product, each term in the expansion is a product obtained by choosing, for each $i = 1, \dots, n$, in the ith factor either 1 or $-\mathbf{1}_{A_i}$. For a specific term, let $J \subset \{1, \dots, n\}$ be the subset consisting of those indices j for which we choose $-\mathbf{1}_{A_j}$. This term then can be written as

$$(-1)^{|J|} \prod_{j\in J} \mathbf{1}_{A_j} = (-1)^{|J|} \mathbf{1}_{\bigcap_{j\in J} A_j}.$$

The product above is the sum of these terms

$$\prod_{i=1}^n \left(1 - \mathbf{1}_{A_i}\right) = \sum_{J\subset\{1,\dots,n\}} (-1)^{|J|} \mathbf{1}_{\bigcap_{j\in J} A_j}.$$

Since $J = \emptyset$ corresponds to the term 1, putting everything together, we finally obtain

$$\mathbf{1}_{\bigcup_{i=1}^n A_i} = 1 - \prod_{i=1}^n \left(1 - \mathbf{1}_{A_i}\right) = \sum_{\emptyset \neq J \subset \{1,\dots,n\}} (-1)^{|J|+1} \mathbf{1}_{\bigcap_{j\in J} A_j}.$$

The claim follows.

[28] Here we use the usual summation notation: If I is a finite set and $A = \{a_i \mid i \in I\} \subset \mathbb{R}$ is a finite set of real numbers, then $\sum_{i\in I} a_i$ stands for the sum of all elements in A. In particular, if $I = \{m, m+1, \dots, n\} \subset \mathbb{Z}$, $m \leq n$, then we set $\sum_{i=m}^n a_i = \sum_{i\in I} a_i = a_m + a_{m+1} + \cdots + a_n$. Replacing the sum with product, we will also use the notation $\prod_{i\in I} a_i$ for the product of all elements in A; and $\prod_{i=m}^n a_i = \prod_{i\in I} a_i = a_m \cdot a_{m+1} \cdot \cdots \cdot a_n$.

In terms of subsets that the indicator functions correspond to, counting the number of elements in each subset, we obtain the following

$$\left|\bigcup_{i=1}^{n} A_i\right| = \sum_{\emptyset \neq J \subset \{1,\ldots,n\}} (-1)^{|J|+1} \left|\bigcap_{j\in J} A_j\right|.$$

This is called the Principle of Inclusion-Exclusion, and it is of paramount importance in combinatorics.

A simple application of the Principle of Inclusion-Exclusion is the following:

Example 0.4.5 How many positive integers ≤ 120 are multiples of 2, 3, or 5?

Let A_1, A_2, resp. A_3 be the set of positive integers ≤ 120 that are multiples of 2, 3, resp. 5. We need to find $|A_1 \cup A_2 \cup A_3|$. We have $|A_1| = 60$, $|A_2| = 40$, $|A_3| = 24$. Moreover, the sets $A_1 \cap A_2$, $A_2 \cap A_3$, resp. $A_3 \cap A_1$, are the sets of multiples of 6, 15, resp. 10, so that we have $|A_1 \cap A_2| = 20$, $|A_2 \cap A_3| = 8$, and $|A_3 \cap A_1| = 12$. Finally, $A_1 \cap A_2 \cap A_3$ is the set of multiples of 30, so that we have $|A_1 \cap A_2 \cap A_3| = 4$. Using the Principle of Inclusion-Exclusion, we obtain

$$|A_1 \cup A_2 \cup A_3| = (60 + 40 + 24) - (20 + 12 + 8) + 4 = 88.$$

Returning to the main line, for sets with infinitely many elements the situation is markedly different.

For example, the set of natural numbers $\mathbb{N}$ has the same cardinality as the set of non-negative integers $\mathbb{N}_0$, even though the former is a proper subset of the latter. A one-to-one correspondence that establishes this is $f : \mathbb{N} \to \mathbb{N}_0$ given by $f(n) = n - 1$, $n \in \mathbb{N}$. We thus have $|\mathbb{N}| = |\mathbb{N}_0|$.

Moreover, we also have $|\mathbb{N}| = |\mathbb{Z}|$. The one-to-one correspondence $f : \mathbb{N} \to \mathbb{Z}$ that establishes this is defined, for $n \in \mathbb{N}$, by

$$f(n) = \begin{cases} n/2 & \text{if } n \text{ is even} \\ (1-n)/2 & \text{if } n \text{ is odd.} \end{cases}$$

Diagrammatically

$$\begin{array}{ccccccccccc} 1 & 2 & 3 & 4 & 5 & 6 & 7 & \ldots & 2n & 2n+1 \\ \updownarrow & \updownarrow & \updownarrow & \updownarrow & \updownarrow & \updownarrow & \updownarrow & & \updownarrow & \updownarrow \\ 0 & 1 & -1 & 2 & -2 & 3 & -3 & \ldots & n & -n \end{array}$$

A set X is called **countable** if $|X| = |\mathbb{N}|$. By the above, we have $|\mathbb{N}| = |\mathbb{N}_0| = |\mathbb{Z}|$; that is, $\mathbb{N}_0$ and $\mathbb{Z}$ are countable sets. In general, any infinite subset of a countable set is countable. Indeed, let $X \subset \mathbb{N}$ be an infinite subset. Since X consist of natural numbers, its elements can be listed in an increasing order as $n_1, n_2, n_3, \ldots, n_k, \ldots$. We let $f : X \to \mathbb{N}$ be defined as $f(n_k) = k$, $k \in \mathbb{N}$. Then f is the desired one-to-one correspondence between X and $\mathbb{N}$.

We now turn to Cartesian products of countable sets. The **Cantor pairing** is a map $C : \mathbb{N}_0 \times \mathbb{N}_0 \to \mathbb{N}_0$ defined by

$$C(m,n) = \frac{(m+n)(m+n+1)}{2} + m, \quad m,n \in \mathbb{N}_0.$$

Using $(m,n) \in \mathbb{N}_0 \times \mathbb{N}_0$ as row-column indices, the first few values of C are as follows:

$$\begin{array}{cccccc} 0 & 1 & 3 & 6 & 10 & \dots \\ 2 & 4 & 7 & 11 & \dots & \\ 5 & 8 & 12 & \dots & & \\ 9 & 13 & \dots & & & \\ 14 & \dots & & & & \end{array}$$

We claim that C is a one-to-one correspondence, so that we have $|\mathbb{N}_0 \times \mathbb{N}_0| = |\mathbb{N}_0|$.

To get a better insight into the properties of C we introduce the **triangular numbers**. For $n \in \mathbb{N}_0$, the nth triangular number is defined by $T_n = n(n+1)/2$. The name comes from the fact that an isosceles triangular array of dots with $n \in \mathbb{N}$ dots in the base, $n-1$ dots in the next level, $n-2$ dots in the next level, etc. and 1 dot in the top $(n-1)$st level, have the total number of dots equal to T_n; that is, we have

$$T_n = \sum_{i=1}^{n} i = 1 + 2 + \cdots + n = \frac{n(n+1)}{2}, \quad n \in \mathbb{N}.$$

For a "Greek proof" of this, stack up rectangles of base lengths $n, n-1, \dots, 2, 1$ and constant height 1 in a staircase pattern with total height n and (cross-sectional) area $1+2+\cdots+n$. Two of these staircases can be joined along their jagged edges to form a rectangle of base length $n+1$ and height n. The formula follows.

Another proof is based on writing the sum $1+2+\cdots+n$ backwards as $n + (n-1) + \cdots + 1$ and adding. Pairing the numbers in the same position we obtain $(1+n) + (2+(n-1)) + \cdots + (n+1)$, the sum of n copies of $(n+1)$, that is, $n(n+1)$.

History

At the age of seven, Carl Friedrich Gauss (1777 – 1855) started elementary school. His teacher, Büttner, and his assistant, Martin Bartels, realized his talent for mathematics early on. One of his early achievements was to discover the (second) proof above in summing up the first 100 natural numbers by doubling and realizing that the sum was 50 pairs of numbers with each pair adding up to 101.

Remark For $m, n \in \mathbb{N}$, the triangular numbers satisfy the following[29]

$$T_{m+n} = T_m + T_n + m \cdot n$$

[29] A numerical special case of the first (T_{12}) was a problem in the American Mathematics Competitions, 2002.

$$T_{m\cdot n} = T_m \cdot T_n + T_{m-1} \cdot T_{n-1},$$

both of which can easily be checked by inspection or computation.

Example 0.4.6 Determine the 1000th term of the sequence $1, 2, 2, 3, 3, 3, 4, 4, 4, 4, 5, \ldots$.

The last time the natural number $n \in \mathbb{N}$ appears in this sequence is $T_n = n(n+1)/2$. For $n = 44$ we have $T_{44} = 990$ and $T_{45} = 1035$. Thus, the 1000th term is 45.

Example 0.4.7 What is the smallest $n \in \mathbb{N}$ such that the sum of 10 consecutive integers starting with n is a perfect square?

We have $n + (n+1) + (n+2) + \cdots + (n+9) = 10n + 9 \cdot 10/2 = 10n + 45 = 5(2n+9) = m^2$ for some $m \in \mathbb{N}$. Hence we must have $2n + 9 = 5k^2$ for some $k \in \mathbb{N}$ **odd**. The smallest odd number to realize this is $k = 3$. This gives $n = 18$.

Returning to our Cantor pairing, we thus have

$$C(m, n) = T_{m+n} + m, \quad m, n \in \mathbb{N}_0.$$

We now show that C is injective. To do this, we first claim that $m + n < m' + n'$ implies $C(m, n) < C(m', n')$. Letting $k = m + n$ and $k' = m' + n'$, $k < k'$, we have

$$\begin{aligned}\max_{k=m+n} C(m, n) = C(k, 0) = \frac{k(k+1)}{2} + k &< \frac{k(k+1)}{2} + k + 1 \\ = \frac{(k+1)(k+2)}{2} &\le \frac{k'(k'+1)}{2} = \min_{k'=m+n} C(m, n).\end{aligned}$$

The claim follows.

Thus, $C(m, n) = C(m', n')$ implies $m + n = m' + n'$. But then $m = m'$ and therefore $n = n'$ also follows. Hence C is injective.

Next we show that C is surjective. Let $t \in \mathbb{N}_0$ and let T_k be the largest triangular number not exceeding t. Let $m = t - T_k \in \mathbb{N}_0$ and $n = k - m$.

We first claim that $n \in \mathbb{N}_0$, that is, $m = t - T_k \le k$. Assume not. If $t - T_k > k$, then $t > T_k + k = k(k+1)/2 + k = (k+1)(k+2)/2 = T_{k+1}$. This means that T_k is not the largest triangular number not exceeding t, a contradiction. The claim follows.

With these choices we have $C(m, n) = T_k + m = t$, $k = m + n$. Surjectivity follows.

Summarizing, we obtain that $C : \mathbb{N}_0 \times \mathbb{N}_0 \to \mathbb{N}_0$ is a one-to-one correspondence; that is, we have $|\mathbb{N}_0 \times \mathbb{N}_0| = |\mathbb{N}_0|$. It follows that, for any countable set X, we have $|X| = |\mathbb{N}_0| = |\mathbb{N}_0 \times \mathbb{N}_0| = |X \times X|$; that is, the Cartesian product $X \times X$ is also countable.

As another consequence, we also have $|\mathbb{N}| = |\mathbb{Q}|$. Indeed, any non-zero rational number $0 \ne q \in \mathbb{Q}$ can be **uniquely** written as an **irreducible** (or reduced) fraction $q = a/b$, $a, b \in \mathbb{Z}$, $a, b \ne 0$, where a and b have no common divisors. This gives an injective map $f : \mathbb{Q} \to \mathbb{Z} \times \mathbb{N}$, $f(\pm a/b) = (\pm a, b)$, $a, b \in \mathbb{N}$ with no common

divisors; and $f(0) = (0, 1)$. Since f is a one-to-one correspondence between $\mathbb{Q}$ and its range in $\mathbb{Z} \times \mathbb{N}$, and $|\mathbb{Z} \times \mathbb{N}| = |\mathbb{N}|$, we see that the set of rational numbers $\mathbb{Q}$ is countable.

Returning to the general setting, let X and Y be sets. We say that X has cardinality **less than or equal** than the cardinality of Y if there is an injective map $f : X \to Y$. We write this as $|X| \leq |Y|$. If, in addition, there is no surjective map $g : X \to Y$, then we say that X has cardinality **strictly less** than the cardinality of Y, and write $|X| < |Y|$.

The Cantor–Schröder–Bernstein Theorem below states that the relation $\leq$ is "antisymmetric" with respect to cardinality[30] in the sense that, for any two sets X and Y, the inequalities $|X| \leq |Y|$ and $|Y| \leq |X|$ imply $|X| = |Y|$.

Cantor-Schröder-Bernstein Theorem *Let X and Y be sets. If there exist injective maps $f : X \to Y$ and $g : Y \to X$, then there is also a bijective map $h : X \to Y$.*

Proof We first prove this statement when $Y \subset X$ and the map $g : Y \to X$ is the inclusion. In this special case we have the injective composition $f : X \to Y \subset X$ (also denoted by f) which can be **iterated**. More precisely, we define the $n$fold composition $f^n = \overbrace{f \circ f \circ \ldots \circ f}^{n} : X \to X$, $n \in \mathbb{N}$, $(f^1 = f)$.[31] We also set $f^0 = \mathrm{id}_X : X \to X$, the identity on X, so that f^n is defined for all $n \in \mathbb{N}_0$.

We now let

$$A = \bigcup_{n \in \mathbb{N}_0} f^n(X \setminus Y).$$

An important property of the subset $A \subset X$ is that $x \in A$ implies $f(x) \in A$. In addition, for $n = 0$ in the union above, we have $X \setminus Y \subset A$. Hence $x \notin A$ implies $x \in Y$.
With these we now define the map $h : X \to Y$ by

$$h(x) = \begin{cases} f(x) & \text{if } x \in A \\ x & \text{if } x \notin A. \end{cases}$$

We claim that h is bijective.

First, we show injectivity. Assume $x, x' \in X$ such that $h(x) = h(x')$. If $x, x' \in A$, then we have $f(x) = h(x) = h(x') = f(x')$. Since f is injective, we obtain $x = x'$. If $x, x' \notin A$, then $x = h(x) = h(x') = x'$ automatically. Finally, if $x \in A$ and $x' \notin A$, then $f(x) \in A$ so that $h(x) = f(x) \neq x' = h(x')$, a contradiction. Injectivity follows.

[30] As noted above, having the same cardinality is an equivalence relation on the class of all sets.

[31] Strictly speaking, we need here Peano's Principle of Induction; see Section 1.3.

Second, we show surjectivity. Let $y \in Y$. If $y \in A$, then $y \in f^n(X \setminus Y)$ for some $n \in \mathbb{N}$ ($n \neq 0$). Hence, there exists $x \in f^{n-1}(X \setminus Y)$ such that $y = f(x) = h(x)$. If $y \notin A$, then, by definition, we have $y = h(y)$. Surjectivity follows.

Summarizing, we proved the theorem for the special case of an injective map $f : X \to Y \subset X$.

Returning to the general setting, let $f : X \to Y$ and $g : Y \to X$ be injective maps. Since the composition of injective maps is injective, we see that $g \circ f : X \to g(Y)$ is injective. Since $g(Y) \subset X$, by what we proved above, we have a bijective map $h : X \to g(Y)$. On the other hand, restricted to its range, $g : Y \to g(Y)$ is certainly bijective, and therefore the inverse $g^{-1} : g(Y) \to Y$ exists and is also bijective. Now, the composition $g^{-1} \circ h : X \to Y$ is a bijective map.

History
The Cantor–Schröder–Bernstein Theorem has a long and interesting history. In 1887 Cantor published the theorem without proof. Around this time Dedekind proved the theorem, but did not publish it, and his proof was discovered only in 1908 by Zermelo (who then published his own proof). In 1896 Ernst Schröder (1841 – 1902) announced the theorem with a sketch proof which was shown to be incorrect. In 1897, Felix Bernstein (1878 – 1956), then a student, presented his proof in Cantor's seminar, and almost simultaneously, Schröder found another proof. Subsequently, Cantor worked on simplifying the proof for years, but always gave full credit to Bernstein. Shortly afterwards, Dedekind, after a visit to Bernstein, came up with his second proof. Finally, note that there is also yet another beautiful proof by the Hungarian mathematician Gyula König (1849 – 1913) published in 1906.

Another quick proof of the countability of $\mathbb{Q}$ using the Cantor–Schröder–Bernstein Theorem is as follows: As before, write every non-zero rational number as an irreducible fraction $q = \pm a/b$, with $a, b \in \mathbb{N}$ having no common divisors. With this, define a map $f : \mathbb{Q} \to \mathbb{Z}$ by $f(\pm a/b) = \pm 2^a \cdot 3^b$; and $f(0) = 0$. Clearly, f is injective. Letting $g : \mathbb{Z} \to \mathbb{Q}$ to be the inclusion, the Cantor–Schröder–Bernstein Theorem implies $|\mathbb{Q}| = |\mathbb{Z}|$.

Remark It is natural to ask whether trichotomy holds for the relation $\leq$; that is, if, for any two sets X and Y, we have $|X| \leq |Y|$ or $|Y| \leq |X|$. The answer is "yes," and trichotomy, in fact, is equivalent to the **Axiom of Choice**.

Given a set X with power set $\mathcal{P}(X)$, we have $|X| \leq |\mathcal{P}(X)|$ since the map $g : X \to \mathcal{P}(X)$ that associates to any $x \in X$ the one-element subset $\{x\} \in \mathcal{P}(X)$ is injective. We now prove a result of Cantor which asserts that $|X| < |\mathcal{P}(X)|$.

Cantor Theorem *For any set X, there is no surjective map $f : X \to \mathcal{P}(X)$.*

Proof Assume that $f : X \to \mathcal{P}(X)$ is a surjective map for some set X. Define $Y = \{x \in X \,|\, x \notin f(x)\}$. Since $Y \subset X$, we have $Y \in \mathcal{P}(X)$. By the assumed surjectivity of f, we have $f(x_0) = Y$ for some $x_0 \in X$. By construction, we have $x_0 \in Y$ if and only if $x_0 \notin f(x_0) = Y$. This is a contradiction.

We will show later (Section 2.2) that the power set $\mathcal{P}(\mathbb{N})$ and the set of real numbers $\mathbb{R}$ have the same cardinality. We thus have $|\mathbb{N}| < |\mathcal{P}(\mathbb{N})| = |\mathbb{R}|$.

Exercises

0.4.1. Let A be the set of all sequences $\mathbb{N} \to \{0, 1\}$. Show that A is uncountable.
0.4.2. Show that the union of countably many sets is countable.
0.4.3. How many passwords of length 8 can be made from the letters a, b, and c such that each occurs at least once?
0.4.4. Let $m < n$, $m, n \in \mathbb{N}$. Determine the number of possible sums of the elements in an m-element subset of $\{1, 2, \ldots, n\}$.
0.4.5. Show that the set of all irrational numbers, $\mathbb{R} \setminus \mathbb{Q}$, is uncountable.[32]

0.5 The Zermelo–Fraenkel Axiomatic Set Theory*

We now turn to a brief account on how naïve set theory can be axiomatized.

In axiomatic set theory the precise meaning of sets and the set membership relation are not addressed; they are primitives. The primary focus is on describing the **properties** of sets and the set membership. This description is given by a set of axioms and statements that can be deduced from the axioms by inference using the rules of logic. The set of axioms should satisfy three criteria: (1) **Consistency**: No statement and its negation are to be deduced; (2) **Credibility**: The axioms and the derived statements should be in accord with the naïve set theory; (3) **Richness**: Statements of the Cantor naïve set theory should be derived as theorems.

History
As noted previously, Cantor recognized that naïve set theory quickly gives birth to paradoxes. The two best known are **Cantor's Paradox** asserting that "the set of all sets" cannot exist; and **Russell's Paradox** (1899/1901) asserting that "the set of all sets that do not contain themselves" cannot exist. Axiomatic set theory was created to avoid these paradoxes.

In the **Zermelo–Fraenkel axiomatic set theory**, termed ZF or ZFC (see the discussion below), all sets are **hereditary** and **well-founded**.

A set is hereditary if all of its elements are also hereditary sets. (For example, the so-called von Neumann ordinals[33] $\emptyset$, $\{\emptyset\}$, $\{\emptyset, \{\emptyset\}\}$, $\{\emptyset, \{\emptyset\}, \{\emptyset, \{\emptyset\}\}\}$, etc., are all hereditary sets.) Thus, there is no difference between "objects" and "sets" as we had in our naïve approach. Any element of a set is also a set, and consequently, there is only one primitive, the **set** itself. This also implies that the single primitive relation $\in$, the **set membership**, usually spelled out as "element/member of" is actually a (binary) relation between sets. Thus, the Zermelo–Fraenkel set theory excludes **urelements**,[34] elements of sets that are not themselves sets.

[32] To show that $\mathbb{R} \setminus \mathbb{Q}$ has the same cardinality as $\mathbb{R}$ is harder, and, by the Cantor–Schröder–Bernstein Theorem, it amounts to construct an injective map of $\mathbb{R}$ to $\mathbb{R} \setminus \mathbb{Q}$.

[33] More about this at the end of this section.

[34] Using the German prefix ur- "primordial."

To define the concept of well-founded sets, we need some preparation. A set X is **transitive** if $x \in X$ and $y \in x$ implies $y \in X$. A set is transitive if and only if $\bigcup X \subset X$, where $\bigcup X$ is the union of all elements of X that are sets (see Section 0.1).

The **transitive closure** of a set X is the smallest transitive set (with respect to the inclusion relation) that contains X. The transitive closure $TC(X)$ of a set X is the union $TC(X) = \bigcup_{n\in\mathbb{N}_0} X_n$, where $X_0 = X$ and $X_{n+1} = \bigcup X_n$, $n \in \mathbb{N}_0$.[35]

Finally, a set X is well-founded if the set membership relation on every non-empty subset of the transitive closure of $TC(X)$ has a minimal element; that is, for any $\emptyset \neq Y \subset TC(X)$, there is $y \in Y$ such that, for all $z \in Y$, we have $z \notin y$. An axiom in the Zermelo–Fraenkel system (the Axiom of Foundation/Regularity; see below) guarantees that all sets are well-founded.

The Zermelo–Fraenkel set theory steers clear from Cantor's and Russel's Paradoxes noted above. (The Axiom of Foundation/Regularity does not allow the existence of a universal set (a set that contains all sets), and the Axiom Schema of Specification/Comprehension avoids Russel's Paradox; see below.)

Finally, typographically, to designate any set the typical practice is to use lowercase letters. Uppercase letters will be used sparingly and mostly in specific situations.

The Zermelo–Fraenkel axioms comprise a system of nine axioms. As we have seen in Section 0.1, a number of constructions in naïve set theory use the vague concept of "predicate" or "property" to be decidable (true or false) for the elements of a given set, by means of which a subset of the set can be defined (consisting of those elements for which the property holds). Zermelo called this property a "definite formula" for all the elements of the given set. As noted above, Fraenkel and Skolem made this vague concept more precise by what is known as a **formula** of ZFC. Before stating the axioms, we briefly elaborate on this.

The language of axiomatic set theory in the framework of first-order predicate calculus[36] has two basic predicates (Boolean-valued functions with range {true, false}); the **equality predicate** $=$, and the **set membership predicate** $\in$.

The basic building blocks of **formulas** are the two **atomic formulas**: $x = y$ and $x \in y$, for any variables x and y.

The atomic formulas are used to build more complex formulas recursively by means of **connectives** and **quantifiers**. Connectives can be used to derive from formulas ϕ and ψ new ones as follows:

$$\phi \wedge \psi \qquad \text{(logical conjunction "and")}$$

$$\phi \vee \psi \qquad \text{(logical disjunction "or")}$$

$$\neg\phi \qquad \text{(logical negation "not")}$$

[35] This definition requires Peano's Principle of Induction; see Section 1.3.

[36] First-order predicate calculus is an assembly of formal systems that allows to use **quantified variables** over nonlogical objects and sentences that contain **variables**.

$$\phi \implies \psi \qquad \text{(implication "implies")}$$
$$\phi \iff \psi \qquad \text{(equivalence "if and only if").}$$

In addition, if x is a variable, then quantifiers can be used to derive new formulas:

$$\forall x\, \phi \qquad \text{(universal quantification "for all")}$$
$$\exists x\, \phi \qquad \text{(existential quantification "there exists").}$$

Formulas are constructed in **finitely many steps** starting with atomic formulas and proceeding with the steps above.

In first-order predicate calculus formulas are allowed to have **free variables**. A variable is free if it occurs in the formula at least once without being introduced by any universal or existential quantifiers. A useful convention is to indicate **all** free variables (or parameters) $p_1, \ldots, p_n$ of a formula by writing $\phi(p_1, \ldots, p_n)$. A formula with no free variables is called a **sentence**.

A formula $\phi(p_1, \ldots, p_n)$ with free variables $p_1, \ldots, p_n$ is often called a **condition** on $p_1, \ldots, p_n$. It attains a meaning only when a **domain of interpretation** is provided which specifies the range of values of the variables and the membership relation amongst them.

Any formula (in the language $\{\in\}$) $\phi(x, p_1, \ldots, p_n)$ defines a **class**:

$$\mathfrak{C} = \{x \mid \phi(x, p_1, \ldots, p_n)\}.$$

A **set** x is a member of the class $\mathfrak{C}$ if and only if $\phi(x, p_1, \ldots, p_n)$.

We say that the class $\mathfrak{C}$ above is **definable** from $p_1, \ldots, p_n$; and simply **definable** if there are no parameters.

Sets are objects that satisfy the Zermelo–Fraenkel system of axioms expounded below. Every **set** x is considered a class definable by the formula $u \in x$; that is, x is identified by the class $\{u \mid u \in x\}$. A class that is not a set is called a **proper class**.

For example, the **universe**, the class of all sets, is the definable class $\mathfrak{V} = \{x \mid x = x\}$. Note that, by Cantor's Paradox, $\mathfrak{V}$ is a proper class (see below).

The classes $\mathfrak{C} = \{x \mid \phi(x, p_1, \ldots, p_n)\}$ and $\mathfrak{D} = \{x \mid \psi(x, q_1, \ldots, q_m)\}$, given by the formulas $\phi(x, p_1, \ldots, p_n)$ and $\psi(x, q_1, \ldots, q_m)$, are **equal**, $\mathfrak{C} = \mathfrak{D}$, if $x \in \mathfrak{C}$ if and only if $x \in \mathfrak{D}$, or equivalently

$$\forall x\, (\phi(x, p_1, \ldots, p_n) \iff \psi(x, q_1, \ldots, q_m)).$$

The class $\mathfrak{C} = \{x \mid \phi(x, p_1, \ldots, p_n)\}$ is a **subclass** of $\mathfrak{D} = \{x \mid \psi(x, q_1, \ldots, q_m)\}$, that is, we have the inclusion $\mathfrak{C} \subset \mathfrak{D}$, if $x \in \mathfrak{C}$ implies $x \in \mathfrak{D}$, or equivalently

$$\forall x\, (\phi(x, p_1, \ldots, p_n) \implies \psi(x, q_1, \ldots, q_m)).$$

The operations of **union**, **intersection**, and **difference** can be naturally defined on classes as follows:

$$\mathfrak{C} \cup \mathfrak{D} = \{x \mid x \in \mathfrak{C} \vee x \in \mathfrak{D}\}$$

$$\mathfrak{C} \cap \mathfrak{D} = \{x \mid x \in \mathfrak{C} \wedge x \in \mathfrak{D}\}$$
$$\mathfrak{C} \setminus \mathfrak{D} = \{x \mid x \in \mathfrak{C} \wedge x \notin \mathfrak{D}\} \quad \notin = \neg \in .$$

Similarly, the **union** of a class $\mathfrak{C}$ is defined as

$$\bigcup \mathfrak{C} = \bigcup \{C \mid C \in \mathfrak{C}\} = \{x \mid \exists\, C\, (x \in C \wedge C \in \mathfrak{C}\}.$$

With these, not striving for minimality, the Zermelo–Fraenkel axioms are as follows:

1. **Axiom of Extensionality**: If two sets have the same elements, then they are equal. This axiom imparts the idea that a set is uniquely determined by its members.

$$\forall x\, \forall y\, [\, \forall z\, (z \in x \iff z \in y) \implies x = y\,]\,.$$

The converse, that is, if two sets are equal, then they have the same elements, is an axiom of predicate calculus. Putting these together gives

$$\forall x\, \forall y\, [\, \forall z\, (z \in x \iff z \in y) \iff x = y\,]\,.$$

2. **Axiom of (Unordered) Pairing**:

$$\forall x\, \forall y\, \exists z\, \forall u\, [u \in z \iff (u = x \vee u = y)].$$

By the Axiom of Extensionality, the set z is unique. We denote $z = \{x, y\}$.

The Axiom of Pairing applied to a set x gives the existence of the singleton $\{x\} = \{x, x\}$. Applying the Axiom of Pairing again, this time to the sets $\{x\}$ and $\{x, y\}$, we see that $\{\{x\}, \{x, y\}\}$ is also a set. Following Kazimierz Kuratowski (1896–1980), the **ordered pair** (x, y) is defined as $(x, y) = \{\{x\}, \{x, y\}\}$. With this, we have

$$(x, y) = (u, v) \iff x = u \wedge y = v.$$

We define ordered triples, quadruples, quintuples, etc. by $(x, y, z) = ((x, y), z)$, $(x, y, u, v) = ((x, y, u), v)$, etc. In general, we define $(x_1, \ldots, x_n)$, $n \in \mathbb{N}$, inductively[37] by

$$(x_1, \ldots, x_n, x_{n+1}) = ((x_1, \ldots, x_n), x_{n+1}).$$

As before, $(x_1, \ldots, x_n) = (y_1, \ldots, y_n)$, $n \in \mathbb{N}$, if and only if $x_1 = y_1, \ldots, x_n = y_n$.

[37] This needs Peano's Principle of Induction; see Section 1.3.

3. **Axiom Schema of Specification/Comprehension**: We have seen that in naïve set theory a set y can be defined as a subset of a given set z with typical element x satisfying a certain condition $\phi(x, p_1, \ldots, p_n)$. As discussed above, in the Zermelo–Fraenkel system of axioms properties are given by formulas.

   Given a formula $\phi(x, p_1, \ldots, p_n)$, we have

$$\forall z\, \forall p_1 \ldots \forall p_n\, \exists y\, \forall x\, [\, x \in y \iff (\, x \in z \wedge \varphi(x, p_1, \ldots p_n)\,].$$

   We denote $y = \{x \in z \mid \phi(x, p_1, \ldots, p_n)\}$.

   Using classes, this axiom can be reformulated in the following form. Let $\mathfrak{C}$ be the class

$$\mathfrak{C} = \{x \mid \phi(x, p_1, \ldots, p_n)\}.$$

   Then, we have

$$\forall z\, \exists y\, (z \cap \mathfrak{C} = y).$$

   This means that the intersection of a class and a set is a set; in particular, a subclass of a set is a set (called a subset).

   A consequence of this is that the intersection and difference of two sets are sets.

   Another consequence is that the universe $\mathfrak{V}$ is a proper class. Otherwise, consider the set $y = \{z \in \mathfrak{V} \mid z \notin z\}$. By definition, $y \in y$ if and only if $y \in \mathfrak{V}$ and $y \notin y$. Since $\mathfrak{V}$ is universal, we have $y \in \mathfrak{V}$, and the last statement reduces to $y \in y$ if and only if $y \notin y$. This is a contradiction.

   Yet another consequence that we note here is that, given a non-empty class of sets $\mathfrak{C}$, the intersection

$$\bigcap \mathfrak{C} = \bigcap \{C \mid C \in \mathfrak{C}\} = \{x \mid \forall\, C \in \mathfrak{C}\, (x \in C)\}.$$

   is a set.

   Remark An axiom schema in mathematical logic generalizes the concept of an axiom. It contains a schematic variable in which countably many subformulas can be substituted. Therefore an axiom schema stands for countably many axioms.

4. **Axiom of Foundation/Regularity**: Every non-empty set contains an element which, as a set, is disjoint from the set itself:

$$\forall x\, [x \neq \emptyset \implies \exists y \in x\, (y \cap x = \emptyset)].$$

   As noted above this axiom implies (almost verbatim) that all sets are well-founded.

In addition, this axiom also implies that there is no set x which is an element of itself, $x \in x$, a property needed for defining the von Neumann ordinal rank in the cumulative hierarchy of the universe (see below).

Indeed, let x be a set and consider the singleton $\{x\}$ which exists by the Axiom of Pairing above. The Axiom of Foundation applied to $\{x\}$ says that this set has an element disjoint from the set itself. But this set contains only one element, x, therefore, x, as a set, must be disjoint from $\{x\}$. In particular, x, the only element of the set $\{x\}$ cannot be contained in x. The statement follows.

5. **Axiom of Union**: For any set of sets $\mathcal{X}$ there exists a set X that contains all the elements that are elements of some member of $\mathcal{X}$:

$$\forall \mathcal{X}\, \exists X\, \forall Y\, \forall x\, [(x \in Y \wedge Y \in \mathcal{X}) \implies x \in X].$$

Given $\mathcal{X}$ and the corresponding X whose existence is guaranteed by this axiom, we use the Axiom of Specification to define

$$\bigcup \mathcal{X} = \{x \in X \mid \exists Y (x \in Y \wedge Y \in \mathcal{X})\}.$$

Let x and y be sets. By the Axiom of Pairing, $\{x, y\}$ is a set, and, by the Axiom of Union, we define the union $x \cup y = \bigcup\{x, y\}$ as a set.

Moreover, we define $\{x_1, \ldots, x_n\}$, $n \in \mathbb{N}$, inductively[38] by $\{x_1, \ldots, x_n, x_{n+1}\} = \{x_1, \ldots, x_n\} \cup \{x_{n+1}\}$.

Finally, if $x_1, \ldots, x_n$, $n \in \mathbb{N}$, are sets, then we define the union

$$x_1 \cup \ldots \cup x_n = \bigcup\{x_1, \ldots, x_n\}$$

as a set.

6. **Axiom of Infinity**:

$$\exists X\, [\emptyset \in X \wedge \forall x \in X\, (x \cup \{x\} \in X)].$$

For a set x, we let $S(x) = x \cup \{x\}$. For $k \in \mathbb{N}_0$, we define the set $S^k(x)$ inductively[39] by $S^0(x) = x$ and $S^{k+1}(x) = S(S^k(x)) = S^k(x) \cup \{S^k(x)\}$. We claim that, for $k \neq l$, $k, l \in \mathbb{N}$, the sets $S^k(x)$ and $S^l(x)$ are different; in particular X whose existence is postulated in the axiom above is **infinite**.[40]

Indeed, assuming $k > l$, and setting $m = k - l \in \mathbb{N}$ and $y = S^l(x)$, we need to show that $y \neq S^m(y)$.

First, $y \subset S^n(y)$ for all $n \in \mathbb{N}_0$. Indeed, for $n = 0$ this is tautology; for $n = 1$, we have $y \subset y \cup \{y\} = S(y)$, and, inductively, $y \subset S^n(y)$ implies $y \subset S^n(y) \cup \{S^n(y)\} = S^{n+1}(y)$.

[38] This needs Peano's Principle of Induction; see Section 1.3.

[39] Once again, as noted above, we use Peano's Principle of Induction here.

[40] X satisfying the Axiom of Infinity above is usually called **inductive**.

Finally, let $z = S^{m-1}(y)$ and apply the Axiom of Foundation to the set $\{z\}$. We obtain that $\{z\}$ has an element disjoint from the set itself. But this set contains only one element, z, therefore, z, as a set, must be disjoint from $\{z\}$. Returning to $z = S^{m-1}(y)$, we obtain $S^{m-1}(y) \cap \{S^{m-1}(y)\} = \emptyset$. On the other hand, $y \subset S^{m-1}(y)$, so that we arrive at $y \cap \{S^{m-1}(y)\} = \emptyset$. This shows that $S^m(y) = S^{m-1}(y) \cup \{S^{m-1}(y)\} \not\subset y$. We obtain $y \neq S^m(y)$. The claim follows.

Note also that this axiom guarantees that there exists at least one set, X. With this the empty set can be defined by $\emptyset = \{x \in X \mid (x \in x) \wedge (x \notin x)\}$. This is usually extracted as the so-called the **Axiom of the Empty Set**. Moreover, by the Axiom of Extensionality, the empty set is unique.

Note finally that a **minimal** infinite set X is the **von Neumann ordinal** ω (see below).

7. **Axiom of the Power Set**: We first define the concept of a subset:

$$(y \subset x) \iff [\forall z\, (z \in y \implies z \in x)].$$

With this the axiom is the following

$$\forall x\, \exists y\, \forall z\, (z \in y \iff z \subset x).$$

We denote $y = \mathcal{P}(x)$, the power set of x.

With these axioms in place, we can prove the existence of the Cartesian product of two sets X and Y as follows. As noted above, the union $X \cup Y$ is a set. Clearly, for $x \in X$ and $y \in Y$, the ordered pair $(x, y) = \{x, \{x, y\}\} \in \mathcal{P}(\mathcal{P}(X \cup Y))$. We define

$$X \times Y = \{u \mid \exists x\, \exists y\, (u = (x, y) \wedge x \in X \wedge y \in Y)\}.$$

Finally, we define $X_1 \times \ldots \times X_n = \{(x_1, \ldots, x_n) \mid x_1 \in X_1 \wedge \ldots \wedge x_n \in X_n\}$ inductively as

$$X_1 \times \ldots \times X_n \times X_{n+1} = (X_1 \times \ldots \times X_n) \times X_{n+1}.$$

In particular, we have $X^n = \overbrace{X \times \ldots \times X}^{n}$.

8. **Axiom Schema of Replacement**: A class R is called a (binary) **relation** if all elements of R are ordered pairs (x, y), where x and y are sets. With the Cartesian square of the universe $\mathfrak{V}^2 = \{z \mid \exists x\, \exists y\, (z = (x, y) \wedge x \in \mathfrak{V} \wedge y \in \mathfrak{V})\}$, we have $R \subset \mathfrak{V}^2$.

Any formula $\phi(x, y, p_1, \ldots, p_n)$ defines a relation:

$$R = \{(x, y) \mid (x, y) \in \mathfrak{V}^2 \wedge \phi(x, y, p_1, \ldots, p_n)\}.$$

A pair (x, y) is a member of the relation if $(x, y) \in R$.

We say that the relation R above is **definable** from $p_1, \ldots, p_n$; and simply definable if the parameters are absent.

The **domain** $\mathcal{D}_R$ and the **range** of a relation are defined by

$$\mathcal{D} = \mathcal{D}_R = \{x \mid \exists y\, (x, y) \in R\} \quad \text{and} \quad \mathcal{R} = \mathcal{R}_R = \{y \mid \exists x\, (x, y) \in R\}.$$

A definable relation (defined by $\phi(x, y, p_1, \ldots, p_n)$) is a **definable function**[41] f if

$$\forall x\, \forall y\, \forall y'\, \forall p_1 \ldots \forall p_n\, [\varphi(x, y, p_1, \ldots, p_n) \wedge \varphi(x, y', p_1, \ldots, p_n) \implies y = y'].$$

The unique y thereby associated with x by f via $\phi(x, y, p_1, \ldots, p_n)$ is denoted by $f(x)$. Indicating the domain and the range, a definable function is usually denoted by $f : \mathcal{D}_f \to \mathcal{R}_f$.

The Axiom Schema of replacement says that, if a function f is definable by a formula $\varphi(x, y, p_1, \ldots, p_n)$, then for any **set** A, there exists a set $B = f(A) = \{f(x) \mid x \in A\}$:

$$\forall x\, \forall y\, \forall z\, \forall p_1 \ldots \forall p_n\, [\varphi(x, y, p_1, \ldots, p_n) \wedge \varphi(x, z, p_1, \ldots, p_n) \implies y = z)$$
$$\implies \forall A\, \exists B\, \forall y\, (y \in B \iff \exists x\, (x \in A \wedge \varphi(x, y, p_1, \ldots, p_n))].$$

Remark The Axiom Schema of Replacement and the Axiom of the Empty Set (which we did not include in the list of axioms) together imply the Axiom Schema of Specification. Indeed, let $\phi(x, p_1, \ldots, p_n)$ be a formula and z a set, and define the function f such that $f(x) = x$ if $\phi(x, p_1, \ldots, p_n)$ is true and $f(x) = u$ if $\phi(x, p_1, \ldots, p_n)$ is false, where $u \in z$ such that $\phi(u, p_1, \ldots, p_n)$ is true. Then the set y guaranteed by the Axiom Schema of Replacement is precisely the set y required in the Axiom Schema of Specification. If u does not exist, then $f(x)$ in the Axiom Schema of Specification is the empty set whose existence is needed here.

Axioms 1–8 define the Zermelo–Fraenkel set theory, ZF, for short.

History

The Axiom Schema of Replacement was not part of the original Zermelo system of axioms published in 1908. This axiom greatly extends the potential of ZF in providing proofs of theorems as well as its strength in consistency. While it appeared around 1917 in the works of the Russian mathematician Dmitry Mirimanoff (1861–1945), it was the publication in 1922 by Fraenkel (announced earlier in the 1921 Jena meeting of the German Mathematical Society) when this axiom took its right place amongst what is now known as ZF, the Zermelo–Fraenkel system of axioms. Skolem also realized the necessity of this axiom later in the same year (announced in the 1922 Helsinki meeting of the Congress of Scandinavian Mathematicians and published in 1930), and his augmented system of axioms also included the von Neumann Axiom of Foundation. The term "replacement" (German "Ersetzungsaxiom") is due to Fraenkel. Originally this was only meant to be tentative until a final formalization of Zermelo's "definite property" could be obtained.

[41] Or a **class function**.

9. Axiom of Choice: For any set of sets $\mathcal{X}$ there is a **choice function** with domain $\mathcal{X}$ and range $\bigcup \mathcal{X}$ that associates to any member x of $\mathcal{X}$ and element contained in x.

$$\forall \mathcal{X} \left[\emptyset \notin \mathcal{X} \implies \exists f \left[\text{Func}(f) \wedge \mathcal{D}_f = \mathcal{X} \wedge \mathcal{R}_f \subset \bigcup \mathcal{X} \wedge \right.\right.$$
$$\left.\left. (\forall x \in \mathcal{X}) \left[f(x) \in x \right] \right] \right],$$

where $\text{Func}(f)$ if and only if f is definable, $\mathcal{D}_f$, resp. $\mathcal{R}_f$, are the domain, resp. range of f.

Adding this axiom to ZF defines ZFC, where C stands for the Axiom of Choice.

Remark Although nowadays most mathematicians accept it, there has been a considerable scrutiny and reluctance to incorporate the Axiom of Choice, AC, to the Zermelo–Fraenkel system. As noted earlier, AC is equivalent to the **Well-Ordering Theorem**, that is, the statement that every set can be well-ordered. But the construction leading to well ordering is non-canonical in the sense that well-ordering cannot be explicitly constructed. For this reason, AC is considered as non-constructive because it postulates the existence of a choice function without actually asserting anything about how this function is to be constructed. In addition, the Axiom of Choice leads to some highly counter-intuitive results.

It is known that the consistency of ZFC cannot be proved within ZFC itself (unless it turns out to be inconsistent). Most mathematicians are confident, however, that the ZFC is consistent since they believe that if the ZFC were inconsistent then it would have been discovered by now. There has been a considerable amount of study targeting independence of each axiom from the others; for example, the Axiom of Foundation/Regularity is known to be independent from the rest of the axioms in ZFC.

A **(von Neumann) ordinal** is a set α such that α is strictly well-ordered with respect to set membership $\in$, and every element of α is also a subset of α, that is, α is transitive. For the strict well-order we will use interchangeably $\in$ and the generic order $<$.

The **non-negative integers** are ordinals. The first few are tabulated here:

$$0 = \{\} = \emptyset$$
$$1 = \{0\} = 0 \cup \{0\} = \{\emptyset\}$$
$$2 = \{0, 1\} = 1 \cup \{1\} = \{\emptyset, \{\emptyset\}\}$$
$$3 = \{0, 1, 2\} = 2 \cup \{2\} = \{\emptyset, \{\emptyset\}, \{\emptyset, \{\emptyset\}\}\}$$
$$4 = \{0, 1, 2, 3\} = 3 \cup \{3\} = \{\emptyset, \{\emptyset\}, \{\emptyset, \{\emptyset\}\}, \{\emptyset, \{\emptyset\}, \{\emptyset, \{\emptyset\}\}\}\}$$
$$5 = \{0, 1, 2, 3, 4\} = 4 \cup \{4\}$$
$$= \{\emptyset, \{\emptyset\}, \{\emptyset, \{\emptyset\}\}, \{\emptyset, \{\emptyset\}, \{\emptyset, \{\emptyset\}\}\}, \{\emptyset, \{\emptyset\}, \{\emptyset, \{\emptyset\}\}\}, \{\emptyset, \{\emptyset\}, \{\emptyset, \{\emptyset\}\}\}\}\}.$$

In general, the successor of an ordinal number α is the ordinal number $\alpha \cup \{\alpha\}$ denoted by $\alpha + 1$. The finite ordinal $n \in \mathbb{N}_0$ is therefore defined by $\{0, 1, 2, \ldots, n-1\}$, in which any element $k = 0, 1, 2, \ldots, n-1$ is identified with the ordinal $\{0, 1, \ldots, k\}$. The first infinite ordinal as a set is $\mathbb{N}_0$, and, as an ordinal, it is denoted by ω. Its successor is the ordinal $\omega + 1 = \{0, 1, 2, \ldots, \omega\}$. The successor of $\omega + 1$ is $\omega + 2 = \omega \cdot 2 = \{0, 1, 2, \ldots, \omega, \omega + 1\}$. The ordinal $\omega + \omega = 2 \cdot \omega$ is the ordinal $\{0, 1, 2, \ldots; \omega, \omega + 1, \omega + 2, \ldots\}$. Then there comes $2 \cdot \omega + 1$, $2 \cdot \omega + 2$, etc. $3 \cdot \omega$. Continuing, we have $4 \cdot \omega$, $5 \cdot \omega$, etc. $\omega \cdot \omega = \omega^2$. The latter is the ordinal $\{n \cdot \omega + m \mid m, n \in \mathbb{N}_0\}$. Continuing further we obtain the ordinals ω^ω, ω^{ω^ω}, etc. (Note that ω^ω is still countable as a set.) The first uncountable ordinal, the ordinal of all countable ordinals is denoted by ω_1.

Returning to the main line, we first claim that every element of an ordinal is an ordinal itself. Indeed, if α is an ordinal and $\beta \in \alpha$, then, as a subset of α, we have $\beta = \{\gamma \in \alpha \mid \gamma \in \beta\} = \{\gamma \in \alpha \mid \gamma < \beta\}$. In other words, an element β in an ordinal α is the set of all elements of α that are (strictly) less than β. Cleary, this implies that β is an ordinal itself.

Next, we claim that if α and β are ordinals, then $\beta \in \alpha$ if and only if $\beta \subset \alpha$ and $\beta \neq \alpha$. Indeed, if $\beta \in \alpha$, then, as we have seen above, $\beta \subset \alpha$. Now, $\beta = \alpha$ cannot happen because $\alpha \in \alpha$ would contradict to the Zermelo–Fraenkel Axiom of Foundation. For the converse, let α and β be ordinals and assume that $\beta \subset \alpha$ is a proper subset. Let $\gamma \in \alpha$ be a minimal element in $\alpha \setminus \beta$. Then we have $\{\xi \in \alpha \mid \xi < \beta\} = \{\xi \in \alpha \mid \xi < \gamma\}$. On the one hand, this is β, and, on the other hand, this is γ. Therefore $\beta = \gamma \in \alpha$.

Finally, we claim that if α and β are ordinals, then either $\alpha \in \beta$ or $\beta \in \alpha$ or $\alpha = \beta$, so that trichotomy holds. The key fact here is that $\alpha \cap \beta$ is an ordinal. Clearly, $\alpha \cap \beta \subset \alpha$ and $\alpha \cap \beta \subset \beta$. Now, proper inclusion cannot be in both relations since then, by the above, we would have $\alpha \cap \beta \in \alpha$ and $\alpha \cap \beta \in \beta$, and this would imply $\alpha \cap \beta \in \alpha \cap \beta$, contradicting the Zermelo–Fraenkel Axiom of Foundation. If $\alpha \cap \beta = \alpha$, then $\alpha \subset \beta$. Thus, either $\alpha = \beta$ or $\alpha \in \beta$. If $\alpha \cap \beta = \beta$, then $\beta \subset \alpha$. Thus, either $\beta = \alpha$ or $\beta \in \alpha$. Trichotomy follows.

As a corollary, we see that an ordinal α is a set whose elements are precisely those ordinals that are strictly less than α itself.

Remark It can be proved that every strictly well-ordered set is order isomorphic to one of the ordinals.

Recall the **universe** $\mathfrak{V}$, the class of all sets. In the so-called **von Neumann universe**, $\mathfrak{V}$ possesses a so-called **cumulative hierarchy** $\mathfrak{V} = \bigcup_\alpha \mathfrak{V}_\alpha$, where the union is over all ordinals α. We call $\mathfrak{V}_\alpha$ stage α, the stage corresponding to the ordinal number α. In stage 0 there are no sets, that is, we have $\mathfrak{V}_0 = \{\}$. In stage 1 there is the empty set $\emptyset$, so that $\mathfrak{V}_1 = \{\emptyset\}$. At each stage of the hierarchy, a set is added if all of its elements appear in previous stages. So, for example, as above in stage 2, the set $\{\emptyset\}$ (with a single element, the empty set) is added, and we have $\mathfrak{V}_2 = \{\emptyset, \{\emptyset\}\}$. In general, stage α is defined by $\mathfrak{V}_\alpha = \bigcup_{\beta<\alpha} \mathcal{P}(\mathfrak{V}_\beta)$. For stage 3, this gives $\mathfrak{V}_3 = \{\emptyset, \{\emptyset\}, \{\{\emptyset\}\}, \{\emptyset, \{\emptyset\}\}\}$, in particular, $|\mathfrak{V}_3| = 4$. Continuing, we have $|\mathfrak{V}_4| = 2^4 = 16$, $|\mathfrak{V}_5| = 2^{16} = 65,536$, $|\mathfrak{V}_6| = 2^{65,536}$ (19,729 decimal digits), etc.

The collection of all sets obtained in this way forms a natural hierarchy. Each set X possesses a unique stage rank, its so-called **birthday**, the smallest ordinal α such that $X \subset \mathfrak{V}_\alpha$.

Exercise

0.5.1. Determine the following sets: $\mathcal{P}(\{\emptyset\})$, $\mathcal{P}(\{\emptyset, \{\emptyset\}\})$, $\mathcal{P}(\mathcal{P}(\emptyset))$, and $\mathcal{P}(\mathcal{P}(\{\emptyset\}))$.

Chapter 1
Natural, Integral, and Rational Numbers

> *"But I will try to show you by means of geometrical proofs, which you will be able to follow, that, of the numbers named by me and given in the work which I sent to Zeuxippus, some exceed not only the number of the mass of sand equal in magnitude to the Earth filled up in the way described, but also that of the mass equal in magnitude to the universe."*
> in The Sand Reckoner by Archimedes of Syracuse.

In this chapter we present a very detailed and slow-paced arithmetic exposition of the natural, integral, and rational number systems. Natural numbers are introduced using Peano's system of axioms. Inherent in the last Peano axiom is his Principle of Induction, one of the fundamental postulates of arithmetic on natural numbers. Among the myriad of applications of this principle, we discuss here the Division Algorithm for Integers along with the greatest common divisor and prime factorization.

To mollify the complexity of the exposition, the longer and more demanding passages are interrupted by reflections back to the past; how ancient Greeks multiplied natural numbers by systematic doubling and halving; and why the concept of negative numbers took almost a millennium, making a circuitous route beginning with China, through the Hellenistic Alexandria, and India, and finally to settle down in its permanent place in European mathematics.

1.1 Natural Numbers

Leopold Kronecker (1823–1891), the 19th century German mathematician, is often quoted saying *"God made the whole numbers, all else is the work of man."* Deviating from the customary translation "natural numbers" of the original German phrase "die ganzen Zahlen" (and not "natürlichen Zahlen"), we insisted here on the literal rendering. This phrase may ambiguously refer to the set of natural numbers $\mathbb{N}$ or to the larger set of integers $\mathbb{Z}$. Kronecker asserts the divinity of these numbers,

G. Toth, *Elements of Mathematics*, Undergraduate Texts in Mathematics,
https://doi.org/10.1007/978-3-030-75051-0_1

Fig. 1.1 The Egyptian God Heh on the back panel of a chair of the pharaoh Tutankhamun (c. 1342 – c. 1325 BCE)

and it actually took over two millennia of "work of man" to gain full understanding of them.

History

People of ancient times seldom dared to grasp the concept of infinity, a characteristic feature of $\mathbb{N}$ (and also $\mathbb{Z}$). Ancient Egyptians used the hieroglyph of the seated man looking at the stars in the sky with upraised arms to designate their largest number customarily translated as "million." This sign is also used for "Heh," the deification of infinity or eternity, literally "endlessness," also depicted as a god crouching on the gold-sign and holding a palm stem in each hand. The base of the stem is usually continuously covered with notches whereby each notch represents one year and the base of the stem may end in a "tadpole," the Egyptian sign for 100,000. The literal meaning of this composition is "millions of years," an ambitious well-wish for long after-life of the king. (See Figure 1.1 with the cartouche to the left of Heh enclosing Tutankhamun's Son of Ra name: "The living image of Amun.")

Mathematicians in ancient India were familiar with large numbers; for example, there is an extant religious sacrificial formula from the Vedic period (c. 800 – c. 500 BCE) invoking powers of ten from 100 to 1,000,000,000,000.

The best recorded ancient treatise of very large numbers is "The Sand Reckoner" by Archimedes who made a brilliant attempt to size up the whole world by counting the amount of grains of sand that could fit into the universe. (See the epigraph above to this chapter.)

It was not until the 19th century, however, that mathematicians realized the need of placing the set of natural numbers $\mathbb{N}$ (and thereby $\mathbb{Z}$ and $\mathbb{Q}$, etc.) to axiomatic foundation. The key feature of the set of natural numbers $\mathbb{N}$ is that it possesses a

successor (or primitive recursion) which, for any natural number $a \in \mathbb{N}$ (however large), provides its successor $\mathcal{S}(a) \in \mathbb{N}$. (As we will see below, for $a \in \mathbb{N}$, the successor $\mathcal{S}(a)$ also provides the initial step in defining addition in $\mathbb{N}$ by declaring $\mathcal{S}(a) = a + 1$.) As recognized by the German mathematician Hermann Grassmann (1809 – 1877) and fully developed by the Italian mathematician Giuseppe Peano, the existence of a successor $\mathcal{S}$ cannot be proved but has to be postulated as an axiom.

A triple $(\mathbb{N}, \mathcal{S}, 1)$ is called a **natural number system** if $\mathbb{N}$ is a set, called the set of natural numbers, $\mathcal{S} : \mathbb{N} \to \mathbb{N}$ is a self-map of $\mathbb{N}$, called the successor, and $1 \in \mathbb{N}$ is a marked element, called "one." The following axioms are required:

(P1) 1 is not in the range of $\mathcal{S}$;
(P2) $\mathcal{S} : \mathbb{N} \to \mathbb{N}$ is injective;
(P3) Let $A \subset \mathbb{N}$ be a subset with the following properties: $1 \in A$ and whenever $a \in A$, then $\mathcal{S}(a) \in A$. Then $A = \mathbb{N}$.

Axiom P3 is called **Peano's Principle of Induction**. This is the most complex and most frequently used axiom. We have already used this a few times in the previous chapter, and it will recur below and in later chapters in various settings.

Remark Revisiting briefly the Zermelo–Fraenkel system of axioms, we recall the first few von Neumann ordinals:

$$0 = \{\} = \emptyset,$$

$$1 = \{0\} = 0 \cup \{0\} = \{\emptyset\},$$

$$2 = \{0, 1\} = 1 \cup \{1\} = \{\emptyset, \{\emptyset\}\},$$

$$3 = \{0, 1, 2\} = 2 \cup \{2\} = \{\emptyset, \{\emptyset\}, \{\emptyset, \{\emptyset\}\}\},$$

$$4 = \{0, 1, 2, 3\} = 3 \cup \{3\} = \{\emptyset, \{\emptyset\}, \{\emptyset, \{\emptyset\}\}, \{\emptyset, \{\emptyset\}, \{\emptyset, \{\emptyset\}\}\}\},$$

$$5 = \{0, 1, 2, 3, 4\} = 4 \cup \{4\},$$

$$= \{\emptyset, \{\emptyset\}, \{\emptyset, \{\emptyset\}\}, \{\emptyset, \{\emptyset\}, \{\emptyset, \{\emptyset\}\}\}, \{\emptyset, \{\emptyset\}, \{\emptyset, \{\emptyset\}\}, \{\emptyset, \{\emptyset\}, \{\emptyset, \{\emptyset\}\}\}\}\}, \quad \text{etc.}$$

We say that a set $\mathcal{I}$ is **inductive** if $0 \in \mathcal{I}$ and, for every $x \in \mathcal{I}$, the successor of x, $\mathcal{S}(x) = x \cup \{x\}$, is also contained in $\mathcal{I}$. Using this, we see that $\mathbb{N}_0$ must be contained in all inductive sets. Hence, we can **define** $\mathbb{N}_0$ as the smallest inductive set:

$$\mathbb{N}_0 = \bigcap_{\mathcal{I} \text{ inductive}} \mathcal{I} = \{n \mid \forall\, \mathcal{I} \text{ inductive } (n \in \mathcal{I})\}.$$

The Axiom of Infinity asserts that there is at least one inductive set. By the Axiom Schema of Specification, $\mathbb{N}_0$ (and hence $\mathbb{N}$) is defined within ZF.

The first question we wish to settle is **unicity** of the natural number system. Clearly, unicity can only be expected up to one-to-one correspondence since the Roman numerals {I, II, III, IV, V, VI, VII, VIII, IX, X, XI, . . .} or the **binary num-**

ber system $\{1, 10, 11, 100, 101, 110, 111, 1000, \ldots\}$ and their declared successors and distinguished elements I and 1 also serve as natural number systems.
With this, the unicity in question can be stated in the following:

Proposition 1.1.1 *Let* $(\mathbb{N}, \mathcal{S}, 1)$ *and* $(\mathbb{N}', \mathcal{S}', 1')$ *be natural number systems. Then there exists a one-to-one correspondence* $f : \mathbb{N} \to \mathbb{N}'$ *such that* $f(1) = 1'$ *and* $f \circ \mathcal{S} = \mathcal{S}' \circ f$.

Proof We define f as follows. Let $f(1) = 1'$, and, if $f(a)$, $a \in \mathbb{N}$, is defined, then we define $f(\mathcal{S}(a)) = \mathcal{S}'(f(a))$. If $D \subset \mathbb{N}$ is the domain of definition of f (that is, $D \subset \mathbb{N}$ is the set of all natural numbers for which f is defined), then $1 \in D$ ($f(1) = 1'$), and, by the above, $a \in D$ ($f(a)$ exists) implies $\mathcal{S}(a) \in D$ ($f(\mathcal{S}(a)) = \mathcal{S}'(f(a))$). By P3, Peano's Principle of Induction, we have $D = \mathbb{N}$. Therefore $f : \mathbb{N} \to \mathbb{N}'$ is a map defined everywhere in $\mathbb{N}$.

Switching the roles of $\mathbb{N}$ and $\mathbb{N}'$, we obtain a map $g : \mathbb{N}' \to \mathbb{N}$ satisfying $g(1') = 1$ and $g(\mathcal{S}'(a')) = \mathcal{S}(g(a'))$, $a' \in \mathbb{N}'$.

Consider the composition $g \circ f : \mathbb{N} \to \mathbb{N}$. We have $(g \circ f)(1) = g(f(1)) = g(1') = 1$, and

$$(g \circ f)(\mathcal{S}(a)) = g(f(\mathcal{S}(a))) = g(\mathcal{S}'(f(a))) = \mathcal{S}(g(f(a))) = \mathcal{S}((g \circ f)(a)), \ a \in \mathbb{N}.$$

Let $I = \{a \in \mathbb{N} \,|\, (g \circ f)(a) = a\}$. Then, we have $1 \in I$, and, by the computation above, $a \in I$ ($(g \circ f)(a) = a$) implies $\mathcal{S}(a) \in I$ ($(g \circ f)(\mathcal{S}(a)) = \mathcal{S}(a)$). By Peano's Principle of Induction again, $I = \mathbb{N}$. We obtain that the composition $g \circ f$ is the **identity** on $\mathbb{N}$. Similarly, $f \circ g$ is also the identity on $\mathbb{N}'$. Thus, f and g are inverses of each other; in particular, $f : \mathbb{N} \to \mathbb{N}'$ is a one-to-one correspondence with the stated properties. The proposition follows.

From now on we denote the set of natural numbers by $\mathbb{N}$ with $1 \in \mathbb{N}$ and successor $\mathcal{S} : \mathbb{N} \to \mathbb{N}$.

By axiom P1, 1 is not in the range of $\mathcal{S}$. It is natural to ask what the range of the successor is. In fact, axioms P1-P3 imply that the range of $\mathcal{S}$ is precisely $\mathbb{N} \setminus \{1\}$. In other words, 1 is the **only** natural number which is not the successor of any natural number; that is, if $a \in \mathbb{N}$ and $a \neq 1$, then $a = \mathcal{S}(b)$ for some $b \in \mathbb{N}$.

Indeed, consider the set

$$A = \{a \in \mathbb{N} \,|\, a = 1 \text{ or } a = \mathcal{S}(b) \text{ for some } b \in \mathbb{N}\}.$$

Then, $1 \in A$ is a tautology. Letting $a \in A$, $\mathcal{S}(a) \in A$ is again a tautology. Thus, by Peano's Principle of Induction, $A = \mathbb{N}$. This means that the range of $\mathcal{S}$ is $\mathbb{N} \setminus \{1\}$.

History
The first complete and precisely formulated set of axioms for the natural number system was published in 1889 by Peano in his *Arithmetices principia, nova methodo exposita*. As noted above, about three decades earlier Grassmann already recognized the two key elements in this system: The role of the successor and the Principle of Induction. Two precursors of Peano's work were by Charles Sanders Peirce (1839 – 1914) in 1881, and by Richard Dedekind in 1888.

Theorem 1.1.1 *The set of natural numbers $\mathbb{N}$ carries the operations of addition $+$ and multiplication $\cdot$, and they satisfy the following properties (for all $a, b, c \in \mathbb{N}$):*

$$a + (b + c) = (a + b) + c \qquad \textit{(associativity of addition)}$$
$$a + b = b + a \qquad \textit{(commutativity of addition)}$$
$$a \cdot (b \cdot c) = (a \cdot b) \cdot c \qquad \textit{(associativity of multiplication)}$$
$$a \cdot b = b \cdot a \qquad \textit{(commutativity of multiplication)}$$
$$a \cdot (b + c) = a \cdot b + a \cdot c \qquad \textit{(distributivity)}.$$

The proof of this theorem will be carried out in several steps.

We first define **addition** in $\mathbb{N}$ by the following:

$$a + 1 = \mathcal{S}(a) \text{ and } a + \mathcal{S}(b) = \mathcal{S}(a + b), \;\; a, b \in \mathbb{N}.$$

To see that these indeed define addition on the entire set of natural numbers $\mathbb{N}$, consider the set

$$A = \{b \in \mathbb{N} \mid a + b \text{ is defined for all } a \in \mathbb{N}\}.$$

The first part of the definition of addition above shows that $1 \in A$, and the second part shows that $b \in A$ implies $\mathcal{S}(b) \in A$. By Peano's Principle of Induction, we have $A = \mathbb{N}$. Hence addition is defined for all natural numbers.

Next we define **multiplication** in $\mathbb{N}$ by the following:

$$a \cdot 1 = a \quad \text{and} \quad a \cdot \mathcal{S}(b) = a \cdot b + a, \;\; a, b \in \mathbb{N}.$$

Using Peano's Principle of Induction again, we see that multiplication is defined for all natural numbers.

We now proceed to show that the operations of addition and multiplication in $\mathbb{N}$ satisfy the properties listed in the theorem above. The proof will be broken up into several propositions. Note that proper sequencing of the statements is important. We begin with **associativity of addition**.

Proposition 1.1.2 *The addition $+$ is associative in $\mathbb{N}$.*

Proof We need to show that $a + (b + c) = (a + b) + c$ holds for all $a, b, c \in \mathbb{N}$. To do this, we consider the set

$$A = \{c \in \mathbb{N} \mid a + (b + c) = (a + b) + c \text{ for all } a, b \in \mathbb{N}\}.$$

We first claim that $1 \in A$. Indeed, using the definition of addition in three different instances, we have

$$a + (b + 1) = a + \mathcal{S}(b) = \mathcal{S}(a + b) = (a + b) + 1.$$

We obtain $1 \in A$.

Next, assuming $c \in A$, that is, $a + (b + c) = (a + b) + c$ for all $a, b \in \mathbb{N}$, we claim that $\mathcal{S}(c) \in A$. To show this, using the definition of addition three times, we calculate

$$a+(b+\mathcal{S}(c)) = a+\mathcal{S}(b+c) = \mathcal{S}(a+(b+c)) = \mathcal{S}((a+b)+c) = (a+b)+\mathcal{S}(c).$$

This shows that $\mathcal{S}(c) \in A$. By Peano's Principle of Induction, we have $A = \mathbb{N}$. This means that associativity holds throughout $\mathbb{N}$. The proposition follows.

Next we show **commutativity of addition**. We begin with a special case.

Lemma *We have $a + 1 = 1 + a$ for all $a \in \mathbb{N}$.*

Proof Let $C = \{a \in \mathbb{N} \mid a + 1 = 1 + a\}$. Note that $1 \in C$ is a tautology. Assuming $a \in C$, that is, $a + 1 = 1 + a$, we want to show that $\mathcal{S}(a) \in C$. We calculate

$$\mathcal{S}(a) + 1 = \mathcal{S}(\mathcal{S}(a)) = \mathcal{S}(a + 1) = \mathcal{S}(1 + a) = 1 + \mathcal{S}(a),$$

where we used the definition of addition three times. This shows that $\mathcal{S}(a) \in C$. By Peano's Principle of Induction, we have $C = \mathbb{N}$. The lemma follows.

Proposition 1.1.3 *The addition $+$ is commutative in $\mathbb{N}$.*

Proof Let $C = \{b \in \mathbb{N} \mid a + b = b + a \text{ for all } a \in \mathbb{N}\}$. By the lemma just proved, we have $1 \in C$. Assume $b \in C$, that is, $a + b = b + a$, for all $a \in \mathbb{N}$. We calculate

$$\begin{aligned} a + \mathcal{S}(b) &= \mathcal{S}(a + b) = \mathcal{S}(b + a) = b + \mathcal{S}(a) \\ &= b + (a + 1) = b + (1 + a) = (b + 1) + a = \mathcal{S}(b) + a, \end{aligned}$$

where we used the definition of addition several times along with Propositions 1.1.2 and the lemma above. This shows that $\mathcal{S}(b) \in C$. By Peano's Principle of Induction, we have $C = \mathbb{N}$. Commutativity of addition follows.

We now interrupt the natural sequence above, and, instead of proving associativity and commutativity of the multiplication, we turn to distributivity. Since we have not shown commutativity of the multiplication, we actually need to distinguish between **left- and right-distributivity** as follows:

$$a \cdot (b + c) = a \cdot b + a \cdot c \text{ and } (a + b) \cdot c = a \cdot c + b \cdot c, \;\; a, b, c \in \mathbb{N}.$$

Proposition 1.1.4 *Left- and right-distributivity hold in $\mathbb{N}$.*

Proof For left-distributivity, we let

$$D = \{c \in \mathbb{N} \mid a \cdot (b + c) = a \cdot b + a \cdot c \text{ for all } a, b \in \mathbb{N}\}.$$

First, $1 \in D$ since

$$a \cdot (b + 1) = a \cdot \mathcal{S}(b) = a \cdot b + a = a \cdot b + a \cdot 1.$$

Second, assume that $c \in D$, that is, $a \cdot (b + c) = a \cdot b + a \cdot c$ for all $a, b \in \mathbb{N}$. We claim that $\mathcal{S}(c) \in D$. This is somewhat more complex compared to the previous computations. We show all details as follows:

$$\begin{aligned}
a \cdot (b + \mathcal{S}(c)) &= a \cdot \mathcal{S}(b + c) && \text{(definition of addition)} \\
&= a \cdot (b + c) + a && \text{(definition of multiplication)} \\
&= (a \cdot b + a \cdot c) + a && \text{(assumption)} \\
&= a \cdot b + (a \cdot c + a) && \text{(associativity of addition)} \\
&= a \cdot b + (a \cdot c + a \cdot 1) && \text{(definition of multiplication)} \\
&= a \cdot b + a \cdot (c + 1) && (1 \in D) \\
&= a \cdot b + a \cdot \mathcal{S}(c) && \text{(definition of addition).}
\end{aligned}$$

This shows that $\mathcal{S}(c) \in D$. By Peano's Principle of Induction, $D = \mathbb{N}$, and left-distributivity follows.

The argument for right-distributivity is similar. We let

$$D = \{c \in \mathbb{N} \mid (a + b) \cdot c = a \cdot c + b \cdot c \text{ for all } a, b \in \mathbb{N}\}.$$

First, $1 \in D$ since

$$(a + b) \cdot 1 = a + b = a \cdot 1 + b \cdot 1.$$

Second, assume that $c \in D$, that is, $(a + b) \cdot c = a \cdot c + b \cdot c$ for all $a, b \in \mathbb{N}$. We need to show that $\mathcal{S}(c) \in D$. This time we give fewer details as follows:

$$\begin{aligned}
(a + b) \cdot \mathcal{S}(c) &= (a + b) \cdot c + (a + b) = (a \cdot c + b \cdot c) + (a + b) \\
&= a \cdot c + (b \cdot c + a) + b = a \cdot c + (a + b \cdot c) + b \\
&= (a \cdot c + a) + (b \cdot c + b) = a \cdot (c + 1) + b \cdot (c + 1) \\
&= a \cdot \mathcal{S}(c) + b \cdot \mathcal{S}(c).
\end{aligned}$$

Hence $\mathcal{S}(c) \in D$. By Peano's Principle of Induction, $D = \mathbb{N}$. Right-distributivity also follows.

After this detour, we return to the original sequence and show **associativity of multiplication**.

Proposition 1.1.5 *The multiplication $\cdot$ is associative in $\mathbb{N}$.*

Proof We need to show that $a \cdot (b \cdot c) = (a \cdot b) \cdot c$ holds for all $a, b, c \in \mathbb{N}$. As usual, we let

$$A = \{c \in \mathbb{N} \mid a \cdot (b \cdot c) = (a \cdot b) \cdot c \text{ for all } a, b \in \mathbb{N}\}.$$

First, $1 \in A$ since

$$a \cdot (b \cdot 1) = a \cdot b = (a \cdot b) \cdot 1.$$

Second, assume $c \in A$, that is, $a \cdot (b \cdot c) = (a \cdot b) \cdot c$ for all $a, b \in \mathbb{N}$. To show that $\mathcal{S}(c) \in A$, we use left-distributivity just proved and calculate

$$\begin{aligned} a \cdot (b \cdot \mathcal{S}(c)) &= a \cdot (b \cdot c + b) = a \cdot (b \cdot c) + a \cdot b \\ &= (a \cdot b) \cdot c + (a \cdot b) = (a \cdot b) \cdot \mathcal{S}(c). \end{aligned}$$

We obtain $\mathcal{S}(c) \in A$. By Peano's Principle of Induction, we have $A = \mathbb{N}$. Associativity of multiplication follows.

Finally, we show **commutativity of multiplication**. First we prove a special case.

Lemma *We have $1 \cdot a = a$ for all $a \in \mathbb{N}$.*

Proof Let $C = \{a \in \mathbb{N} \mid 1 \cdot a = a\}$. Clearly, $1 \in C$. Assuming $a \in C$, that is, $1 \cdot a = a$, using left-distributivity, we have

$$1 \cdot \mathcal{S}(a) = 1 \cdot (a + 1) = 1 \cdot a + 1 \cdot 1 = a + 1 = \mathcal{S}(a).$$

Thus, $\mathcal{S}(a) \in C$. By Peano's Principle of Induction, we have $C = \mathbb{N}$. The lemma follows.

Proposition 1.1.6 *The multiplication $\cdot$ is commutative in $\mathbb{N}$.*

Proof We let $C = \{b \in \mathbb{N} \mid a \cdot b = b \cdot a \text{ for all } a \in \mathbb{N}\}$. By the lemma just proved, we have $a \cdot 1 = a = 1 \cdot a$, $a \in \mathbb{N}$, so that $1 \in C$. We now assume that $b \in C$, that is, $a \cdot b = b \cdot a$ for all $a \in \mathbb{N}$, and show that $\mathcal{S}(b) \in C$. We calculate

$$a \cdot \mathcal{S}(b) = a \cdot b + a = b \cdot a + a = b \cdot a + 1 \cdot a = (b + 1) \cdot a = \mathcal{S}(b) \cdot a,$$

where we used the previous proposition and right-distributivity asserted by Proposition 1.1.4. Thus, we have $\mathcal{S}(b) \in C$. By Peano's Principle of Induction, we have $C = \mathbb{N}$. The proposition follows.

Summarizing, we accomplished our aim; Propositions 1.1.2–1.1.6 show that the addition and the multiplication are associative and commutative, and they are connected through distributivity. Theorem 1.1.1 follows.

Next we turn to the **cancellation law for addition**.

Proposition 1.1.7 *For $a, b, c \in \mathbb{N}$, the equality $a + c = b + c$ implies $a = b$.*

Proof Let

$$C = \{c \in \mathbb{N} \mid a + c = b + c \text{ for } a, b \in \mathbb{N} \text{ implies } a = b\}.$$

We first claim that $1 \in C$. Indeed, if $a + 1 = b + 1$ for some $a, b \in \mathbb{N}$, then $\mathcal{S}(a) = \mathcal{S}(b)$. By Axiom P2, injectivity of the successor $\mathcal{S}$, we have $a = b$, and the claim follows. Second, assume that $c \in C$, and show that $\mathcal{S}(c) \in C$. Let $a + \mathcal{S}(c) = b + \mathcal{S}(c)$ for some $a, b \in \mathbb{N}$. This means that $a + (c + 1) = b + (c + 1)$. By associativity of the addition, this rewrites as $(a + c) + 1 = (b + c) + 1$. Since $1 \in C$, it follows that $a + c = b + c$. Now, by assumption, $c \in C$, so that $a = b$ follows.

Remark It is natural to expect the cancellation law for multiplication: For $a, b, c \in \mathbb{N}$, the equality $a \cdot c = b \cdot c$ implies $a = b$. We defer the proof of this after the study of the natural ordering on $\mathbb{N}$.

Addition in $\mathbb{N}$ defines an **ordering** of the natural numbers: For $a, b \in \mathbb{N}$, we define $a < b$ (or $b > a$) if $b = a + c$ for some $c \in \mathbb{N}$.

We claim that $<$ is a **strict total order** on $\mathbb{N}$.

Transitivity is a consequence of associativity of the addition. Indeed, if $a < b$ and $b < c$, then $b = a + d$ and $c = b + e$ for some $d, e \in \mathbb{N}$. Therefore, we have $c = b + e = (a + d) + e = a + (d + e)$, and $a < c$ follows.

Trichotomy is asserted in the following:

Proposition 1.1.8 *For any $a, b \in \mathbb{N}$, exactly one of the following is true: $a < b$, $a = b$, and $a > b$.*

The proof is preceded by the following.

Lemma *For any $a, b \in \mathbb{N}$, we have $a \neq a + b$.*

Proof We let

$$A = \{a \in \mathbb{N} \mid a \neq a + b \text{ for all } b \in \mathbb{N}\}.$$

First, $1 \in A$ since, by P1, we have $1 \neq \mathcal{S}(b) = b + 1 = 1 + b$ for any $b \in \mathbb{N}$. Second, assume that $a \in A$, that is, we have $a \neq a + b$ for all $b \in \mathbb{N}$. We claim that $\mathcal{S}(a) \in A$, that is, $\mathcal{S}(a) \neq \mathcal{S}(a) + b$ for all $b \in \mathbb{N}$. Since $\mathcal{S}(a) + b = b + \mathcal{S}(a) = \mathcal{S}(b + a) = \mathcal{S}(a + b)$, we need to show that $\mathcal{S}(a) \neq \mathcal{S}(a + b)$ for all $b \in \mathbb{N}$. By P2, this is equivalent to $a \neq a + b$ for all $b \in A$, which was our assumption. The lemma follows.

Corollary *If $ab = 1$ for some $a, b \in \mathbb{N}$, then $a = b = 1$.*

Proof Assuming that this is false, there exist $a, b \in \mathbb{N}$ such that $a \neq 1 \neq b$ and $ab = 1$. Then $a = \mathcal{S}(c)$ and $b = \mathcal{S}(d)$ for some $c, d \in \mathbb{N}$. Hence, we have

$$ab = \mathcal{S}(c)\mathcal{S}(d) = \mathcal{S}(c)d + \mathcal{S}(c) = \mathcal{S}(\mathcal{S}(c)d + c) = 1.$$

This contradicts to P1.

With these preparations, we are now ready to prove **trichotomy**.

Proof of Proposition 1.1.8 The fact that the conditions $a < b$, $a = b$, and $a > b$, $a, b \in \mathbb{N}$, are mutually exclusive follows from the lemma above. Indeed, if $a < b$, then $a + c = b$ for some $c \in \mathbb{N}$. This implies that $a \neq b$ (since otherwise we would have $a = b = a + c$) and that $a \not> b$ (since otherwise $a = b + d$ for some $d \in \mathbb{N}$, so that we would have $a = b + d = (a + c) + d = a + (c + d)$). The other cases are similar.

It remains to prove that **one** of the conditions $a < b$, $a = b$, and $a > b$, $a, b \in \mathbb{N}$, always holds. To show this, we fix $a \in \mathbb{N}$ and define

$$T_a = \{b \in \mathbb{N} \mid a < b \text{ or } a = b \text{ or } a > b\}.$$

First, $1 \in T_a$. Indeed, if $a = 1$, then this is a tautology ($1 = 1$). If $a \neq 1$, then a is in the range of the successor $\mathcal{S}$, so that $a = \mathcal{S}(c)$ for some $c \in \mathbb{N}$. We thus have $a = \mathcal{S}(c) = c + 1 = 1 + c$. This gives $1 < a$. Hence $1 \in T_a$ holds.

Second, let $b \in T_a$, so that $a < b$ or $a = b$ or $a > b$. Accordingly, we distinguish three cases.

Case I. $a < b$. We have $b = a + c$ for some $c \in \mathbb{N}$. Using this, we calculate $\mathcal{S}(b) = \mathcal{S}(a + c) = a + \mathcal{S}(c)$ so that $a < \mathcal{S}(b)$. This gives $\mathcal{S}(b) \in T_a$.

Case II. $a = b$. We have $\mathcal{S}(b) = b + 1 = a + 1$ so that $a < \mathcal{S}(b)$. This gives $\mathcal{S}(b) \in T_a$.

Case III. $a > b$. We have $a = b + c$ for some $c \in \mathbb{N}$. If $c = 1$, then $a = b + 1 = \mathcal{S}(b)$, so that $\mathcal{S}(b) \in T_a$. If $c \neq 1$, then c is in the range of the successor $\mathcal{S}$, so that $c = \mathcal{S}(d)$ for some $d \in \mathbb{N}$. This gives

$$a = b + c = b + \mathcal{S}(d) = b + (d + 1) = b + (1 + d) = (b + 1) + d = \mathcal{S}(b) + d.$$

This means that $\mathcal{S}(b) < a$, and in particular, we have $\mathcal{S}(b) \in T_a$.

Summarizing, in all three cases, we have $\mathcal{S}(b) \in T_a$. By Peano's Principle of Induction, we obtain $T_a = \mathbb{N}$. Thus, for any $a, b \in \mathbb{N}$, we have $a < b$ or $a = b$ or $a > b$. The proposition follows.

The strict total ordering $<$ defines a total ordering $\leq$ on $\mathbb{N}$ in the usual way: $a \leq b$ (or $b \geq a$), $a, b \in \mathbb{N}$, if $a = b$ or $a < b$. As discussed in Section 0.2, the total ordering means that $\leq$ is **transitive**, **antisymmetric**, and **total**.

We now derive the **cancellation law for multiplication**. Actually, we can state somewhat more as follows:

Proposition 1.1.9 *Let $a, b, c \in \mathbb{N}$. We have $a < b$ if and only if $a \cdot c < b \cdot c$.*

Proof Assume $a < b$. Then we have $a + d = b$ for some $d \in \mathbb{N}$. Using distributivity, we obtain $(a + d) \cdot c = a \cdot c + d \cdot c = b \cdot c$. This gives $a \cdot c < b \cdot c$. Conversely, assume $a \cdot c < b \cdot c$. By trichotomy, we have $a < b$ or $a = b$ or $a > b$. First, $a = b$ cannot hold since then $a \cdot c = b \cdot c$ would follow, a contradiction. Second, $a > b$ cannot hold since, by what we just proved, $a \cdot c > b \cdot c$ would follow, a contradiction. Thus, we obtain $a < b$. The proposition follows.

$1 \in \mathbb{N}$ is the **multiplicative identity** in the sense that $1 \cdot a = a$ for all $a \in \mathbb{N}$. It is the unique natural number with this property; that is, if $1' \in \mathbb{N}$ such that $1' \cdot a = a$ for all (actually some) $a \in \mathbb{N}$, then $1' = 1$. Indeed, this follows from Proposition 1.1.9 above. If $1' \cdot a = a = 1 \cdot a$, $a \in \mathbb{N}$, then the cancellation law for multiplication gives $1' = 1$.

We now show that $\mathbb{N}$ is **well-ordered** with respect to $\leq$.

Theorem 1.1.2 *Any non-empty set $A \subset \mathbb{N}$ has an infimum $\inf A \in A$, a unique least element with respect to the total ordering $\leq$.*

Proof First, a least element of a non-empty subset of $\mathbb{N}$ must be unique. Indeed, if $a \in A$ and $a' \in A$ both are least elements, then we have simultaneously $a \leq a'$ and $a' \leq a$. Since $\mathbb{N}$ is totally ordered, $a = a'$ follows.

To show the existence of the least element, let

$$L = \{a \in \mathbb{N} \mid \text{any subset } A \subset \mathbb{N} \text{ with } a \in A \text{ has a least element}\}.$$

The statement of the theorem amounts to showing that $L = \mathbb{N}$.

First, we claim that 1 is the least element of the whole $\mathbb{N}$. This will imply $1 \in L$.

Indeed, if $1 \neq a \in \mathbb{N}$, then $a = \mathcal{S}(b)$ for some $b \in \mathbb{N}$. By Proposition 1.1.8, we have $1 < a$ or $1 > a$. The second inequality is impossible since $1 > a = \mathcal{S}(b)$ implies $1 = \mathcal{S}(b) + c = \mathcal{S}(b + c)$ for some $c \in \mathbb{N}$, and this contradicts to P1. The claim follows.

Second, assume that $a \in L$, and show that $\mathcal{S}(a) \in L$. Let $A \subset \mathbb{N}$ be such that $\mathcal{S}(a) \in A$. We need to show that A has a least element. We may assume that $a \notin A$ since otherwise A has a least element by assumption ($a \in L$).

We now extend the set A to the set $B = A \cup \{a\} \subset \mathbb{N}$. Since $a \in B$ (and $a \in L$), B has a **least element**, b, say. Since $a \in B$, we have $b \leq a$, or equivalently, we have two cases: $a > b$ and $a = b$.

The first case implies that $a \neq b$ (see the lemma after Proposition 1.1.8), so that we have $b \in A$. Since b was a least element of B, we see that b is also a least element of the smaller set A. In this case we are done.

We are left with the second case $a = b$. We claim that $\mathcal{S}(b)$ is a least element of A. Letting $c \in A$ be a general element, we need to show that $\mathcal{S}(b) \leq c$. Now, in addition to $a = b \leq c$, we also know that $b = a \notin A$ while $c \in A$. This means that $b \neq c$, and consequently, we have the sharp inequality $b < c$. Let $d \in \mathbb{N}$ such that $b + d = c$. If $d = 1$, then $b + 1 = \mathcal{S}(b) = c$ (in particular, $\mathcal{S}(b) \leq c$), and we are done. If $d \neq 1$, then $d = \mathcal{S}(e)$ for some $e \in \mathbb{N}$, and hence $b + \mathcal{S}(e) = \mathcal{S}(b + e) = \mathcal{S}(b) + e = c$. This gives $\mathcal{S}(b) < c$, and we are done in this case as well. Thus, $\mathcal{S}(b)$ is the least element of A.

Summarizing, we obtain that $a \in L$ implies $\mathcal{S}(a) \in L$. By Peano's Principle of Induction, $L = \mathbb{N}$. The theorem follows.

Corollary *Any non-empty set $A \subset \mathbb{N}$ which is bounded above has a supremum $\sup A \in A$, a unique greatest element with respect to the total ordering $\leq$.*

Proof Let $\emptyset \neq A \subset \mathbb{N}$ be bounded above. If A has an upper bound which belongs to A, then it is the supremum of A and we are done. Otherwise, let $B \subset \mathbb{N}$ be the set of all upper bounds of A which do not belong to A. By Theorem 1.1.2, $b = \inf B$ exists and it is an element of B. Now, $b \neq 1$ since A is non-empty (and 1 is the least element of the whole $\mathbb{N}$). Let $a \in \mathbb{N}$ such that $\mathcal{S}(a) = b$. We claim that a is an upper bound for A.

Indeed, if $c \in A$, then $c < b$ so that $c + d = b$ for some $d \in \mathbb{N}$. If $d = 1$, then $\mathcal{S}(c) = c + 1 = b = \mathcal{S}(a)$, so that, by P2, we have $c = a$; in particular, $c \leq a$. If $d \neq 1$, then $d = \mathcal{S}(e)$ for some $e \in \mathbb{N}$. Then $\mathcal{S}(c + e) = c + \mathcal{S}(e) = c + d = b = \mathcal{S}(a)$. By P2 again, we have $c + e = a$; in particular, $c < a$. The claim follows.

Now, if $a \notin A$, then, by definition, $a \in B$. This cannot happen since $\mathcal{S}(a) = b = \inf B$. Hence $a \in A$. Therefore, $a = \sup A$, and the corollary follows.

Remark The existence of the least element in any subset of $\mathbb{N}$, Theorem 1.1.2, actually **implies** P3, Peano's Principle of Induction, provided we assume that the range of the successor is all $\mathbb{N}$ but 1.

Indeed, assume that any non-empty subset of $\mathbb{N}$ has a least element. Proceeding in the contrapositive, assume that P3 fails; that is, there exists a **proper** subset $A \subset \mathbb{N}$, $A \neq \mathbb{N}$, such that $1 \in A$, and whenever $a \in A$, then $\mathcal{S}(a) \in A$. By assumption, the (non-empty) complement $B = \mathbb{N} \setminus A$ has a least element $b \in B$, say. Since $1 \in A$, we have $b \neq 1$; in particular, b is in the range of the successor, $b = \mathcal{S}(a) = a + 1$ for some $a \in \mathbb{N}$. Now, $a \notin B$ since $a < b$ and b was a least element in B. Hence $a \in A$. By assumption, $\mathcal{S}(a) = b \in A$, which contradicts to $b \in B$.

Example 1.1.1 Let $m \in \mathbb{N}$ be even, and define

$$A_m = \{n \in \mathbb{N} \,|\, n^2 + 2m \cdot n \text{ is a perfect square}\}.$$

We claim that A_m is bounded above, and $\sup A_m = (m/2)^2 - m + 1$.

Indeed, let $n \in A_m$ so that $n^2 + 2m \cdot n = l^2$ for some $l \in \mathbb{N}$, and let $k = n + m \in \mathbb{N}$. We have[1] $k^2 = (n + m)^2 = n^2 + 2n \cdot m + m^2 = l^2 + m^2$. Hence $(k - l)(k + l) = k^2 - l^2 = m^2$. Now, $k - l$ and $k + l$ have the same parity (since they differ by the even number $2l$) so that they both must be even (since m is even). Thus, we have

$$\frac{k-l}{2} \cdot \frac{k+l}{2} = \left(\frac{m}{2}\right)^2, \quad \frac{k-l}{2}, \frac{k+l}{2} \in \mathbb{N}.$$

Combining this with

$$k = \frac{k-l}{2} + \frac{k+l}{2},$$

we see that the largest value of k (and hence the largest value of $n = k - m \in A_m$) occurs when $(k - l)/2 = 1$. This and $(k + l)/2 = (m/2)^2$ give $k = (m/2)^2 + 1$. We obtain that $n = (m/2)^2 - m + 1$ is the largest number within A_m.

[1] Thus, (l, m, k) is a Pythagorean triple (see Section 5.7), but we will not need this fact.

Another corollary of the fact that $\mathbb{N}$ is well-ordered is the **Archimedean Property** as follows:

Corollary *Let $a, b \in \mathbb{N}$. Then there exists $n \in \mathbb{N}$ such that $b \leq na$.*

Proof Assume that, for some $a, b \in \mathbb{N}$, we have $b > na$ **for all** $n \in \mathbb{N}$. We reformulate this by saying that the set

$$A = \{na \mid n \in \mathbb{N}\}$$

is bounded above (since b is an upper bound). By the previous corollary, $Na = \sup A \in A$ for some $N \in \mathbb{N}$. Then $Na + a = (N + 1)a \in A$, but $Na + a > Na = \sup A$, a contradiction.

History
The name "Archimedean Property" is a misnomer; it was attributed to Archimedes of Syracuse by Otto Stolz (1842 – 1905) in the 1880s since it appears as Axiom V in Archimedes' work *On the Sphere and Cylinder*. This property also appears in Euclid's Elements as Definition 4: *"Magnitudes are said to have a ratio to one another which can, when multiplied, exceed one another."* Archimedes himself attributed this property to Eudoxus of Cnidus (c. 390 – c. 337 BCE).

Returning to the main line, from now on, for our natural number system $\mathbb{N}$, we adopt the **Hindu-Arabic numeral system**, a positional decimal numerical system, and abandon the use of the successor $\mathcal{S}$ that proved to be so useful in this section. The term **positional** refers to the use of the same glyph for different orders of magnitude, and **decimal** refers to base ten (or denary) arithmetic. The glyphs are the **digits** 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, and the base ten is written as 10. The orders of magnitude are the powers of ten, and the notation uses specific positions for each power: units, tens, hundreds, thousands, ten thousands, etc. The position of each digit within a given number stands for the digit multiplied by the power of ten corresponding to the position of the digit. The powers of ten are lined up sequentially from right to left; each position is ten times the value of the position to its immediate right. Displaying a natural number using positional decimals is called the **decimal representation** of that number.

For example, on September 16, 2018, the US National Debt in decimal representation was

$$\$\,21,432,542,252,109 = \$\,2 \cdot 10^{13} + 1 \cdot 10^{12} + 4 \cdot 10^{11} + 3 \cdot 10^{10} + 2 \cdot 10^{9}$$
$$+5 \cdot 10^{8} + 4 \cdot 10^{7} + 2 \cdot 10^{6} + 2 \cdot 10^{5} + 5 \cdot 10^{4} + 2 \cdot 10^{3} + 1 \cdot 10^{2} + 9.$$

Remark In the last expression, we used (the first time) powers of 10; for example, $10^2 = 100$, $10^3 = 1{,}000$, $10^4 = 10{,}000$, $10^5 = 100{,}000$, etc. Strictly speaking, they are defined **inductively** (that is, using Peano's Principle of Induction). We let $10^1 = 10$, and, assuming that 10^n, $n \in \mathbb{N}$, is defined, we set $10^{n+1} = 10 \cdot 10^n$. Note that powers of other bases, such as 2 and 3 (see below), can be defined analogously.

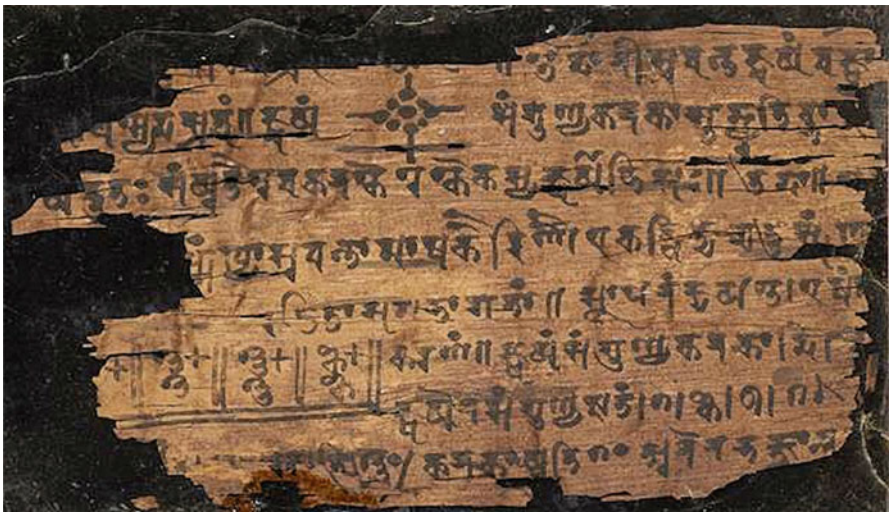


Fig. 1.2 A leaf from the Bakhshali manuscript showing 0 in the bottom register seventh from the right; National Geographic.

History

The abstract mathematical concept of numbers should not be confused with **numerals**, symbols that are used to represent them. The first ciphered numeral system was invented by the ancient Egyptians who used strokes for units and different signs for 10 (hobble without cross-bar), 100 (coil of rope), 1,000 (lotus plant), 10,000 (finger), 100,000 (tadpole), etc. The ancient Greeks used letters from the Ionian and Doric alphabets to denote their numerals while the Romans used combinations of letters from the Roman alphabet.

The Hindu-Arabic numeral system was originally invented by Indian mathematicians between the 1st and 4th centuries. The Indian mathematician Bhāskara I (c. 600 – c. 680 CE) wrote numbers in the Hindu decimal system; and, in addition, he was also the first who used circle for zero. Although disputed of its age (224 – 383 to 885 – 993 CE by carbon dating extremely fragile parts), the allegedly oldest Indian document, the Bakhshali manuscript (written on 70 leaves of birch barks), contains the sequence of Hindu numerals[2] 1–9 and also a small circle for zero. (See Figure 1.2 of the leaf that contains the Hindu numerals.) Note that much earlier Archimedes of Syracuse in "the Sand Reckoner" (see the epigraph to this chapter) invented a decimal positional system which was based on 10^8. It is also worth noting that the Roman and Chinese numerals, even though based on powers of ten, are non-positional numeral systems.

Around the 9th century the original Hindu numeral system was introduced to the Islamic world by the Arabic mathematicians Muhammad ibn Mūsā Al-Khwārizmī (c. 780 – 850) in his book *On the Calculation with Hindu Numerals* (c. 825) and Al-Kindi (801 – 873) in his four-volume book *On the Use of the Hindu Numerals* (c. 830). The Roman system dominated Europe until the late Middle Ages (late 14th century) when it was replaced by the far superior Arabic numeral system. The reason of superiority of the Arabic numerals lies in its positional nature with principal role played by the symbol for zero.

[2]In contrast, for a theory advocating the Chinese origin of the Hindu-Arabic numerals, see the works of Lam Lay Yong of the National University of Singapore.

Finally, note that the ten glyphs of the Hindu-Arabic numerals were originally Brahmi numerals (3rd century BCE); the glyphs 1, 2, 3, 4, 5, 6, 7, 8, 9, 0 used today are Latin/Roman script numbers. The exact origins of these glyphs are clear only for the first three: 1, 2, 3 (especially 3, for example, on the Bakhshali manuscript) corresponding to the Roman *I*, *II*, *III* and also the horizontal bars of the Chinese versions.

Example 1.1.2 Fill in the following eight squares with the numbers 1, 2, 3, 4, 5, 6, 7, 8 as digits (each used only once) to obtain valid equations: $\square\square = \square \times \square$ and $\square\square = \square \times \square$. Show that the solution is essentially unique up to commutativity.

Notice first that, since there are four even numbers, at least one of the single digits of the two double digit numbers must be even. Now, a simple inspection gives $12 = 3 \times 4$ and $56 = 7 \times 8$, as well as unicity.

We close this section by a brief description of how the ancient Greeks performed multiplication by ingeniously halving and doubling the factors of a product.

Example 1.1.3 Suppose we want to multiply two natural numbers a and b. We will do this by systematically halving a and, at the same time, doubling b. We define the sequence $a_0, a_1, a_2, \ldots$ as follows: $a_0 = a$, $a_1 = [a_0/2]$, $a_2 = [a_1/2], \ldots$; in general, $a_{n+1} = [a_n/2]$, $n \in \mathbb{N}_0$. (Here, for $c \in \mathbb{R}$, $[c]$ denotes the **greatest integer** $\leq c$; that is, for $n \leq c < n+1$ with $n \in \mathbb{N}_0$, we have $[c] = n$.) Clearly, this sequence has a last non-zero member where we stop. Now, the **product** ab is the **sum** of those iterated doubles $2^n b$ of b for which a_n is **odd**.

It is not difficult to see why this procedure gives the correct answer for the product. At the nth stage of the process, if a_n is even, then $a_n = 2[a_n/2] = 2a_{n+1}$ so that transferring the 2 factor to b does not change the product from the nth stage to the $(n+1)$th: $a_n \cdot 2^n b = a_{n+1} \cdot 2 \cdot 2^n b = a_{n+1} \cdot 2^{n+1} b$. If, on the other hand, a_n is odd, then $a_n = 2[a_n/2] + 1 = 2a_{n+1} + 1$ so that $a_n \cdot 2^n b = (2a_{n+1} + 1) \cdot 2^n b = a_{n+1} \cdot 2 \cdot 2^n b + 2^n b = a_{n+1} \cdot 2^{n+1} b + 2^n b$. The extra terms $2^n b$ that pile up for every odd a_n (including the last one which is 1) therefore give the product.

As a specific example, let $a = 18$ and $b = 27$. We tabulate the halves of a and the doubles of b as follows:

2^n	a_n	$2^n b$
1	18	27
2	9	54
2^2	4	108
2^3	2	216
2^4	1	432

The odd a's in the sequence are 9 and 1. The corresponding b's are 54 and 432. The sum of these, $54 + 432 = 486$, is the product $18 \cdot 27$.

It is enlightening to observe[3] that the powers 2^n corresponding to the odd a_n add up to a. In our example, 9 and 1 correspond to 2 and 2^4 with sum $2 + 2^4 = 18$.

[3]The author is indebted to one of the reviewers for pointing this out.

With this, we have $18 \cdot 27 = (2 + 2^4) \cdot 27 = 2 \cdot 27 + 2^4 \cdot 27 = 54 + 432 = 486$. Thus, in general, this method amounts to write a as a sum of powers of 2 and use distributivity, along with systematic doubling, to obtain $a \cdot b$.

Exercises

1.1.1. Let $2 \leq n \in \mathbb{N}$. Define the successor $\mathcal{S}$ on the set $\mathbb{N}_n = \{1, 2, \ldots, n\}$ by $\mathcal{S}(m) = m + 1$, $1 \leq m < n$, $m \in \mathbb{N}_n$, and (a) $\mathcal{S}(n) = 1$; (b) $\mathcal{S}(n) = 2$; (c) $\mathcal{S}(n) = n$. Which Peano axiom(s) fail in (a)-(c)?

1.1.2. Given $a \in \mathbb{N}$, show that there is no natural number $b \in \mathbb{N}$ such that $a < b < \mathcal{S}(a) = a + 1$.

1.1.3. Show that

$$1 + 3 + \cdots + (2n - 1) = n^2, \quad n \in \mathbb{N}.$$

1.2 Integers

Despite Kronecker's assertion of divinity, the set of **natural** numbers $\mathbb{N}$ does not have an **additive identity**, a specific number that, when added to a number, does not change the number itself. This specific number is **zero**, 0, which belongs to the set of integers $\mathbb{Z}$ but not to $\mathbb{N}$. (This deficiency is the reason why some mathematicians count 0 as a natural number.)

Still staying with addition, the set of integers $\mathbb{Z}$ possesses another useful property. Any integer a has an **additive inverse**, a number that, when added to a, gives 0. The additive inverse of a is its **negative** $-a$, and the stated property can be written as $a + (-a) = -a + a = 0$.

History

Negative numbers appeared first in one of the earliest Chinese texts in mathematics, *The Nine Chapters on the Mathematical Art*, composed by scholars during the 10th – 2nd century BCE. Note that Chapter 8 of this work uses Gaussian elimination predating Carl Friedrich Gauss almost two millennia. The positive and negative numbers are represented by red and black counting rods, respectively. Using these methods the Chinese were able to solve simultaneous equations with negative coefficients and negative roots.

An early theory of linear and quadratic equations was developed by the Hellenistic mathematician Diophantus of Alexandria (c. 200/214 – 284/298 CE) and the Indian mathematician Brahmagupta (597 – 668 CE); although in his *Arithmetica*, Diophantus claimed an equation equivalent to $4x + 20 = 0$ as being absurd since it has negative solution.

In the 7th century CE, there was a widespread use of negative numbers in India to represent debt. In his work *Brahma–Sphuta–Siddhanta* (c. 628 CE), Brahmagupta gave general rules of operations involving zero and negative numbers.

Surprisingly, European mathematicians resisted using negative numbers; for example, in the study of cubic equations contained in his *Ars Magna*, Gerolamo Cardano (1501 – 1576) refused to move a (linear) term with positive coefficient to the other side of the equation.

Finally, it was Gottfried Wilhelm Leibniz who considered the set of negative numbers as an "integral" part of his infinitesimal calculus.

We now proceed to construct the set of integers $\mathbb{Z}$ from $\mathbb{N}$. We represent each **integer** by a pair (a, b) of natural numbers, an element of the Cartesian product $\mathbb{N} \times \mathbb{N}$. (Intuitively, we think of (a, b) as being "$a - b$.") The pair $(a, b) \in \mathbb{N} \times \mathbb{N}$ represents the same integer as the pair $(a', b') \in \mathbb{N} \times \mathbb{N}$ if and only if $a + b' = a' + b$. (Continuing with our intuition, the last equality is equivalent to "$a - b = a' - b'$.") To put this into a precise framework, we define a relation $\sim$ on $\mathbb{N} \times \mathbb{N}$ such that, for $(a, b), (a', b') \in \mathbb{N} \times \mathbb{N}$, we have $(a, b) \sim (a', b')$ if $a + b' = a' + b$.

We first claim that $\sim$ is an **equivalence relation** on $\mathbb{N} \times \mathbb{N}$. The properties of reflexivity, $(a, b) \sim (a, b)$, $(a, b) \in \mathbb{N} \times \mathbb{N}$, and symmetry, $(a, b) \sim (a', b')$ implies $(a', b') \sim (a, b)$, $(a, b), (a', b') \in \mathbb{N} \times \mathbb{N}$, are tautologies. Finally, transitivity, $(a, b) \sim (a', b')$ and $(a', b') \sim (a'', b'')$ imply $(a, b) \sim (a'', b'')$, $(a, b), (a', b'), (a'', b'') \in \mathbb{N} \times \mathbb{N}$, also follows since adding $a + b' = a' + b$ and $a' + b'' = a'' + b'$ and using the cancellation law in $\mathbb{N}$ (Proposition 1.1.7 of the previous section) along with commutativity and associativity of the addition, $a + a' + b' + b'' = a' + a'' + b + b'$ gives $a + b'' = a'' + b$.

The equivalence relation $\sim$ partitions $\mathbb{N} \times \mathbb{N}$ into equivalence classes. We define the **set of integers** as the quotient $\mathbb{Z} = \mathbb{N} \times \mathbb{N}/\sim$, the set of equivalence classes in $\mathbb{N} \times \mathbb{N}$ by the equivalence relation $\sim$. In other words, by an **integer**, an element of $\mathbb{Z}$, we mean an equivalence class in $\mathbb{N} \times \mathbb{N}$ via $\sim$.

We now define the operations of **addition** $+$ and **multiplication** $\cdot$ in $\mathbb{Z}$ in terms of representatives as

$$(a, b)+(c, d)=(a+c, b+d) \text{ and } (a, b)\cdot(c, d)=(ac+bd, ad+bc),\ \ a, b, c, d \in \mathbb{N}.$$

We need to show that these operations are well-defined in $\mathbb{Z}$; that is, the definitions do not depend on the representatives chosen. We let $(a, b), (a', b'), (c, d), (c', d') \in \mathbb{N} \times \mathbb{N}$ such that $(a, b) \sim (a', b')$ and $(c, d) \sim (c', d')$. By the definition of the equivalence, these give the pair of equations $a+b' = a'+b$ and $c+d' = c'+d$. Adding, we have $a+c+b'+d' = a'+c'+b+d$. We rewrite this in terms of the equivalence relation $\sim$ as $(a+c, b+d) \sim (a'+c', b'+d')$ and, in terms of the addition, as $(a, b) + (c, d) \sim (a', b') + (c', d')$. Hence, addition is well-defined in $\mathbb{Z}$.

Returning to our pair of equations above, we have

$$(a+b')c+(a'+b)d+a'(c+d')+b'(c'+d)=(a'+b)c+(a+b')d+a'(c'+d)+b'(c+d').$$

Expanding (using distributivity in $\mathbb{N}$) and using the cancellation law for addition in $\mathbb{N}$, the "hybrid terms" cancel, and we obtain $ac + bd + a'd' + b'c' = a'c' + b'd' + ad + bc$. We rewrite this in terms of the equivalence relation $\sim$ as $(ac + bd, ad + bc) \sim (a'c' + b'd', a'd' + b'c')$ and, in terms of the multiplication, as $(a, b) \cdot (c, d) \sim (a', b') \cdot (c', d')$.

Thus, multiplication in $\mathbb{Z}$ is well-defined.

Next, we claim that the operations of addition and multiplication in $\mathbb{Z}$ satisfy the same properties as those for $\mathbb{N}$. In other words, we claim that the **addition is associative and commutative** and the **cancellation law** holds, **multiplication is associative and commutative**, and they are connected through **distributivity**.

Commutativity of addition and multiplication, the cancellation law for addition, and associativity of addition follow immediately from the definitions even on the level of representatives of the equivalence classes (in $\mathbb{N} \times \mathbb{N}$). Distributivity and associativity also follow by simple and direct computations (once again on the level of representatives).

The structure of the quotient $\mathbb{Z} = \mathbb{N} \times \mathbb{N}/\sim$ is actually fairly simple as each equivalence class carries a **unique** representative. To get to this, we let $(a, b) \in \mathbb{N} \times \mathbb{N}$ be given and seek a unique representative within the equivalence class of (a, b). Using trichotomy, we split the treatment into three cases: $a > b$, $a = b$, and $a < b$.

If $a > b$, then $a = b + c$ for some $c \in \mathbb{N}$. Then $(c + 1, 1) \sim (a, b)$ and $(c + 1, 1)$ is unique within the equivalence class as its second component is 1, the least natural number. The entire equivalence class is $\{(c + d, d) \in \mathbb{N} \times \mathbb{N} \mid d \in \mathbb{N}\}$.

If $a = b$, then the unique representative of the corresponding equivalence class is $(1, 1)$, and the whole equivalence class is $\{(d, d) \in \mathbb{N} \times \mathbb{N} \mid d \in \mathbb{N}\}$.

If $a < b$, then $a + c = b$ for some $c \in \mathbb{N}$. Then $(1, c + 1) \sim (a, b)$ and $(1, c + 1)$ is unique within the equivalence class as its first component is 1. The entire equivalence class is $\{(d, c + d) \in \mathbb{N} \times \mathbb{N} \mid d \in \mathbb{N}\}$.

We summarize these as follows:

$$(a, b) \sim \begin{cases} (c + 1, 1) \text{ if } a > b \text{ with } a = b + c \\ (1, 1) \text{ if } a = b \\ (1, c + 1) \text{ if } a < b \text{ with } a + c = b. \end{cases}$$

Thus, by trichotomy, $\{(c+1, 1) \mid c \in \mathbb{N}\} \cup \{(1, 1)\} \cup \{(1, c+1) \mid c \in \mathbb{Z}\}$ is a **complete set of representatives** of the equivalence classes.

Remark Although we have been pursuing here an algebraic approach, using the plane $\mathbb{R}^2$, the following simple geometric picture emerges (see Figure 1.3). The set $\mathbb{N} \times \mathbb{N}$ is a positive integer lattice in $\mathbb{R}^2$; the equivalence classes under $\sim$ are equidistantly spaced along the lines $y = x - c$, $c \in \mathbb{Z}$, and the unique representatives above are equidistantly lined up in the "perimeter" of the lattice along the two half-lines $y = 1$, $x \geq 1$ and $x = 1$, $y \geq 1$.

We now verify that $\mathbb{Z}$ is an **extension** of $\mathbb{N}$. To do this, we define the map $\iota : \mathbb{N} \to \mathbb{Z}$ such that, for $c \in \mathbb{N}$, the range $\iota(c)$ is the equivalence class of $(c + 1, 1)$ in $\mathbb{Z}$. Since, for $c \neq d$, $c, d \in \mathbb{N}$, we have $(c + 1, 1) \not\sim (d + 1, 1)$, we immediately see that ι is an injective map. Using the definitions of addition and multiplication, for $c, d \in \mathbb{N}$, we have

$$(c + 1, 1) + (d + 1, 1) = (c + d + 2, 2) \sim (c + d + 1, 1)$$

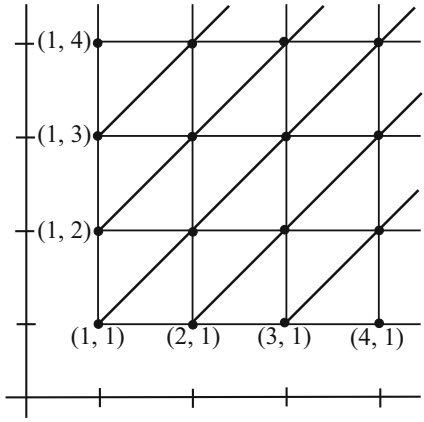


Fig. 1.3 The integers as the quotient $\mathbb{N} \times \mathbb{N}/\sim$.

$$\begin{aligned}(c+1,1)\cdot(d+1,1) &= ((c+1)(d+1)+1,(c+1)+(d+1))\\ &= (cd+c+d+2,c+d+2)\sim(cd+1,1).\end{aligned}$$

Taking equivalence classes, these give

$$\iota(c+d)=\iota(c)+\iota(d) \text{ and } \iota(c\cdot d)=\iota(c)\cdot\iota(d),\quad c,d\in\mathbb{N}.$$

These show that the embedding ι can be used to **identify** $\mathbb{N}$ with its range in $\mathbb{Z}$ under ι, and, under this identification, the arithmetic operations performed in $\mathbb{N}$ are the same as those in $\mathbb{Z}$. In what follows, for $c\in\mathbb{N}$, the equivalence class of $(c+1,1)\in\mathbb{Z}$ will also be **denoted** by the single letter c. In other words, we identify $\mathbb{N}$ with its range under ι in $\mathbb{Z}$ and write $c\in\mathbb{N}\subset\mathbb{Z}$ in place of the equivalence class of $(c+1,1)$ in $\mathbb{Z}$.

Recall that a complete set of representatives of the equivalence classes in $\mathbb{Z}$ is $\{(c+1,1)\,|\,c\in\mathbb{N}\}\cup\{(1,1)\}\cup\{(1,c+1)\,|\,c\in\mathbb{N}\}$. Continuing with simplifying the notation, we denote by 0, the zero, the equivalence class of $(1,1)$, and, for $c\in\mathbb{N}$, we denote by $-c$, the negative of c, the equivalence class of $(1,c+1)$. With these, we have

$$\mathbb{Z}=\{c\,|\,c\in\mathbb{N}\}\cup\{0\}\cup\{-c\,|\,c\in\mathbb{N}\}=\{0,\pm1,\pm2,\pm3,\ldots\}.$$

The justification for these notations is as follows.

First, for $c\in\mathbb{N}$, we have $(c+1,1)+(1,1)=(c+2,2)\sim(c+1,1)$, and this gives $c+0=0+c=c$, $c\in\mathbb{N}$. Similarly, $(1,c+1)+(1,1)=(2,c+2)\sim(1,c+1)$, yielding $-c+0=0+(-c)=-c$. Since obviously $0+0=0$, these can be compactly expressed as

$$a+0=0+a=a,\quad a\in\mathbb{Z},$$

where we used the single letter a ($= \pm c$, $c \in \mathbb{N}$, or 0) as a generic integer in $\mathbb{Z}$. This shows that 0 is an **additive identity** in $\mathbb{Z}$.

Remark The additive identity is unique. Indeed, if $0'$ is another additive identity, then we have $0 = 0 + 0' = 0' + 0 = 0'$.

Second, according to our conventions, the equivalence class of (-1) is represented by $(1, 2)$, and, for $c \in \mathbb{N}$, we have $(1, 2) \cdot (c + 1, 1) = (c + 3, 2c + 3) \sim (1, c + 1)$. Taking equivalence classes, we obtain $-c = (-1) \cdot c$, $c \in \mathbb{N}$. Since $(1, 2) \cdot (1, 2) = (5, 4) \sim (2, 1)$, we have $(-1)^2 = (-1) \cdot (-1) = 1$, so that $-(-c) = c$, $c \in \mathbb{N}$. We now define $-a = (-1) \cdot a$, $a \in \mathbb{Z}$, the **negative** of a. With this we have $-(-a) = a$, $a \in \mathbb{Z}$. Moreover, commutativity and associativity of the multiplication in $\mathbb{Z}$ give

$$(-a) \cdot b = a \cdot (-b) = -(a \cdot b), \quad a, b \in \mathbb{Z}.$$

Third, for $c \in \mathbb{N}$, we have $(c + 1, 1) + (1, c + 1) = (c + 2, c + 2) \sim (1, 1)$ so that $c + (-c) = -c + c = 0$, $c \in \mathbb{N}$. Since $-(-c) = c$, $c \in \mathbb{N}$, these give

$$a + (-a) = -a + a = 0, \quad a \in \mathbb{Z}.$$

We obtain that $-a$ is an **additive inverse** of $a \in \mathbb{Z}$.

Remark Since the cancellation law holds for addition, the additive inverse is unique.

Fourth, $(c+1, 1) \cdot (1, 1) = (c+2, c+2) \sim (1, 1)$, $c \in \mathbb{N}$, gives $c \cdot 0 = 0 \cdot c = 0$. This immediately generalizes to

$$a \cdot 0 = 0 \cdot a = 0, \quad a \in \mathbb{Z}.$$

The converse of this also holds, and it is an important tool in factoring to be discussed later. We have

$$a \cdot b = 0, a, b \in \mathbb{Z} \;\Rightarrow\; a = 0 \text{ or } b = 0.$$

Indeed, for the contrapositive statement, we may assume $a, b \in \mathbb{N}$, and then clearly $a \cdot b \in \mathbb{N}$.

Remark The cancellation law for multiplication in $\mathbb{Z}$ says that, for $a, b, c \in \mathbb{Z}$, if $a \neq 0$ and $a \cdot b = a \cdot c$, then $b = c$. This is the direct consequence of the property above. The detailed steps of the proof are as follows:

$$\begin{aligned} a \cdot b &= a \cdot c && \text{(assumption)} \\ a \cdot b + (-(a \cdot c)) &= 0 && \text{(additive inverse)} \\ a \cdot b + a \cdot (-c) &= 0 && \text{(multiplicative property)} \\ a \cdot (b + (-c)) &= 0 && \text{(distributivity)} \end{aligned}$$

$$\begin{aligned} b + (-c) &= 0 && (a \neq 0) \\ b &= c && \text{(additive inverse).} \end{aligned}$$

Fifth, since 1, represented by (2, 1), is the multiplicative identity in $\mathbb{N}$, we also have

$$a \cdot 1 = 1 \cdot a = a, \ \ a \in \mathbb{Z}.$$

Recall from the beginning of this section that we intuitively thought of the pair $(a, b) \in \mathbb{N} \times \mathbb{N}$ to represent "$a - b$." To complete the circle, note that we have $(a, b) \sim (a + 2, b + 2) = (a + 1, 1) + (1, b + 1)$, so that the equivalence class of (a, b) is given by $a + (-b)$ confirming our intuition. As a final simplification, from now on, we denote $a + (-b)$ by $a - b$, $a, b \in \mathbb{Z}$.

A natural ordering $<$ on the integers $\mathbb{Z}$ is given as follows: For $a, b \in \mathbb{Z}$, we define $a < b$ (or $b > a$) if $b - a \in \mathbb{N}$.

We quickly observe that the ordering $<$ on $\mathbb{Z}$ is an extension on the ordering in $\mathbb{N}$. This is because, according to our earlier definition, $a < b$, $a, b \in \mathbb{N}$, if $a + c = b$ for some $c \in \mathbb{N}$, and this is equivalent to $(c =) b - a \in \mathbb{N}$.

As expected, the extension $<$ remains a **strict total order** on the set of integers $\mathbb{Z}$.

Transitivity is clear. Trichotomy means that, for any $a, b \in \mathbb{Z}$, exactly one of the following holds: $a < b$, $a = b$, and $a > b$. Indeed, letting $c = b - a \in \mathbb{Z}$, exactly one of the following holds: $c \in \mathbb{N}$, $c = 0$, and $-c \in \mathbb{N}$. These three cases give $a < b$, $a = b$, and $a > b$, respectively.

We call an integer $c \in \mathbb{Z}$ **positive** if $c > 0$ and **negative** if $c < 0$. Clearly, $c \in \mathbb{Z}$ is positive if and only if $c \in \mathbb{N}$, and $c \in \mathbb{Z}$ is negative if and only if $-c \in \mathbb{N}$. Moreover, in $\mathbb{Z}$, the usual arithmetic properties hold: For any $a, b, c \in \mathbb{Z}$, (1) $a < b$ implies $-a > -b$; (2) $a < b$ implies $a + c < b + c$; (3) $a < b$ implies $ac < bc$ if $c > 0$; (4) $a < b$ implies $ac > bc$ if $c < 0$; etc.

As usual, we also define $a \leq b$ (or $b \geq a$), $a, b \in \mathbb{Z}$, if $a = b$ or $a < b$. Equivalently, $a \leq b$ if and only if $c = b - a$ is either a natural number or zero.

The set of integers $\mathbb{Z}$ with $\leq$ is a **totally ordered set**; it is transitive, antisymmetric, and total.

Finally, $\mathbb{Z}$ has the Least Upper Bound Property (Section 0.2). A stronger statement is the following:

Proposition 1.2.1 *If a non-empty set $A \subset \mathbb{Z}$ is bounded above, then* sup A *exists, and it is attained in A. If A is bounded below, then* inf A *exists, and it is attained in A.*

Proof It is enough to prove one of these statements. Assume that the non-empty set $A \subset \mathbb{Z}$ is bounded below, and let $b \in \mathbb{Z}$ be a lower bound. Consider the set $B = \{a - b + 1 \,|\, a \in A\} \subset \mathbb{Z}$. For $a \in A$, we have $a \geq b$ so that $a - b + 1 > 0$. We obtain $B \subset \mathbb{N}$. Since $\mathbb{N}$ is well-ordered, we know that inf $B \in B$ exists. Clearly, inf $A =$ inf $B + b - 1$.

Remark $\mathbb{Z}$ is **not** well-ordered with respect to its natural ordering. Indeed, as a consequence of the Archimedean Property, the set of all negative integers $-\mathbb{N} = \{-n \in \mathbb{Z} \mid n \in \mathbb{N}\}$ does not have a lower bound.

We introduce the **absolute value** $|\cdot| : \mathbb{Z} \to \mathbb{N}_0$. For $a \in \mathbb{Z}$, we define

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

Proposition 1.2.2 *Let $c \in \mathbb{N}$. Then, for $a \in \mathbb{Z}$, we have $-c \leq a \leq c$ if and only if $|a| \leq c$. The same holds for strict inequalities.*

Proof Let $a \in \mathbb{N}_0$, that is, $a = |a|$. Then, for $c \in \mathbb{N}$, we have $a \leq c$ if and only if $|a| \leq c$, while $-c \leq a$ obviously holds.

Let $-a \in \mathbb{N}_0$, that is, $-a = |a|$. Then $-c \leq a$ if and only if $-a \leq c$ if and only if $|a| \leq c$, while $a \leq c$ obviously holds. The proposition follows.

Corollary *We have*

$$||a| - |b|| \leq |a + b| \leq |a| + |b|, \quad a, b \in \mathbb{Z}.$$

Remark The second inequality is usually called the **triangle inequality** based on its generalization to $\mathbb{R}^2$. We will discuss this later.

Proof We first show the second inequality. By the previous proposition, we have $-|a| \leq a \leq |a|$ and $-|b| \leq b \leq |b|$. Adding, we obtain $-(|a| + |b|) \leq a + b \leq |a| + |b|$. Again by this proposition, we have $|a + b| \leq |a| + |b|$.

The first inequality is a special case of the second. Indeed, we have $|a| = |(a + b) - b| \leq |a + b| + |b|$ and hence $|a| - |b| \leq |a + b|$. Switching the roles of a and b, Make a line space here.

Finally, note that the **decimal representation** of natural numbers naturally extends to that of integers. For a negative integer $a \in \mathbb{Z}$, we take the decimal representation of the natural number $-a \in \mathbb{N}$ and insert a negative sign in front of the representation. (The decimal representation of zero is 0 itself.)

Exercises

1.2.1. Derive the identity $-(a - b) = b - a, a, b \in \mathbb{Z}$.
1.2.2. Show that the equation $1 - a = a, a \in \mathbb{Z}$, has no solution.

1.3 The Division Algorithm for Integers

The **division algorithm for integers** states that upon dividing an integer $n \in \mathbb{Z}$ by a **non-zero** integer $0 \neq d \in \mathbb{Z}$, we obtain an integral **quotient** $q \in \mathbb{Z}$ and an integral **remainder** $r \in \mathbb{Z}$:

$$\frac{n}{d} = q + \frac{r}{d}.$$

The remainder r is always **non-negative** and satisfies the inequality

$$0 \leq r < |d|.$$

As we will see below, n and d uniquely determine the quotient q and the remainder r. The number n that we start with is called the **dividend**, and the non-zero number d that n is divided by is called the **divisor**. Since we wish to stay in the realm of the integers, the equation above is usually written in the form of an equality of integers.

Division Algorithm (Integers) *For any $n, d \in \mathbb{Z}$, $d \neq 0$, there exist unique $q, r \in \mathbb{Z}$ such that*

$$n = q \cdot d + r, \quad 0 \leq r < |d|.$$

Proof Let $n, d \in \mathbb{Z}$, $d \neq 0$.

We first show existence. By changing the sign of the quotient q if needed, we may assume that the divisor d is positive. Let

$$A = \{n - q \cdot d \,|\, q \in \mathbb{Z} \text{ such that } n - q \cdot d \geq 0\}.$$

For $q = -|n| \in \mathbb{Z}$, we have $n - qd = n + |n|d \geq n + |n| \geq 0$ so that $A \neq \emptyset$. Since A is bounded below by zero, by Proposition 1.2.1, it has a unique infimum which is attained: $r = \inf A \geq 0$ with $r = n - qd$ for some $q \in \mathbb{Z}$. We claim that $r < d$. Indeed, if $r \geq d$, then $n - (q+1)d = n - qd - d = r - d \geq 0$, and this contradicts to the minimality of r. Existence follows.

It remains to show uniqueness of q, r. Assume $n = qd + r = q'd + r'$ with $0 \leq r, r' < |d|$. These give $(q - q')d = r' - r$. Assuming $q \neq q'$, we have $|d| \leq |d||q - q'| = |r' - r| < |d|$, a contradiction. Hence $q = q'$ and also $r = r'$.

Given $n \in \mathbb{Z}$ and $0 \neq d \in \mathbb{Z}$, we say that d **divides** n, written as $d \,|\, n$, if $n = qd$ for some $q \in \mathbb{Z}$. In other words, d divides n if, upon division by d, we have zero remainder, $r = 0$.

Let $a, b \in \mathbb{Z}$ with at least one of them non-zero. The **greatest common divisor** of a and b, written as $\gcd(a, b)$, is a **natural number** $d \in \mathbb{N}$ such that (1) $d \,|\, a$ and $d \,|\, b$, and (2) $e \,|\, a$ and $e \,|\, b$, $e \in \mathbb{N}$, imply $e \,|\, d$.

In other words, $\gcd(a, b)$ is a common divisor of a and b, and any common divisor $e \in \mathbb{N}$ of a and b also divides $\gcd(a, b)$.

Example 1.3.1 Let $a \in \mathbb{N}_0$, and consider the infinite sequence $(a_n)_{n\in\mathbb{N}_0}$ defined by $a_n = a^2 + n^2$, $n \in \mathbb{N}$.[4] Letting $d_n = \gcd(a_n, a_{n+1})$, $n \in \mathbb{N}$, show that $\max_{n\in\mathbb{N}} d_n = 4a^2 + 1$.

We have

$$\begin{aligned} d_n &= \gcd(a_n, a_{n+1}) = \gcd(a^2 + n^2, a^2 + (n+1)^2) \\ &= \gcd(a^2 + n^2, a^2 + n^2 + 2n + 1) = \gcd(a^2 + n^2, 2n + 1) \\ &= \gcd(4(a^2 + n^2), 2n + 1) = \gcd(4a^2 + 1 + 4n^2 - 1, 2n + 1) \\ &= \gcd(4a^2 + 1 + (2n - 1)(2n + 1), 2n + 1) = \gcd(4a^2 + 1, 2n + 1), \end{aligned}$$

where multiplying by 4 is allowed since $2n + 1$ is odd. The example follows.

Proposition 1.3.1 *Let $a, b \in \mathbb{Z}$ with at least one of them non-zero. Then $\gcd(a, b)$ is the unique infimum of the set*

$$A_{a,b} = \{m \cdot a + n \cdot b \,|\, m, n \in \mathbb{Z} \text{ such that } m \cdot a + n \cdot b > 0\}.$$

In particular, $\gcd(a, b)$ exists, and it is also unique.

Proof Letting $m = a$ and $n = b$, we have $ma + nb = a^2 + b^2 > 0$ since at least one of a or b is non-zero. This shows that $A_{a,b}$ is non-empty.

By Theorem 1.1.2, $d = \inf A_{a,b}$ exists, and it is attained in $A_{a,b}$; that is, we have $d = ma + nb$ for some $m, n \in \mathbb{Z}$.

We first prove that d is a common divisor of a and b. Due to symmetry, we only need to show $d \,|\, a$. Using the division algorithm, we have $a = qd + r$ with $0 \le r < d$. We calculate

$$r = a - qd = a - q(ma + nb) = (1 - qm)a - qnb.$$

Since this is a remainder, we have $r \ge 0$. If $r > 0$, then we have $r \in A_{a,b}$. This is a contradiction since $r < d$ and d is the infimum in $A_{a,b}$. Thus, $r = 0$ and so $d \,|\, a$.

Second, if $e \in \mathbb{N}$ is a common divisor of a and b, then it is clearly also a common divisor of $d = ma + nb$. Existence of the greatest common divisor follows.

For unicity, assume that $d \in \mathbb{N}$ and $d' \in \mathbb{N}$ are both greatest common divisors of a and b. Then, we have $d \,|\, d'$ and $d' \,|\, d$; that is, we have $d' = ed$ and $d = e'd'$ for some $e, e' \in \mathbb{N}$. These give $d = e'd' = e'e'd$ so that, canceling, we obtain $ee' = 1$. By Corollary to Proposition 1.1.8, we get $e = e' = 1$. Thus, $d = d'$, and unicity follows.

[4] A special case ($a = 10$) was a problem in the American Invitational Mathematics Examination, 1983.

Two integers a and b with at least one of them non-zero are called **relatively prime** if $\gcd(a, b) = 1$. In other words, a and b have no common divisors (other than ± 1).

Corollary *Let $a, b, d \in \mathbb{Z}$, $d \neq 0$. If $d \mid ab$ and $\gcd(d, a) = 1$ (d and a are relatively prime), then $d \mid b$.*

Proof The condition on the greatest common divisor implies that $md + na = 1$ for some $m, n \in \mathbb{Z}$. Multiplying through b, we obtain $mdb + nab = b$. Thus, if $d \mid ab$, then $d \mid mdb + nab = b$. The corollary follows.

A natural number $p \geq 2$ is called a **prime** if the only natural number that divides p is 1 and the number p itself. We denote by Π the set of all primes:

$$\begin{aligned}\Pi = \{&2, 3, 5, 7, 11, 13, 17, 19, 23, 29, 31, 37, 41, 43, 47, 53, 59, 61, 67, 71, 73, 79, 83, 89\\ &97, 101, 103, 107, 109, 113, 127, 131, 137, 139, 149, 151, 157, 163, 167, 173,\\ &179, 181, 191, 193, 197, 199, 211, 223, 227, 229, 233, 239, 241, 251, 257, \ldots,\\ &2^{57,885,161} - 1, \ldots, 2^{74,207,281} - 1, \ldots, 2^{77,232,917} - 1, \ldots, 2^{82,589,933} - 1, \ldots\},\end{aligned}$$

where we indicated the largest known primes as of January 2020 (the previous two were discovered in 2014 and 2017, respectively).

An integer $n \in \mathbb{Z}$, $n \neq 0, \pm 1$, is called a **composite number** if $|n|$ is **not** a prime.

We now digress from the main line momentarily and introduce another form of Peano's Principle of Induction that we will need frequently in the sequel.

Let $(P_n)_{n \in \mathbb{N}} = (P_1, P_2, P_3, \ldots)$ be an infinite sequence of **statements**. Assume that P_1 holds, and, for any $n \in \mathbb{N}$, whenever P_n holds then so does P_{n+1}. Then the **Principle of Mathematical Induction** asserts that P_n holds **for all** $n \in \mathbb{N}$.

The Principle of Mathematical Induction is a simple consequence of Peano's Principle of Induction. Indeed, let $A = \{n \in \mathbb{N} \mid P_n \text{ is valid}\}$. Then the assumptions on the sequence $(P_1, P_2, P_3, \ldots)$ in the Principle of Mathematical Induction above translate to the following: $1 \in A$, and whenever $n \in A$, then we also have $n + 1 \in A$. By Peano's Principle of Induction, we have $A = \mathbb{N}$. This simply means that P_n is valid for all $n \in \mathbb{N}$.

We use the Principle of the Mathematical Induction, or **induction**, for short, when we need to prove infinitely many statements $P_1, P_2, P_3, \ldots$ at the same time.

The proof of P_1 is called the **initial step**, and "P_n implies P_{n+1}," $n \in \mathbb{N}$, is called the **general induction step**. In the latter, P_n is called the **induction hypothesis**, and the general induction step is often written symbolically as $n \Rightarrow n + 1$.

Example 1.3.2 In Example 0.3.1, we claimed (without proof) that if $f : X \to Y$ is an injective map between finite sets X and Y of the same cardinality, $|X| = |Y|$, then f must be surjective. We now show this by induction on the number of elements $n = |X| = |Y|$.

The initial step when both X and Y are singletons is clear. For the general induction step $n \Rightarrow n + 1$, assume that the statement holds for $n = |X| = |Y|$. Let $f : X \to Y$ be an injective map between sets X and Y with cardinality

$n+1 = |X| = |Y|$. Let $x_0 \in X$, and denote $y_0 = f(x_0) \in Y$. Finally, let $X_0 = X \setminus \{x_0\}$ and $Y_0 = Y \setminus \{y_0\}$. Since $f : X \to Y$ is injective, there is no other element in X than x_0 that maps to y_0. This means that f can be restricted to an injective map $f_0 : X_0 \to Y_0$, $f_0(x) = f(x)$, $x \in X_0$. Since $|X_0| = |Y_0| = n$, the induction hypothesis applies, and we obtain that $f_0 : X_0 \to Y_0$ is surjective. Since $f(X) = f_0(X_0) \cup \{f(x_0)\} = Y_0 \cup \{y_0\} = Y$, we obtain that $f : X \to Y$ is surjective as well. The statement follows.

In the Principle of Mathematical Induction, the general induction step may be modified to the effect that whenever $P_1, P_2, \ldots, P_n$ hold then so does P_{n+1}. This may be indicated in writing as $1, 2, \ldots, n \Rightarrow n+1$. This is sometimes called the strong form of the Principle of Mathematical Induction. This is a misnomer since this variant is actually equivalent to the original form of the Principle of Mathematical Induction. To see this, given an infinite sequence of statements $P_1, P_2, P_3, \ldots$, one needs to define, for any $n \in \mathbb{N}$, the statement $Q_n = P_1 \wedge P_2 \wedge \cdots \wedge P_n$, the logical conjunction of $P_1, P_2, \ldots, P_n$ (that is, Q_n is true if and only if $P_1, P_2, \ldots, P_n$ are all true).

Finally, note that the Principle of Mathematical Induction does not necessarily have to start at $n = 1$ since the indices can be shifted so that the index of the initial step becomes 1.

We motivate the next result by a simple question: What is the smallest $n \in \mathbb{N}$ such that $120n$ is a perfect square? To answer this, we write $120 = 2^3 \cdot 3 \cdot 5$ and realize that we need to make the exponents even. With this, we have $n = 2 \cdot 3 \cdot 5 = 30$, so that $120n = 2^4 \cdot 3^2 \cdot 5^2 = (2^2 \cdot 3 \cdot 5)^2 = 60^2$.

Fundamental Theorem of Arithmetic *Any integer $a \geq 2$ is either a prime number itself, or it can be written as a product of primes uniquely up to order of the factors.*

Proof This is an example for an induction which starts at $a = 2$. This initial step is obviously true since $a = 2$ is a prime. To perform the general induction step, we use the second version $2, 3, 4, \ldots n \Rightarrow n+1$ as follows. Assume that the statement is true for $a = 2, 3, 4, \ldots, n$. Consider $n+1$. If $n+1$ is a prime, then we are done. If $n+1$ is not a prime then, by definition, n has a divisor n_1 satisfying $1 < n_1 < n+1$. Then $n+1 = n_1 \cdot n_2$, where $n_2 = (n+1)/n_1$ also satisfies $1 < n_2 < n+1$. By the induction hypothesis, both n_1 and n_2 are primes or products of primes. Thus, $n+1 = n_1 \cdot n_2$ is also a product of primes. Finally, unicity of the prime factors follows directly from Corollary to Proposition 1.3.1 since distinct primes are relatively prime.

Example 1.3.3 For what $n \in \mathbb{N}$ is the integer $n^4 - 360n^2 + 400$ a prime?

We will write this expression as a product of two integers.[5] The crux is to use $400 = 20^2$ to calculate

[5]Note that the typical trick of letting $m = n^2$ does not work since, in terms of m, the expression above gives $m^2 - 360m + 400 = (m-180)^2 - 32000$, but the constant 32000 is **not** a perfect square.

$$n^4 - 360n^2 + 400 = n^4 - 360n + 20^2 = (n^2 + 20)^2 - 400n^2$$
$$= (n^2 + 20)^2 - (20n)^2 = (n^2 - 20n + 20)(n^2 + 20n + 20).$$

Since the first factor on the right-hand side is smaller than the second, in order for the original expression to be a prime, we must have $n^2 - 20n + 20 = 1$. This gives

$$n^2 - 20n + 19 = (n - 1)(n - 19) = 0.$$

Thus, we obtain $n = 1$ and $n = 19$. Finally, $1^4 - 360 \cdot 1^2 + 400 = 41$ and $19^4 - 360 \cdot 19^2 + 400 = 361^2 - 360 \cdot 361 + 400 = 361 + 400 = 761$, and both of these are primes.

Example 1.3.4 Show that, for any $n \in \mathbb{N}$, the numbers $n^3 + 2n$ and $n^4 + 3n^2 + 1$ are relatively prime.[6]

We have $n^3 + 2n = n(n^2 + 2)$ and $n^4 + 3n^2 + 1 = n^2(n^2 + 2) + n^2 + 1$. If a prime $p \in \mathbb{N}$ divides both, then it would also divide $n^2 + 1$. This, however, is relatively prime to $n(n^2 + 2)$.

Beyond their obvious use in simple arithmetic in simplifying fractions, the **greatest common divisor** plays a fundamental role in mathematics, notably in number theory. Recall that, when dealing with fractions, we call a fraction a/b, $a, b \in \mathbb{Z}$, $b \neq 0$, **irreducible** (or reduced) if a and b are relatively prime.

If the prime factorizations of a and b are known, then the greatest common divisor $\gcd(a, b)$ is easy to obtain; we first collect only the common prime factors, then raise each to the lower power that the prime factor participates in either of the factorizations, and finally create a product with these prime powers.

For example, to calculate the greatest common divisor gcd(17640, 3300), we first use the prime factorizations $17640 = 2^3 \cdot 3^2 \cdot 5 \cdot 7^2$ and $3300 = 2^2 \cdot 3 \cdot 5^2 \cdot 11$. Comparing, we arrive at is $\gcd(17640, 3300) = 2^2 \cdot 3 \cdot 5 = 60$. Thus, the fraction $3300/17640$ can be divided through 60 to obtain the irreducible fraction $55/294$.

Prime factorization works very well for small numbers, but it is very inefficient for large values. In rare cases, some clever shortcuts sporadically show up in mathematical contests as in the following:

Example 1.3.5 What is the prime factorization of the number 3, 374, 784?

Observe that $150^3 = 3,375,000$ so that $3,374,784 = 150^3 - 6^3 = 6^3(25^3 - 1)$ (or better yet, note that 3, 374, 784 is divisible by $2^3 \cdot 3^3 = 6^3$). In addition, we have $25^3 - 1 = (25 - 1)(25^2 + 25 + 1) = 24 \cdot 651 = 2^3 \cdot 3^2 \cdot 7 \cdot 31$. Putting all these together, we obtain $3,374,784 = 2^6 \cdot 3^5 \cdot 7 \cdot 31$.

A much more efficient method of finding the greatest common divisor is the **Euclidean algorithm**. This is based on the principle that $\gcd(a, b)$ divides any linear combination $ma + nb$ with $m, n \in \mathbb{Z}$.

[6] An equivalent version of this was a USSR Mathematics Olympiad problem.

To perform the Euclidean algorithm, we may assume that $a > b \geq 0$, start by dividing a by b and obtain the first remainder r_1. We then divide b by r_1, and obtain the second remainder $r_2 < r_1$. We continue this process until we obtain a zero remainder. The **last non-zero divisor** is then equal to $\gcd(a, b)$. We tabulate this process as follows:

$$\begin{aligned}
a &= bq_1 + r_1, \quad 0 \leq r_1 < b \\
b &= r_1 q_2 + r_2, \quad 0 \leq r_2 < r_1 \\
r_1 &= r_2 q_3 + r_3, \quad 0 \leq r_3 < r_2 \\
r_2 &= r_3 q_4 + r_4, \quad 0 \leq r_4 < r_3 \\
&\dots \qquad\qquad \dots \qquad\quad \dots \\
r_{n-3} &= r_{n-2} q_{n-1} + r_{n-1}, \quad 0 \leq r_{n-1} < r_{n-2} \\
r_{n-2} &= r_{n-1} q_n .
\end{aligned}$$

We set the indices here such that $r_n = 0$, so that $\gcd(a, b) = r_{n-1}$.

It is not difficult to see why the Euclidean algorithm gives the greatest common divisor. Recall that $\gcd(a, b)$ is defined by the following: The greatest common divisor is a common divisor of a and b, and any common divisor of a and b divides $\gcd(a, b)$.

With this, we proceed as follows. By the last equation, r_{n-1} divides r_{n-2}. Using this and the next-to-last equation, we see that r_{n-1} also divides r_{n-3}. Proceeding inductively, and working backwards, we see that r_{n-1} divides both r_2 and r_1 and hence also divides b (second equation), but then it must divide a (first equation).

Thus $\gcd(a, b)$ is a common divisor of a and b.

If d is a divisor of a and b, then, by the first equation, it must divide r_1. By the second equation, d then must divide r_2. Proceeding inductively and moving forward, we see that d must divide r_{n-1}.

Thus r_{n-1} is the greatest common divisor of a and b.

Example 1.3.6 Let $a, b \in \mathbb{N}$. Show that the fraction

$$\frac{a(b+1)n + (b+2)}{abn + (b+1)}, \quad n \in \mathbb{N},$$

is irreducible.[7]

We use the Euclidean algorithm as follows:

$$a(b+1)n + (b+2) = (abn + (b+1)) \cdot 1 + (an + 1)$$

[7]Specific cases of this and other variants abound in various mathematical contests; see, for example, the irreducibility of the fraction $(21n + 4)/(14n + 3)$ ($a = 7, b = 2$) in the International Mathematical Olympiad, 1959.

$$abn + (b+1) = (an+1)\cdot b + 1$$
$$an + 1 = 1\cdot(an+1).$$

Thus, $\gcd(a(b+1)n + (b+2), abn + (b+1)) = 1$.

Example 1.3.7 For what prime numbers p do we have solutions $a, b, c, d \in \mathbb{N}$ of the system of equations $a^5 = b^4$, $c^3 = d^2$, $c - a = p$?[8]

Assume we that have a solution $a, b, c, d \in \mathbb{N}$ for a prime p. By the Fundamental Theorem of Arithmetic, in the prime factorization of the number $a^5 = b^4$, all exponents are divisible by 4 and 5 and hence also divisible by 20 (since $\gcd(4, 5) = 1$). This gives $a^5 = b^4 = n^{20}$ for some $n \in \mathbb{N}$. Similarly, $c^3 = d^2 = m^6$ for some $m \in \mathbb{N}$. The solutions of the first two equations of the system can therefore be written as $a = n^4$, $b = n^5$, $c = m^2$, $d = m^3$, $m, n \in \mathbb{N}$. The third equation gives $c - a = m^2 - n^4 = (m - n^2)(m + n^2) = p$. Since p is a prime, we must have $m - n^2 = 1$, and hence $m + n^2 = p$. Solving these, we obtain $m = (p+1)/2$ and $n^2 = (p-1)/2$. The first equation gives $p \neq 2$ (since every prime number beyond 2 is odd). The second equation gives $p = 2n^2 + 1$, $n \in \mathbb{N}$. Summarizing, we obtain that there is a unique solution of the system if and only if the prime p is of the form $2n^2 + 1$ for some $n \in \mathbb{N}$, and then the solution is $a = n^4$, $b = n^5$, $c = (n^2+1)^2$, $d = (n^2+1)^3$.

Note that primes of the form $p = 2n^2 + 1$ abound,[9] e.g., 19, 73, 163, 883, 1153, 1459, 1801, 2179, 2593, 3529, 4051, 8713, 10369, 11251, 15139, 17299, 18433, 19603, etc.

We finish this section by a somewhat challenging and computational example.[10]

Example 1.3.8 Let $f_n = 1! + 2! + \cdots + n!$, $n \in \mathbb{N}$. Find the smallest prime number p such that $p | f_{p-1}$ and $p^2 \nmid f_{p^2-1}$.

Note that, for $m \leq n$, $m, n \in \mathbb{N}$, we obviously have $m | n!$. Hence, the inductive definition $f_n = f_{n-1} + n!$, $2 \leq n \in \mathbb{N}$, shows that, for the prime p as above, we have $p | f_n$ for $p \leq n \in \mathbb{N}$ and $p^2 \nmid f_n$ for $p^2 \leq n \in \mathbb{N}$.

To begin with, we calculate the first few values as follows: $f_1 = 1$, $f_2 = 3$, $f_3 = 3^2$, $f_4 = 3\cdot 11$, $f_5 = 3^2\cdot 17$, $f_6 = 3^2\cdot 97$, $f_7 = 3^4\cdot 73$, $f_8 = 3^2\cdot 11\cdot 467$, $f_9 = 3^2\cdot 131\cdot 347$, $f_{10} = 3^2\cdot 11\cdot 40787$, where we displayed the prime decompositions. Since f_n, $n \in \mathbb{N}$, is odd, the first possible prime is $p = 3$. Not only do we have $3 | f_2$ but also $3^2 | f_8$. For the next two primes, we have $5 \nmid f_4$ and $7 \nmid f_6$.

[8] A special numerical case of this problem ($p = 19$) was in the American Invitational Mathematics Examination, 1985.

[9] It is a yet unsolved conjecture of Hardy that there are infinitely many primes of the form $an^2 + bn + c$, where $a, b, c \in \mathbb{N}$ do not have common divisors, $a > 0$, (at least) one of the numbers $a + b$ and c is odd, and $b^2 - 4ac$ is not a perfect square. See Hardy, G.H., Wright, E.M., *An Introduction to the Theory of Numbers*, 5th ed. Oxford: Clarendon Press, New York, 1979.

[10] Our approach is elementary, and, for some computations, a computer algebra system is recommended.

For the next prime $p = 11$, we have $11|f_{10}$. Finally, the use of a computer algebra system shows that $121 = 11^2 \not| f_{120}$. (Note that the natural number $f_{120}/11$ has 198 digits.) Thus, $p = 11$ is the smallest prime sought.

Actually, some more work gives a clearer picture. By the above, for $1 \leq n < 11$, we have $11|f_n$ if and only if $n = 4, 8, 10$. As for the square, once again a computer algebra system gives that, for $11 \leq n < 11^2$, we have $11^2|f_n$ if and only if $n = 12, 20$. Note that the prime decompositions of the these exceptional cases are $f_{12} = 3^2 \cdot 11^2 \cdot 23 \cdot 20879$ and $f_{20} = 3^2 \cdot 11^2 \cdot 53 \cdot 67 \cdot 662348503367$.

Remark A brief overview of the previous example shows that $f_n, n \in \mathbb{N}$, is a perfect square if and only if $n = 1, 3$. (A perfect square cannot be (exactly) divisible by a prime, that is, divisible by that prime but not divisible by its square.) This corollary is, however, much simpler and follows directly by observing that the last (ones) digit of $f_n, 4 \leq n \in \mathbb{N}$, is 3 (since, for $5 \leq n \in \mathbb{N}$, the factorial $n!$ ends with 0), whereas a square could only have possible last digits as $1, 4, 5, 6, 9, 0$ with 3 missing.

Exercises

1.3.1. Find the smallest prime number that is the sum of two different prime numbers and, at the same time, it is also the sum of three different prime numbers.

1.3.2. Let p and q be primes with $p > q \geq 5$. Show that 24 divides $p^2 - q^2$.

1.3.3. Find the largest prime factor of the number $2^{18} - 64$.

1.3.4. How many natural numbers ≤ 400 are relatively prime to 400?

1.3.5. Let $n \in \mathbb{N}$ and D_n be the set of positive divisors of n. Determine D_8, D_{12}, D_{30} and $D_8 \cap D_{12}$ and $D_{12} \cap D_{30}$.

1.3.6. Show the following properties of the greatest common divisor (with all arguments in $\mathbb{N}$):

$$\begin{aligned}
\gcd(na, nb) &= \gcd(a, b) \\
\gcd(a + nb, b) &= \gcd(a, b) \\
\gcd(a, \gcd(b, c)) &= \gcd(\gcd(a, b), c) \\
\gcd(a_1, a_1) = 1 &\Rightarrow \gcd(a_1 a_2, b) = \gcd(a_1, b)\gcd(a_2, b) \\
\gcd(a, bc) = 1 &\Leftrightarrow \gcd(a, b) = 1 \text{ and } \gcd(a, c) = 1.
\end{aligned}$$

1.4 Rational Numbers

The number 1 is natural and it is the **multiplicative identity** for both $\mathbb{N}$ and $\mathbb{Z}$; that is, when multiplied by another number, it leaves that number unchanged.

No other natural number ($\neq 1$) or integer ($\neq \pm 1$) has a **multiplicative inverse**, a number that, when multiplied by the number, produces 1. (See Corollary to Proposition 1.1.8.)

This deficiency is remedied by introducing the set $\mathbb{Q}$ of **rational numbers** as in Section 0.1:

$$\mathbb{Q} = \left\{ \frac{a}{b} \,\middle|\, a, b \in \mathbb{Z} \text{ and } b \neq 0 \right\}.$$

The construction of $\mathbb{Q}$ is, in principle, similar to the construction of the set of integers $\mathbb{Z}$ with focusing on the multiplicative structure instead of the additive structure. The main idea is that a fraction a/b with $a, b \in \mathbb{Z}$ and $b \neq 0$ is determined by the pair $(a, b) \in \mathbb{Z} \times \mathbb{Z}$ consisting of the numerator and the denominator. Obviously, for any $c \in \mathbb{Z}$ and $c \neq 0$, the pair (ac, bc) represents the same fraction as (a, b).

This gives the construction of the set of rational numbers $\mathbb{Q}$ from $\mathbb{Z}$ as follows. We represent each rational number by a pair (a, b) of integers with $b \neq 0$, an element of the Cartesian product $\mathbb{Z} \times \mathbb{Z}^\sharp$, where $\mathbb{Z}^\sharp = \mathbb{Z} \setminus \{0\}$ is the set of **nonzero** integers. The pair (a, b) represents the same rational number as (a', b') if and only if $a \cdot b' = a' \cdot b$. We therefore introduce the relation $\sim$ on $\mathbb{Z} \times \mathbb{Z}^\sharp$ by setting $(a, b) \sim (a', b')$, $(a, b), (a', b') \in \mathbb{Z} \times \mathbb{Z}^\sharp$, if $a \cdot b' = a' \cdot b$.

We first claim that $\sim$ is an equivalence relation on $\mathbb{Z} \times \mathbb{Z}^\sharp$. Reflexivity, $(a, b) \sim (a, b)$, and symmetry, $(a, b) \sim (a', b')$ implies $(a', b') \sim (a, b)$, are tautologies. Transitivity, $(a, b) \sim (a', b')$ and $(a', b') \sim (a'', b'')$ imply $(a, b) \sim (a'', b'')$, also follows since $a \cdot b' = a' \cdot b$ and $a' \cdot b'' = a'' \cdot b'$ imply $aa' \cdot b'b'' = a'a'' \cdot b'b$, and, by the cancellation law for multiplication, we obtain $a \cdot b'' = a'' \cdot b$. (Note that $a' = 0$ implies $a = a'' = 0$.)

The equivalence relation $\sim$ partitions $\mathbb{Z} \times \mathbb{Z}^\sharp$ into equivalence classes. We define the **set of rational numbers** as the quotient $\mathbb{Q} = \mathbb{Z} \times \mathbb{Z}^\sharp / \sim$, the set of equivalence classes in $\mathbb{Z} \times \mathbb{Z}^\sharp$ by the equivalence relation $\sim$.

We now define the operations of addition $+$ and multiplication $\cdot$ on $\mathbb{Q}$ in terms of representatives as

$$(a, b)+(c, d) = (ad+bc, bd) \text{ and } (a, b)\cdot(c, d) = (ac, bd), \quad a, c \in \mathbb{Z},\ b, d \in \mathbb{Z}^\sharp.$$

We first need to show that these operations are compatible with the equivalence relation $\sim$, that is, if $(a, b) \sim (a', b')$ and $(c, d) \sim (c', d')$, $(a, b), (a', b'), (c, d), (c', d') \in \mathbb{Z} \times \mathbb{Z}^\sharp$, then $(a, b) + (c, d) \sim (a', b') + (c', d')$ and $(a, b) \cdot (c, d) \sim (a', b') \cdot (c', d')$. The assumptions translate into the pair of equations $ab' = a'b$ and $cd' = c'd$. Multiplying the first equation by dd', the second by bb', and adding, we obtain $(ad + bc)b'd' = (a'd' + b'c')bd$. This gives $(a, b) + (c, d) \sim (a', b') + (c', d')$ as stated. Returning to this pair of equations, multiplying, we obtain $acb'd' = a'c'bd$. This gives $(a, b) \cdot (c, d) \sim (a', b') \cdot (c', d')$.

Compatibility just proved means that the operations of addition $+$ and multiplication $\cdot$ given above define addition and multiplication on the quotient $\mathbb{Q} = \mathbb{Z} \times \mathbb{Z}^\sharp / \sim$.

The set of integers $\mathbb{Z}$ can naturally be embedded into $\mathbb{Q}$ via the injective map $\iota : \mathbb{Z} \to \mathbb{Q}$ defined by associating with the integer $a \in \mathbb{Z}$ the equivalence class of $(a, 1)$. For $a, b \in \mathbb{Z}$, we have $(a, 1)+(b, 1) = (a+b, 1)$ and $(a, 1)\cdot(b, 1) = (ab, 1)$. This shows that the embedding ι is compatible with the additions and multiplications in $\mathbb{Z}$ and $\mathbb{Q}$. From now on, we identify $\mathbb{Z}$ with its range in $\mathbb{Q}$ under ι and say that the set of rational numbers $\mathbb{Q}$ is an **extension** of $\mathbb{Z}$.

Addition and multiplication of rational numbers are both **associative** and **commutative**, and the two operations are connected through **distributivity**. The equivalence class of $(0, 1)$ (corresponding to $0 \in \mathbb{Z}$) is the **additive identity**. Any element has an **additive inverse**; the additive inverse of the equivalence class of $(a, b) \in \mathbb{Z} \times \mathbb{Z}^\sharp$ is the equivalence class of $(-a, b) \sim (a, -b) \in \mathbb{Z} \times \mathbb{Z}^\sharp$. The equivalence class of $(1, 1)$ (corresponding to $1 \in \mathbb{Z}$) is a (unique) **multiplicative identity**. Every non-zero equivalence class has a **multiplicative inverse**; the multiplicative inverse of $(a, b) \in \mathbb{Z}^\sharp \times \mathbb{Z}^\sharp$ is the equivalence class of $(b, a) \in \mathbb{Z}^\sharp \times \mathbb{Z}^\sharp$.

These statements follow directly from the definitions of the addition and multiplication. We give the details for distributivity which is the least trivial. Letting $(a, b), (c, d), (e, f) \in \mathbb{Z} \times \mathbb{Z}^\sharp$, we calculate

$$\begin{aligned}(a, b) \cdot ((c, d) + (e, f)) &= (a, b) \cdot (cf + de, df) = (a(cf + de), bdf) \\ &\sim ((ac)(bf) + (bd)(ae),(bd)(bf)) = (ac, bd)+(ae, bf) \\ &= (a, b) \cdot (c, d) + (a, b) \cdot (e, f).\end{aligned}$$

Taking equivalence classes, distributivity follows.

The properties of addition and multiplication listed above are expressed compactly by saying that the set of rational numbers forms a **field**.

The natural **ordering** $<$ on the set of rational numbers $\mathbb{Q}$ is given as follows: Let $q, r \in \mathbb{Q}$ and choose representatives $(a, b) \in q$, $(c, d) \in r$, $(a, b), (c, d) \in \mathbb{Z} \times \mathbb{Z}^\sharp$, such that $b, d > 0$. (This can always be done since $(a, b) \sim (-a, -b)$ and $(c, d) \sim (-c, -d)$.) We then define $(a, b) < (c, d)$ (or $(c, d) > (a, b)$) if $ad < bc$.

If $(a', b') \sim (a, b)$ and $(c', d') \sim (c, d)$, $(a', b'), (c', d') \in \mathbb{Z} \times \mathbb{Z}^\sharp$, with $b', d' > 0$, then multiplying $ad < bc$ by $b'd' > 0$, we obtain $(ab')(dd') < (cd')(bb')$. Using $ab' = a'b$ and $cd' = c'd$, this gives $(a'b)(dd') < (c'd)(bb')$, or equivalently, $(a'd')(bd) < (b'c')(bd)$. Canceling $bd > 0$, we obtain $a'd' < b'c'$. Thus, the ordering $<$ is well-defined on the equivalence classes, and thereby it defines an ordering on $\mathbb{Q}$.

Note that this ordering is clearly an extension of the earlier ordering $<$ on $\mathbb{Z}$ since $(a, 1) < (b, 1)$, $a, b \in \mathbb{Z}$, if and only if $a < b$.

The relation $<$ is a **strict total order** on the set of rational numbers $\mathbb{Q}$. To show transitivity, let $(a, b) < (c, d)$ and $(c, d) < (e, f)$ with $b, d, f > 0$. We have $ad < bc$ and $cf < de$. Hence, $adf < bcf < bde$, so that $af < be$, and $(a, b) < (e, f)$ follows.

For trichotomy, let $q, r \in \mathbb{Q}$, and show that exactly one of the following holds: $q < r$, $q = r$, and $q > r$. Indeed, as before, letting $(a, b) \in q$, $(c, d) \in r$,

$(a, b), (c, d) \in \mathbb{Z} \times \mathbb{Z}^\sharp$, such that $b, d > 0$; exactly one of the following holds: $ad < bc$, $ad = bc$, and $ad > bc$. These give the respective cases.

We call a rational number $q \in \mathbb{Q}$ **positive** if $q > 0$ and **negative** if $q < 0$. Clearly, $q \in \mathbb{Q}$ is positive if and only if $-q \in \mathbb{Q}$ is negative. Moreover, in $\mathbb{Q}$, the usual arithmetic properties hold: For any $q, r, s \in \mathbb{Q}$, (1) $q < r$ implies $-q > -r$; (2) $q < r$ implies $q + s < r + s$; (3) $q < r$ implies $qs < rs$ if $s > 0$; (4) $q < r$ implies $qs > rs$ if $s < 0$; etc.

In addition to the strict total order and cancellation law for addition, $\mathbb{Q}$ also has the following property: $q > 0$ and $r > 0$, $q, r \in \mathbb{Q}$, imply $q \cdot r > 0$. We express this by saying that the set of rational numbers $\mathbb{Q}$ is an **ordered field** with respect to the order relation $<$. As direct consequences, we obtain the following: (1) The cancellation law for multiplication for inequalities: $q \cdot s < r \cdot s, q, r, s \in \mathbb{Q}$, implies $q < r$ if $s > 0$ and $q > r$ if $s < 0$; (2) If $q \neq 0, q \in \mathbb{Q}$, then $q^2 > 0$; in particular, $1 > 0$; and (3) $0 < q < r$ implies $0 < 1/q < 1/r$.

We also define $q \leq r$ (or $r \geq q$), $q, r \in \mathbb{Q}$, if $q = r$ or $q < r$. Equivalently, $q \leq r$ if and only if $r - q$ is either positive or zero.

The set of rational numbers $\mathbb{Q}$ with $\leq$ is a **totally ordered field**; that is, $\leq$ is transitive and antisymmetric and satisfies the property of totality. These are easy consequences of the properties of the strict ordering $<$ above.

From now on, we adopt the customary notation for rational numbers as fractions; that is, we denote the equivalence class of $(a, b) \in \mathbb{Z} \times \mathbb{Z}^\sharp$ by the fraction a/b with the understanding that the fraction $(ac)/(bc)$ is the same as a/b. As before, the fraction $a/1$ then becomes a. Note that the multiplicative inverse is usually called the **reciprocal**; that is, we have $1/(a/b) = b/a, a, b \in \mathbb{Z}^\sharp$.

History

One of the earliest attestations of fractions is in the pair of **Akhmin wooden tablets** dated in the early Middle Kingdom of ancient Egypt (c. 1950 BCE). It contains multiplication problems involving reciprocals of integers such as $1/3, 1/7, 1/10, 1/11, 1/13$. The fractions are written in ancient Egyptian fashion using parts of the Eye of Horus. The best known ancient Egyptian record of mathematics is the **Rhind Mathematical Papyrus** (dated c. 1550 – 1650 BCE) itself a copy of an earlier Berlin Papyrus and other texts. It contains an extensive list of computations with fractions including fractions of type $2/n$, $n = 3, 4, \ldots, 101$, and equations how to decompose them into sums of reciprocals of natural numbers, such as $2/15 = 1/10 + 1/30, \ldots, 2/101 = 1/101 + 1/202 + 1/303 + 1/606$. In addition, it contains a list of how to multiply different fractions by the expression $1 + 1/2 + 1/4 = 7/4$.

The beginning of basic arithmetic involving integers and fractions can be found in the works of the Indian mathematicians Aryabhatta (476 – 550 CE), Brahmagupta (c. 628 CE) and Bhāskara II (1114 – c. 1185). For example, the Bakhshali document contains the so-called Rule of Three (still used sporadically today in secondary education), the solution of the equation $c/x = a/b$ as $x = bc/a$.

Note finally that the horizontal fraction bar first appears in the works of the Muslim mathematician Al-Hassār (12th century CE) from Fez, Morocco.

Example 1.4.1 A point $(a, b) \in \mathbb{R}^2$ in the plane is called an **integer point** if $(a, b) \in \mathbb{Z} \times \mathbb{Z}$. In this example, we consider integer points in the plane whose coordinates are relatively prime:[11] $(a, b) \in \mathbb{Z} \times \mathbb{Z}$, $(a, b) \neq (0, 0)$, $\gcd(a, b) = 1$.

[11] These are precisely the **visible points** (from the origin); that is, the line segment with end-points $(0, 0)$ and (a, b) contains no other integer points. We will not need this geometric interpretation.

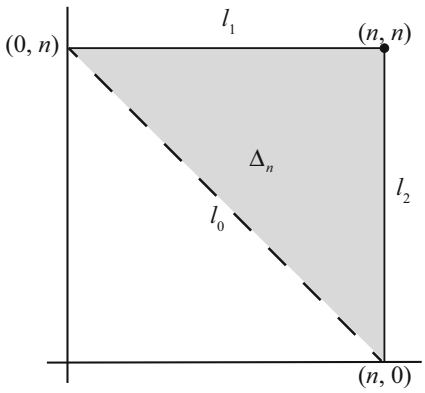


Fig. 1.4 The triangle Δ_n.

Let $n \in \mathbb{N}$. First, for $(a, b) \in \mathbb{N}_0 \times \mathbb{N}_0$, with $a + b = n$, we have $\gcd(a, b) = \gcd(a, n) = \gcd(n, b)$; in particular, we have

$$\gcd(a, b) = 1 \quad \Leftrightarrow \quad \gcd(a, n) = 1 \text{ and } \gcd(n, b) = 1.$$

Moreover, $a + b = n$ also gives

$$\frac{1}{a \cdot b} = \frac{1}{a \cdot n} + \frac{1}{b \cdot n}, \quad (a, b) \in \mathbb{N} \times \mathbb{N}.$$

Second, let l_0 be the line segment with end-points $(n, 0)$ and $(0, n)$; and l_1, resp. l_2, the line segments with end-points (n, n) and $(0, n)$ resp. $(n, 0)$. Summing up the fractions in the identity above in the respective line segments, we obtain

$$\sum_{\substack{(a,b)\in l_0 \\ \gcd(a,b)=1}} \frac{1}{a \cdot b} = \sum_{\substack{(a,n)\in l_1 \\ \gcd(a,n)=1}} \frac{1}{a \cdot n} + \sum_{\substack{(n,b)\in l_2 \\ \gcd(b,n)=1}} \frac{1}{b \cdot n}.$$

(We use here the one-to-one correspondences $(a, b) \leftrightarrow (a, n) \leftrightarrow (n, b)$, $a + b = n$, $a, b \in \mathbb{N}_0$.)

Finally, let $\Delta_n \subset \mathbb{R}^2$, $n \in \mathbb{N}$, denote the (solid) triangle with vertices $(n, 0)$, $(0, n)$, and (n, n) (see Figure 1.4). Clearly, the line segments l_0, l_1, and l_2 are the sides of Δ_n. We now agree that, for Δ_n, the sides l_1 and l_2 are counted in, but the side l_0 is counted out. In other words, we define

$$\Delta_n = \{(x, y) \in \mathbb{R}^2 \mid 0 < x, y \leq n < x + y\}.$$

Note also that in this example, we will use some simple geometric concepts such as lines, line segments, etc. For a detailed account on these, see Section 5.5.

We now claim

$$\sum_{\substack{(a,b)\in\Delta_n \\ \gcd(a,b)=1}} \frac{1}{a\cdot b} = 1,$$

independent of $n \in \mathbb{N}$.

To prove this, let S_n, $n \in \mathbb{N}$, denote the sum on the left-hand side. We proceed by induction with respect to $n \in \mathbb{N}$.

Clearly, $S_1 = 1$. (The only point competing in the sum is $(1, 1)$. Note also that, in S_2, the competing points are $(1, 2)$ and $(2, 1)$, giving $1/2 + 1/2 = 1$; and, in S_3, the competing points are $(1, 3)$, $(2, 3)$, $(3, 2)$, $(3, 1)$, giving $1/3+1/6+1/6+1/3 = 1$.)

For the general induction step $n - 1 \Rightarrow n$, $2 \leq n \in \mathbb{N}$, comparing the triangles Δ_{n-1} and Δ_n, we have

$$S_n - S_{n-1} = \sum_{\substack{(a,n)\in l_1 \\ \gcd(a,n)=1}} \frac{1}{a\cdot n} + \sum_{\substack{(n,b)\in l_2 \\ \gcd(b,n)=1}} \frac{1}{b\cdot n} - \sum_{\substack{(a,b)\in l_0 \\ \gcd(a,b)=1}} \frac{1}{a\cdot b} = 0,$$

where we used the result in the second step above. Hence $S_n = 1$ for all $n \in \mathbb{N}$. The claim follows.

The **absolute value** can naturally be extended from integers to rational numbers: $|a/b| = |a|/|b|$, $(a, b) \in \mathbb{Z} \times \mathbb{Z}^\sharp$. The analogues of Proposition 1.2.2 and the subsequent corollary at the end of Section 1.2 hold with almost verbatim proofs.

Proposition 1.4.1 *Let $0 \leq q \in \mathbb{Q}$. For $r \in \mathbb{Q}$, we have $-q \leq r \leq q$ if and only if $|r| \leq q$. The same holds for strict inequalities.*

Corollary *We have*

$$||q| - |r|| \leq |q + r| \leq |q| + |r|, \quad q, r \in \mathbb{Q}.$$

Finally, we show that the **Archimedean Property** holds for rational numbers.

Proposition 1.4.2 *Let $0 < q, r \in \mathbb{Q}$. Then there exists $n \in \mathbb{N}$ such that $r \leq nq$.*

Proof Taking common denominators, we can write $q = a/c$ and $r = b/c$, $a, b, c \in \mathbb{N}$. The Archimedean Property for natural numbers asserts the existence of $n \in \mathbb{N}$ such that $b \leq na$. Dividing by c, the proposition follows.

Corollary *We have*

$$\inf\left\{\frac{1}{n}\,\middle|\, n \in \mathbb{N}\right\} = 0.$$

Proof Zero is obviously a lower bound for all $1/n$, $n \in \mathbb{N}$. We claim that it is the greatest lower bound. Let $0 < \epsilon \in \mathbb{Q}$. By the Archimedean Property above, there exists $n \in \mathbb{N}$ such that $1/\epsilon \leq n$, or equivalently, we have $1/n \leq \epsilon$. Thus, no positive $\epsilon \in \mathbb{Q}$ can be a lower bound for all $1/n$, $n \in \mathbb{N}$. The corollary follows.

Exercises

1.4.1. Use Peano's Principle of Induction to derive the formula

$$\left(1 - \frac{1}{4}\right)\left(1 - \frac{1}{9}\right) \cdots \left(1 - \frac{1}{n^2}\right) = \frac{n+1}{2n}, \quad 2 \leq n \in \mathbb{N}.$$

1.4.2. In the *Jade Mirror of the Four Unknowns* written by Zhu Shijie (1249–1314), the following equality is given without proof:

$$1 + 8 + 30 + 80 + \cdots + \frac{n^2(n+1)(n+2)}{3!} = \frac{n(n+1)(n+2)(n+3)(4n+1)}{5!}, \quad n \in \mathbb{N}.$$

Prove this equality using Peano's Principle of Induction.

1.4.3. Show that, for all $n \in \mathbb{N}$, we have[12]

$$\frac{1}{n+1} + \frac{1}{n+2} + \cdots + \frac{1}{2n} \geq \frac{1}{2}.$$

1.4.4. Let $0 < a, b \in \mathbb{Q}$. Show that $\sqrt{a} + \sqrt{b} \in \mathbb{Q}$ implies $\sqrt{a}, \sqrt{b} \in \mathbb{Q}$.

1.4.5. In 1637, Fermat jotted down the following in a margin of his copy of Diophantus' Arithmetica: "It is impossible to write a cube [of a natural number] as a sum of two cubes [of natural numbers], a fourth power as a sum of fourth powers, and, in general, any power beyond the second as a sum of two similar powers."[13] Show that, for any rational number $0 < q < 1$, the number $\sqrt[3]{1 - q^3}$ is not rational. Generalize this to an arbitrary exponent $n \geq 3$.

[12] In Example 10.5.2, we will show $\lim_{n\to\infty}\left(\frac{1}{n+1} + \frac{1}{n+2} + \cdots + \frac{1}{2n}\right) = \ln 2$.

[13] An early proof of this for cubes was given by Euler using complex arithmetic. For any exponent, this is the famous Fermat's Last Theorem proved by Andrew Wiles in 1995.

Chapter 2
Real Numbers

"A person who can solve[1] $x^2 - 92 \cdot y^2 = 1$ *within a year is a mathematician."*
in Brahma-Sphuta-Siddhanta
by Brahmagupta (c. 628 CE)

With the rational number system $\mathbb{Q}$ in place, leaning back to the past, we begin this chapter by showing how the dialogue between Theaetetus and Socrates leads naturally to Dedekind's original proof of irrationality of the square root of a non-square natural number. As an immediate byproduct, this implies that the Least Upper Bound Property fails in $\mathbb{Q}$. Another advantage of this proof is that it leads directly to the concept of Dedekind cuts, and thereby to Dedekind's construction of the real number system. Using Dedekind cuts offers a quick and easy proof of the Least Upper Bound Property in this model of the real number system.

Dedekind's proof naturally raises the problem of rational approximations of the square root of a non-square natural number. In view of later applications, we make a short detour here to the related Pell's equation and its solution by Brahmagupta's identity. We close our study of the Dedekind model of the real number system by introducing exponentiation with integer exponents, and deriving the corresponding Bernoulli inequality. This opens the first opportunity to present a whole cadre of contest problems some of which are on Olympiad level.

Working with the Dedekind model of the real number system is cumbersome, and not well suited to do analysis, however. We therefore build another model of the real number system via Cauchy sequences. Once again, we choose a slow-paced approach, and first treat the real numbers naïvely as infinite decimals. Meshing well with this, we introduce and treat limits of (numerical) sequences by the least strenuous path, through suprema and infima.[2] Cauchy sequences also

[1] This is a specific Pell's equation, and x and y are meant to be natural numbers. The **smallest** solution turns out to be $(x, y) = (1151, 120)$. See Section 2.1 for a quick solution.

[2] Plurals of supremum and infimum; note that the plurals "supremums" and "infimums" are also widely used.

G. Toth, *Elements of Mathematics*, Undergraduate Texts in Mathematics,
https://doi.org/10.1007/978-3-030-75051-0_2

open the way to introduce a few fundamental methods and ideas of analysis: the Monotone Convergence Theorem and the Bolzano-Weierstrass Theorem. The material here is developed enough to present many challenging problems inspired by past mathematical Olympiads. As before, whenever an opportunity arises, we also ease up the complexity of the presentation by showing, for example, irrationality of $\sqrt{2}$ by origami.

At the end of this chapter, we take an optional short detour to discuss the pigeonhole principle, the Dirichlet approximation, and an elementary proof of the Equidistribution Theorem. This technically more demanding section can be skipped at the first reading.

2.1 Real Numbers via Dedekind Cuts

In the previous chapter we constructed the rational number system $\mathbb{Q}$, and showed that it is a totally ordered field.

The next step is to investigate whether the Least Upper Bound Property holds in $\mathbb{Q}$. In other words, if a non-empty set $A \subset \mathbb{Q}$ is bounded above does sup A exist **in** $\mathbb{Q}$? As we have seen in Proposition 1.2.1, this holds for $\mathbb{Z}$. We will show below that the Least Upper Bound Property **fails** in $\mathbb{Q}$. This is a major deficiency of the field of rational numbers.

To begin with we derive the following elementary fact: Given $n \in \mathbb{N}$, for a positive rational number $q \in \mathbb{Q}$, we have $q^2 = n$ if and only if n is a **perfect square**; that is $n = a^2$ for some $a \in \mathbb{N}$ (and $q = a$).

We first give a proof of this following Dedekind. The starting point of his proof is based on a dialogue between Theaetetus and Socrates in Plato's Theaetetus (650 BCE).

History

Excerpt from Plato's Theaetetus:[3]

"Theaetetus: Theodorus was writing out for us something about roots, such as the roots of three or five, showing that they are incommensurable by the unit: he selected other examples up to seventeen - there he stopped. Now as there are innumerable roots, the notion occurred to us of attempting to include them all under one name or class.

Socrates: And did you find such a class?

Theaetetus: I think that we did; but I should like to have your opinion.

Socrates: Let me hear.

Theaetetus: We divided all numbers into two classes: those which are made up of equal factors multiplying into one another, which we compared to square figures and called square or equilateral numbers; - that was one class.

Socrates: Very good.

Theaetetus: The intermediate numbers, such as three and five, and every other number which is made up of unequal factors, either of a greater multiplied by a less, or of a less multiplied by a greater, and when regarded as a figure, is contained in unequal sides; - all these we compared to oblong figures, and called them oblong numbers.

[3]Translated by Benjamin Jowett.

Socrates: Capital; and what followed?
Theaetetus: The lines, or sides, which have for their squares the equilateral plane numbers, were called by us lengths or magnitudes; and the lines which are the roots of (or whose squares are equal to) the oblong numbers, were called powers or roots; the reason of this latter name being, that they are commensurable with the former i.e., with the so-called lengths or magnitudes not in linear measurement, but in the value of the superficial content of their squares; and the same about solids."

Dedekind's proof is to show that if $n \in \mathbb{N}$ is not a perfect square then there is no positive rational number $q \in \mathbb{Q}$ such that $q^2 = n$. It starts with assuming that $n \in \mathbb{N}$ is an "oblong number" (as in the excerpt above), that is, n is **not** a perfect square. The concept of an oblong number being "intermediate" is interpreted as follows: There exists $m \in \mathbb{N}$ such that

$$m^2 < n < (m+1)^2.$$

We need to show that $n = q^2$ cannot hold for **any** rational number $q \in \mathbb{Q}$. Let $q = a/b$, $a, b \in \mathbb{N}$, $b \neq 0$, with $a \in \mathbb{N}$ **minimal** as the positive numerator in the fractional representation of q. This gives

$$a^2 - nb^2 = 0.$$

We need to show that these two conditions lead to contradiction. We have

$$m^2a^2 < na^2 = n^2b^2 < (m+1)^2a^2$$

which gives

$$ma < nb < (m+1)a.$$

Similarly, we have

$$m^2b^2 < nb^2 = a^2 < (m+1)^2b^2$$

which gives

$$mb < a < (m+1)b.$$

We now define $a' = nb - ma$ and $b' = a - mb$. By the inequalities above, we have $0 < a' < a$ and $0 < b' < b$. We now calculate

$$\begin{aligned} a'^2 - nb'^2 &= (nb - ma)^2 - n(a - mb)^2 \\ &= n^2b^2 + m^2a^2 - na^2 - nm^2b^2 \\ &= (m^2 - n)(a^2 - nb^2) = 0. \end{aligned}$$

This gives $n = q^2 = (a'/b')^2$, and hence $q = a'/b'$. This contradicts the minimality of a since $0 < a' < a$. The statement follows.

Remark A simpler proof uses divisibility properties of the integers. As before, we let $q = a/b$, $a, b \in \mathbb{N}$, $b \neq 0$, and assume that the fraction a/b is irreducible; that is a and b are relatively prime.

The equation $(a/b)^2 = n$ gives $nb = a^2/b$. Since the left-hand side is an integer, we see that b divides a^2. We claim that $b = 1$. If not then, by the Fundamental Theorem of Arithmetic (Section 1.3), b must have a prime divisor p. Now, p divides a^2 so that it must also divide a itself (Corollary to Proposition 1.3.1). Hence p is a common divisor of a and b which is a contradiction since we assumed that a/b was irreducible. Thus $b = 1$, and the equality above reduces to $n = a^2$. Our statement follows again.

We now continue to follow Dedekind, and use this statement just proved to show that the Least Upper Bound Property fails in $\mathbb{Q}$.

Assume that $n \in \mathbb{N}$ is **not** a perfect square, and define

$$R_n = \{q \in \mathbb{Q} \,|\, q < 0 \text{ or } q^2 < n\} \quad \text{and} \quad S_n = \{q \in \mathbb{Q} \,|\, q \geq 0 \text{ and } q^2 > n\}.$$

Since n is not a perfect square, by the above, $q^2 \neq n$, that is, $q^2 < n$ or $q^2 > n$, for all rational numbers $q \in \mathbb{Q}$. This shows that, $R_n \cup S_n = \mathbb{Q}$ and $R_n \cap S_n = \emptyset$. It is easy to see that any element in R_n is a lower bound for S_n, and any element in S_n is an upper bound for R_n.

We now claim that neither $\sup R_n$ nor $\inf S_n$ exist **within** $\mathbb{Q}$.

Let $q \in \mathbb{Q}$, and define

$$q' = \frac{q(q^2 + 3n)}{3q^2 + n} \in \mathbb{Q}.$$

We calculate

$$q' - q = \frac{q(q^2 + 3n)}{3q^2 + n} - q = q\left(\frac{q^2 + 3n}{3q^2 + n} - 1\right) = \frac{2q(n - q^2)}{3q^2 + n}.$$

Moreover, we have[4]

$$\begin{aligned} q'^2 - n &= \frac{q^2(q^2 + 3n)^2}{(3q^2 + n)^2} - n = \frac{q^2(q^2 + 3n)^2 - n(3q^2 + n)^2}{(3q^2 + n)^2} \\ &= \frac{q^6 - 3nq^4 + 3n^2q^2 - n^3}{(3q^2 + n)^2} = \frac{(q^2 - n)^3}{(3q^2 + n)^2}. \end{aligned}$$

[4] As before, here and in the sequel we will use basic algebraic identities without explicit references. These will be treated in detail in Chapter 6.

Now, let $0 < q \in \mathbb{Q}$, and construct $q' \in \mathbb{Q}$ as above. We claim that q cannot be the supremum of R_n nor the infimum of S_n.

Since $R_n \cup S_n = \mathbb{Q}$, we have $q \in R_n$ or $q \in S_n$.

If $q \in R_n$ then $q^2 < n$. By the two computations above, we have $q' - q > 0$ and $q'^2 - n < 0$. These give $q' > q$ and $q' \in R_n$. Thus, q cannot be the supremum of R_n, and obviously it cannot be the infimum of S_n (as it does not belong to S_n).

If $q \in S_n$ then $q^2 > n$. By the two computations above, we have $q' - q < 0$ and $q'^2 - n > 0$. These give $q' < q$ and $q' \in S_n$. Thus, q cannot be the infimum of S_n, and it cannot be the supremum of R_n.

We conclude that no **rational number** can be sup R_n or inf S_n; these do not exist in $\mathbb{Q}$. Hence, $\mathbb{Q}$ does not have the Least Upper Bound Property.

The subset $R_n \subset \mathbb{Q}$ (or the pair (R_n, S_n)), $n \in \mathbb{N}$, is called a Dedekind cut. In Dedekind's approach the square root $\sqrt{n}$, for $n \in \mathbb{N}$ not a perfect square, as an "irrational number" **is** (given by) the Dedekind cut R_n. This is the starting point of Dedekind's constructive approach to the real number system $\mathbb{R}$.

We are now ready to introduce the general concept of a Dedekind cut:

A **proper** subset $R \subset \mathbb{Q}$, $\emptyset \neq R \neq \mathbb{Q}$, is called a **Dedekind cut** if it satisfies the following properties:

(D1) For every $q \in R$ and $q' \in R^c = \mathbb{Q} \setminus R$, we have $q < q'$;
(D2) For every $q \in R$, there exists $q' \in R$ such that $q < q'$.

We will also use use the equivalent forms of (D1) and (D2) as follows:

(D3) If $q \in R$ and $q' < q$, $q' \in \mathbb{Q}$, then $q' \in R$;
(D4) If $q \in \mathbb{Q}$ is such that $q' < q$ for all $q' \in R$ then $q \in R^c$.

Remark Here and below the complement is always taken with respect to the universal set $\mathbb{Q}$, the set of rational numbers. Oftentimes, in particular in Dedekind's original work, a Dedekind cut is defined as a pair (R, R^c) of complementary subsets of $\mathbb{Q}$. Some authors define the Dedekind cuts using (D3) and (D4).

A Dedekind cut is called a **real number**. The **set of real numbers** is denoted by $\mathbb{R}$. Henceforth we will use the terms "R is a Dedekind cut" and "$R \in \mathbb{R}$" alternatively.

History
The term "real number" as an antonym to "imaginary number" is due to Descartes who introduced them to describe real roots of polynomials as opposed to imaginary ones.

Given a rational number $q \in \mathbb{Q}$, we let

$$Q_q = \{q' \in \mathbb{Q} \,|\, q' < q\}.$$

The proper subset $Q_q \subset \mathbb{Q}$, $q \in \mathbb{Q}$, satisfies (D1), and also (D2) (for $q' < q$ we have $q' < (q + q')/2 < q$, $q' \in \mathbb{Q}$). Hence Q_q, $q \in \mathbb{Q}$, is a Dedekind cut. Clearly, sup $Q_q = q$. We call Q_q, $q \in \mathbb{Q}$, the **rational Dedekind cut** defined by q.

Conversely, if $R \in \mathbb{R}$ is a Dedekind cut such that $\sup R = q \in \mathbb{Q}$ exists then $R = Q_q$. (Indeed, by (D2), we have $q \notin R$, and, by (D3), we have $Q_q \subset R$. Since $\sup R = q$, by (D1), we obtain $Q_q = R$.)

Associating to $q \in \mathbb{Q}$ the rational Dedekind cut $Q_q \in \mathbb{R}$ gives rise to an embedding of the set of rational numbers $\mathbb{Q}$ into $\mathbb{R}$.

Due to its frequent occurrence, the Dedekind cut Q_0, the set of all negative rational numbers, will be denoted by O.

If a Dedekind cut $R \subset \mathbb{Q}$ does not have a supremum in $\mathbb{Q}$ then $R \neq Q_q$ for all $q \in \mathbb{Q}$. In this case $R \in \mathbb{R}$ is called an **irrational number**. The example at the end of the previous section shows that, for $n \in \mathbb{N}$, the Dedekind cut R_n is rational if and only if n is a perfect square. Since there are infinitely many natural numbers that are not perfect squares (such as primes) we obtain infinitely many irrational numbers.

A natural **ordering** $<$ on the set of Dedekind cuts $\mathbb{R}$ is given setting $R < S$, $R, S \in \mathbb{R}$, if $R \subset S$ and $R \neq S$. Note that this ordering is an extension of the strict total order $<$ on the set of rational numbers $\mathbb{Q}$ since $Q_{q'} < Q_q$ if and only if $q' < q$, $q, q' \in \mathbb{Q}$.

We now claim that $<$ is a **strict total order** on $\mathbb{R}$.

Transitivity is obvious. For trichotomy, let $R, S \in \mathbb{R}$ such that $R \neq S$. Then, one of the differences, $R \setminus S$ or $S \setminus R$, is non-empty. Without loss of generality, we may assume $R \setminus S \neq \emptyset$ (since otherwise we interchange R and S). Let $q \in R \setminus S$. Since $q \in R$, by (D3), we have $Q_q \subset R$. Since $q \in S^c$, by (D1), q is an upper bound for S, and hence $S \subset Q_q$. These give $S \subset R$ and $S \neq R$. We obtain $S < R$. Trichotomy follows.

We can also define $R \leq S$ (or $S \geq R$), $R, S \in \mathbb{R}$, if $R \subset S$. The set of real numbers $\mathbb{R}$ with $\leq$ is a **totally ordered set**, that is, $\leq$ is transitive, antisymmetric and total (see Section 0.2).

We now show that, unlike its rational predecessor $\mathbb{Q}$, the set of real numbers $\mathbb{R}$ has the **Least Upper Bound Property**; that is, a subset bounded above has a supremum in $\mathbb{R}$.

Theorem 2.1.1 *If a non-empty set $\mathcal{R} \subset \mathbb{R}$ is bounded above then* $\sup \mathcal{R}$ *exists in* $\mathbb{R}$.

Proof Consider the set

$$\bigcup \mathcal{R} = \bigcup_{R \in \mathcal{R}} R \subset \mathbb{Q},$$

the union of all Dedekind cuts in $\mathcal{R}$. If a Dedekind cut $S \in \mathbb{R}$ is an upper bound for $\mathcal{R} \subset \mathbb{R}$ then $R \subset S$ for all $R \in \mathcal{R}$, so that we have $\bigcup \mathcal{R} \subset S$. Since this holds for all upper bounds $S \in \mathbb{R}$ of $\mathcal{R}$, the union $\bigcup \mathcal{R}$ will be the least upper bound once we show that it is a Dedekind cut.

Clearly, $\bigcup \mathcal{R}$ is non-empty, and also proper since the complement S^c of any upper bound S of $\mathcal{R}$ is disjoint from $\bigcup \mathcal{R}$.

For (D1), let $q \in \bigcup \mathcal{R}$ and $q' \in (\bigcup \mathcal{R})^c$. The first relation means that $q \in R$ for a specific $R \in \mathcal{R}$. Since $(\bigcup \mathcal{R})^c = \bigcap_{R' \in \mathcal{R}} (R')^c$ (De Morgan's identity), the second

relation means that $q' \in (R')^c$ for all $R' \in \mathcal{R}$; in particular, $q' \in R^c$. Since R is a Dedekind cut, we obtain $q < q'$.

For (D2), let $q \in \bigcup \mathcal{R}$. As before, we have $q \in R$ for a specific $R \in \mathcal{R}$. Since R is a Dedekind cut, there exists $q' \in R$ such that $q < q'$. Since $R \subset \bigcup \mathcal{R}$, we obtain $q' \in \bigcup \mathcal{R}$. Thus (D2) holds. We obtain that $\bigcup \mathcal{R}$ is a Dedekind cut.
The theorem follows.

It is customary to extend the real number system $\mathbb{R}$ by the symbols $\pm\infty$ with the understanding that $-\infty < R < \infty$ for any real number $R \in \mathbb{R}$. With this, if a non-empty set $\mathcal{A} \subset \mathbb{R}$ is not bounded above then we write $\sup \mathcal{A} = \infty$, and if $\mathcal{A}$ is not bounded below then we write $\inf \mathcal{A} = -\infty$.

We now turn to the **arithmetic** properties of $\mathbb{R}$. We define the operation of **addition** by setting

$$R + S = \{q + r \,|\, q \in R,\ r \in S\}, \quad R, S \in \mathbb{R}.$$

We proceed to show that $R + S$, $R, S \in \mathbb{R}$, is a Dedekind cut, so that the operation of addition is well-defined on $\mathbb{R}$.

Let $R, S \in \mathbb{R}$. Clearly, $R + S \subset \mathbb{Q}$ is non-empty, and it is also proper since the sum of upper bounds for R and S is an upper bound for $R + S$.

For (D1), let $q + r \in R + S$, $q \in R$, $r \in S$, and $s \in (R + S)^c$. Since $s \neq q + r$, we have $s < q + r$ or $s > q + r$. We claim that the first inequality cannot happen. Indeed, $s < q + r$ implies $s - q < r$ so that $s - q \in S$. Thus, $s = q + (s - q) \in R + S$, a contradiction. Thus, $q + r < s$, and (D1) follows.

For (D2), let $q \in R$ and $r \in S$. By (D2) applied to R and S resp., there exist $q' \in R$ and $r' \in S$ such that $q < q'$ and $r < r'$. Hence, we have $q + r < q' + r' \in R + S$, and (D2) follows. Thus, $R + S$ is a Dedekind cut.

Note that the operation of addition on Dedekind cuts is an extension of the addition in $\mathbb{Q}$ since $Q_q + Q_r = Q_{q+r}$, $q, r \in \mathbb{Q}$.

It is clear that the operation of addition is **commutative** and **associative**.

We now claim that $O = Q_0$ is the **additive identity**:

$$R + O = R, \quad R \in \mathbb{R}.$$

Indeed, recalling that O is the set of negative rational numbers, for $q \in R$ and $q' \in O$, we have $q + q' < q$, so that, by (D1), $q + q' \in R$. This gives the inclusion $R + O \subset R$. For the reverse inclusion, let $q \in R$. By (D2), there exists $q' \in R$ such that $q < q'$. We have $q = q' + (q - q') \in R + O$. This gives $R \subset R + O$. The claim follows.

Remark Note that O as an additive identity is **unique**. Indeed, if $O' \in \mathbb{R}$ is any additive identity then we have $O' = O' + O = O + O' = O$.

Before we proceed any further, we show an important and crucial property of the Dedekind cut to be used in the sequel:

Proposition 2.1.1 *Let $R \in \mathbb{R}$ be a Dedekind cut. (a) For every $0 < \epsilon \in \mathbb{Q}$ there exists $q \in R$ such that $q + \epsilon \in R^c$. (b) Let $R > O$. Then, for every $1 < a \in \mathbb{Q}$, there exists $0 < q \in R$ such that $q \cdot a \in R^c$.*

Proof Assume that part (a) of the proposition is false. This means that there exists $0 < \epsilon \in \mathbb{Q}$ such that, for **any** $q \in R$, we have $q + \epsilon \in R$. A simple use of Peano's Principle of Induction shows that $q + n\epsilon \in R$ for **any** $n \in \mathbb{N}$. (Indeed, the general induction step is given by $q + (n+1)\epsilon = (q + n\epsilon) + \epsilon$.) Now let $q \in R$ and $r \in R^c$ so that $q < r$. By the Archimedean Property of $\mathbb{Q}$ (Proposition 1.4.2), there exists $n \in \mathbb{N}$ such that $0 < r - q < n\epsilon$. This gives $r < q + n\epsilon \in R$. This contradicts to (D1). Part (a) of the proposition follows.

Assume that part (b) of the proposition is false. This means that there exists $1 < a \in \mathbb{Q}$ such that, for **any** $0 < q \in R$ we have $q \cdot a \in R$. Once again a simple use of Peano's Principle of Induction shows that $q \cdot a^n \in R$ for **any** $n \in \mathbb{N}$. Now let $0 < q \in R$ (q exists since $R > O$), and $r \in R^c$ so that $q < r$. By the Archimedean Property of $\mathbb{Q}$ again, there exists $n \in \mathbb{N}$ such that $0 < r/q < a^n$. This gives $r < q \cdot a^n \in R$. This contradicts to (D1). Part (b) of the proposition follows.

We now introduce the **negative** of a Dedekind cut $R \in \mathbb{R}$ as

$$-R = \{q \in \mathbb{Q} \mid -q > r \text{ for some } r \in R^c\}.$$

We claim that, for $R \in \mathbb{R}$, $-R \subset \mathbb{Q}$ is a Dedekind cut. Since R^c is non-empty, so is $-R$. The complement $(-R)^c$ is the set of all rational numbers $q' \in \mathbb{Q}$ such that $-q' \leq r'$ for **all** $r' \in R^c$. Since R^c is bounded below (by any element in R), we see that $(-R)^c$ s also non-empty.

For (D1), let $q \in -R$ and $q' \in (-R)^c$. Then $-q > r$ for some $r \in R^c$, and $-q' \leq r'$ for all $r' \in R^c$. Hence, we have $-q > r \geq -q'$ so that $q < q'$.

For (D2), let $q \in -R$ with $-q > r \in R^c$. Let $q' = (q - r)/2 \in \mathbb{Q}$. We have $-q' = (r - q)/2 > r$ so that $q' \in -R$. We also have $q < (q - r)/2 = q'$, so that (D2) follows.

Summarizing, we obtain that the negative is well defined in $\mathbb{R}$.

For rational Dedekind cuts, we have

$$-Q_q = Q_{-q}, \quad q \in \mathbb{Q}.$$

Indeed, using the fact that Q_q^c is the set of rational numbers $\geq q$, we calculate

$$\begin{aligned} -Q_q &= \{q' \in \mathbb{Q} \mid -q' > r \text{ for some } r \in Q_q^c\} \\ &= \{q' \in \mathbb{Q} \mid -q' > q\} \\ &= \{q' \in \mathbb{Q} \mid q' < -q\} = Q_{-q}. \end{aligned}$$

We now claim that the negative is the **additive inverse**:

$$R + (-R) = O, \quad R \in \mathbb{R}.$$

To show this, we first note that $q \in R$ and $q' \in -R$ with $-q' > r \in R^c$ imply $q + q' < q - r < 0$, so that $R + (-R) \subset O$. Conversely, let $s \in O$, a negative rational number. We apply Proposition 2.1.1 above to $0 < \epsilon = -s/2 \in \mathbb{Q}$ to obtain $q \in R$ such that $q + \epsilon = q - s/2 \in R^c$. Letting $q' = s - q \in \mathbb{Q}$, we have $-q' = q - s > q - s/2 \in R^c$ so that $q' \in -R$. With this, we have $q + q' = q + (s - q) = s \in O$. Thus, we have $O \subset R + (-R)$, and the claim follows.

Remark The additive inverse is unique. Indeed, if $R + R' = O$, $R, R' \in \mathbb{R}$, then we have $R' = R' + O = R' + (R + (-R)) = (R' + R) + (-R) = (R + R') + (-R) = O + (-R) = -R$.

Using the additive inverse property just proved, we obtain the **cancellation law for addition**: $A + C = B + C$, $A, B, C \in \mathbb{R}$, implies $A = B$. (Indeed, add $-C$ to both sides of the first equation and use associativity.) This, in turn, also gives $-(-A) = A$, $A \in \mathbb{R}$ (since $A + (-A) = (-A) + (-(-A)) = O$).

The sum and the negative satisfy the usual properties with respect to the order relation: $A < B$ implies $-B < -A$ and $A + C < B + C$ for any $C \in \mathbb{R}$. In particular, we call $A \in \mathbb{R}$ **positive** if $A > O$, and this holds if and only if $-A$ is **negative**, that is, $-A < O$.

Before turning to the multiplicative structure of $\mathbb{R}$, we introduce the **absolute value** of a Dedekind cut $R \in \mathbb{R}$ as

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

As before (Section 1.4), we have the usual properties of the absolute value. For $R \in \mathbb{R}$, we have $|-R| = |R|$ and $R \leq |R|$. In addition, if $0 \leq C \in \mathbb{R}$ then $-C \leq R \leq C$ if and only if $|R| \leq C$. Consequently, the **triangle inequality** holds:

$$||R| - |S|| \leq |R + S| \leq |R| + |S|, \quad R, S \in \mathbb{R}.$$

We now procceed to discuss the multiplicative structure of $\mathbb{R}$. We first define the **product** of non-negative Dedekind cuts $R, S \geq 0$ as

$$R \cdot S = \{q \cdot r \,|\, 0 \leq q \in R,\ 0 \leq r \in S\} \cup O.$$

Note that, if $R = O$ or $S = O$ then $R \cdot S = O$ (since the first set in the union above is empty). To show that $R \cdot S$ is a Dedekind cut we may therefore assume that $R, S > 0$, that is, we have $O \subset R \cap S$ and $R \neq O \neq S$.

Clearly, $R \cdot S$ is non-empty (since it contains O). Let $q' \in R^c$ and $r' \in S^c$. Then, by (D1) (applied to R and S), for **any** $0 \leq q \in R$ and $0 \leq r \in S$, we have $0 \leq q < q'$ and $0 \leq r < r'$, so that $0 \leq q \cdot r < q' \cdot r'$. Hence we have $q' \cdot r' \in (R \cdot S)^c$; in particular, $(R \cdot S)^c$ is non-empty.

For (D1), let $s \in (R \cdot S)^c$. Then, $s > 0$, and we need to show that $q \cdot r < s$ for all $0 < q \in R$ and $0 < r \in S$. Assume not. If $0 < s \leq q \cdot r$ for some $0 < q \in R$ and $0 < r \in S$ then $0 < s/q \leq r$, so that, by (D3), $0 < s/q \in S$. Hence $s = q \cdot s/q \in R \cdot S$, a contradiction. Thus (D1) follows.

For (D2), we let $0 \leq q \in R$ and $0 \leq r \in S$ and find $0 \leq q' \in R$ and $0 \leq r' \in S$ such that $q < q'$ and $r < r'$. Then we have $q \cdot r < q' \cdot r'$ and (D2) follows.

We conclude that the product $R \cdot S$, $R, S \in \mathbb{R}$, is a Dedekind cut.

As a byproduct, we also obtain that $R, S \geq O$, $R, S \in \mathbb{R}$, imply $R \cdot S \geq O$ with $R \cdot S = O$ if and only if $R = O$ or $S = O$.

We now extend the definition of the **product** to all Dedekind cuts $R, S \in \mathbb{R}$ using the absolute value as

$$R \cdot S = \begin{cases} -(R \cdot |S|) & \text{if } R \geq O \text{ and } S < O \\ -(|R| \cdot S) & \text{if } R < O \text{ and } S \geq O \\ |R| \cdot |S| & \text{if } R, S < O. \end{cases}$$

It follows immediately that the product of any Dedekind cuts is a Dedekind cut, so that multiplication is well-defined in $\mathbb{R}$.

Commutativity and **associativity** of the multiplication and **distributivity** follow directly from the definitions, first for non-negative Dedekind cuts, and then extended to all Dedekind cuts via the identity $-(-R) = R$, $R \in \mathbb{R}$, established earlier.

The fact that $Q_1 = \{q \in \mathbb{Q} \,|\, q < 1\}$, henceforth denoted by I, is the **multiplicative identity** also follows directly from the definitions:

$$R \cdot I = R, \quad R \in \mathbb{R}.$$

The existence of **multiplicative inverse** needs some elaboration.

First, for $R > O$, $R \in \mathbb{R}$, we define the multiplicative inverse of R by

$$R^{-1} = \{0 < q \in \mathbb{Q} \,|\, 1/q > r \text{ for some } r \in R^c\} \cup O \cup \{0\}.$$

For $R < O$, we define

$$R^{-1} = -|R|^{-1}.$$

By trichotomy, R^{-1} is now defined for all Dedekind cuts $R \neq O$.

Remark The definition of R^{-1} is analogous to that of $-R$ replacing the additive structure with the multiplicative structure.

Given $O \neq R \in \mathbb{R}$, we now need to show that R^{-1} is a Dedekind cut. We may assume $R > O$. Clearly R^{-1} is non-empty. The complement $(R^{-1})^c$ consists of all positive rational numbers $0 < q' \in \mathbb{Q}$ such that $1/q' \leq r'$ for **all** $r' \in R^c$. Since R^c is bounded below (by any element in R), we see that $(-R)^c$ is also non-empty.

For (D1), let $q \in R^{-1}$ and $q' \in (R^{-1})^c$. If $q \leq 0$ then $q < q'$ holds automatically since $q' > 0$. If $q > 0$ then $1/q > r$ for some $r \in R^c$. Since $1/q' \leq r'$ for all $r' \in R^c$, we have $1/q > r \geq 1/q'$, so that $q < q'$. (D1) follows.

For (D2), let $q \in R^{-1}$. We may assume $q > 0$ since $R > O$. We have $1/q > r$ for some $r \in R^c$. Let $q' \in \mathbb{Q}$ be defined by $q' = 2/(r + 1/q)$. We have $1/q' = (r + 1/q)/2 > r$ so that $q' \in R^{-1}$. We also have $q < 2/(r + 1/q) = q'$, so that (D2) follows.

Thus, R^{-1} is a Dedekind cut, and we conclude that the multiplicative inverse is well-defined in $\mathbb{R}$. Note also that $R > 0$ implies $R^{-1} > 0$.

As an easy consequence of the definitions, for rational Dedekind cuts, we have

$$Q_q^{-1} = Q_{1/q}, \quad 0 \neq q \in \mathbb{Q}.$$

Finally, we need to show that the multiplicative inverse of a non-zero Dedekind cut $R \in \mathbb{R}$, $R \neq O$, is R^{-1} defined above; that is, we have

$$R \cdot R^{-1} = I, \quad R \in \mathbb{R}.$$

First, let $R > O$. Combining the definitions of the product and the multiplicative inverse, we have

$$R \cdot R^{-1} = \{q \cdot q' \mid 0 < q \in R, 0 < q' \in \mathbb{Q} \text{ such that } 1/q' > r \text{ for some } r \in R^c\} \cup O \cup \{0\}.$$

To begin with, we note that $0 < q \in R$ and $0 < q' \in \mathbb{Q}$ with $1/q' > r \in R^c$ imply $q \cdot q' < q/r < 1$ since, by (D1), we have $q < r$. Thus, we have $R \cdot R^{-1} \subset I$.

Conversely, let $0 < s \in I$, that is, $s \in \mathbb{Q}$ is a rational number with $0 < s < 1$. We now apply part (b) of Proposition 2.1.1 for $a = 2/(s + 1) > 1$ to obtain $0 < q \in R$ such that $qa \in R^c$. Let $q' = s/q \in \mathbb{Q}$. We then have $1/q' = q/s > 2q/(s + 1) = qa \in R^c$. Therefore $q \cdot q' = s \in R \cdot R^{-1}$. We obtain $I \subset R \cdot R^{-1}$.

Combining these, we obtain that R^{-1} is the multiplicative inverse of R.

For $R < O$, we have $R^{-1} = -|R|^{-1} < 0$. Using this we compute

$$R \cdot R^{-1} = |R| \cdot |R^{-1}| = |R| \cdot |R|^{-1} = I,$$

where the last equality is by the previous step. The multiplicative inverse property above now follows in general.

Simple consequences of the existence of the multiplicative inverse are: (1) The **cancellation law for multiplication**: $R \cdot T = S \cdot T$, $R, S, T \in \mathbb{R}$, $T \neq O$, implies $R = S$; (2) Uniqueness of the multiplicative inverse: $R \cdot R' = I$, $R, R' \in \mathbb{R}$, implies $R' = R^{-1}$; (3) $(R^{-1})^{-1} = R$, $R \in \mathbb{R}$; (4) No zero divisors: $R \neq O \neq S$ imply $R \cdot S \neq O$.

With this we finished proving that the set of Dedekind cuts $\mathbb{R}$ forms a **field**, and it is the extension of the field of rational numbers $\mathbb{Q}$.

In addition, $\mathbb{R}$ is a **totally ordered field** with respect to the order relation $<$ extended from that of $\mathbb{Q}$; that is, $<$ is a strict total order on $\mathbb{R}$ with cancellation law for addition, and $R > O$ and $S > O$, $R, S \in \mathbb{R}$, imply $R \cdot S > O$. As direct consequences, we obtain: (1) The cancellation law for multiplication for inequalities: $R \cdot T < S \cdot T$ implies $R < S$ if $T > O$, and $R > S$ if $T < O$; (2) If $R \neq O$, $R \in \mathbb{R}$, then $R^2 > O$; in particular, $I > O$; (3) $O < R < S$ imply $O < S^{-1} < R^{-1}$.

Note that the symbols $\pm\infty$ introduced previously conform with the usual arithmetic properties; for example, we have $r \pm \infty = \pm\infty$, $r \in \mathbb{R}$; $r \cdot (\pm\infty) = \pm\infty$ if $0 < r \in \mathbb{R}$, $r \cdot (\pm\infty) = \mp\infty$ if $0 > r \in \mathbb{R}$, etc.

As shown earlier, $\mathbb{R}$ also has the Least Upper Bound Property: A subset bounded above, resp. below, has supremum, resp. infimum, attained in $\mathbb{R}$. We briefly refer to this property as **(Dedekind) completeness** of $\mathbb{R}$. We also say that $\mathbb{R}$ is a **complete ordered field**.

Dedekind's construction at the beginning of this section shows that, for $n \in \mathbb{N}$ not a perfect square, $R_n = \{q \in \mathbb{Q} \,|\, q < 0 \text{ or } q^2 < n\}$ is a Dedekind cut. Moreover, we claim

$$R_n^2 = R_n \cdot R_n = \{q \cdot r \,|\, 0 \leq q, r \in \mathbb{Q}, q^2 < n, r^2 < n\} \cup O = Q_n,$$

where $Q_n = \{q \in \mathbb{Q} \,|\, q < n\}$ is the Dedekind cut corresponding to the rational (actually natural) number $n \in \mathbb{N}$.

Indeed, if $q^2 < n$ and $r^2 < n$, $0 \leq q, r \in \mathbb{Q}$, then we have $(q \cdot r)^2 = q^2 \cdot r^2 < n^2$. This gives $q \cdot r < n$. We obtain $R_n^2 \subset Q_n$. For the converse, assume $0 < s \in Q_n$; that is, we have $0 < s < n$, $s \in \mathbb{Q}$. We let $0 < \epsilon = (n - s)/(2n + 1) \in \mathbb{Q}$ and apply part (a) of Proposition 2.1.1 to obtain $q \in R_n$ such that $q + \epsilon \in R_n^c = S_n$. We thus have $q^2 < n < (q + \epsilon)^2 = q^2 + 2q\epsilon + \epsilon^2$. Since $\epsilon = (n - s)/(2n + 1) < n/(2n + 1) < 1$ and $q < n$ (as $q^2 < n \leq n^2$), we obtain

$$0 < n - q^2 < 2q\epsilon + \epsilon^2 < 2q\epsilon + \epsilon = (2q + 1)\epsilon = (2q + 1)\frac{n - s}{2n + 1} < n - s.$$

This gives $s < q^2 \in R_n^2$. Since R_n^2 is a Dedekind cut, we obtain $s \in R_n^2$. Thus, $Q_n \subset R_n^2$, and the claim follows.

In what follows we will usually denote generic real numbers, the elements of $\mathbb{R}$, by lower case letters of the English alphabet.[5] We also think of the natural embedding of $\mathbb{Q}$ to $\mathbb{R}$ as identification, and write $q \in \mathbb{Q}$ for the Dedekind cut Q_q. In addition, we write 0 (zero) for O, and 1 (one) for I. Finally, in $\mathbb{R}$ we use customary notations such as $r - s$ for $r + (-s)$, $1/r$ for r^{-1}, etc.

By the discussion above, for $n \in \mathbb{N}$ not a perfect square, we denote $\sqrt{n} = R_n \in \mathbb{R}$, and then we have $(\sqrt{n})^2 = n$. If $n = a^2$, $a \in \mathbb{N}_0$, is a perfect square then we define $\sqrt{n} = \sqrt{a^2} = a$. (This includes $\sqrt{0} = 0$.) With this, the square root of any non-negative integer is defined in $\mathbb{R}$. This can easily be extended to square roots of

[5] We will also use Greek letters especially in trigonometry; see Chapter 11.

non-negative rational numbers, and with some additional work, to square roots of non-negative real numbers. We do not pursue this approach here as more advanced methods will be given later to define the mth root, $m \in \mathbb{N}$, of real numbers.

For future purposes, we now briefly digress from the main line of our study, and venture out to a related subject: rational approximations of square roots of natural numbers. To motivate this, we return to triangular numbers $T_n = \sum_{i=1}^{n} i = n(n+1)/2$, $n \in \mathbb{N}$, introduced in Section 0.4. We ask the following question: When is a triangular number a perfect square?

A quick inspection shows that the first four perfect square triangular numbers are $T_1 = 1^2$, $T_8 = 6^2$, $T_{49} = 35^2$, and $T_{288} = 204^2$.

The key to understand how to construct these numbers lies in **Pell's equation**

$$x^2 - d \cdot y^2 = 1.$$

Here $d \in \mathbb{N}$ is a **given** non-square natural number, and the solution amounts to finding all pairs $(x, y) \in \mathbb{N} \times \mathbb{N}$ (for this common d) such that the equation is satisfied.

History

The history of Pell's equation is circuitous and goes back to antiquities, due to the fact that the ratio x/y of a solution is a **rational approximation** of $\sqrt{d}$. The special case $d = 2$ was well-known to the Pythagoreans (c. 600 – 500 BCE). Later Archimedes posed and studied problems essentially equivalent to solving Pell's equation for $d = 3$, e.g. he found the rational approximation $1351/780$ of $\sqrt{3}$.

The first breakthrough in solving Pell's equation appeared in Brahmagupta's *Brahma-Sphuta-Siddhanta* (Chapter 18). (See the epitaph of this chapter.) He found an inductive method of constructing an infinite sequence of solutions starting from a given one (or two). His method is based on the so-called Brahmagupta identity; see below. The first general method of solving Pell's equation was given by Bhāskara II around 1150.

In the Western hemisphere, Pell's equation has been rediscovered in the 17th century by Fermat and the English mathematicians John Wallis (1616 – 1703) and Lord William Brounckner (1620 – 1684). Finally, Lord Brounckner's solution was mistakenly attributed by the famous Swiss mathematician Leonhard Euler (1707 – 1783) to John Pell (1611 – 1685) who translated an algebra book from German to English with a discussion on this solution.

The **Brahmagupta identity** alluded to above is the following

$$(x^2 - d \cdot y^2)(u^2 - d \cdot v^2) = (ux + dvy)^2 - d \cdot (vx + uy)^2.$$

The validity of this identity is a straightforward computation.[6] Its significance lies in the simple consequence that if (x, y) and (u, v) are two (not necessarily distinct) solutions of Pell's equation (for a given d) then a new solution is $(ux+dvy, vx+uy)$ (for the same d).

More precisely, given $d \in \mathbb{N}$, a pair $(u, v) \in \mathbb{N} \times \mathbb{N}$ is called the **fundamental solution** for Pell's equation if it is a solution with the smallest $u \in \mathbb{N}$. Then **all** solutions of Pell's equation form an infinite sequence of pairs $(x_k, y_k) \in \mathbb{N} \times \mathbb{N}$,

[6] Here and in the sequel we assume familiarity with basic algebraic computations, and defer a thorough treatment to Chapter 6.

$k \in \mathbb{N}_0$, which, starting with $(x_0, y_0) = (u, v)$, is defined inductively ($k \Rightarrow k+1$)[7] by

$$(x_{k+1}, y_{k+1}) = (ux_k + dvy_k, vx_k + uy_k), \quad k \in \mathbb{N}_0.$$

Remark We briefly indicate how to solve Brahmagupta's equation $x^2 - 92 \cdot y^2 = 1$ in the epitaph for this chapter above. First, we reduce the equation to $x^2 - 23 \cdot z^2 = 1$, with $z = 2y \in \mathbb{N}$ even. The fundamental solution $(u, v) = (24, 5)$ (with $v = z = 5$ odd) of this latter equation can be quickly found since $24^2 - 23 \cdot 5^2 = 24^2 - (24 - 1)(24 + 1) = 24^2 - (24^2 - 1) = 1$. Now, we use the inductive formula above with $(x_0, y_0) = (u, v) = (24, 5)$ to obtain $(x_1, y_1) = (24 \cdot 24 + 23 \cdot 5 \cdot 5, 5 \cdot 24 + 24 \cdot 5) = (1151, 240)$. This gives[8] $(x, y) = (1151, 120)$.

Returning to our triangular numbers, assume $T_n = m^2$, for some $m, n \in \mathbb{N}$. Hence $n(n+1) = 2m^2$, or equivalently, $(2n+1)^2 - 2 \cdot (2m)^2 = 1$. We see that $T_n = m^2$, $m, n \in \mathbb{N}$, if and only if $(x, y) = (2n+1, 2m)$ is a solution to Pell's equation ($d = 2$)

$$x^2 - 2 \cdot y^2 = 1.$$

Since $(3, 2)$ is obviously the fundamental solution, the discussion above (with $u = 3$ and $v = 2$) gives all solutions in the form of the infinite sequence of pairs $(x_k, y_k) \in \mathbb{N} \times \mathbb{N}$, $k \in \mathbb{N}_0$, $(x_0, y_0) = (3, 2)$, defined inductively by

$$(x_{k+1}, y_{k+1}) = (3x_k + 4y_k, 2x_k + 3y_k), \quad k \in \mathbb{N}_0.$$

The first four tems of this sequence[9] are $(3, 2)$, $(17, 12)$, $(99, 70)$, $(577, 408)$. Note that a simple induction shows that the first coordinate x_k is always odd, and the second y_k is even.

Finally, since $x = 2n + 1$, we see that if T_n, $n \in \mathbb{N}$, is a perfect square then the next is $T_{3n+1+\sqrt{8n(n+1)}}$. This gives all the triangular numbers that are perfect squares in the form of the infinite sequence $\{T_{n_k}\}_{k \in \mathbb{N}_0}$, $T_{n_0} = 1$, defined inductively by

---

[7] This statement can be proved by considering the **convergents** (initial segments) of the **continued fraction expansion** for the irrational number $\sqrt{d}$. This goes beyond the scope of our discussion, and, in specific examples in the sequel, we will always tacitly assume that the infinite sequence we obtain from the fundamental solution by induction gives **all** the solutions.

[8] The reader versed in number theory may observe the continued fraction expansion $\sqrt{23} = 4 + \cfrac{1}{1+\cfrac{1}{3+\cfrac{1}{1+\cfrac{1}{8+\cfrac{1}{1+\cfrac{1}{3+\cfrac{1}{1+\dots}}}}}}}$ with period four in 1, 3, 1, 8. The 8th convergent $4 + \cfrac{1}{1+\cfrac{1}{3+\cfrac{1}{1+\cfrac{1}{8+\cfrac{1}{1+\cfrac{1}{3+\cfrac{1}{1+0}}}}}}}$ is 1151/240.

[9] Note the continued fraction expansion $\sqrt{2} = 1 + \cfrac{1}{2+\cfrac{1}{2+\cfrac{1}{2+\cfrac{1}{2+\dots}}}}$ and its convergents $1, 3/2, 7/5, 17/12, 41/29, 99/70, 239/169, 577/408, \dots$.

$$n_{k+1} = 3n_k + 1 + \sqrt{8n_k(n_k+1)}, \quad k \in \mathbb{N}_0.$$

Using this, a more extended initial list of perfect square triangular numbers is: $T_1 = 1^2$, $T_8 = 6^2$, $T_{49} = 35^2$, $T_{288} = 204^2$, $T_{1681} = 1189^2$, $T_{9800} = 6930^2$, $T_{57121} = 40391^2$, $T_{332928} = 235416^2$, etc.

We now return to the main line, and note that completeness of $\mathbb{R}$ implies that the Archimedean property holds in $\mathbb{R}$:

Theorem 2.1.2 *Let $0 < r, s \in \mathbb{R}$. Then there exists $n \in \mathbb{N}$ such that $s \leq nr$.*

Proof Assume the Archimedean Property fails. This means that there exist $0 < r, s \in \mathbb{R}$ such that $nr \leq s$ for all $n \in \mathbb{N}$. The set $A = \{nr \mid n \in \mathbb{N}\}$ is therefore bounded above (with s as an upper bound). Let $s_0 = \sup A \in \mathbb{R}$. Since s_0 is the least upper bound for A, the real number $s_0 - r < s_0$ is **not** an upper bound for A. This means that $nr > s_0 - r$ holds for some $n \in \mathbb{N}$. Thus, we have $s_0 < (n+1)r$, a contradiction.

We now turn to the definition and properties of **non-negative integral exponents** of real numbers.

Let $0 \neq a \in \mathbb{R}$. We define the **powers** a^n, $n \in \mathbb{N}_0$, **inductively** as follows. For $n = 0$, we set $a^0 = 1$. Assuming that a^n is defined for $n \in \mathbb{N}_0$, we let $a^{n+1} = a \cdot a^n$. By Peano's Principle of Induction, a^n is defined for all $n \in \mathbb{N}_0$.

For $m, n \in \mathbb{N}_0$ and $0 \neq a, b \in \mathbb{R}$, counting factors in strings in various exponential expressions, we obtain the following identities:

$$a^{m+n} = a^m \cdot a^n, \quad (a^n)^m = a^{m \cdot n}, \quad (a \cdot b)^n = a^n \cdot b^n.$$

These identities can be established by simple induction. We prove the first formula by induction with respect to $m \in \mathbb{N}_0$. The formula obviously holds for $m = 0$. For the general induction step $m \Rightarrow m + 1$, we calculate

$$a^{m+1} \cdot a^n = (a \cdot a^m) \cdot a^n = a \cdot (a^m \cdot a^n) = a \cdot a^{m+n} = a^{m+n+1}.$$

The first formula follows.

Similarly, the second formula obviously holds for $m = 0$, and, for the general induction step $m \Rightarrow m + 1$, we calculate

$$(a^n)^{m+1} = a^n \cdot (a^n)^m = a^n \cdot a^{mn} = a^{n+mn} = a^{n(m+1)}.$$

The proof of the last formula is simple.

Example 2.1.1 Let $2 \leq n \in \mathbb{N}$. Show that the number $2^{2(2n-1)} + 1$ is composite.

We add and subtract a suitable power of 2 and calculate as follows

$$\begin{aligned} 2^{2(2n-1)} + 1 &= 2^{2(2n-1)} + 2 \cdot 2^{2n-1} + 1 - 2^{2n} \\ &= \left(2^{2n-1} + 1\right)^2 - \left(2^n\right)^2 \end{aligned}$$

$$= \left(2^{2n-1} - 2^n + 1\right)\left(2^{2n-1} + 2^n + 1\right).$$

The example follows.

For $n = 2, 3, 4, 5, 6$, the example above gives

$$\begin{aligned}
2^6 + 1 &= (2^3 - 2^2 + 1)(2^3 + 2^2 + 1) = 5 \cdot 13 \\
2^{10} + 1 &= (2^5 - 2^3 + 1)(2^5 + 2^3 + 1) = 5^2 \cdot 41 \\
2^{14} + 1 &= (2^7 - 2^4 + 1)(2^7 + 2^4 + 1) = 5 \cdot 29 \cdot 113 \\
2^{18} + 1 &= (2^9 - 2^5 + 1)(2^9 + 2^5 + 1) = 5 \cdot 13 \cdot 37 \cdot 109 \\
2^{22} + 1 &= (2^{11} - 2^6 + 1)(2^{11} + 2^6 + 1) = 5 \cdot 397 \cdot 2113,
\end{aligned}$$

where the final equalities are the prime factorizations.[10]

Example 2.1.2 Which is bigger 33^{17} or 15^{20}?

We first notice that $33 > 2^5$ and $15 < 2^4$. With these we calculate

$$33^{17} > (2^5)^{17} = 2^{85},$$

whereas

$$15^{20} < 16^{20} = (2^4)^{20} = 2^{80}.$$

Thus, we have $33^{17} > 15^{20}$.

Example 2.1.3 For $n \in \mathbb{N}$, which is bigger, n^2 or 2^n?

We begin to evaluate a few cases: $1^2 < 2^1$ ($n = 1$), $2^2 = 2^2$ ($n = 2$), $3^2 > 2^3$ ($n = 3$), $4^2 = 2^4$ ($n = 4$), and $5^2 < 2^5$ ($n = 5$). Based on these, we claim

$$n^2 < 2^n, \quad 5 \leq n \in \mathbb{N}.$$

We show this by induction[11] with respect to $5 \leq n \in \mathbb{N}$. The case $n = 5$ has just been listed above. For the general induction step $n \Rightarrow n + 1$, we calculate

$$2^{n+1} = 2 \cdot 2^n > 2n^2 > n^2 + 2n + 1 = (n + 1)^2,$$

where the last inequality is because $n^2 > 2n + 1$, $3 \leq n \in \mathbb{N}$. The claim follows.

A somewhat more involved estimate (to be used in the sequel) is contained in the following:

[10] The factorization of $2^{22} + 1$ was a problem in the $MA\Theta$ National Convention, 1991.

[11] This is an example for an induction that starts at $n = 5$.

Example 2.1.4 We have

$$n^{n+1} > (n+1)^n, \quad 3 \le n \in \mathbb{N}.$$

To show this we use induction with respect to $3 \le n \in \mathbb{N}$. For $n = 3$, we have $3^4 = 81 > 64 = 4^3$. (Note that the inequality fails for $n = 1, 2$.) For the general induction step $n - 1 \Rightarrow n$, we assume that

$$(n-1)^n > n^{n-1}$$

holds for some $4 \le n \in \mathbb{N}$. (The shift in the value of n to $n - 1$ is of technical convenience.) In the next steps we will use all the identities of exponentiation above. We multiply both sides by $n(n+1)^n$, and obtain

$$n(n-1)^n(n+1)^n = n(n^2-1)^n > n^n(n+1)^n,$$

where we used $(n-1)(n+1) = n^2 - 1$. This gives

$$n^{2n+1} = n \cdot (n^2)^n > n(n^2-1)^n > n^n(n+1)^n.$$

Dividing by n^n, we obtain the desired inequality stated above. The induction is complete and the inequality follows.

Example 2.1.5 For $n \in \mathbb{N}$, define the **finite** sequence a_n inductively as follows. Let $a_1 = (1, 1)$, and construct a_{n+1} from a_n by inserting the sum between any two consecutive terms in a_n as a new term. We thus have $a_2 = (1, 2, 1)$, $a_3 = (1, 3, 2, 3, 1)$, $a_4 = (1, 4, 3, 5, 2, 5, 3, 2, 1)$, etc. Let t_n, resp. s_n, $n \in \mathbb{N}$, be the number of terms, resp., the sum of all terms of a_n. Determine t_n and s_n, $n \in \mathbb{N}$.

We have $t_1 = 2$ and $t_{n+1} = t_n + (t_n - 1) = 2t_n - 1$, $n \in \mathbb{N}$ (as there are $t_n - 1$ "gaps" between the consecutive terms in the sequence a_n). Letting $t'_n = t_n - 1$, $n \in \mathbb{N}$, we obtain $t'_1 = 1$ and $t'_{n+1} = 2t'_n$, $n \in \mathbb{N}$. This is the inductive formula for the powers of 2 (with the exponent shifted), so that we obtain $t'_n = 2^{n-1}$, $n \in \mathbb{N}$. Playing this back to the original sequence, we get $t_n = 2^{n-1} + 1$, $n \in \mathbb{N}$. As for the sum, we have $s_1 = 2$ and $s_{n+1} = s_n + (2s_n - 2) = 3s_n - 2$, $n \in \mathbb{N}$ (as each term in the sequence a_n has two neighbors except the two 1's at the end). Letting $s'_n = s_n - 1$, $n \in \mathbb{N}$, we obtain $s'_1 = 1$ and $s'_{n+1} = 3s'_n$, $n \in \mathbb{N}$. This is the inductive formula for the powers of 3 (with the exponent shifted), so that we obtain $s'_n = 3^{n-1}$, $n \in \mathbb{N}$. Playing this back to the original sequence, we get $s_n = 3^{n-1} + 1$, $n \in \mathbb{N}$.

Example 2.1.6 Find all natural numbers $a, b, c \in \mathbb{N}$, $a < b$, such that $2^a + 2^b$ and $2^a + 2^b + 2^c$ are both perfect squares.[12]

[12]The special case $2^8 + 2^{11} = 48^2$ and finding $c \in \mathbb{N}$ was a problem in the Hungarian Olympiad, 1981. (See the case $k = 4$ above.)

Setting $2^a + 2^b = u^2$ and $2^a + 2^b + 2^c = v^2$, $u, v \in \mathbb{N}$, we have $2^c = v^2 - u^2 = (v-u)(v+u)$. This gives

$$v - u = 2^{k+1} \quad \text{and} \quad v + u = 2^{l+1}, \ k < l, \ k, l \in \mathbb{N},$$

where $c = k + l + 2$. (Note that $k = -1, 0$ cannot happen.) Solving, we obtain

$$v = 2^l + 2^k \quad \text{and} \quad u = 2^l - 2^k.$$

Returning to the beginning of the problem, using $a < b$, we obtain

$$2^a + 2^b = 2^a \left(2^{b-a} + 1\right) = u^2 = \left(2^l - 2^k\right)^2 = 2^{2l} - 2 \cdot 2^l \cdot 2^k + 2^{2k}$$
$$= 2^{2k} \left(2^{2(l-k)} - 2^{l-k+1} + 1\right).$$

Comparing, we obtain $a = 2k$, and hence

$$2^{b-a} = 2^{2(l-k)} - 2^{l-k+1} = 2^{l-k+1} \left(2^{l-k-1} - 1\right).$$

This holds if and only if $l - k - 1 = 1$, or equivalently, $l = k + 2$. With this we also obtain $b - a = l - k + 1 = 3$.

Summarizing, we obtain

$$a = 2k, \quad b = 2k + 3, \quad c = 2k + 4, \ k \in \mathbb{N}.$$

Note the first few cases, $k = 1, 2, 3, 4$, as follows

$$2^2 + 2^5 = 6^2 \quad 2^2 + 2^5 + 2^6 = 10^2$$
$$2^4 + 2^7 = 12^2 \quad 2^4 + 2^7 + 2^8 = 12^2$$
$$2^6 + 2^9 = 24^2 \quad 2^6 + 2^9 + 2^{10} = 40^2$$
$$2^8 + 2^{11} = 48^2 \quad 2^8 + 2^{11} + 2^{12} = 80^2.$$

The concept of power a^n, $n \in \mathbb{N}_0$, can be extended to negative integral exponents in a straightforward manner requiring that the identities should hold in the extended range. Setting $m + n = 0$ in the first exponentiation identity, and using $a^0 = 1$, we see that we must define

$$a^{-m} = \frac{1}{a^m}, \quad m \in \mathbb{N}.$$

It is an easy case-by-case verification that the identities above hold for the extended range $m, n \in \mathbb{Z}$. In addition, we also have the new identity

$$a^{m-n} = \frac{a^m}{a^n}, \quad m, n \in \mathbb{Z}.$$

In the next example we briefly return to base 10 arithmetic.

Example 2.1.7 Show that, for $n \in \mathbb{N}$, we have

$$\overbrace{11\ldots1}^{n}\overbrace{22\ldots2}^{n} = \overbrace{33\ldots3}^{n}(\overbrace{33\ldots3}^{n}+1).$$

where the overbraces indicate the number of occurrences of the respective digits.

The crux is to write the number with n repeated digits $d \in \{1, 2, \ldots, 9\}$ as

$$\overbrace{dd\ldots d}^{n} = d\frac{10^n - 1}{9}.$$

With this, we have

$$\begin{aligned}\overbrace{11\ldots1}^{n}\overbrace{22\ldots2}^{n} &= \frac{10^n-1}{9}\cdot 10^n + 2\frac{10^n-1}{9} = \frac{10^n-1}{3}\cdot\frac{10^n+2}{3}\\ &= \frac{10^n-1}{3}\cdot\left(\frac{10^n-1}{3}+1\right) = \overbrace{33\ldots3}^{n}(\overbrace{33\ldots3}^{n}+1).\end{aligned}$$

The example follows.

History

The term **power** that we use nowadays is attributed to Euclid of Alexandria (c. 300 BCE). The power a^2 is called the **square** of a because it represents the area of a square with side length a. Similarly, a^3, the **cube** of a, represents the volume of a cube with edge length a. The first recorded use of the identities of natural exponents was by Archimedes who established the identity $10^{m+n} = 10^m \cdot 10^n$, $m, n \in \mathbb{N}$. The term **exponent** is attributed to Michael Stifel (1487 – 1567) in 1544. The term **theory of indices** (the theory of exponentiation) had a long and widespread use since its introduction by Samuel Jeake (1623 – 1690). The first modern notation for exponents was introduced by Descartes in his work *La géometrie* (published in 1637). It is an interesting fact that Isaac Newton (1642 – 1727) and some of his contemporaries used Descartes' power notation only for exponents greater than or equal to 3. For quadratic terms such as a^2 and b^2 they wrote $a \cdot a$ and $b \cdot b$.

Example 2.1.8 For $n = 1, 2, 3, 4$, calculate the number

$$2^{2^n} + 1.$$

What can be conjectured about these numbers?

We have

$$2^{2^1}+1=2^2+1=5,$$
$$2^{2^2}+1=2^4+1=17,$$
$$2^{2^3}+1=2^8+1=257,$$
$$2^{2^4}+1=2^{16}+1=65,537.$$

We notice that 5, 17, and 257 are primes. It turns out that 65, 537 is also a prime. For this we need only to make sure that it has no prime divisors up to 257 (since $257^2 = 66,049 > 65,537$). See Section 1.3 for a list of primes up to 257.

History
Fermat conjectured that the numbers $2^{2^n}+1$ are primes for all $n \in \mathbb{N}$. Because of this, they are called **Fermat numbers**. In 1732 Euler discovered that the number

$$2^{2^5}+1=2^{32}+1=4,294,967,297=641\cdot 6,700,417$$

is composite.
Beyond the ones given above, it is not known how many Fermat numbers are primes. This is an important problem not only in number theory but also in geometry, since Gauss showed that, for p a prime, a regular p-sided polygon is constructible by straightedge (unmarked ruler) and compass if and only if the pth Fermat number $2^{2^p}+1$ is a prime.

The next two problems are of related genre, still concerning large powers of small numbers.

Example 2.1.9 [13] Determine the prime factorization of the number $2^{18}+1$.
We have

$$2^{18}+1=\left(2^9\right)^2+2\cdot 2^9+1-2\cdot 2^9=\left(2^9+1\right)^2-\left(2^5\right)^2$$
$$=\left(2^9+2^5+1\right)\left(2^9-2^5+1\right)=545\cdot 481.$$

Now, a simple inspection gives $545 = 5 \cdot 109$ and $481 = 13 \cdot 37$. With these, we finally arrive at $2^{18}+1 = 5 \cdot 13 \cdot 37 \cdot 109$.

Example 2.1.10 What is the largest exponent $m \in \mathbb{N}$ such that 2^m divides $3^{2^{18}}-1$?
We have

$$3^{2^{18}}-1=3^{2\cdot 2^{17}}-1=\left(3^{2^{17}}\right)^2-1=\left(3^{2^{17}}+1\right)\left(3^{2^{17}}-1\right).$$

This factorization can be repeated inductively, and we obtain

[13] Many variants of this are used in mathematical contests and preparations; see for example the prime factorization of the number $2^{22}+1$ in the $MA\Theta$ National Convention, 1991.

$$3^{2^{18}} - 1 = \left(3^{2^{17}} + 1\right)\left(3^{2^{16}} + 1\right) \cdots \left(3^{2^2} + 1\right)\left(3^2 + 1\right)(3 + 1) \cdot 2.$$

We now use the simple fact that, for any odd number $a \in \mathbb{N}$, the square $a^2 + 1$ is 2 times an odd number. (Indeed, writing $a = 2k + 1$, $k \in \mathbb{N}_0$, we have $a^2 + 1 = (2k + 1)^2 + 1 = 4k^2 + 4k + 2 = 2(2k(k + 1) + 1)$.) We apply this to each factor of the product above with $a = 3^{2^l}$, $l = 0, 1, 2, \ldots, 16$, except the last two, and obtain that each is a single multiple of 2 (times an odd number). Counting all the 2's, we get $m = 17 + 2 + 1 = 20$.

We close this section with the **Bernoulli inequality**. It will be of paramount importance in our subsequent study.

Bernoulli Inequality (Integral Exponent) *Let $-1 < r \in \mathbb{R}$. Then, for any $n \in \mathbb{N}_0$, we have*

$$(1 + r)^n \geq 1 + nr.$$

Sharp inequality holds for $r \neq 0$ and $n \geq 2$.

Proof We use induction with respect to $n \in \mathbb{N}_0$.

The initial step is obvious, since, by definition, we have $(1 + r)^0 = 1$.

For the general induction step $n \Rightarrow n + 1$, we assume that the inequality above holds, and show that it also holds for $n + 1$. We calculate

$$\begin{aligned}(1 + r)^{n+1} &= (1 + r)(1 + r)^n \geq (1 + r)(1 + nr) \\ &\geq 1 + (n + 1)r + nr^2 \geq 1 + (n + 1)r.\end{aligned}$$

The induction is complete, and the inequality follows. The sharp inequality is clear for $r \neq 0$ and $n = 2$, and therefore, by induction, for $n \geq 2$.

History

The inequality above appeared in the treatise *Positiones Arithmeticae de Seriebus Infinitis* published in 1689 by Jacob Bernoulli (1655–1705), and it was subsequently named after him. The primary authorship is disputed by J.E. Hofman who states the following: "Bernoulli ist durchaus nicht der Erfinder dieser Ungleichung, hat sie jedoch vermutlich nicht direkt aus Sluse, sondern auf dem Umweg über I. Barrow (1630–1677)." (See Formula (4,12) on p. 177 in *Über die Exercitatio Geometrica des M. A. Ricci*, Centaurus, Vol. 9, Issue 3 (1963) 139–193.) The inequality is then somewhat older and is probably due to René-François de Sluse (1622–1685) published in his 1668 work *Mesolabum*, Chapter IV *De maximis & minimis*.

Corollary 1 *Let $1 < a \in \mathbb{R}$ and $0 < s \in \mathbb{R}$. Then there exists $n \in \mathbb{N}$ such that $s \leq a^n$.*

Proof By the Archimedean Property for real numbers, Theorem 2.1.2, we have $s \leq n(a - 1)$ for some $n \in \mathbb{N}$. We now use the Bernoulli inequality (for $a = s + 1$) as follows:

$$s \leq n(a - 1) < 1 + n(a - 1) \leq a^n.$$

The corollary follows.

Corollary 2 *For* $1 < a \in \mathbb{R}$*, we have*

$$\inf\left\{\frac{1}{a^n}\,\middle|\, n \in \mathbb{N}\right\} = 0.$$

Remark According to the synthetic approach, the real number system is defined as a **complete ordered field** via a set of axioms. By what we discussed above, this means that the real number system is a set $\mathbb{R}$ equipped with two (binary) operations, called addition $+$ and multiplication $\cdot$, and a (binary) relation $\leq$ with respect to which $\mathbb{R}$ is a **totally ordered field**. In addition, $\mathbb{R}$ must be complete.

These axioms are **categorical** in the sense that there is an explicitly constructible model for these axioms (usually, but not always, from $\mathbb{Q}$, like in our case as the set of Dedekind cuts), and the axioms can be proved as theorems in these models. Moreover, any two such models are **isomorphic**; that is, there is a one-to-one correspondence between them which respects the field operations and the order.

While the axioms for an ordered field are fairly transparent (and have been discussed for $\mathbb{Q}$ and $\mathbb{R}$), the axiom of completeness takes various, sometimes inequivalent, forms. In our construction of real numbers via Dedekind cuts we used the Least Upper Bound Property which, in synthetic approach, takes the form of an axiom. In Section 2.3 we will introduce another concept of completeness via Cauchy sequences.

Exercises

2.1.1. Solve for $x \in \mathbb{R}$:

$$\left(\frac{x+|x|}{2}\right)^2 + \left(\frac{x-|x|}{2}\right)^2 = x^2.$$

2.1.2. Solve the inequality $x \leq |x - x^2|$, $x \in \mathbb{R}$.

2.1.3. Let $r_1, r_2, \ldots, r_{2n} \in \mathbb{R}$, $n \in \mathbb{N}$, be $2n$ real numbers such that $r_1 \leq r_2 \leq \ldots \leq r_{2n}$. For what $r \in \mathbb{R}$ do we have the least value of the expression

$$|r - r_1| + |r - r_2| + \cdots + |r - r_{2n}|?$$

2.1.4. Which is bigger $\sqrt{101} - \sqrt{100}$ or $1/20$?

2.1.5. Derive the following identity:

$$\sqrt{a + b + 2\sqrt{ab}} = \sqrt{a} + \sqrt{b}, \quad 0 \leq a, b \in \mathbb{R}.$$

2.1.6. Let $a, b \in \mathbb{N}$, $1 \leq a \leq b \leq 100$. For what values of a, b is $\sqrt{\sqrt{a} + \sqrt{b}}$ integral?

2.1.7. Arrange the numbers $\sqrt{2}$, $\sqrt[3]{3}$, and $\sqrt[4]{4}$ in increasing order.

2.1.8. Let $n \in \mathbb{N}$. Calculate

$$\left[\left(\sqrt{2}+1\right)^n + \left(\sqrt{2}-1\right)^n\right]^2 - \left[\left(\sqrt{2}+1\right)^n - \left(\sqrt{2}-1\right)^n\right]^2.$$

2.1.9. For $n \in \mathbb{N}$, derive the following divisibility properties (a) $3 \mid 2^{2n} - 1$; (b) $9 \mid 4^{3n} - 1$.

2.1.10. Let $a^m = b^n$, $a, b \in \mathbb{N}$, with $m, n \in \mathbb{N}$ relatively prime, $\gcd(m, n) = 1$. Show that $a = u^n$ and $b = u^m$ for some $u \in \mathbb{N}$.

2.1.11. Let $0 < a, b \in \mathbb{R}$. Show that

$$\frac{a^n + b^n}{2} \geq \left(\frac{a+b}{2}\right)^n, \quad n \in \mathbb{N}.$$

2.1.12. Solve Pell's equation $x^2 - d \cdot y^2 = 1$ if $d + 2$ is a perfect square. (Note the special case $d = 23$ in Section 2.1.)

2.1.13. Find all $n \in \mathbb{N}$ such that $5^n > n!$.

2.2 Infinite Decimals as Real Numbers

In the previous section we constructed the field of real numbers $\mathbb{R}$ as the set of Dedekind cuts of the set of rational numbers $\mathbb{Q}$. We showed that $\mathbb{R}$ is an extension field of $\mathbb{Q}$, and that it is a (Dedekind) complete ordered field with respect to its natural order $<$. The latter means that it has the Least Upper Bound Property: Any subset bounded above, resp. below, assumes its supremum, resp. infimum, in $\mathbb{R}$.

Although representing real numbers by Dedekind cuts is elegant and **unique** (that is, by definition, to any real number there corresponds a unique Dedekind cut), in computations they are oftentimes cumbersome; consider, for example, the definition of the square root of an integer given at the previous section.

The question therefore naturally arises: How to represent a real number in a simpler and more transparent, preferably algebraic way? The key to this is to consider the decimal representation of rational numbers.

The decimal representation of integers naturally extends to **decimal representation of rational numbers** by introducing the concept of **decimal fraction**. A decimal fraction is a quotient of two integers in which the denominator is a power of 10. Even though they are quotients of integers, decimal fractions are written in decimal notation rather than as fractions. This is done by discarding the denominator and retaining the numerator only, inserting the **decimal separator** into the numerator at the position from the right corresponding to the exponent of the

power of ten of the denominator, and filling the possible gap with zeros if necessary. The decimal separator is the dot "." in the US, and the comma "," in Europe.

For example, the **universal gravitational constant** can be written (in SI units) as

$$G = \frac{667408(31)}{10^{18}} \frac{m^3}{kg \cdot s^2} = 0.0000000000667408(31) \frac{m^3}{kg \cdot s^2}$$

with standard uncertainty in parentheses.

History
The earliest appearance of decimal fractions were in China at the end of the 4th century BCE. The Chinese also compiled the first decimal multiplication table made from bamboo strips around 305 BCE. The use of the decimal numbers then spread to the Middle East and subsequently to Europe.

If the denominator of a rational number $q = a/b$, $a, b \in \mathbb{Z}$, $b \neq 0$, has only 2 and 5 as prime divisors then the conversion of q to a decimal representation is particularly simple. Letting $b = 2^k \cdot 5^l$, $0 \leq k, l \in \mathbb{Z}$, we have

$$q = \frac{a}{b} = \frac{a}{2^k \cdot 5^l} = \frac{2^l \cdot 5^k \cdot a}{10^{k+l}}.$$

As specific examples, we have

$$\frac{1}{2} = 0.5, \quad \frac{1}{5} = 0.2, \quad \frac{1}{4} = 0.25, \quad \frac{1}{25} = 0.04, \quad \text{etc.}$$

In these cases the rational number can be written as a **single** decimal fraction.

In general, converting a rational number into decimal representation is done by the **long division algorithm**.

If $q = a/b$ is a positive rational number with $a, b \in \mathbb{N}$ then, dividing a by b, each decimal in the decimal representation of q is obtained by multiplying the remainder of the previous step by 10 and dividing it by b to get the new remainder. (Here and in what follows, for simplicity, may restrict ourselves to positive rational numbers since the decimal representation of a negative rational number $q \in \mathbb{Q}$ is the negative of the decimal representation of $-q$.)

During the conversion we may end up with an infinite sequence of nonzero remainders, and therefore the corresponding rational number is written as a sum of **infinitely many** decimal fractions, or an **infinite decimal representation**. The simplest example is

$$\frac{1}{3} = \frac{3}{10} + \frac{3}{10^2} + \frac{3}{10^3} + \cdots = 0.333\ldots$$

During the long division of a by b the remainders are between 0 and $b - 1$, and therefore this process necessarily repeats itself. We conclude that a rational number

$q = a/b$ has a decimal representation which ends with a string of decimals (which may be a single zero) repeated indefinitely. (For this reason this string of decimals is sometimes called the **repetend**.)

Summarizing, the decimal representation of a rational number either **repeats** (infinitely) or **terminates** (by zero). Clearly, the latter happens if and only if the denominator b of the rational number $q = a/b$ has only 2 and 5 as prime divisors.

The converse of the statement above is also true: If an infinite decimal representation **ends** with a string of decimals repeated indefinitely then this represents a rational number.

To show this, we start with an infinite repeating decimal representation. As always, it starts with an integer (which may be zero), and, by assumption, after a string of "irregular decimals," it ends with a repetend, a string of decimals $d_1 d_2 \dots d_k$ with $k \geq 1$, repeated indefinitely.

To simplify matters, we make two adjustments. First, we can multiply the decimal representation by a suitable power of 10 to move the irregular string to the left of the decimal point after which the repetition pattern would start immediately:

$$a.d_1 d_2 \dots d_k d_1 d_2 \dots d_k d_1 d_2 \dots d_k \dots = a.\overline{d_1 d_2 \dots d_k}.$$

Here we used the customary notation of placing a bar over the repetend, the group of k digits $d_1 d_2 \dots d_k$ which are repeated indefinitely. Since we want to deduce that this number is rational, the initial multiplication by a power of 10 does not change this. Second, we can also make the "integral part" zero by subtracting $a.0$ since, once again, rationality is not affected by subtracting an integer such as a.

All in all, we can now study the reduced form

$$0.d_1 d_2 \dots d_k d_1 d_2 \dots d_k d_1 d_2 \dots d_k \dots = 0.\overline{d_1 d_2 \dots d_k}.$$

The crux is to understand what fractions create repeating decimal patterns. The simplest repeating pattern is easy to find:

$$\frac{1}{9} = 0.1111111111 \dots = 0.\overline{1}$$

If we multiply both sides by a single digit integer $d_1 \in \{1, 2, \dots, 8, 9\}$ then we obtain the repeating pattern

$$\frac{d_1}{9} = 0.d_1 d_1 d_1 d_1 d_1 d_1 \dots = 0.\overline{d_1}$$

Remark Letting $d_1 = 9$, we obtain $1 = 0.999999\dots = 0.\overline{9}$. On the other hand, 1 has the obvious decimal representation $1 = 1.000000\dots$. We see that the decimal representation of rational numbers is **not unique**. More generally, a decimal representation of a rational number with a tail of infinitely repeating nines

represents the same rational number as the finite decimal representation obtained by deleting this tail and moving up the digit before the tail by one unit. Note finally that this is the only exception; otherwise each rational number has a unique decimal representation.

Next, the simplest double digit pattern is

$$\frac{10}{99} = 0.1010101010\ldots = 0.\overline{10}$$

As before, multiplying both sides with a double digit integer, we obtain

$$\frac{d_1d_2}{99} = 0.d_1d_2d_1d_2d_1d_2\ldots = 0.\overline{d_1d_2}$$

The general case now follows easily. We have

$$\frac{d_1d_2\ldots d_k}{10^k-1} = 0.d_1d_2\ldots d_kd_1d_2\ldots d_kd_1d_2\ldots d_k\ldots = 0.\overline{d_1d_2\ldots d_k}$$

where we replaced the string of k digits of 9 with $10^k - 1$.

Notice that we not only obtained our original statement, but also found a constructive way to obtain any rational number from its decimal representation.

Example 2.2.1 [14] Consider the repeating decimal

$$0.c_1\ldots c_jd_1\ldots d_kd_1d_2\ldots d_kd_1d_2\ldots d_k\ldots = 0.c_1\ldots c_j\overline{d_1\cdots d_k}$$

where $j \geq 1$; that is, there is at least one decimal digit before the repeating part. Represent this as a simple fraction a/b, where a and b have no common divisors. Show that b is divisible by 2 or 5 (or both).

Using our formula for the reduced repeating decimal above, we calculate

$$\begin{aligned}
0.c_1\ldots c_jd_1\ldots d_kd_1d_2\ldots d_k\ldots &= 0.c_1\ldots c_j + 0.\overbrace{00\ldots0}^{j}\overline{d_1\ldots d_k}\\
&= \frac{c_1\ldots c_j}{10^j} + \frac{0.\overline{d_1\ldots d_k}}{10^j} = \frac{c_1\ldots c_j}{10^j} + \frac{d_1\ldots d_k}{10^j(10^k-1)}\\
&= \frac{c_1\ldots c_j(10^k-1)+d_1\ldots d_k}{10^j(10^k-1)} = \frac{c_1\ldots c_j10^k+d_1\ldots d_k-c_1\ldots c_j}{10^j(10^k-1)}\\
&= \frac{c_1\ldots c_jd_1\ldots d_k-c_1\ldots c_j}{10^j(10^k-1)}.
\end{aligned}$$

[14] Although fairly well-known, this problem was in the USA Mathematical Olympiad, 1988, with the specific illustrative example $0.01136363636\ldots = 1/88$.

The decimal representation of the numerator in the last fraction cannot end with j consecutive zeros since $c_1 \dots c_j$ is different from the repeating group $d_1 \cdots d_k$. Hence, upon reducing this fraction to a simple fraction, in the denominator a factor of 2 or 5 survives. The example follows.

We now return to our original question of algebraic representation of real numbers. We consider a Dedekind cut $r \in \mathbb{R}$ which we assume to be positive $r > 0$. This means that $0 \subset r$ is a proper subset. (Recall that we identified 0 with the Dedekind cut of all negative rational numbers.) We now define an infinite sequence of rational numbers all **contained in** r in the following form:

$$\begin{aligned} r_0 &= a \\ r_1 &= a + \frac{d_1}{10} \\ r_2 &= a + \frac{d_1}{10} + \frac{d_2}{10^2} \\ &\dots \\ r_n &= a + \frac{d_1}{10} + \frac{d_2}{10^2} + \frac{d_3}{10^3} + \cdots + \frac{d_n}{10^n} \qquad \dots \end{aligned}$$

where $d_1, d_2, d_3, \dots, d_n, \dots \in \{0, 1, 2, \dots, 9\}$. We choose the first member $a \in \mathbb{N}_0$ to be the largest non-negative integer contained in r. Proceeding **inductively**, assume that r_n in the form above has been chosen. Then choose $r_{n+1} = r_n + d_{n+1}/10^{n+1}$ with the largest $d_{n+1} \in \{0, 1, 2, \dots, 9\}$ contained in r. By Peano's Principle of Induction, r_n is defined for all $n \in \mathbb{N}_0$.

The partial sums above form an infinite sequence of rational numbers which is increasing:

$$r_0 \le r_1 \le r_2 \le r_3 \le \cdots \le r_n \le \cdots \le r.$$

We want to estimate how close the individual members of this sequence are to each other. Letting $1 \le m < n$ first, we have

$$\begin{aligned} r_n - r_m &= \frac{d_{m+1}}{10^{m+1}} + \cdots + \frac{d_n}{10^n} \le \frac{9}{10^{m+1}} + \cdots + \frac{9}{10^n} = \frac{10-1}{10^{m+1}} + \cdots + \frac{10-1}{10^n} \\ &= \left(\frac{1}{10^m} - \frac{1}{10^{m+1}}\right) + \cdots + \left(\frac{1}{10^{n-1}} - \frac{1}{10^n}\right) = \frac{1}{10^m} - \frac{1}{10^n} < \frac{1}{10^m} \end{aligned}$$

since in the last sum all but the first and last terms cancel.[15] This gives the general estimate

[15] These sums are called **telescopic**. More about them later.

$$|r_n - r_m| \leq \frac{1}{10^{\min(m,n)}}, \quad m, n \in \mathbb{N}_0.$$

An important **second** sequence is the following

$$s_0 = r_0 + 1, s_1 = r_1 + \frac{1}{10}, s_2 = r_2 + \frac{1}{10^2}, \ldots s_n = r_n + \frac{1}{10^n}, \ldots$$

Here s_n is obtained from r_n by increasing the last digit by 1. By construction, all members of this sequence belong to the complement r^c of the Dedekind cut $r \in \mathbb{R}$.

This infinite sequence is decreasing

$$r \leq \cdots \leq s_n \leq \cdots \leq s_3 \leq s_2 \leq s_1 \leq s_0.$$

Putting these two sequences together, we obtain

$$r_0 \leq r_1 \leq r_2 \leq r_3 \leq \cdots \leq r_n \leq \cdots \leq r \leq \cdots \leq s_n \leq \cdots \leq s_3 \leq s_2 \leq s_1 \leq s_0.$$

The crux is that we have

$$s_n - r_n = \frac{1}{10^n}, \quad n \in \mathbb{N}_0,$$

so that, by monotonicity, in general, we have the estimate

$$0 \leq s_n - r_m \leq \frac{1}{10^{\min(m,n)}}, \quad n, m \in \mathbb{N}_0.$$

Since $r_m \leq s_n$ for **all** $m, n \in \mathbb{N}_0$, we have $\sup_{m\in\mathbb{N}_0} r_m \leq \inf_{n\in\mathbb{N}_0} s_n$. We claim that equality holds. Otherwise, we let $\epsilon = \inf_{n\in\mathbb{N}_0} s_n - \sup_{m\in\mathbb{N}_0} r_m > 0$. Using Corollary 2 to the Bernoulli Inequality in the previous section, we can choose $k \in \mathbb{N}_0$ such that $1/10^k < \epsilon$. This contradicts to the estimate above for $k = \min(m, n)$. The claim follows.

We obtain

$$\sup_{m\in\mathbb{N}_0} r_m = r = \inf_{n\in\mathbb{N}_0} s_n.$$

As a byproduct, we see that, for $q \in \mathbb{Q}$, we have $q < r$ if and only if there exists $n \in \mathbb{N}_0$ such that $q \leq r_n$. Thus, the infinite sequence $(r_0, r_1, r_2, \ldots)$ recovers the Dedekind cut r uniquely. Since the sequences $(r_0, r_1, r_2, \ldots)$ and $(s_0, s_1, s_2, \ldots)$ mutually determine each other, the latter sequence also recovers r.

The entire sequence $(r_0, r_1, r_2, \ldots)$ can be compactly expressed as the **infinite decimal**

$$a.d_1d_2d_3\ldots,$$

where $a \in \mathbb{N}_0$ and $d_1, d_2, d_3, \ldots$ are the **decimal digits**, each ranging from 0 to 9. We now declare this to be our algebraic representation of the Dedekind cut r as a real number. Finally, recall that r was assumed to be **positive**; otherwise we can perform the analysis above for $-r$ and revert to the original r at the end.

Using powers of 10, the decimal representation of the real number r can be written as the **infinite sum**

$$r = a + \frac{d_1}{10} + \frac{d_2}{10^2} + \frac{d_3}{10^3} + \cdots + \frac{d_n}{10^n} + \cdots, \ 0 \le d_n \le 9, \ n = 0, 1, 2, 3, \ldots$$

This way, we can recover the sequence $r_0, r_1, r_2, \ldots$ as **partial sums** of the infinite sequence.

Example 2.2.2 For $r = 1$, we have $r_n = (10^n - 1)/10^n = 1 - 1/10^n$ and $s_n = 1, n \in \mathbb{N}_0$. In decimal representation, the first sequence is $(0, 0.9, 0.99, 0.999, 0.9999, \ldots)$, and the second is the constant sequence $(1, 1, 1, \ldots)$. They both determine the number 1.

We now take a short detour and discuss the ancient example of the irrational number $\sqrt{2}$ that arises in geometry.

In ancient times mathematicians defined $\sqrt{2}$ geometrically (and naïvely) as the side length of a square whose area is equal to 2. For a more explicit and geometrically equivalent interpretation, they also knew that $\sqrt{2}$ was also the diagonal of the unit square.

For a geometric proof of this equivalence due to the Babylonians (and simpler than using the Pythagorean Theorem) take a square of side length 2, and inscribe into this another (diamond shaped) square whose vertices are the midpoints of the sides. Since the entire square has side length of 2, its area is equal to 4. By cutting off the four corners, this square is reduced to half. It follows that the area of the (diamond shaped) middle square is 2, and therefore its side length must be $\sqrt{2}$. But each of the four sides is also the diagonal of one of the four unit squares that make up the entire square.

Arithmetically (and again naïvely), $\sqrt{2}$ can be defined as the **number** whose square is 2. This definition is naïve because the ancients did not define what kind of a number $\sqrt{2}$ was, let alone how to multiply it by itself.

History

A Babylonian clay tablet (c. 1800 – c. 1600 BCE) shows an approximation of $\sqrt{2}$ as $1; 24, 51, 10 = 1 + 24/60 + 51/60^2 + 10/60^3$ in sexagesimal arithmetic (which the Babylonians used) which in base 10 arithmetic corresponds to $30547/21600 = 1.41421\overline{296296}$. (See Figure 2.1.) This is correct up to 5 decimal places. In the figure this number is in the middle row. The side length of the square in the tablet is chosen to be the sexagesimal 30. This, multiplied by the approximation of $\sqrt{2}$ above gives $30 \cdot (1 + 24/60 + 51/60^2 + 10/60^3) = 42 + 25/60 + 35/60^2$. This latter number is in the bottom register given in sexagesimal digits as 42; 25, 35.

As shown in the Rhind Mathematical Papyrus, the ancient Egyptians extracted square roots by an inverse proportion method. In ancient India square root of two is first attested in the *Baudhayana Sulba Sutra* (c. 800 – c. 500 BCE) from the Vedic period as $\sqrt{2} = 1 + 1/3 + 1/(3 \cdot 4) - 1/(3 \cdot 4 \cdot 34) = 1.4142156$ correct up to 5 decimal places. The ancient Greeks who associated algebraic terms to geometric objects, such as length, perimeter, area, etc., and have thereby created **Geometric**

Fig. 2.1 Babylonian Clay Tablet showing an approximation of $\sqrt{2}$ in sexagesimal digits, Yale Babylonian Collection, YBC 7289.

Algebra, had no difficulty in **accepting** $\sqrt{2}$ as a number. The trouble or the shock (as some say) came when they tried to incorporate this number into what has been hitherto their number system, the set of rational numbers $\mathbb{Q}$. It is quite possible that the discovery of irrationality of $\sqrt{2}$ was made by one of the Pythagoreans. It is widely held but strongly disputed that it was Hippasus of Metapontum who was subsequently drowned at sea as a punishment from gods for revealing this secret. The handful of ancient texts that relate this story either do not mention Hippasus' name, or they say that the discovery revealed was something else (how to inscribe a dodecahedron into a sphere). Very little is known about Hippasus' life in general.

There is a simple but somewhat unusual proof of irrationality of $\sqrt{2}$ by playing **origami** as follows. (See Figure 2.2.) Assume

$$\sqrt{2} = \frac{a}{b},$$

where $a, b \in \mathbb{N}$. This means that $a^2 = 2b^2$ so that, by the Pythagorean Theorem, a square paper of side length b has diagonal length a. Fold a corner of the square along the angular bisector of a side and the adjacent diagonal. The right angle at the corner is folded to another right angle with one side being part of the diagonal. The adjacent right angle on this diagonal is the right angle in an isosceles triangle with side lengths $a - b$ and hypotenuse $b - (a - b) = 2b - a$.
Applying the Pythagorean Theorem again, we have

$$\sqrt{2} = \frac{2b - a}{a - b}.$$

Since $a > b$, we have $a > 2b - a > 0$ (and also $b > a - b$). This folding process now can be repeated for the square paper of side length $a - b$ and diagonal length

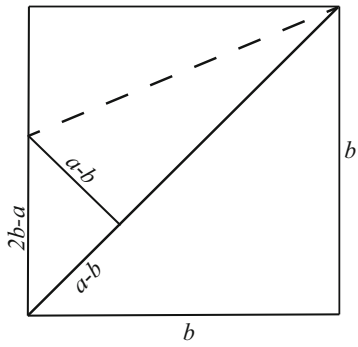


Fig. 2.2 Irrationality of $\sqrt{2}$ by origami.

$2b - a$. Since the lengths are natural numbers and strictly decreasing, repeating this process indefinitely, we obtain a strictly decreasing sequence of infinitely many natural numbers. This contradicts to the fact that $\mathbb{N}$ is well-ordered. Thus $\sqrt{2}$ is irrational.

There is a simple arithmetic process, called the **shifting square root algorithm**, that constructs the **infinite decimal representation** of $\sqrt{2}$ digit-by-digit. The first 60 digits of the decimal representation of $\sqrt{2}$ are:

$$\sqrt{2}=1.41421356237309504880168872420969807856967187537694807317667 9\ldots$$

Remark The shifting square root algorithm, at least in principle, is akin to the long division of polynomials. It is very cumbersome, and will not be discussed here. On the other hand, there are several much more efficient computational methods, such as **Newton's Method** (which, in this case, reduces to the so-called **Babylonian Method**), that provide fast algorithms to find inductively an infinite sequence of rational numbers whose members approximate the square root of a natural number (in particular $\sqrt{2}$) to arbitrary precision. For example, depending on the computer and the algorithm that we use, we can calculate a large (but finite) number of decimals in the decimal representation of $\sqrt{2}$. (A record of 200,000,000,000 digits was achieved by Shigeru Kondo in 2006.) We will give a detailed account on the Babylonian Method in Section 5.4.

There is no repeating pattern in the decimal representation of $\sqrt{2}$ above as it is irrational. Due to the irregularity in the decimal representation, beyond the inductive algorithms noted above, there is no known explicit formula that gives all the decimal digits of $\sqrt{2}$ instantaneously. Note, however, that, in view of Peano's Principle of Induction, an inductive algorithm is all that we need for the existence of $\sqrt{2}$ as a real number.

We finish this section by returning to **cardinality**, and show what we claimed at the end of Section 0.4 without proof: The set of real numbers $\mathbb{R}$ has the same cardinality as the power set $\mathcal{P}(\mathbb{N})$; that is, we have $|\mathbb{R}| = |\mathcal{P}(\mathbb{N})|$.

By the Cantor-Schröder-Bernstein Theorem (Section 0.4) it is enough to construct injective maps $\mathbb{R} \to \mathcal{P}(\mathbb{N})$ and $\mathcal{P}(\mathbb{N}) \to \mathbb{R}$.

To construct the first injective map, note that the representation of real numbers as **Dedekind cuts** (wich are subsets of $\mathbb{Q}$) automatically gives an injective map $\mathbb{R} \to \mathcal{P}(\mathbb{Q})$. Moreover, we have $|\mathbb{Q}| = |\mathbb{N}|$, and hence $|\mathcal{P}(\mathbb{Q})| = |\mathcal{P}(\mathbb{N})|$. Composing this injective map with a bijection $\mathcal{P}(\mathbb{Q}) \to \mathcal{P}(\mathbb{N})$ gives the desired first injective map $\mathbb{R} \to \mathcal{P}(\mathbb{N})$.

To construct the second injective map, we first note that, according to our discussion in Section 0.3, there is a natural bijection between the power set $\mathcal{P}(\mathbb{N})$ and the **set of all indicator functions** $\chi : \mathbb{N} \to \{0, 2\}$ (where we moved up the range value 1 to 2 for technical convenience). To an indicator function $\chi : \mathbb{N} \to \{0, 2\}$ on $\mathbb{N}$ we associate the unique real number in the interval $[0, 1]$ in **ternary** (base 3) expansion $\sum_{n=1}^{\infty} \chi(n)/3^n$. (The missing 1 in the range of the indicator function, and base 3 are chosen to avoid non-uniqueness with expansions terminating in an infinite string of 2's.) This association clearly gives rise to an injective map $\mathcal{P}(\mathbb{N}) \to \mathbb{R}$. Our claim now follows.

Exercises

2.2.1. Find the rational number as a fraction of two integers from the given decimal representations:

$$(a)\ 0.27272727\ldots = 0.\overline{27}$$

$$(b)\ 879.561561561561\ldots = 879.\overline{561}$$

$$(c)\ 923.51510832832832832\ldots = 923.51510\overline{832}.$$

2.2.2. Calculate $\sqrt{0.000244140625}$.

2.2.3. For what exponent $n \in \mathbb{N}$ do we have $1.001^n > 50$?

2.3 Real Numbers via Cauchy Sequences

The sequences of rational numbers $(r_0, r_1, r_2, r_3, \ldots)$ and $(s_0, s_1, s_2, s_3, \ldots)$ that define the Dedekind cut $r \in \mathbb{R}$ through a common infinite decimal introduced in the previous section are sequences with special properties. In this section we define and study their common generalization, the concept of Cauchy sequence. We start with a bit more general setting than necessary and introduce some terminology and notation to be used in the sequel.

Let A be a set. A **sequence** (of elements) in A is a map $a : \mathbb{N}_0 \to A$. Letting $a_n = a(n)$, $n \in \mathbb{N}_0$, the entire sequence a can be depicted by listing the values in sequential order

$$a = (a_0, a_1, a_2, a_2, \ldots) = (a_n)_{n\in\mathbb{N}_0} = (a_n)_{n=0}^{\infty},$$

where the last two are customary notations. Note that $\mathbb{N}_0$ can be replaced by $\mathbb{N}$, or by any **countable** set. (See also Example 0.3.4.)

Our main interest in this section will be **real sequences** $a : \mathbb{N}_0 \to \mathbb{R}$, sequences of real numbers (with $A = \mathbb{R}$). If the range of a real sequence a is contained in the set of rational numbers $\mathbb{Q}$ then we say that $a : \mathbb{N}_0 \to \mathbb{Q}$ is a **rational sequence**.

A real sequence a is **bounded above** (resp. **bounded below**) if the range of a (in $\mathbb{R}$) is bounded above (resp. bounded below). The sequence a is called **bounded**, if it is bounded above and below, or equivalently, if the range of a is contained in a finite interval, $[-c, c]$, $c > 0$, say; or equivalently, $a : \mathbb{N}_0 \to [-c, c] \subset \mathbb{R}$, that is $|a_n| \le c$ for all $n \in \mathbb{N}_0$.

Note that the **sum** and **product** of real sequences are defined using the addition and multiplication in the range $\mathbb{R}$. More specifically, if $a, b : \mathbb{N}_0 \to \mathbb{R}$ are real sequences then we define the **sum** $a + b : \mathbb{N}_0 \to \mathbb{R}$, resp. **product** $a \cdot b : \mathbb{N}_0 \to \mathbb{R}$, by $(a+b)_n = a_n + b_n$, resp. $(a \cdot b)_n = a_n b_n$, $n \in \mathbb{N}_0$. Note that the sum and product of rational sequences are rational.

For $c \in \mathbb{R}$, the **constant sequence** $c : \mathbb{N}_0 \to \mathbb{R}$ is the sequence whose elements are all equal to c; that is, $c_n = c$ for all $n \in \mathbb{N}_0$. By the above, the product of a constant sequence c and a real sequence a is ca, the constant multiple of a by c. In particular, the negative of a is defined by $-a = (-1)a$.

An interesting simple example of sequences with repeating pattern is the following:

Example 2.3.1 [16] Let $(a_n)_{n\in\mathbb{N}_0}$ be a sequence of positive real numbers such that any non-initial member is the product of its two neighbors. Show that the sequence is repeating with period six.

For $n \in \mathbb{N}$, we calculate

$$a_{n+3} = \frac{a_{n+2}}{a_{n+1}} = \frac{a_{n+1}/a_n}{a_{n+1}} = \frac{1}{a_n}$$

and hence

$$a_{n+6} = \frac{1}{a_{n+3}} = \frac{1}{1/a_n} = a_n.$$

Periodicity with period six follows.

The principal definition of this section is the following:

A real sequence $a : \mathbb{N}_0 \to \mathbb{R}$ is called a **Cauchy sequence** if

$$\inf_{N\in\mathbb{N}_0} \sup_{m,n\ge N} |a_n - a_m| = 0.$$

[16] This is a well-known problem in mathematical contest preparation. A special numerical case of this was a problem in the American Mathematics Competitions, 2006.

This definition holds in a model of real number system with the **Least Upper Bound Property** (such as our Dedekind complete $\mathbb{R}$) since we used the concepts of infimum and supremum. In this definition of a (rational) Cauchy sequence the infimum being zero means that for any $0 < \epsilon (\in \mathbb{Q})$ there exists $N \in \mathbb{N}$ such that $\sup_{m,n\geq N} |a_n - a_m| < \epsilon$ (that is, no positive ϵ can be a lower bound). Equivalently:

For any $0 < \epsilon (\in \mathbb{Q})$, there exists $N \in \mathbb{N}_0$ such that $|a_n - a_m| < \epsilon$ for all $n, m \geq N$.

This is the customary definition of a (rational) Cauchy sequence. Albeit less compact, this equivalent formulation does not need the Least Upper Bound Property, the existence of suprema and infima, and is thereby sometimes preferable.

First, we show that Cauchy sequences must be **bounded** (without the use of the Least Upper Bound Property).

Indeed, for $\epsilon = 1$, there exists $N \in \mathbb{N}_0$ such that $|a_n - a_m| < 1$ for all $m, n \geq N$. Thus, by the triangle inequality, we have

$$|a_n| - |a_m| \leq ||a_n| - |a_m|| \leq |a_n - a_m| < 1, \quad m, n \geq N.$$

Setting $m = N$, this gives

$$|a_n| \leq 1 + |a_N|, \quad n \geq N.$$

Joining the first N terms of the sequence, we obtain

$$|a_n| \leq \max(|a_0|, |a_1|, \ldots, |a_{N-1}|, 1 + |a_N|), \quad n \in \mathbb{N}_0.$$

Since the right-hand side of this inequality is a fixed number c (independent of $n \in \mathbb{N}_0$), the entire Cauchy sequence is contained in the interval $[-c, c]$. Boundedness follows.

Notice that if $a : \mathbb{N}_0 \to \mathbb{Q}$ is a rational Cauchy sequence then the upper bound c above is also rational.

Remark By boundedness and the triangle inequality again, the suprema in the definition of Cauchy sequence are all attained. Indeed, for all $N \in \mathbb{N}_0$, we have

$$\sup_{n,m\geq N} |a_n - a_m| \leq \sup_{n,m\geq N} (|a_n| + |a_m|) \leq 2c.$$

Second, we observe the obvious fact that the suprema $\sup_{m,n\geq N} |a_n - a_m|$ are **decreasing** with respect to $N \in \mathbb{N}_0$, that is, we have

$$\sup_{m,n\geq M} |a_n - a_m| \geq \sup_{m,n\geq N} |a_n - a_m|, \quad M < N, \quad M, N \in \mathbb{N}_0.$$

In particular, for **any** $M \in \mathbb{N}_0$, we have

$$\inf_{N\geq M} \sup_{m,n\geq N} |a_n - a_m| = \inf_{N\in\mathbb{N}_0} \sup_{m,n\geq N} |a_n - a_m|.$$

As a byproduct, we see that a Cauchy sequence stays Cauchy if finitely many members are altered or deleted.

The defining condition for a Cauchy sequence cannot be replaced by the condition $\inf_{N\in\mathbb{N}_0} \sup_{n\geq N} |a_{n+1} - a_n| = 0$. In other words it is not enough to require that the **consecutive** members of the sequence get progressively small. The following example shows this.

Example 2.3.2 Let $a : \mathbb{N}_0 \to \mathbb{R}$ be the real sequence defined by $a_n = \sqrt{n}$, $n \in \mathbb{N}_0$. We calculate

$$|a_{n+1} - a_n| = \sqrt{n+1} - \sqrt{n} = \frac{(\sqrt{n+1} - \sqrt{n})(\sqrt{n+1} + \sqrt{n})}{\sqrt{n+1} + \sqrt{n}}$$
$$= \frac{(n+1) - n}{\sqrt{n+1} + \sqrt{n}} = \frac{1}{\sqrt{n+1} + \sqrt{n}} < \frac{2}{\sqrt{n}},$$

where in the last inequality we need to restrict to $n \in \mathbb{N}$. With this, we have

$$0 \leq \inf_{N\in\mathbb{N}_0} \sup_{n\geq N} |a_{n+1} - a_n| \leq \inf_{N\in\mathbb{N}} \sup_{n\geq N} (2/\sqrt{n}) = 2 \inf_{N\in\mathbb{N}} (1/\sqrt{N}) = 0,$$

where, in the last equality, we used the Archimedean Property. (If, for some $0 < \epsilon$, we had $1/\sqrt{N} \geq \epsilon$ for all $N \in \mathbb{N}$, then we would have $N \leq 1/\epsilon^2$ for all $N \in \mathbb{N}$, a contradiction.)

On the other hand, this sequence a cannot be Cauchy since it is not bounded. This is yet another application of the Archimedean Property.

We now return to infinite decimals discussed in the previous section, and make the crucial observation that the sequence of partial sums $(r_n)_{n\in\mathbb{N}_0}$ of an infinite decimal r is a **rational Cauchy sequence**. Using the notations there, this follows as

$$0 \leq \inf_{N\in\mathbb{N}_0} \sup_{m,n\geq N} |r_n - r_m| \leq \inf_{N\in\mathbb{N}_0} \sup_{m,n\geq N} \frac{1}{10^{\min(m,n)}} = \inf_{N\in\mathbb{N}_0} \frac{1}{10^N} = 0,$$

where, in the last equality, we used the second corollary to the Bernoulli inequality for $a = 10$ (Section 2.1).

In the previous section we also saw that, for the sequence of partial sums $(r_n)_{n\in\mathbb{N}_0}$ constructed from a Dedekind cut $r \in \mathbb{R}$, we have $\sup_{n\in\mathbb{N}_0} r_n = r$. We now generalize this to (Cauchy) sequences by introducing the concept of **limit**.

Let $a : \mathbb{N}_0 \to \mathbb{R}$ be a **bounded** real sequence. The **limit inferior**, resp. **limit superior**, of the sequence a are defined as

$$\liminf_{n\to\infty} a_n = \sup_{N\in\mathbb{N}_0} \inf_{n\geq N} a_n, \quad \text{resp.} \quad \limsup_{n\to\infty} a_n = \inf_{N\in\mathbb{N}_0} \sup_{n\geq N} a_n.$$

For any $M, N \in \mathbb{N}_0$ with $K = \max(M, N)$, we have

$$\inf_{n \geq M} a_n \leq \inf_{n \geq K} a_n \leq \sup_{n \geq K} a_n \leq \sup_{n \geq N} a_n.$$

Taking the supremum for $M \in \mathbb{N}_0$ (resp. infimum for $N \in \mathbb{N}_0$) of the left-hand side (resp. right-hand side), we obtain

$$\liminf_{n \to \infty} a_n \leq \limsup_{n \to \infty} a_n.$$

If **equality** holds with common value L then we say that the real sequence a **converges** to the **limit** L, and we write

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

It follows directly from the definitions that, for $0 \neq c \in \mathbb{R}$, we have

$$\begin{aligned}
\limsup_{n \to \infty} (ca_n) &= c \limsup_{n \to \infty} a_n, \quad c > 0; \\
\liminf_{n \to \infty} (ca_n) &= c \liminf_{n \to \infty} a_n, \quad c > 0; \\
\limsup_{n \to \infty} (ca_n) &= c \liminf_{n \to \infty} a_n, \quad c < 0,
\end{aligned}$$

and therefore

$$\lim_{n \to \infty} (ca_n) = c \lim_{n \to \infty} a_n, \quad c \in \mathbb{R},$$

provided that the limits exist.[17]

Another direct consequence is **monotonicity** of the limit superior and limit inferior, and therefore also the limit:

If $a, b : \mathbb{N}_0 \to \mathbb{R}$ are real sequences such that $a_n \leq b_n$ for all $n \in \mathbb{N}_0$, then we have

$$\liminf_{n \to \infty} a_n \leq \liminf_{n \to \infty} b_n \quad \text{and} \quad \limsup_{n \to \infty} a_n \leq \limsup_{n \to \infty} b_n,$$

and therefore

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

provided that the limits exist.

It is customary to extend the definition of limit superior and limit inferior to unbounded real sequences. If a real sequence $a : \mathbb{N}_0 \to \mathbb{R}$ is **not** bounded above then we set $\limsup_{n \to \infty} a_n = \infty$. If $a : \mathbb{N}_0 \to \mathbb{R}$ is **not** bounded below then we set

[17] The existence of the limit on one side implies the existence of the other.

$\liminf_{n\to\infty} a_n = -\infty$. For consistency, we adjoin $\pm\infty$ to $\mathbb{R}$ to form the extended real number system $\overline{\mathbb{R}} = \mathbb{R} \cup \{\pm\infty\}$.

Example 2.3.3 Let $b, c \in \mathbb{R}$, and define the real sequence $a : \mathbb{N}_0 \to \mathbb{R}$ by

$$a_n = \frac{b+c}{2} + (-1)^n \frac{b-c}{2}, \quad n \in \mathbb{N}_0.$$

This sequence is alternating between the two values b and c; that is, we have $a = (b, c, b, c, \ldots)$. We obtain $\limsup_{n\to\infty} a_n = \max(b, c)$ and $\liminf_{n\to\infty} a_n = \min(b, c)$. The sequence is converges if and only if $b = c$ (to this common value).

For the next example, recall from Section 0.4 that the **factorial** of a natural number $n \in \mathbb{N}$, denoted by $n!$, is the product of all natural numbers less than equal to n. The inductive definition of the factorial is as follows: $1! = 1$ and $(n+1)! = (n+1) \cdot n!$, $n \in \mathbb{N}$. We also set $0! = 1$ and this defines the factorial of all non-negative integers.

Example 2.3.4 Let p_n denote the n**th prime number** (Section 1.3). We claim

$$\limsup_{n\to\infty} (p_{n+1} - p_n) = \infty.$$

Indeed, this follows directly from the fact that, for any $2 \leq k \in \mathbb{N}$, the $k-1$ **consecutive** natural numbers $k! + 2, k! + 3, \ldots, k! + k$ are all composite numbers (by the definition of the factorial).

Example 2.3.5 A pair $(p, p') \in \mathbb{N} \times \mathbb{N}$ consisting of two prime numbers p, p', $p < p'$, is called a **twin prime** if $p' - p = 2$. For example

$$(3, 5), (5, 7), (11, 13), (17, 19), (29, 31), (41, 43), (59, 61), (71, 73),$$
$$(101, 103), (107, 109), (137, 139), \ldots$$
$$(2996863034895 \cdot 2^{1290000} - 1,\ 2996863034895 \cdot 2^{1290000} + 1), \ldots$$

are twin primes (where the last twin prime in the list was discovered in September, 2016). The twin primes become increasingly rare.

The yet unsolved **twin prime conjecture** states that there are infinitely many twin primes. Using the limit inferior, the twin prime conjecture can be stated as

$$\liminf_{n\to\infty} (p_{n+1} - p_n) = 2.$$

A deep result of number theory asserts[18] that

$$\liminf_{n\to\infty} (p_{n+1} - p_n) \leq 246.$$

[18] For the original article, see Yitang Zhang, *Bounded gaps between primes*, Annals of Mathematics, 179 (3) (2014), 1121–1174. For an introduction, see Lin, T., *After prime proof, an unlikely star rises*, Quanta Magazine, April 2 (2015).

Remark How many "triplet primes" are there? To be precise, a triplet $(p, p', p'') \in \mathbb{N} \times \mathbb{N} \times \mathbb{N}$ consisting of three prime numbers $p, p', p'', p < p' < p''$, is called a **triplet prime** if $p' - p = p'' - p' = 2$.

Clearly $(3, 5, 7)$ is a triplet prime. We claim that there are no more triplet primes. Indeed, if (p, p', p'') is a triplet prime other than $(3, 5, 7)$ then $p = 3n + 1$ or $p = 3n + 2$ for some $n \in \mathbb{N}$. In the first case $p' = 3n + 3 = 3(n + 1)$, and in the second $p'' = 3n + 6 = 3(n + 2)$, both composite numbers.

Our definition of convergence has the obvious advantage that we do not need to know **a priori** the value of the limit L of a convergent sequence $(a_n)_{n\in\mathbb{N}_0}$; we simply need to calculate the limit inferior and the limit superior (which may not be finite) and compare. Nevertheless, there is an equivalent formulation of convergence which, albeit involves the value of the limit explicitly, is useful in many instances in calculations.

We state this as a follows:

Proposition 2.3.1 *A real sequence $(a_n)_{n\in\mathbb{N}_0}$ is convergent to $L \in \mathbb{R}$ if and only if we have*

$$\inf_{N\in\mathbb{N}_0} \sup_{n\geq N} |a_n - L| = 0.$$

Proof Denote $\underline{L} = \liminf_{n\to\infty} a_n$ and $\overline{L} = \limsup_{n\to\infty} a_n$. We have $\underline{L} \leq \overline{L}$ with equality if and only if $(a_n)_{n\in\mathbb{N}_0}$ is convergent to the common value. Consider first

$$\underline{L} = \liminf_{n\to\infty} a_n = \sup_{N\in\mathbb{N}_0} \inf_{n\geq N} a_n.$$

By definition, for any $\epsilon > 0$, the real number $\underline{L} - \epsilon$ cannot be an upper bound for all the infima on the right-hand side, so that there exists $M \in \mathbb{N}_0$ such that $\inf_{n\geq M} a_n > \underline{L} - \epsilon$. Similarly, for the limit superior, for the given $\epsilon > 0$, there exists $N \in \mathbb{N}_0$ such that $\sup_{n\geq N} a_n < \overline{L} + \epsilon$. Setting $K = \max(M, N)$, we combine these as

$$\underline{L} - \epsilon < \inf_{n\geq M} a_n \leq \inf_{n\geq K} a_n \leq \sup_{n\geq K} a_n \leq \sup_{n\geq N} a_n < \overline{L} + \epsilon.$$

Assume now that the limit exists: $\underline{L} = \overline{L} = L$. Then, by the above, for every $\epsilon > 0$, there exists $K \in \mathbb{N}_0$ such that

$$L - \epsilon < \inf_{n\geq K} a_n \leq \sup_{n\geq K} a_n < L + \epsilon,$$

or equivalently, we have $L - \epsilon < a_n < L + \epsilon, n \geq K$. We rewrite these inequalities as $\sup_{n\geq K} |a_n - L| < \epsilon$. Since $\epsilon > 0$ is arbitrary, this gives

$$\inf_{K\in\mathbb{N}_0} \sup_{n\geq K} |a_n - L| = 0.$$

The converse follows by retracing the steps above.

The condition of convergence in Proposition 2.3.1 is a compact reformulation of the customary definition of convergence; namely, $\lim_{n\to\infty} a_n = L$ if:

For every $0 < \epsilon$ there exists $N \in \mathbb{N}_0$ such that $|a_n - L| < \epsilon$ for all $n \geq N$.

Note that this definition does not use the Least Upper Bound Property, the existence of suprema and infima. Notice also that this definition can be restricted to rational sequences verbatim with $0 < \epsilon \in \mathbb{Q}$ and $L \in \mathbb{Q}$.

In our study a primary role will be played by **null-sequences**, (real or rational) sequences that converge to zero. For now, we only need the following simple facts: (1) The sum of two null-sequences is a null-sequence; and (2) The product of a null-sequence and a bounded sequence is a null-sequence.

Indeed, let $u, v : \mathbb{N}_0 \to \mathbb{R}$ be null-sequences and $a : \mathbb{N}_0 \to [-c, c] \subset \mathbb{R}$ a bounded sequence (with bound $c > 0$). Given $0 < \epsilon$, choose $M, N \in \mathbb{N}_0$ such that $|u_n| < \epsilon/2$ for $n \geq M$, and $|v_n| < \epsilon/2$ for $n \geq N$. Then, by the triangle inequality, for $n \geq \max(M, N)$, we have $|u_n + v_n| \leq |u_n| + |v_n| < \epsilon/2 + \epsilon/2 = \epsilon$, and the first statement follows. For the second statement, given $0 < \epsilon$, choose $N \in \mathbb{N}_0$ such that $|u_n| < \epsilon/c$ for $n \geq N$. Then, for $n \geq N$ again, we have $|a_n u_n| \leq c \cdot \epsilon/c = \epsilon$, and the second statement also follows.

Finally, for a real sequence $a : \mathbb{N}_0 \to \mathbb{R}$, we define the **absolute value** $|a| : \mathbb{N}_0 \to \mathbb{R}$ by $|a|_n = |a_n|$, $n \in \mathbb{N}_0$. As a consequence of the triangle inequality, the absolute value of a Cauchy sequence is a Cauchy sequence. Moreover, we have the obvious fact that a real sequence u is a null-sequence if and only if $|u|$ is a null-sequence.

We now discuss the special case of monotonic sequences. A real sequence $a : \mathbb{N}_0 \to \mathbb{R}$ is called **increasing** (resp. **decreasing**) if $m < n$, $m, n \in \mathbb{N}_0$, implies $a_m \leq a_n$ (resp. $a_m \geq a_n$). The sequence a is called **monotonic** if it is increasing or decreasing. Replacing the inequality signs by strict inequalities, we obtain the concepts of **strictly increasing** and **strictly decreasing** sequences.

Next, we discuss two classical monotonic sequences.

A real sequence $a : \mathbb{N}_0 \to \mathbb{R}$ is called **arithmetic** if there exists $d \in \mathbb{R}$ such that $a_{n+1} = a_n + d$ for all $n \in \mathbb{N}_0$. The real number d is called the **difference** of the arithmetic sequence. By induction, the general term of an arithmetic sequence is $a_n = a_0 + nd$, $n \in \mathbb{N}_0$.

Example 2.3.6 [19] Let $a : \mathbb{N} \to \mathbb{R}$ be an arithmetic sequence with difference 1. (Note the change in the index.) Given $n \in \mathbb{N}$, if $a_1 + a_2 + a_3 + \cdots + a_{2n} = A$ find $a_2 + a_4 + a_6 + \cdots + a_{2n}$ in terms of A.

We have $a_{2n-1} = a_{2n} - 1$, $n \in \mathbb{N}$. Using this, we have

$$\begin{aligned} A &= a_1 + a_2 + a_3 + a_4 + \cdots + a_{2n-1} + a_{2n} \\ &= (a_2 - 1) + a_2 + (a_4 - 1) + a_4 + \cdots + (a_{2n} - 1) + a_{2n} \\ &= 2(a_2 + a_4 + \cdots + a_{2n}) - n. \end{aligned}$$

This gives $a_2 + a_4 + \cdots + a_{2n} = (A + n)/2$.

[19] A special case of this problem was in the American Invitational Mathematics Examination, 1984.

Example 2.3.7 [20] Let $a : \mathbb{N} \to \mathbb{R}$ be an arithmetic sequence with difference d. Given $n \in \mathbb{N}$, if $a_1+a_2+a_3+\cdots+a_n = A$ and $a_{n+1}+a_{n+2}+a_{n+3}+\cdots+a_{2n} = B$, find d in terms of A and B.

Taking the difference of the two equations, after grouping, we find

$$(a_{n+1} - a_1) + (a_{n+2} - a_2) + (a_{n+3} - a_3) + \cdots + (a_{2n} - a_n) = B - A.$$

Now, notice that each difference in the parentheses on the left-hand side is equal to nd. We obtain $n^2 d = B - A$, and hence $d = (B - A)/n^2$.

In Section 2.1 we showed that, for $n \in \mathbb{N}$, the square root $\sqrt{n}$ is a rational number if and only if n is a perfect square. We use this in the following:

Example 2.3.8 Let $n_1, n_2, n_3 \in \mathbb{N}$ distinct, and asssume that the linear relation

$$c_1\sqrt{n_1} + c_2\sqrt{n_2} + c_3\sqrt{n_3} = 0$$

holds for some non-zero **rational** coefficients $0 \neq c_1, c_2, c_3 \in \mathbb{Q}$. Then the products n_1n_2, n_2n_3, and n_3n_1 must be perfect squares.

The equality above holds if $\sqrt{n_1}$, $\sqrt{n_2}$, $\sqrt{n_3}$ are members of an arithmetic sequence; and thereby the same conclusion holds. In particular, the square roots of three distinct primes cannot participate in an arithmetic sequence.

By symmetry, it is enough to show that n_1n_2, say, is a perfect square. Rearranging and squaring, we get

$$c_1^2 n_1 + c_2^2 n_2 + 2c_1c_2\sqrt{n_1n_2} = c_3^2 n_3.$$

This gives

$$\sqrt{n_1n_2} = \frac{c_3^2 n_3 - c_1^2 n_1 - c_2^2 n_2}{2c_1c_2} \in \mathbb{Q},$$

a rational number. By the above, n_1n_2 must be a perfect square. The first statement follows.

To show the second statement, assume that $\sqrt{n_1}$, $\sqrt{n_2}$, $\sqrt{n_3}$ participate in an arithmetic sequence with difference $d \in \mathbb{R}$. Then we have

$$\sqrt{n_1} = \sqrt{n_3} + a_1 d \quad \text{and} \quad \sqrt{n_2} = \sqrt{n_3} + a_2 d, \;\; a_1 \neq a_2, \; 0 \neq a_1, a_2 \in \mathbb{Z}.$$

Eliminating d, we obtain the linear relation

$$a_2\sqrt{n_1} - a_1\sqrt{n_2} + (a_1 - a_2)\sqrt{n_3} = 0$$

with non-zero with integer cofficients. The second statement follows.

[20] A special numerical case of this problem was in the American Mathematics Competition, 2002.

A real sequence $a : \mathbb{N}_0 \to \mathbb{R}$ is called **geometric** if there exists $r \in \mathbb{R}$ such that $a_{n+1} = r \cdot a_n$ for all $n \in \mathbb{N}_0$. The real number r is called the **quotient** or **ratio** of the geometric sequence. By induction, the general term of a geometric sequence is $a_n = a_0 \cdot r^n$, $n \in \mathbb{N}_0$.

For the question of convergence we can discard the initial term and set $a_0 = 1$. For $r \in \mathbb{R}$, we thus consider the geometric sequence $(r^n)_{n\in\mathbb{N}}$.

We first let $r \geq 0$. Since $r^{n+1} - r^n = r^n(r-1)$, $n \in \mathbb{N}$, the sequence is decreasing for $0 \leq r < 1$, and increasing for $r > 1$ (and constant 1 for $r = 1$). By the two corollaries of the Bernoulli inequality in Section 2.1, we have

$$\lim_{n\to\infty} r^n = \begin{cases} 0, & \text{if} \quad 0 \leq r < 1 \\ 1, & \text{if} \quad r = 1 \\ \infty, & \text{if} \quad r > 1. \end{cases}$$

For $r < 0$, we have

$$r^n = (-|r|)^n = (-1)^n |r|^n, \quad n \in \mathbb{N}.$$

Splitting the sequence into two subsequences according to the parity of $n \in \mathbb{N}$ (even-odd), we obtain

$$\liminf_{n\to\infty} r^n = -\lim_{n\to\infty} |r|^n = \begin{cases} 0, & \text{if} \quad -1 < r < 0 \\ -1, & \text{if} \quad r = -1 \\ -\infty, & \text{if} \quad r < -1. \end{cases}$$

and

$$\limsup_{n\to\infty} r^n = \lim_{n\to\infty} |r|^n = \begin{cases} 0, & \text{if} \quad -1 < r < 0 \\ 1, & \text{if} \quad r = -1 \\ \infty, & \text{if} \quad r < -1. \end{cases}$$

We conclude that, the sequence is not convergent for $r \leq -1$. Putting together the remaining case ($-1 < r < 0$) with the case of positive quotient ($0 \leq r < 1$), we obtain

$$\lim_{n\to\infty} r^n = 0, \quad |r| < 1.$$

Example 2.3.9 [21] In an increasing sequence of four positive integers, the first three terms form an arithmetic sequence with difference d, the last three terms form a geometric sequence, and the first and fourth terms differ by Δ. Show that $\Delta/4 < d < \Delta/3$.

[21] This example is inspired by a problem in the American Invitational Mathematics Examination, 2003.

According to the conditions, the four positive integers are

$$a,\ a+d,\ a+2d,\ \frac{(a+2d)^2}{a+d},\quad a,d\in\mathbb{N}.$$

We have

$$\frac{(a+2d)^2}{a+d}-a=\Delta.$$

Eliminating the denominator, expanding, and simplifying, this last condition gives $3ad+4d^2=\Delta a+\Delta d$, or equivalently

$$d(4d-\Delta)=a(\Delta-3d).$$

This shows that $4d-\Delta$ and $\Delta-3d$ must have the same sign. Clearly, the negative sign is not realized. Therefore, we have $4d-\Delta>0$ and $\Delta-3d>0$. These give $\Delta/4<d<\Delta/3$.

Example 2.3.10 Find a positive integer $M\in\mathbb{N}$ such that the sum of the arithmetic sequence $12, 14, 16, \ldots, M$ is a perfect square.

The general element of the sequence is $a_k=12+2(k-1)=2k+10$, $k\in\mathbb{N}$ (since the difference $d=2$). The sum of the first $n\in\mathbb{N}$ elements is $\sum_{k=1}^{n}(2k+10)=n(n+1)+10n=n^2+11n$, where we used $\sum_{k=1}^{n}k=n(n+1)/2$ (Section 0.4). For this to be a perfect square, we need $n^2+11n=m^2$ to hold for some $m\in\mathbb{N}$. Since this does not factor well among the integers, we use the standard trick[22] to multiply through by 4. We obtain $4n^2+44n=4m^2$, and hence $(2n+11)^2=4m^2+121$. Equivalently, we have $(2n+11)^2-(2m)^2=(2n+2m+11)(2n-2m+11)=121$. Since $121=11^2$, the only way the last factorization could hold is $2n+2m+11=121$ and $2n-2m+11=1$. Solving, we obtain $n=25$, $m=30$, and hence $M=2n+10=60$.

The following important result is a consequence of the Least Upper Bound Property of the real number system $\mathbb{R}$:

Monotone Convergence Theorem *If $a:\mathbb{N}_0\to\mathbb{R}$ is an increasing (resp. decreasing) sequence which is bounded above (resp. below) then*

$$\lim_{n\to\infty}a_n=\sup_{n\in\mathbb{N}_0}a_n\quad\textit{resp.}\quad\lim_{n\to\infty}a_n=\inf_{n\in\mathbb{N}_0}a_n.$$

Proof It is enough to prove the first statement. Letting $\sup_{n\in\mathbb{N}_0}a_n=L$, since a is increasing, we have $\sup_{n\geq N}a_n=L$ for all $N\in\mathbb{N}_0$. Thus, for the limit superior, we obtain $\limsup_{n\to\infty}a_n=\inf_{N\in\mathbb{N}_0}L=L$. For the limit inferior, again since a is

[22] There are several mathematical contest problems that center around this trick, e.g. to solve $n^2+p\cdot n=m^2$, $m,n\in\mathbb{N}$ (and also for $\mathbb{Z}$), where $3\leq p\in\mathbb{N}$ is a given prime. The method above gives $n=((p-1)/2)^2$.

increasing, we have $\liminf_{n\to\infty} a_n = \sup_{N\in\mathbb{N}_0} \inf_{n\geq N} a_n = \sup_{N\in\mathbb{N}_0} a_N = L$. The theorem follows.

Remark As the proof above shows, the Monotone Convergence Property (the statement in the theorem above) is a special case of the Least Upper Bound Property. Actually the two properties can be shown to be **equivalent**. This means that, in an axiomatic development of the real number system, the Monotone Convergence Property can be used as an **axiom**, and the Least Upper Bound Property and thereby completeness of $\mathbb{R}$ follow from this.

As an immediate application, the sequence of partial sums $(r_n)_{n\in\mathbb{N}_0}$ of a Dedekind cut r (and also the sequence $(s_n)_{n\in\mathbb{N}_0}$) are convergent: $\lim_{n\to\infty} r_n = \lim_{n\to\infty} s_n = r$.

To what extent are monotonic sequences special among all real sequences? To answer this question we define the concept of **subsequence** of a real sequence. Let $a : \mathbb{N}_0 \to \mathbb{R}$ be a real sequence. A real sequence $b : \mathbb{N}_0 \to \mathbb{R}$ is called a subsequence of a if there exists a strictly increasing map $\iota : \mathbb{N}_0 \to \mathbb{N}_0$ such that $b = a \circ \iota$. Given $a = (a_0, a_1, a_2, a_3, \ldots) = (a_n)_{n\in\mathbb{N}_0} = (a_n)_{n=0}^{\infty}$, letting $n_k = \iota(k)$, $k \in \mathbb{N}_0$, we have $b_k = a_{n_k}$, $k \in \mathbb{N}_0$, and we obtain

$$b = (b_k)_{k\in\mathbb{N}_0} = (b_0, b_1, b_2, b_3, \ldots) = (a_{n_0}, a_{n_1}, a_{n_2}, a_{n_3}, \ldots) = (a_{n_k})_{k\in\mathbb{N}_0}.$$

We now state a simple but important property of real sequences:

Proposition 2.3.2 *Any real sequence has a monotonic subsequence.*

Proof We present here the classical proof. Let $a : \mathbb{N}_0 \to \mathbb{R}$ be a real sequence. We call an element a_m, $m \in \mathbb{N}_0$, a **peak** if, for all $m \leq n$, we have $a_m \geq a_n$.

If a has infinitely many peaks, $a_{n_0}, a_{n_1}, a_{n_2}, \cdots$, say, then, by definition, we have $a_{n_0} \geq a_{n_1} \geq a_{n_2} \geq \cdots$. Therefore, the sequence of peaks forms an infinite decreasing subsequence of a.

We may therefore assume that a has only finitely many (possibly no) peaks. Let $n_0 \in \mathbb{N}_0$ be such that a_n is **not a peak** for all $n \geq n_0$. Since a_{n_0} is not a peak, for some $n_1 > n_0$ we have $a_{n_0} < a_{n_1}$. Proceeding inductively, assume that we have $n_0 < n_1 < \cdots < n_k$ such that $a_{n_0} < a_{n_1} < a_{n_2} < \cdots < a_{n_k}$. Since a_{n_k} is not a peak, for some $n_{k+1} > n_k$ we have $a_{n_k} < a_{n_{k+1}}$. By Peano's Principle of Induction, the (strictly) increasing subsequence $(a_{n_k})_{k\in\mathbb{N}_0}$ has been defined. The proposition follows.

If $a : \mathbb{N}_0 \to \mathbb{R}$ is a bounded real sequence then, by the above, it has a monotonic subsequence. Being part of the original bounded sequence, it is necessarity bounded. By the Monotone Convergence Theorem, it then converges. We obtain the following:

Bolzano–Weierstrass Theorem *Any bounded real sequence subconverges; that is, it has a monotonic subsequence.*

Remark In axiomatic development of the real number system the Bolzano-Weierstrass Property stated above is equivalent to the Monotone Convergence Property, and thereby to the Least Upper Bound Property.

History
The Bolzano-Weierstrass Theorem was first proved by Bolzano in 1817 as a preparatory lemma to his proof of the Intermediate Value Theorem (to be treated here later). As noted previously, Bolzano's results were not known in mathematical circles; in fact, most were published posthumously in 1851. Around 1867, Karl Weierstrass (1815 – 1897), recognizing the significance of this result, proved this theorem again.

Remark If a real sequence is monotonic and subconverges then the sequence itself converges. Indeed, assume that $(a_n)_{n\in\mathbb{N}}$ is increasing, and has a convergent subsequence $(a_{n_k})_{k\in\mathbb{N}}$ with $\lim_{k\to\infty} a_{n_k} = L$. We need to show that $\sup_{n\in\mathbb{N}} a_n \leq L$. Assume not. Then there exists $N \in \mathbb{N}$ such that $a_n > L$ for $n \geq N$. By monotonicity, $a_{n_k} > L$ for $n_k \geq N$. Since this holds for infinitely many values of $k \in \mathbb{N}$, this is a contradiction.

We have now come to the main point of our discussion of Cauchy sequences in our model of the real number system $\mathbb{R}$ via Dedekind cuts. Completeness of $\mathbb{R}$ (the Least Upper Bound Property) implies the following:

Proposition 2.3.3 *A real sequence is Cauchy if and only if it is convergent.*

Proof First, assume that $a : \mathbb{N}_0 \to \mathbb{R}$ is convergent: $\lim_{n\to\infty} a_n = L$. By the triangle inequality, we have

$$|a_n - a_m| = |(a_n - L) - (a_m - L)| \leq |a_n - L| + |a_m - L|, \quad m, n \in \mathbb{N}_0.$$

Let $M, N \in \mathbb{N}_0$ with $K = \max(M, N)$. The inequality above gives

$$\sup_{m,n\geq K} |a_n - a_m| \leq \sup_{n\geq N} |a_n - L| + \sup_{m\geq M} |a_m - L|.$$

Taking the infimum on the left-hand side we obtain

$$0 \leq \inf_{K\in\mathbb{N}_0} \sup_{m,n\geq K} |a_n - a_m| \leq \sup_{n\geq N} |a_n - L| + \sup_{m\geq M} |a_m - L|.$$

The infimum on the left-hand side is now constant, independent of M and N, so that we can take the infima of the two terms on the right-hand side separately as

$$0 \leq \inf_{K\in\mathbb{N}_0} \sup_{m,n\geq K} |a_n - a_m| \leq \inf_{N\in\mathbb{N}_0} \sup_{n\geq N} |a_n - L| + \inf_{M\in\mathbb{N}_0} \sup_{m\geq M} |a_m - L|.$$

By Proposition 2.3.1 of this section, the two terms on the right-hand side vanish. Hence the left-hand side must also vanish. Thus, a is a Cauchy sequence.

To prove the converse statement, assume that $a : \mathbb{N}_0 \to \mathbb{R}$ is a Cauchy sequence. Since a is bounded, by the Bolzano-Weierstrass Theorem, it has a subsequence $(b_k)_{k\in\mathbb{N}_0} = (a_{n_k})_{k\in\mathbb{N}_0}$ convergent to a limit L, say; that is, we have $\lim_{k\to\infty} b_k = \lim_{k\to\infty} a_{n_k} = L$.

We now claim that L is also the limit of the original Cauchy sequence a.
To begin with, for $n, k \in \mathbb{N}_0$, the triangle inequality gives

$$|a_n - L| \leq |a_n - a_{n_k}| + |a_{n_k} - L|.$$

Let $M, N \in \mathbb{N}_0$ with $K = \max(M, N)$. The inequality above gives

$$0 \leq \inf_{K \in \mathbb{N}_0} \sup_{n \geq K} |a_n - L| \leq \sup_{n \geq K} |a_n - L| \leq \sup_{n, n_k \geq N} |a_n - a_{n_k}| + \sup_{n_k \geq M} |a_{n_k} - L|.$$

The infimum on the left-hand side does not depend on M and N, so that we can take the infima on M and N separately, and obtain

$$0 \leq \inf_{K \in \mathbb{N}_0} \sup_{n \geq K} |a_n - L| \leq \inf_{N \in \mathbb{N}_0} \sup_{n, n_k \geq N} |a_n - a_{n_k}| + \inf_{M \in \mathbb{N}_0} \sup_{n_k \geq M} |a_{n_k} - L|.$$

Since the sequence a is Cauchy, we have

$$\inf_{N \in \mathbb{N}_0} \sup_{n, n_k \geq N} |a_n - a_{n_k}| \leq \inf_{N \in \mathbb{N}_0} \sup_{m, n \geq N} |a_n - a_m| = 0.$$

Since the subsequence $(b_k)_{k \in \mathbb{N}_0} = (a_{n_k})_{k \in \mathbb{N}_0}$ converges to L, we also have

$$\inf_{M \in \mathbb{N}_0} \sup_{n_k \geq M} |a_{n_k} - L| = 0.$$

These give

$$\inf_{K \in \mathbb{N}_0} \sup_{n \geq K} |a_n - L| = 0.$$

Thus, $\lim_{n \to \infty} a_n = L$, and the proposition follows.

Remark Note that Cauchy Completeness (the property that every Cauchy sequence is convergent) is implied by but not equivalent to the Bolzano-Weierstrass Property, the Least Upper Bound Property, etc. The two properties become equivalent if we assume the Archimedean Property.

Our construction of the real number system was based on Dedekind cuts of the set of rational numbers $\mathbb{Q}$. With this $\mathbb{R}$ is Dedekind complete; that is, it satisfies the Least Upper Bound Property or any other equivalents, as noted above.

Another model of $\mathbb{R}$, due to Georg Cantor, is based on extending $\mathbb{Q}$ by adjoining "limits" interpreted as **rational Cauchy sequences**. With this $\mathbb{Q}$ will have a **Cauchy complete** extension $\mathbb{R}$, another model of the real number system, in which any real Cauchy sequence converges.

In what follows, we now give a detailed account on Cantor's construction.

First, we need to recall the customary definition of a Cauchy sequence (without the use of the Least Upper Bound Property):

A rational sequence $a : \mathbb{N}_0 \to \mathbb{Q}$ is a **Cauchy sequence** if, for every rational $0 < \epsilon \in \mathbb{Q}$, there exists $N \in \mathbb{N}_0$ such that $|a_n - a_m| < \epsilon$ for $m, n \geq N$.

Staying within $\mathbb{Q}$, note also that a rational sequence $a : \mathbb{N}_0 \to \mathbb{Q}$ is said to **converge** to a **rational** number $L \in \mathbb{Q}$ if, for every rational $0 < \epsilon \in \mathbb{Q}$, there exists $N \in \mathbb{N}_0$ such that $|a_n - L| < \epsilon$ for $n \geq N$. In particular, for $L = 0$, the concept of **rational null-sequence** is defined.

We let $\mathfrak{C}$ denote the set of all **rational Cauchy sequences**.

Since we have seen that the (monotonic) Cauchy sequence $(r_n)_{n\in\mathbb{N}_0}$ uniquely defines the Dedekind cut $r \in \mathbb{R}$, we would like to **define** a real number r as the rational Cauchy sequence $(r_n)_{n\in\mathbb{N}_0}$. An immediate problem in this approach is non-uniqueness; for example, the sequence $(s_n)_{n\in\mathbb{N}_0}$ (and many others) also "define" the same Dedekind cut $r \in \mathbb{R}$.

Example 2.3.11 The sequences $r, s : \mathbb{N}_0 \to \mathbb{R}$ defined by $r_n = (10^n - 1)/10^n = 1 - 1/10^n$ and $s_n = 1, n \in \mathbb{N}_0$ are rational Cauchy sequences. In decimal representation, the first sequence is $(0, 0.9, 0.99, 0.999, 0.9999, \ldots)$, and the second is the constant sequence $(1, 1, 1, 1, \ldots)$. They both converge to the number 1. In particular, $s - r$ is a null-sequence.

We therefore need to partition $\mathfrak{C}$ into classes of Cauchy sequences in which every class consists of Cauchy sequences that define the same "limit."

This is done by introducing a relation $\sim$ on $\mathfrak{C}$ as follows: For $a, a' \in \mathfrak{C}$, we let $a \sim a'$ if $a - a'$ is a (rational) null-sequence.

A simple consequence of the Cauchy property and the definition of the relation $\sim$ is the following: For $a \in \mathfrak{C}$, let $a' \in \mathfrak{C}$ be obtained from a by **altering** or **deleting finitely many** elements. Then, we have $a \sim a'$.

Indeed, if a' is obtained from a by altering finitely many elements then there exists $N \in \mathbb{N}_0$ such that $a_n = a'_n$ for $n \geq N$. Hence, $a_n - a'_n = 0$ for $n \geq N$; and $a - a'$ is obviously a null-sequence.

If a' is obtained from a by deleting finitely many elements then there exist $k, N \in \mathbb{N}_0$ such that $a'_n = a_{n+k}$, $n \geq N$. Since a is a Cauchy sequence, we have $\lim_{n\to\infty} |a_n - a'_n| = \lim_{n\to\infty} |a_n - a_{n+k}| = 0$. Thus, $a \sim a'$ follows.

We now claim that $\sim$ is an equivalence relation on $\mathfrak{C}$, and thereby partitions $\mathfrak{C}$ into the desired equivalence classes.

Reflexivity: For $a \in \mathfrak{C}$, $a - a = 0$ is the constant zero sequence, thereby a null-sequence. Symmetry: For $a, a' \in \mathfrak{C}$, if $a - a'$, is a null-sequence then so is $a' - a = -(a - a')$. Transitivity: For $a, a', a'' \in \mathfrak{C}$, if $a - a'$ and $a' - a''$ are null-sequences then so is their sum $a - a'' = (a - a') + (a' - a'')$.

The equivalence relation $\sim$ on $\mathfrak{C}$ partitions $\mathfrak{C}$ into mutually disjoint equivalence classes. An equivalence class is called a **real number**. The quotient $\mathbb{R} = \mathfrak{C}/\sim$ is Cantor's definition of the **set of real numbers**.

To keep the notation simple we will not introduce a specific notation for the equivalence classes, and we will state most of the properties of the quotient $\mathbb{R} = \mathfrak{C}/\sim$ in terms of representatives of the equivalence classes (making sure that the statements themselves are valid up to the equivalence relation $\sim$).

We now claim that the **addition** and the **multiplication** of sequences in $\mathfrak{C}$ are compatible with the equivalence relation, and thereby give rise to the operations of addition and multiplication in $\mathbb{R} = \mathfrak{C}/\sim$.

We need to show that, for $a, a', b, b' \in \mathfrak{C}$, the relations $a \sim a'$ and $b \sim b'$ imply $a + b \sim a' + b'$ and $a \cdot b \sim a' \cdot b'$.

Indeed, for addition, we have

$$(a + b) - (a' + b') = (a - a') + (b - b').$$

The right-hand side is the sum of two null-sequences, and therefore it is a null-sequence. Thus, $a + b \sim a' + b'$ follows.

For multiplication, we first write

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

Now recall that b and a' are bounded since they are Cauchy sequences. As the product of a bounded sequence and a null-sequence is a null-sequence, on the right-hand side we have the sum of two null-sequences; therefore the sum itself is also a null-sequence. Thus, $a \cdot b \sim a' \cdot b'$ follows.

We conclude that the addition and the multiplication are well-defined in $\mathbb{R} = \mathfrak{C}/\sim$.

Since addition and multiplication are both **associative** and **commutative** and they are connected through **distributivity** even on the level of rational Cauchy sequences, it follows that these rules hold in $\mathbb{R} = \mathfrak{C}/\sim$.

For $q \in \mathbb{Q}$, the **constant** rational sequence $q = (q, q, q, \ldots)$ is obviously a (rational) Cauchy sequence: $q \in \mathfrak{C}$. Moreover, if $q \neq q'$, $q, q' \in \mathbb{Q}$, the constant rational sequences q and q' are inequivalent: $q \not\sim q'$. Associating to a rational number the equivalence class of its constant sequence gives rise to an embedding of $\mathbb{Q}$ into $\mathbb{R}$. Clearly, this embedding respects the operations of addition and multiplication. From now on we identify $\mathbb{Q}$ with its range in $\mathbb{R}$, the field of rational numbers $\mathbb{Q}$.

By definition, the equivalence class of the constant zero sequence $0 \in \mathfrak{C}$ consists of all (rational) null-sequences, and it is the **additive identity**: For $a \in \mathfrak{C}$, we have $a + 0 = a$. For a rational Cauchy sequence $a \in \mathfrak{C}$, the **additive inverse** or **negative** of the equivalence class of a is given by the equivalence class of $-a = (-1)a \in \mathfrak{C}$: For $a \in \mathfrak{C}$, we have $a + (-a) = 0$.

Similarly, the equivalence class of the constant sequence $1 \in \mathfrak{C}$ is the **multiplicative identity**: For $a \in \mathfrak{C}$, we have $1 \cdot a = a$.

The multiplicative inverse (of non-zero real numbers) needs some elaboration. We first introduce a convenient terminology. We say that a rational Cauchy sequence $a \in \mathfrak{C}$ is **bounded away from zero** if:

There exists $0 < \epsilon \in \mathbb{Q}$ and $N \in \mathbb{N}_0$ such that $a_n > \epsilon$ for $n \geq N$.

We first claim that this property is additive in the sense that if $a, b \in \mathfrak{C}$ are both bounded away from zero then so is their sum $a + b \in \mathfrak{C}$.

Indeed, we have $0 < \delta, \epsilon \in \mathbb{Q}$ and $M, N \in \mathbb{N}_0$ such that $a_n > \delta$ for $n \geq M$, and $b_n > \epsilon$ for $n \geq N$. Then, for $n \geq \max(M, N)$, we have $a_n + b_n > \delta + \epsilon$. The claim follows.

Next, we note that the property of bounded away from zero remains unchanged if we add a null-sequence. Indeed, let $a \in \mathfrak{C}$ be a rational Cauchy sequence bounded away from zero, and $b \in \mathfrak{C}$ a null sequence. Let $0 < \epsilon \in \mathbb{Q}$ and $N \in \mathbb{N}_0$ such that $a_n > \epsilon$ for $n \geq N$. Then, choose $M \in \mathbb{N}_0$ such that $|b_n| < \epsilon/2$ for $n \geq M$. Then, for $n \geq \max(M, N)$, we have

$$a_n + b_n \geq a_n - |b_n| > \epsilon - \epsilon/2 = \epsilon/2.$$

Thus, the sequence $a + b$ is bounded away from zero, and the statement follows.

The importance of this concept is shown by the following: For a rational Cauchy sequence $a \in \mathfrak{C}$, we have $a \not\sim 0$ if and only if $|a|$ is bounded away from zero.

The "if" part is clear (since a rational Cauchy sequence whose absolute value is bounded away from zero cannot be a null-sequence). For the "only if" part, first note that a rational Cauchy sequence $a \in \mathfrak{C}$ does not converge to zero if there exists $0 < \epsilon \in \mathbb{Q}$ such that, for all $k \in \mathbb{N}_0$, we have **some** $n_k \geq k$ with $|a_{n_k}| \geq \epsilon$. On the other hand, since a is a Cauchy sequence, there exists $N \in \mathbb{N}_0$ such that $|a_m - a_n| < \epsilon/2$ for $m, n \geq N$. Since $\lim_{k\to\infty} n_k = \infty$, we can choose $k_0 \in \mathbb{N}$ such that $n_{k_0} \geq N$. For $n \geq N$, we calculate

$$|a_n| = |a_{n_{k_0}} - (a_{n_{k_0}} - a_n)| \geq |a_{n_{k_0}}| - |a_{n_{k_0}} - a_n| > \epsilon - \epsilon/2 = \epsilon/2.$$

The "only if" part now follows.

Let $a \in \mathfrak{C}$ be a rational Cauchy sequence, and assume that the equivalence class of a in $\mathbb{R} = \mathfrak{C}/\sim$ is non-zero. On the level of sequences this means that $a \not\sim 0$. By the above, $|a|$ is bounded away from zero, and hence there exist $0 < \epsilon \in \mathbb{Q}$ and $N \in \mathbb{N}_0$ such that $|a_n| > \epsilon$ for $n \geq N$.

Define the sequence $a^{-1} = 1/a : \mathbb{N}_0 \to \mathbb{Q}$ such that $(a^{-1})_n = 1/a_n$ if $a_n \neq 0$, and $(a^{-1})_n = a_n = 0$ otherwise. By definition, for $n \geq N$, we have $a_n \neq 0$, and therefore $(a \cdot a^{-1})_n = 1$. We see that $a \cdot a^{-1}$ is in the same equivalence class as the multiplicative identity 1 since it differs from the constant sequence $(1, 1, 1, \ldots)$ only in the first N terms.

We need to show that this construction of the multiplicative inverse depends only on the equivalence classes; that is, for $a, a' \in \mathfrak{C}$, the relations $a \sim a'$ and $a, a' \not\sim 0$ imply $a^{-1} \sim a'^{-1}$.

Indeed, choose $0 < \epsilon, \epsilon' \in \mathbb{Q}$ and $N, N' \in \mathbb{N}_0$ such that $|a_n| > \epsilon$ for $n \geq N$, and $|a'_n| > \epsilon'$ for $n \geq N'$. We then have

$$0 \leq \limsup_{n\to\infty} |a_n^{-1} - a_n'^{-1}| = \limsup_{n\to\infty} \frac{|a_n - a'_n|}{|a_n||a'_n|} \leq \frac{\limsup_{n\to\infty} |a_n - a'_n|}{\epsilon\epsilon'},$$

where we used the monotonicity property of the limit superior. If $a \sim a'$ then the right-hand side is zero, and $a^{-1} \sim a'^{-1}$ follows.

With this we define the multiplicative inverse of the (non-zero) equivalence class of $a \in \mathfrak{C}$, $a \not\sim 0$, as the equivalence class of $a^{-1} = 1/a \in \mathfrak{C}$. With the additive and multiplicative inverses now in place, it now follows that $\mathfrak{C} = \mathbb{R}$ is a **field** (with respect to the operations of addition and multiplication).

A natural order in $\mathbb{R} = \mathfrak{C}/\sim$ is given as follows. For rational sequences $a, b : \mathbb{Q} \to \mathbb{Q}$, we define $a < b$ if $b - a$ is bounded away from zero.

Once again, we need to show that $<$ gives rise to a relation on the equivalence classes; that is, for $a' \sim a$ and $b' \sim b$, the relation $a < b$ implies $a' < b'$. This, however, follows writing

$$b' - a' = (b' - b) + (b - a) + (a - a')$$

and noting that (adding) the null-sequences $a - a'$ and $b' - b$ do not change the property of $b - a$ being bounded away from zero.

We conclude that $<$ depends on the equivalence classes only, and thereby defines a relation $<$ on $\mathbb{R}$. As usual, we call the equivalence class of $a \in \mathfrak{C}$ **positive** if $a > 0$ (a is bounded away from zero) and **negative** if $-a > 0$ ($-a$ is bounded away from zero).

We claim that $<$ is a **strict total order**; that is $<$ is transitive and trichotomous.

For transitivity, we let $a, b, c \in \mathfrak{C}$ be three rational Cauchy sequences such that $a < b$ and $b < c$. These mean that $b - a$ and $c - b$ are bounded away from zero. Hence the sum $(c - b) + (b - a) = c - a$ is also bounded away from zero. Thus, $a < c$, and transitivity follows.

For trichotomy, assume that $a \in \mathfrak{C}$ is a rational Cauchy sequence representing a non-zero equivalence class: $a \not\sim 0$. This means the existence of $0 < \epsilon \in \mathbb{Q}$ and $N \in \mathbb{N}_0$ such that $|a_n| > \epsilon$ for $n \geq N$. On the other hand, since a is Cauchy, there exists $M \in \mathbb{N}_0$ such that $|a_n - a_m| < \epsilon$ for $n \geq M$. Putting these together, we either have $a_n > \epsilon$ for all $n \geq \max(M, N)$, or $-a_n > \epsilon$ for all $n \geq \max(M, N)$. In the first case the equivalence class of a is positive, in the second, it is negative. Trichotomy follows.

We conclude that $<$ is a strict total order on $\mathbb{R}$.

Finally, it is routine to check that the cancellation laws for inequalities hold. With these, it follows that $\mathbb{R}$ is a **totally ordered field**.

In the next step we show that the Archimedean Property holds in $\mathbb{R}$.[23] As usual, we formulate this in terms of (rational) Cauchy sequences, representatives of the respective equivalence classes.

Proposition 2.3.4 *Let $0 < a, b \in \mathfrak{C}$. Then there exists $m \in \mathbb{N}(\subset \mathbb{Q})$ such that $b < ma$.*

Proof Since $a > 0$, there exists $0 < \epsilon_0 \in \mathbb{Q}$ and $N_0 \in \mathbb{N}_0$ such that $\epsilon_0 < a_n$ for $n \geq N_0$. Since b is a Cauchy sequence, it is bounded with a rational upper bound $0 < q_0 \in \mathbb{Q}$; that is, we have $b_n \leq q_0$ for all $n \in \mathbb{N}_0$.

[23] Strictly speaking, we do not need this as it will follow from the Least Upper Bound Property to be proved below. For completeness, we include this here as a separate proposition, however.

Applying the Archimedean Property **in** $\mathbb{Q}$, for $0 < \epsilon_0, q_0 \in \mathbb{Q}$, we have $q_0 \leq (m_0 - 1)\epsilon_0$ for some $m_0 \in \mathbb{N}$. (The shift in the multiple of ϵ_0 is of technical convenience.)

The Archimedean Property to be proved (in $\mathbb{R}$) states that $0 < ma - b$ for some $m \in \mathbb{N}$, that is, the Cauchy sequence $ma - b$ is bounded away from zero.

Assume that the Archimedean Property does not hold for $0 < a, b \in \mathfrak{C}$. This means that, for every $m \in \mathbb{N}$, for every $0 < \epsilon \in \mathbb{Q}$ and for every $k \in \mathbb{N}_0$, there exists $n_k \geq k$ such that $ma_{n_k} - b_{n_k} \leq \epsilon$.

Now letting $m = m_0$ and $\epsilon = \epsilon_0$, for $n_k \geq k$, $k \in \mathbb{N}_0$, we calculate

$$m_0 a_{n_k} \leq b_{n_k} + \epsilon_0 \leq q_0 + \epsilon_0 \leq m_0 \epsilon_0.$$

Thus, $a_{n_k} \leq \epsilon_0$ for $k \in \mathbb{N}_0$. Since $\lim_{k\to\infty} n_k = \infty$, this contradicts to $\epsilon_0 < a_n$ for $n \geq N_0$. The proposition follows.

The Archimedean Property for $\mathbb{R}$ has an important consequence usually termed as the **density** of the rational numbers among the reals:

Corollary *Given $a, b \in \mathfrak{C}$ such that $a < b$, there exists $q \in \mathbb{Q}$ such that $a < q < b$.*

Proof We may assume $a > 0$ since the remaining cases can be treated similarly. Since $0 < b - a$, there exist $0 < \epsilon \in \mathbb{Q}$ and $N \in \mathbb{N}_0$ such that $\epsilon < b_n - a_n$ for $n \geq N$. Letting $q_0 = \epsilon/2 \in \mathbb{Q}$, we have $0 < \epsilon/2 < b - a - q_0$ for $n \geq N$. This gives $q_0 < b - a$.

Let $A = \{n \in \mathbb{N} \,|\, a < nq_0\}$. By the Archimedean Property of $\mathbb{R}$ just proved, the set A is non-empty. Since $\mathbb{N}$ is well-ordered, there exists $n_0 = \inf A$. We have $a < n_0 q_0$ and n_0 is the smallest natural number with this property.

We claim that $n_0 q_0 < b$. Assume not: $n_0 q_0 \geq b$. Combining this with $q_0 < b - a$, we have

$$a = b + (a - b) < b - q_0 \leq n_0 q_0 - q_0 = (n_0 - 1)q_0.$$

This contradicts to the minimal choice of n_0 as the infimum of the set A. Letting $q = n_0 q_0 \in \mathbb{Q}$, the corollary follows.

As the final task to finish the construction of the Cauchy real number system $\mathbb{R} = \mathfrak{C}/\sim$ we need to show the Least Upper Bound Property.

Proposition 2.3.5 *In $\mathbb{R} = \mathfrak{C}/\sim$ the Least Upper Bound Property holds.*

Proof Let $A \subset \mathbb{R}$ be a non-empty subset, and assume that it is bounded above by the equivalence class of a rational Cauchy sequence $c \in \mathfrak{C}$. Since c, as a sequence, is bounded, there is a (constant) rational sequence $q_0 \in \mathbb{Q}$ such that $c < q_0$. This means that the equivalence class of q_0 is also an upper bound for A.

Let $a \in \mathfrak{C}$ such that the equivalence class of a belongs to A. (Since A is non-empty, a exists.) Since a is a rational Cauchy sequence, it is bounded from below. Choose a (constant) rational sequence $p_0 \in \mathbb{Q}$ such that $p_0 < a$. Then the equivalence class of p_0 is **not** an upper bound for A.

Proceeding inductively, assume that, for $n \in \mathbb{N}_0$, the (constant) rational sequences $p_n, q_n \in \mathbb{Q}$ have been chosen such that the equivalence class of q_n is an upper bound for A while the equivalence class of p_n is not.

Consider the (constant) rational sequence $m_n = (p_n + q_n)/2 \in \mathbb{Q}$, the arithmetic mean of p_n and q_n. If the equivalence class of m_n is an upper bound for A then we define $p_{n+1} = p_n$ and $q_{n+1} = m_n$. If the equivalence class of m_n is not an upper bound for A then we define $p_{n+1} = m_n$ and $q_{n+1} = q_n$. By Peano's Principle of Induction, $p_n, q_n \in \mathbb{Q}$ are defined for all $n \in \mathbb{N}_0$. Again by induction, $(p_n)_{n\in\mathbb{N}_0}$ is an increasing sequence of rational numbers whose equivalence classes are not upper bounds for A, and $(q_n)_{n\in\mathbb{N}_0}$ is a decreasing sequence of rational numbers whose equivalence classes are upper bounds for A. In addition, $p_n < q_n$, $n \in \mathbb{N}_0$ (since the equivalence class of q_n is an upper bound for A while that of p_n is not), and we have

$$q_{n+1} - p_{n+1} = \frac{q_n - p_n}{2} > 0, \quad n \in \mathbb{N}_0.$$

As an easy induction shows, we have

$$q_n - p_n = \frac{q_0 - p_0}{2^n}, \quad n \in \mathbb{N}_0.$$

We claim that $(p_n)_{n\in\mathbb{N}_0}$ and $(q_n)_{n\in\mathbb{N}_0}$ are (rational) **Cauchy sequences**.

To show this, we first note that, by construction, we have

$$p_{n+1} - p_n \leq \frac{q_n - p_n}{2} = \frac{q_0 - p_0}{2^{n+1}}, \quad n \in \mathbb{N}_0.$$

We now claim that, for $m \leq n$ ($m, n \in \mathbb{N}_0$), we have

$$p_n - p_m \leq (q_0 - p_0)\left(\frac{1}{2^m} - \frac{1}{2^n}\right).$$

We show this by induction with respect to $n\ (\geq m)$. For $n = m$ both sides of the inequality are zero. For the general induction step $(m \leq)\ n \Rightarrow n + 1$, we calculate

$$\begin{aligned} p_{n+1} - p_m &\leq (p_{n+1} - p_n) + (p_n - p_m) \leq \frac{q_0 - p_0}{2^{n+1}} + (q_0 - p_0)\left(\frac{1}{2^m} - \frac{1}{2^n}\right) \\ &= (q_0 - p_0)\left(\frac{1}{2^m} - \frac{1}{2^n} + \frac{1}{2^{n+1}}\right) = (q_0 - p_0)\left(\frac{1}{2^m} - \frac{1}{2^{n+1}}\right). \end{aligned}$$

The claim follows.

Now, let $0 < \epsilon \in \mathbb{Q}$. We use the second corollary to the Bernoulli inequality (Section 2.1) to find $N \in \mathbb{N}_0$ such that

$$\frac{1}{2^N} < \frac{\epsilon}{2(q_0 - p_0)}.$$

With this, for $m, n \geq N$, we have

$$|p_n - p_m| \leq (q_0 - p_0)\left|\frac{1}{2^m} - \frac{1}{2^n}\right| \leq (q_0 - p_0)\frac{2}{2^{\min(m,n)}} \leq (q_0 - p_0)\frac{2}{2^N} < \epsilon.$$

Thus, $(p_n)_{n\in\mathbb{N}_0}$ is a (rational) Cauchy sequence. We denote this by $r = (p_n)_{n\in\mathbb{N}_0} \in \mathfrak{C}$.

Similar computation show that $(q_n)_{n\in\mathbb{N}_0}$ is also a Cauchy sequence, denoted by $s = (q_n)_{n\in\mathbb{N}_0} \in \mathfrak{C}$. Now the difference $s - r$ is the null-sequence $(q_n - p_n)_{n\in\mathbb{N}_0} = ((q_0 - p_0)/2^n)_{n\in\mathbb{N}_0}$. We obtain $r \sim s$, so that, in $\mathbb{R} = \mathfrak{C}/\sim$ they define the same equivalence class. We claim that this equivalence class is the least upper bound of A.

First, we show that the equivalence class of the decreasing sequence $s = (q_n)_{n\in\mathbb{N}_0}$ is an **upper bound** for A. Assume not. Then there exists $a \in \mathfrak{C}$ whose equivalence class is an element of A such that $s < a$. This means that $a - s$ is bounded away from zero, that is, there exist $0 < \epsilon \in \mathbb{Q}$ and $N \in \mathbb{N}_0$ such that $\epsilon < a_n - q_n$ for $n \geq N$.

Now, $s \in \mathfrak{C}$ is a Cauchy sequence, so (for our ϵ) there exists $M \in \mathbb{N}_0$ such that $|q_m - q_n| < \epsilon$ for $m, n \geq M$. Combining these, we have

$$q_m - q_n \leq |q_m - q_n| < \epsilon < a_n - q_n, \quad m, n \geq K = \max(M, N).$$

This gives $q_m < a_n$, $m, n \geq K$. Now, we **fix** $m \geq K$, consider $q_m \in \mathbb{Q} \subset \mathfrak{C}$ as the constant (rational) sequence, and compare it with the Cauchy sequence $a \in \mathfrak{C}$. By the inequality above, $a < q_m$ cannot happen. On the other hand, $q_m < a$ cannot happen either since the equivalence class of q_m is an upper bound for A. By trichotomy, we obtain $q_m \sim a$.

Since s is a decreasing sequence it therefore must become constant after the Kth term. (Otherwise, for some $k \in \mathbb{N}$, we would have $q_{m+k} < q_m < a_n$ for $n \geq K$, and this (with $0 < \epsilon = q_m - q_{m+k}$) would imply $q_{m+k} < a$, a contradiction again.) We obtain $s \sim q_m \sim a$, a contradiction again to the original assumption $s < a$.
Summarizing, we obtain that the equivalence class of s is an upper bound for A.

Second, we need to show that the equivalence class of the increasing sequence $r = (p_n)_{n\in\mathbb{N}_0}(\sim s)$ is the **least** upper bound for A. The argument is similar to the above in the use of r (instead of s). Assume not. Then there exists $t \in \mathfrak{C}$ whose equivalence class is an upper bound for A such that $t < r$. This means that $r - t$ is bounded away from zero, that is, there exist $0 < \epsilon \in \mathbb{Q}$ and $N \in \mathbb{N}_0$ such that $\epsilon < p_n - t_n$ for $n \geq N$.

Now, $r \in \mathfrak{C}$ is a Cauchy sequence, so (for our ϵ) there exists $M \in \mathbb{N}_0$ such that $|p_n - p_m| < \epsilon$ for $m, n \geq M$. Combining these, we have

$$p_n - p_m \leq |p_n - p_m| < \epsilon < p_n - t_n, \quad m, n \geq K = \max(M, N).$$

This gives $t_n < p_m$, $m, n \geq K$. Now, we **fix** $m \geq K$, consider $p_m \in \mathbb{Q} \subset \mathfrak{C}$ as the constant (rational) sequence, and compare it with the Cauchy sequence $t \in \mathfrak{C}$. By the inequality above, $p_m < t$ cannot happen. On the other hand, $t < p_m$ cannot happen either since the equivalence class of p_m is not an upper bound for A (and that of t is). By trichotomy, we obtain $p_m \sim t$.

Since r is an increasing sequence it therefore must become constant after the Kth term. (Otherwise, for some $k \in \mathbb{N}$, we would have $t_n < p_m < p_{m+k}$ for $n \geq K$, and this (with $0 < \epsilon = p_{m+k} - p_m$) would imply $t < p_{m+k}$, a contradiction again.) We obtain $r \sim p_m \sim t$, a contradiction to the original assumption $t < r$.

Thus, the equivalence class of $r \sim s$ is the least upper bound for A, and the theorem follows.

This completes Cantor's construction of the real number system by Cauchy sequences. Since this model is a **complete ordered field** containing $\mathbb{Q}$ as a subfield, all the statements at the beginning of this section apply. More specifically, in this model the Least Upper Bound Property holds, and therefore a sequence is convergent (to a real number) if and only if it is a Cauchy sequence, and the Monotone Convergence Theorem and the Bolzano-Weierstrass Theorem are valid.

The Cantor model $\mathbb{R}_C$ and Dedekind model $\mathbb{R}_D$ of the real number system are isomorphic in the sense that there is a one-to-one correspondence between them which respects the field operations and the order. (Here we used subscript to distinguish between the two models.) As alluded to above, the isomorphism is given by associating to a Dedekind cut, an element of $\mathbb{R}_D$, the equivalence class of either of the rational Cauchy sequences $(r_n)_{n\in\mathbb{N}_0} \in \mathfrak{C}$ or $(s_n)_{n\in\mathbb{N}_0} \in \mathfrak{C}$ (constructed in Section 2.2) up to null-sequences.

Exercises

2.3.1. For a real sequence a, define $A \subset \mathbb{R}$ to be the set of limits of all convergent subsequences of a. Show that $\limsup_{n\to\infty} a_n = \sup A$ and $\liminf_{n\to\infty} a_n = \inf A$.

2.3.2. Let $(a_n)_{n\in\mathbb{N}}$ be a sequence of positive terms. Show that

$$\liminf_{n\to\infty} \frac{1}{a_n} = \frac{1}{\limsup_{n\to\infty} a_n}.$$

2.3.3. Let $(a_n)_{n\in\mathbb{N}_0}$ be a sequence defined inductively by $a_0 = 1$ and $a_n = \sqrt{1 + a_{n-1}}$, $n \in \mathbb{N}$. (Thus, we have $a_n = \sqrt{1 + \sqrt{1 + \cdots \sqrt{1 + \sqrt{2}}}}$ with n nested square roots.) Use the Monotone Convergence Theorem to show that $\lim_{n\to\infty} a_n = (1 + \sqrt{5})/2$. (Note that $\tau = (1 + \sqrt{5})/2$ is the golden number; see Example 3.1.2.)

2.3.4. Let $(a_n)_{n\in\mathbb{N}}$ be a sequence such that $\{a_n \,|\, n \in \mathbb{N}\} = (0,1) \cap \mathbb{Q}$ (as sets). ($(a_n)_{n\in\mathbb{N}}$ exists since $\mathbb{Q}$ is countable.) Find $\liminf_{n\to\infty} a_n$ and $\limsup_{n\to\infty} a_n$.

2.3.5. Let $(a_n)_{n\in\mathbb{N}}$ be a bounded real sequence. Show that there exist convergent subsequences $(a_{m_k})_{k\in\mathbb{N}}$ and $(a_{n_l})_{l\in\mathbb{N}}$ such that $\lim_{k\to\infty} a_{m_k} = \liminf_{n\to\infty} a_n$ and $\lim_{l\to\infty} a_{n_l} = \limsup_{n\to\infty} a_n$.

2.4 Dirichlet Approximation and Equidistribution*

We have seen that the set of rational numbers $\mathbb{Q}$ is a **dense** subset of the set of real numbers $\mathbb{R}$ (Corollary to Proposition 2.3.4). In other words, any irrational number can be approximated by rational numbers up to arbitrary precision.

In this section we will look at this approximation more closely, find approximating fractions in specific forms, and give a quantitative measure of the density of the approximating rationals through the Equidistibution Theorem.

For the next theorem, recall from Example 1.1.3 that, for $a \in \mathbb{R}$, $[a]$ denotes the greatest integer $\le a$. In addition, we introduce here the **fractional part** $\{a\}$ of $a \in \mathbb{R}$ defined by $\{a\} = a - [a]$. The definitions imply $0 \le \{a\} < 1$ and $\{a + n\} = \{a\}$, $a \in \mathbb{R}$, $n \in \mathbb{Z}$. Moreover, we have

$$\{a\} + \{-a\} = \begin{cases} 0 \text{ if } a \in \mathbb{Z}, \\ 1 \text{ if } a \notin \mathbb{Z}. \end{cases}$$

Dirichlet Approximation Theorem *Let $\alpha \in \mathbb{R}$ and $n \in \mathbb{N}$. Then there exist $p \in \mathbb{Z}$ and $q \in \mathbb{N}$, $q < n$, such that*

$$|q\alpha - p| < \frac{1}{n}.$$

Proof We may assume $0 < \alpha \in \mathbb{R}$. Consider the $n + 1$ numbers

$$\{k\alpha\} \in [0,1), \quad k = 0, 1, \ldots, n.$$

Subdivide the interval $[0, 1)$ into n disjoint subintervals as

$$[0,1) = \bigcup_{m=1}^{n} \left[\frac{m-1}{n}, \frac{m}{n}\right).$$

By the **Pigeonhole Principle** there must be two numbers $\{i\alpha\}$ and $\{j\alpha\}$, $i > j$, say, in the same subinterval $[(m-1)/n, m/n)$, say. Since the length of each subinterval is $1/n$, we obtain

$$|\{i\alpha\} - \{j\alpha\}| < \frac{1}{n}, \quad i > j.$$

Using the definition of the fractional part and rearranging, this gives

$$|(i-j)\alpha - ([i\alpha] - [j\alpha])| < \frac{1}{n}, \quad i > j.$$

Letting $p = [i\alpha] - [j\alpha] \in \mathbb{Z}$ and $q = i - j \in \mathbb{N}$, $q < n$, the theorem follows.

History
Johann Peter Gustav Lejeune Dirichlet (1805 – 1859) used the pigeonhole principle first around 1834 as a counting argument (as in the proof above) to prove the approximation theorem named after him. This principle, termed by him as "Schubfachprinzip" (in German) or "Principe de tiroirs" (in French) (drawer/shelf principle) has many interpretations, and curious applications. For example, it has been noted that, since the average number of hairs on a person's head is less than the total population of London, there must be at least two people there with the same number of hairs on their heads. Since Dirichlet's father was a postmaster the term pigeonhole principe may even be historically accurate alluding to a post office having furniture with many pigeonholes bulging with sorted letters.

A simple application of the pigeonhole principle is the following:

Example 2.4.1 Show that, among five distinct real numbers, there are always two a and b, say, that satisfy the inequality $|a - b| < |1 + ab|$.

To prove this, subdivide the set of real numbers into four intervals as follows

$$\mathbb{R} = (-\infty, -1] \cup (-1, 0) \cup [0, 1) \cup [1, \infty).$$

By the pigeonhole principle, among the five given real numbers, there are two, a and b, say, that are contained in one of the four intervals above. Since the inequality stays the same by taking the opposites $-a$ and $-b$, there are only two cases to consider: $a, b \in [0, 1)$ and $a, b \in [1, \infty)$. Since the inequality is unchanged by taking non-zero reciprocals $1/a$ and $1/b$, ($|1/a - 1/b| < |1/a \cdot 1/b - 1|$), we may assume that $a, b \in [0, 1)$ or $a, b \in (0, 1]$. This final case, however, is obvious since $|a - b| < 1 \leq |1 + ab|$.

We now make a short detour here, and give a brief description of yet another model of the real number systems, the **Eudoxus** reals $\mathbb{R}_E$.

Our starting point is Euclid's Elements:[24]

History
Excerpt from Euclid's *Elements* (Book V, Definition 5):[25]
"Magnitudes are said to be *in the same ratio*, the first to the second and the third to the fourth,

[24]The material here follows closely the beginning of the paper: Athan, R.D., (2004) *The Eudoxus Real Numbers*, arXiv:math/0405454.

[25]The excerpt quoted here is from the translation by Sir Thomas L. Heath of the Greek text of J.L. Heiberg (1854 – 1928) and H. Menge, from *Euclidis opera omnia*, 8 vols & supplement, in Greek, Teubner, Leipzig, 1883–1916. Edited by J.L. Heiberg and H. Menge.

when, if any equimultiples whatever are taken of the first and third, and any equimultiples whatever of the second and the fourth, the former equimultiples alike exceed, are alike equal to, or alike fall short of, the latter equimultiples respectively taken in corresponding order."

As widely accepted, Euclid describes here the work of Eudoxus of Cnidus, and asserts that two ratios $a \div b$ and $c \div d$ are equal if, for all $m, n \in \mathbb{N}$, the relations $ma > nb$ and $mc > nd$ are simultaneously true or false, and similarly for equalities and for reverse inequalities.

History

De Morgan's interpretation[26] of Eudoxus above is as follows:
Consider an infinite equidistantly spaced railings of a fence in front of another infinite equidistant colonnade. If the distance between consecutive railings is $0 < \alpha \in \mathbb{R}$, and the distance between consecutive columns is unity, then, riding along the fence[27] and counting columns, for $k \in \mathbb{N}$, we denote the number of columns to the left or aligned to the kth railing by $c_k \in \mathbb{N}$, the sequence $(c_k)_{k\in\mathbb{N}}$ will "represent" the real number α.

De Morgan's interpretation of the real number $0 < \alpha \in \mathbb{R}$ simply means that $k\alpha = c_k + \{k\alpha\}$ with the greatest integer $c_k = [k\alpha] \in \mathbb{N}_0$, and the fractional part $0 \le \{k\alpha\} < 1, k \in \mathbb{N}$.

The Dirichlet Approximation Theorem above asserts that the positive integral multiples $k\alpha$ get arbitrarily close to integers (that is, to columns).

For the construction of the Eudoxus reals, we are interested in the **arithmetic properties** of the sequence of integers $c_k = [k\alpha], k \in \mathbb{N}$.

A simple computation gives

$$c_{j+k} = c_j + c_k + \{j\alpha\} + \{k\alpha\} - \{(j+k)\alpha\}, \quad j, k \in \mathbb{N}.$$

Note that, if $\alpha \in \mathbb{N}$ then the three fractional parts on the right-hand side are zero.

This motivates the following definition: A **slope**[28] is a map $c : \mathbb{Z} \to \mathbb{Z}$, $c_k = c(k), k \in \mathbb{Z}$, such that the set $\{c_{j+k} - c_j - c_k \mid j, k \in \mathbb{Z}\}$ is **finite**. We denote the set of slopes by $\mathfrak{S}$.

The operation of **addition** $+$ on $\mathfrak{S}$ is defined naturally by $(c + c')_k = c_k + c'_k$, $k \in \mathbb{Z}$, where $c, c' \in \mathfrak{S}$. Similarly, the operation of **multiplication** $\cdot$ on $\mathfrak{S}$ is given by composition: $(c \cdot c')_k = (c \circ c')(k) = c_{c'_k}, k \in \mathbb{Z}$, where $c, c' \in \mathfrak{S}$.

Finally, two slopes $c, c' \in \mathfrak{S}$ are called **equivalent**, written as $c \sim c'$, if the set $\{c_k - c'_k \mid k \in \mathbb{Z}\}$ is **finite**.

With these definitions in place, it can be proved that $\sim$ is an equivalence relation on $\mathfrak{S}$, and it is compatible with the addition and multiplication.

Finally, with a considerably more work,[29] it can be shown that the quotient space $\mathfrak{S}/\sim$ is a complete totally ordered field (with respect to a natural oder). This is the

[26] See the commentary by Heath ibid.

[27] As we are in the 19th century.

[28] Also called almost homomorphism (of $\mathbb{Z}$).

[29] See Athan, R.D. ibid.

Eudoxus real number system $\mathbb{R}_E$. Note the special feature of this model that it is constructed directly from the integers, bypassing the rational numbers.

We now return to the main line, and note a direct consequence of the Dirichlet Approximation Theorem.

Corollary *Given $\alpha \in \mathbb{R}$, there exists a rational number $p/q \in \mathbb{Q}$, $p \in \mathbb{Z}$, $q \in \mathbb{N}$, such that*

$$\left|\alpha - \frac{p}{q}\right| < \frac{1}{q^2}.$$

We call the rational number $p/q \in \mathbb{Q}$, $p \in \mathbb{Z}$, $q \in \mathbb{N}$, in the corollary a **Dirichlet approximation** of α. Clearly, we may assume that p/q is an irreducible fraction, that is, p and q have no common divisors. We denote by D_α, $\alpha \in \mathbb{R}$, the set of all Dirichlet approximations $p/q \in \mathbb{Q}$, $p \in \mathbb{Z}$, $q \in \mathbb{N}$, of α.

Assume that $\alpha \in \mathbb{R}$ has Dirichlet approximations p/q, $p'/q \in D_\alpha$ with the **same denominator**. We claim that $p' = p$ if $q \neq 1$, and $p' = p$ or $p' = p \pm 1$, if $q = 1$.

Indeed, we have

$$-\frac{1}{q^2} < \alpha - \frac{p}{q} < \frac{1}{q^2} \quad \text{and} \quad -\frac{1}{q^2} < \frac{p'}{q} - \alpha < \frac{1}{q^2}.$$

Adding, and simplifying, we obtain

$$|p' - p| < \frac{2}{q}.$$

If $2 \leq q \in \mathbb{N}$ then $p' = p$ holds. If $q = 1$ then $|p' - p| \leq 1$, and hence $p' = p$ or $p' = p \pm 1$. The claim follows. In all cases, there are at most two possibilities for a Dirichlet approximation with the same denominator.

Proposition 2.4.1 *A rational number has only finitely many Dirichlet approximations. An irrational number has infinitely many Dirichlet approximations.*

Proof First, in the rational case, we may assume $0 < \alpha = a/b \in \mathbb{Q}$, $a, b \in \mathbb{N}$. Assuming $p/q \in D_{a/b}$ such that $p/q \neq a/b$, we have

$$\frac{1}{bq} \leq \frac{|aq - bp|}{bq} = \left|\frac{a}{b} - \frac{p}{q}\right| < \frac{1}{q^2}.$$

This gives $(0 <) q < b$. This means that, for Dirichlet approximations, there are at most $b - 1$ available denominators. Since each denominator can have at most two numerators, we get $|D_{a/b}| \leq 2(b-1)+1 = 2b-1$ (including $p/q = a/b \in D_{a/b}$). The first statement of the theorem follows.

Let $\alpha \in \mathbb{R} \setminus \mathbb{Q}$ be irrational, and assume that D_α is finite. Since α is irrational and D_α is finite, we have a positive minimum $0 < \min_{p/q \in D_\alpha} |q\alpha - p|$ (which is attained). Let $n \in \mathbb{N}$ such that $1/n < \min_{p/q \in D_\alpha} |q\alpha - p|$. By the Dirichlet

Approximation Theorem, there exist $p_0 \in \mathbb{Z}$ and $q_0 \in \mathbb{N}$, $q_0 < n$, such that $|q_0\alpha - p_0| < 1/n$. Hence

$$\left|\alpha - \frac{p_0}{q_0}\right| < \frac{1}{nq_0} < \frac{1}{q_0^2},$$

and we obtain $p_0/q_0 \in D_\alpha$. This contradicts to the minimal choice of $n \in \mathbb{N}$. Thus, D_α cannot be finite.

Since there are only two choices for a denominator of a Dirichlet approximation, it follows that, for $\alpha \in \mathbb{R}$ is irrational, there are Dirichlet approximations $p/q \in D_\alpha$ with p and q relatively prime such that the denominator q is **arbitrarily large**.

Equidistribution Theorem *Let $\alpha \in \mathbb{R} \setminus \mathbb{Q}$ be irrational, and $0 \le a < b \le 1$. Then, we have*

$$\lim_{n\to\infty} \frac{|\{0 \le j < n \mid \{j\alpha\} \in [a, b]\}|}{n} = b - a,$$

where the numerator of the fraction counts the number of times when $\{j\alpha\}$, $j = 0, 1, \ldots, n-1$, falls into the interval $[a, b]$.

History
The Equidistribution Theorem was proved independently by Hermann Weyl, Wacław Sierpiński and Piers Bohn in 1909–1910. Many variants have been derived since then, and it is still a very active area of research.

We begin the proof with the following:

Lemma *Let $\alpha \in \mathbb{R}\setminus\mathbb{Q}$ be irrational, and assume that $p/q \in D_\alpha$ (p and q relatively prime) is a Dirichlet approximation of α. Then, for every integer $0 \le i < q$, there exits a unique integer $0 \le j < q$ such that*

$$\{j\alpha\} \in \left(\frac{i}{q}, \frac{i+1}{q}\right].$$

Proof We may assume $\alpha > p/q$. (If $\alpha < p/q$ then $-\alpha > (-p)/q$ and $\{j(-\alpha)\} = 1 - \{j\alpha\}$, $j \in \mathbb{Z}$.)

Since $0 < \alpha - p/q < 1/q^2$, we have

$$0 < j\alpha - \frac{jp}{q} < \frac{1}{q}, \quad 0 \le j < q.$$

The division algorithm gives $jp = q_j \cdot q + r_j$, $0 \le r_j < q$. Substituting and rearranging, we obtain

$$\frac{r_j}{q} < [j\alpha] - q_j + \{j\alpha\} < \frac{r_j + 1}{q}, \quad 0 \le r_j < q.$$

Since $[j\alpha] - q_j$ is an integer, it must be zero. We finally get

$$\frac{r_j}{q} < \{j\alpha\} < \frac{r_j + 1}{q}, \quad 0 \le r_j < q,\ 0 \le j < n.$$

The correspondence $j \mapsto r_j$, $j = 0, \ldots, q-1$, defines a self-map of the set $\{0, \ldots, q-1\}$. Once we show that this map is a bijection, the lemma follows. Since the ambient set is finite, it is enough to show injectivity. (See Example 1.3.2.) Assume $r_j = r_k$, $0 \le j \le k < q$. Going back to the division algorithm, $jp = q_j q + r_j$ and $kp = q_k q + r_k$ give $(k-j)p = (q_k - q_j)q$. In particular, q divides $(k-j)p$. Since p and q are relatively prime, q must divide $k - j$. Since $0 \le k - j < q$ this is possible only if $j = k$. Thus, injectivity, and therefore surjectivity hold. The proof is complete.

Proof of the Equidistribution Theorem Let $\alpha \in \mathbb{R} \setminus \mathbb{Q}$ be irrational.

Let $0 < \epsilon \in \mathbb{R}$, and choose a Dirichlet approximation $p/q \in D_\alpha$ with p and q (≥ 2) relatively prime, such that $2/q < \epsilon/3$. (This choice is possible since, as a consequence of the Dirichlet Approximation Theorem, for α irrational, there exist Dirichlet approximations of aribitrarily large denominators.) Let $N \in \mathbb{N}$ such that $q/N < \epsilon/3$. Finally, using the division algorithm, let $n = vq + r$ with $0 \le r < q$.

As a first step, we clearly have

$$|\{0 \le j < n \,|\, \{j\alpha\} \in [a,b]\}| \ge \sum_{u=1}^{v} \left|\{(u-1)q \le k < uq \,\middle|\, \{k\alpha\} \in [a,b]\}\right|.$$

We claim

$$\left|\{(u-1)q \le k < uq \,\middle|\, \{k\alpha\} \in [a,b]\}\right| \ge q(b-a) - 2, \quad u = 1, \ldots, v.$$

First, we show this for $u = 1$:

$$\left|\{0 \le j < q \,\middle|\, \{j\alpha\} \in [a,b]\}\right| \ge q(b-a) - 2$$

(we switched back to j from k).

This is a direct consequence of the previous lemma. Split the interval $[0, 1)$ into q disjoint subintervals

$$[0,1) = \bigcup_{i=0}^{q-1} \left[\frac{i}{q}, \frac{i+1}{q}\right).$$

According to the lemma, the numbers $\{j\alpha\}$, $0 \le j < q$, are equidistributed in this splitting; each subinterval contains exactly one of these numbers. The interval $[a, b]$ **completely** contains at least $q(b-a) - 2$ of these subintervals, and the deduction of 2 corresponds to the mismatch of the end-points of $[a, b]$ with those of the subintervals. Thus, the lower estimate follows for $u = 1$.

We now let $u = 1, \dots, v$ be arbitrary. We let $k = j + (u-1)q$, $0 \le j < q$, so that $(u-1)q \le k < uq$. With these, we calculate

$$\begin{aligned}\{k\alpha\} &= \{(j + (u-1)q)\alpha\} = \{j\alpha + (u-1)q\alpha\} \\ &= j\alpha + (u-1)q\alpha - [j\alpha + (u-1)q\alpha] \\ &= \{j\alpha\} + (u-1)q\alpha + [j\alpha] - [j\alpha + (u-1)q\alpha] \\ &= \{\{j\alpha\} + (u-1)q\alpha\}.\end{aligned}$$

By the previous lemma, we see that the numbers $\{k\alpha\}$, $(u-1)q \le k < uq$, are equidistributed in the splitting of $[0, 1)$ into the subintervals $[i/q, (i+1)/q)$, $i = 0, \dots, q-1$, **translated by the constant** $(u-1)q\alpha$ (and with the interval falling to the end-points of $[0, 1)$ possibly split). As before, the interval $[a, b]$ **completely** contains at least $q(b-a) - 2$ of these subintervals, so that we have the same lower estimate claimed above.

Continuing our lower estimate, we have

$$\begin{aligned}|\{0 \le j < n \mid \{j\alpha\} \in [a, b]\}| &\ge \sum_{u=1}^{v} \left|\{(u-1)q \le k < uq \;\middle|\; \{k\alpha\} \in [a, b]\}\right| \\ &\ge \sum_{u=1}^{v} (q(b-a) - 2) = v(q(b-a) - 2) \\ &= n(b-a) - r(b-a) - 2v.\end{aligned}$$

Due to our choices, for $n \ge N$, we have $r/n < q/n \le q/N < \epsilon/3$ and $2v/n \le 2/q < \epsilon/3$. Since $b - a < 1$, we thus obtain

$$\frac{|\{0 \le j < n \mid \{j\alpha\} \in [a, b]\}|}{n} \ge b - a - \frac{2\epsilon}{3} \ge b - a - \epsilon.$$

The upper estimate is similar. Since $n = vq + r < (v+1)q$, we have

$$\begin{aligned}|\{0 \le j < n \mid \{j\alpha\} \in [a, b]\}| &\le \sum_{u=1}^{v+1} \left|\{(u-1)q \le k < uq \;\middle|\; \{k\alpha\} \in [a, b]\}\right| \\ &\le \sum_{u=1}^{v+1} (q(b-a) + 2) = (v+1)(q(b-a) + 2) \\ &= n(b-a) + (q-r)(b-a) + 2(v+1).\end{aligned}$$

We have $(q-r)(b-a)/n \le q/N < \epsilon/3$ and $2(v+1)/n = 2v/n + 2/n \le 2/q + 2/N \le 2/q + q/N < \epsilon/3 + \epsilon/3 = 2\epsilon/3$. With these, we obtain

$$\frac{|\{0 \le j < n \mid \{j\alpha\} \in [a,b]\}|}{n} < b - a + \epsilon.$$

Summarizing, we obtain that, for every $0 < \epsilon \in \mathbb{R}$, there exists $N \in \mathbb{N}$ such that, for $n \ge N$, we have

$$\left| \frac{|\{0 \le j < n \mid \{j\alpha\} \in [a,b]\}|}{n} - (b-a) \right| < \epsilon.$$

The Equidistribution Theorem follows.

Exercises

2.4.1. Let $\alpha \in \mathbb{R} \setminus \mathbb{Q}$ be irrational. Show that

$$\limsup_{n\to\infty} \{n\alpha\} = 1 \quad \text{and} \quad \liminf_{n\to\infty} \{n\alpha\} = 0,$$

where $\{\cdot\}$ denotes the fractional part.

2.4.2. Let $n \in \mathbb{N}$ and $\triangle \subset \mathbb{R}^2$ an equilateral triangle with side length n. Show that if a subset $A \subset \triangle$ consists of more than n^2 elements then there are (at least) two points in A with distance ≤ 1.

Chapter 3
Rational and Real Exponentiation

"The sum of an infinite series whose final term vanishes (meaning that $\lim_{n\to\infty} a_n = 0$ *for a series* $\sum_{n=0}^{\infty} a_n$*) is perhaps infinite, perhaps finite."*
in the Ars Conjectandi by
Jacob Bernoulli (1655–1705)

The main purpose of this chapter is to give a detailed treatise on powers with rational and real exponents. We begin with a preparatory section on the arithmetic properties of the limit inferior and limit superior and (thereby) the limit. The Fibonacci sequence, the geometric and p-series, and some of their contest level offsprings serve here as illustrations. The core material of this chapter proves the existence of roots of (positive) real numbers paving the way to rational exponentiation and the Bernoulli inequality for rational exponents. The latter is then used to establish (the existence of) powers with real exponents and thereby the extension of the Bernoulli inequality to real exponents. The text is accompanied here with a large variety of illustrative examples of classical limits. From the myriad topics on powers, we discuss linear independence of fractional exponents of integers due to Besicovitch, the Young inequality, some sharp estimates on the p-series, equiconvergence through the Cauchy condensation test, power sums, and the lesser known method of (arithmetic) means. A short section on logarithms along with a few contest level problems is followed by a final section on the Stolz–Cesàro Theorems. These tools complete an arsenal to tackle a large number of sophisticated limits. Several methods developed here will recur later in more complex settings.

3.1 Arithmetic Properties of the Limit

In this preparatory section, we return to our real sequences. The limit superior and limit inferior have simple arithmetic properties. For $a, b : \mathbb{N}_0 \to \mathbb{R}$, we have

$$\liminf_{n\to\infty} a_n + \liminf_{n\to\infty} b_n \le \liminf_{n\to\infty}(a_n+b_n) \le \limsup_{n\to\infty}(a_n+b_n) \le \limsup_{n\to\infty} a_n + \limsup_{n\to\infty} b_n.$$

G. Toth, *Elements of Mathematics*, Undergraduate Texts in Mathematics,
https://doi.org/10.1007/978-3-030-75051-0_3

Example 3.1.1 Let $0 < r \in \mathbb{R}$. Given a sequence $c : \mathbb{N}_0 \to \mathbb{R}$, we define the sequence $c^r : \mathbb{N}_0 \to \mathbb{R}$ by

$$c_0^r = c_0 \quad \text{and} \quad c_n^r = c_n - r \cdot c_{n-1}, \ n \in \mathbb{N}.$$

Clearly, if c is a null-sequence, then, for any $r \in \mathbb{R}$, the sequence c^r is also null.

We are interested in the converse. We note first that if $1 < r \in \mathbb{R}$, then the geometric sequence given by $c_n = r^n$, $n \in \mathbb{N}_0$, diverges whereas $c_n^r = r^n - r \cdot r^{n-1} = 0$, $n \in \mathbb{N}$. Moreover, we have seen (Example 2.3.2) that, even though the sequence given by $c_n = \sqrt{n}$, $n \in \mathbb{N}_0$ is divergent, we have $\lim_{n\to\infty} \left(\sqrt{n} - \sqrt{n-1}\right) = 0$ $(r = 1)$.

These examples show that the converse that we seek cannot hold for $r \geq 1$.

We now claim that, given $0 < r < 1$, for every real sequence c, we have

$$\lim_{n\to\infty} c_n^r = \lim_{n\to\infty} (c_n - r \cdot c_{n-1}) = 0 \quad \Rightarrow \quad \lim_{n\to\infty} c_n = 0.$$

To show this, let c be a real sequence, and set

$$\underline{L} = \liminf_{n\to\infty} c_n \leq \limsup_{n\to\infty} c_n = \overline{L}.$$

Assume that we have

$$\lim_{n\to\infty} (c_n - r \cdot c_{n-1}) = 0.$$

We calculate

$$\begin{aligned} \overline{L} = \limsup_{n\to\infty} c_n &= \limsup_{n\to\infty} ((c_n - r \cdot c_{n-1}) + r \cdot c_{n-1}) \\ &\leq \limsup_{n\to\infty} (c_n - r \cdot c_{n-1}) + \limsup_{n\to\infty} (r \cdot c_{n-1}) = r\overline{L}. \end{aligned}$$

Since $0 < r < 1$, we obtain $\overline{L} \leq 0$.

On the other hand, we have

$$\begin{aligned} \underline{L} = \liminf_{n\to\infty} c_n &= \liminf_{n\to\infty} ((c_n - r \cdot c_{n-1}) + r \cdot c_{n-1}) \\ &\geq \liminf_{n\to\infty} (c_n - r \cdot c_{n-1}) + \liminf_{n\to\infty} (r \cdot c_{n-1}) = r\underline{L}. \end{aligned}$$

Since $0 < r < 1$, we obtain $\underline{L} \geq 0$.

Combining these, we obtain

$$0 \leq \underline{L} \leq \overline{L} \leq 0.$$

This gives $\underline{L} = \overline{L} = 0$. The example follows.

Proposition 3.1.1 *Let $a, b : \mathbb{N}_0 \to \mathbb{R}$ be real sequences, and assume that a is bounded and $\lim_{n\to\infty} b_n$ exists. Then, we have*

$$\limsup_{n\to\infty}(a_n+b_n)=\limsup_{n\to\infty} a_n+\lim_{n\to\infty} b_n \quad \textit{and} \quad \liminf_{n\to\infty}(a_n+b_n)=\liminf_{n\to\infty} a_n+\lim_{n\to\infty} b_n.$$

In particular, if $\lim_{n\to\infty} a_n$ and $\lim_{n\to\infty} b_n$ both exist, then so does $\lim_{n\to\infty}(a_n + b_n)$, and we have

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

Proof First, since the sequence b is convergent, we have

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

Second, writing $a = (a + b) - b$, we have

$$\begin{aligned}\limsup_{n\to\infty} a_n &= \limsup_{n\to\infty}((a_n + b_n) - b_n) \le \limsup_{n\to\infty}(a_n + b_n) + \limsup_{n\to\infty}(-b_n)\\ &= \limsup_{n\to\infty}(a_n + b_n) + \lim_{n\to\infty}(-b_n) = \limsup_{n\to\infty}(a_n + b_n) - \lim_{n\to\infty} b_n.\end{aligned}$$

Hence

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

The first formula in the proposition follows. The proof of the second formula is analogous.

Proposition 3.1.2 *Let $a, b : \mathbb{N}_0 \to \mathbb{R}$ be real sequences, and assume that a is bounded and $\lim_{n\to\infty} b_n$ exists and is non-negative. Then we have*

$$\limsup_{n\to\infty}(a_n \cdot b_n) = \limsup_{n\to\infty} a_n \cdot \lim_{n\to\infty} b_n \quad \textit{and} \quad \liminf_{n\to\infty}(a_n \cdot b_n) = \liminf_{n\to\infty} a_n \cdot \lim_{n\to\infty} b_n.$$

In particular, if $\lim_{n\to\infty} a_n$ and $\lim_{n\to\infty} b_n$ both exist, then so does $\lim_{n\to\infty}(a_n \cdot b_n)$, and we have

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

Proof If $\lim_{n\to\infty} b_n = 0$, then b is a null-sequence. In the previous section we showed, that the product of a bounded sequence and a null-sequence is a null-sequence. Thus, $a \cdot b$ is a null-sequence, and the two formulas obviously hold.

We may therefore assume that $c = \lim_{n\to\infty} b_n > 0$. We write $a \cdot b = ca + a(b - c)$ and observe that $b - c$ is a null-sequence. Therefore, by boundedness of a, the product $a(b - c)$ is also a null-sequence. We now use Proposition 3.1.1 to obtain

$$\limsup_{n\to\infty}(a_n b_n) = \limsup_{n\to\infty}(ca_n + a_n(b_n - c)) = \limsup_{n\to\infty}(ca_n)$$
$$= c\limsup_{n\to\infty} a_n = \limsup_{n\to\infty} a_n \cdot \lim_{n\to\infty} b_n.$$

The first formula in the proposition follows. The proof of the second formula is analogous.

Remark By simple induction, we have

$$\lim_{n\to\infty} a_n^m = \left(\lim_{n\to\infty} a_n\right)^m, \quad m \in \mathbb{N},$$

provided that $\lim_{n\to\infty} a_n$ exists.

Proposition 3.1.3 *Let $a, b : \mathbb{N}_0 \to \mathbb{R}$ be real sequences. Assume that $\lim_{n\to\infty} a_n$ exists, the sequence b consists of non-zero elements, and $\lim_{n\to\infty} b_n$ exists and is non-zero. Then, we have*

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

Proof We first claim that the sequence $1/b = (1/b_n)_{n\in\mathbb{N}_0}$ is bounded.

Since b is not a null-sequence, $|b|$ is bounded away from zero, that is, there exists $0 < \epsilon$ and $N \in \mathbb{N}_0$ such that $\epsilon < |b_n|$ for $n \geq N$. Thus, we have $|1/b_n| < 1/\epsilon$ for $n \geq N$. Adjoining the first N elements, we obtain

$$|1/b_n| \leq \max\left(|1/b_0|, |1/b_1|, \ldots, |1/b_{N-1}|, 1/\epsilon\right), \quad n \in \mathbb{N}_0.$$

Boundedness of the sequence $1/b$ follows.

Since a convergent sequence is bounded (as it is Cauchy) and the product of bounded sequences is bounded, we obtain that the sequence a/b is also bounded. We now apply Proposition 3.1.2 to the product $a = (a/b) \cdot b$. We have

$$\lim_{n\to\infty} a_n = \lim_{n\to\infty}((a_n/b_n)\cdot b_n) = \limsup_{n\to\infty}((a_n/b_n)\cdot b_n) = \limsup_{n\to\infty}(a_n/b_n)\cdot \lim_{n\to\infty} b_n.$$

The same holds if lim sup is replaced by lim inf. The proposition follows.

Proposition 2.3.2 asserts that every real sequence has a monotonic subsequence. Non-monotonic (convergent) sequences, however, arise naturally. Going beyond trivial examples such as $((-1)^n/n)_{n\in\mathbb{N}}$, we introduce here the **Fibonacci sequence** whose consecutive ratios form a rational sequence. As we will see below, this latter sequence splits into an increasing and a decreasing subsequence both converging to the same irrational number.

Example 3.1.2 The sequence of Fibonacci numbers is defined inductively as

$$F_0 = 0, \quad F_1 = 1, \quad \text{and} \quad F_{n+1} = F_n + F_{n-1}, \ \text{for } n \in \mathbb{N}.$$

History
We noted previously the Italian mathematician Leonardo Pisano Bigollo, "Fibonacci" for short. He is best known for the sequence above named after him; this sequence is contained in his book *Liber Abaci* (1202). Note, however, that the Fibonacci sequence was known to Indian mathematicians around the 6th century. Fibonacci introduced this sequence as the pattern of growth of a population of pairs of rabbits. He assumed that every time a new pair (male and female) of rabbits is born, they mature in a month and then produce another pair of rabbits. If no rabbits ever die, he asked: How many pairs of rabbits will there be after a year?
The inductive definition of the sequence can clearly be seen here. If F_n is the number of pairs of rabbits at the end of the nth month, then F_{n+1} is equal to the new pairs F_{n-1} plus F_n, the number of pairs existing at the end of the nth month.

The Fibonacci numbers satisfy many identities. (See the exercises at the end of this section.) For our purposes, we only need **Cassini's Identity**

$$F_{n+1}F_{n-1} - F_n^2 = (-1)^n, \quad n \in \mathbb{N}.$$

We show this by using Peano's Principle of Induction with respect to $n \in \mathbb{N}$.

For the initial step, $n = 1$, we have $F_2F_0 - F_1^2 = -1$ and the identity holds.

For the general induction step $n \Rightarrow n + 1$, we assume that Cassini's identity is valid for n, start with this, and calculate

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

The last equality is Cassini's identity for $n + 1$. The general induction step is completed, and the identity follows.

Cassini's identity has many interesting consequences. First, we define the ratio

$$r_n = \frac{F_{n+1}}{F_n}, \quad n \in \mathbb{N}.$$

Dividing both sides of Cassini's identity (for n) by $F_{n-1}F_n$, we obtain the 1-step difference formula

$$r_n - r_{n-1} = \frac{(-1)^n}{F_{n-1}F_n}.$$

Moving up the value of n by one, we have

$$r_{n+1} - r_n = \frac{(-1)^{n+1}}{F_nF_{n+1}}.$$

Adding these two, we obtain the 2-step difference formula

$$r_{n+1} - r_{n-1} = \frac{(-1)^n}{F_n}\left(\frac{1}{F_{n-1}} - \frac{1}{F_{n+1}}\right) = \frac{(-1)^n}{F_n}\,\frac{F_{n+1} - F_{n-1}}{F_{n-1}F_{n+1}} = \frac{(-1)^n}{F_{n-1}F_{n+1}},$$

where we used the defining equation $F_{n+1} = F_n + F_{n-1}$.

For $n = 2k$ even, this gives $r_{2k+1} - r_{2k-1} = 1/(F_{2k-1}F_{2k+1}) > 0$, and hence $r_{2k-1} < r_{2k+1}$. For $n = 2k+1$ odd, we have $r_{2k+2} - r_{2k} = -1/(F_{2k}F_{2k+2}) < 0$, and hence $r_{2k+2} < r_{2k}$. Returning to the original 1-step difference formula above, $n = 2k+2$ gives $r_{2k+2} - r_{2k+1} = 1/(F_{2k+1}F_{2k+2}) > 0$.

Putting all these together, we arrive at

$$r_{2k-1} < r_{2k+1} < r_{2k+2} < r_{2k}, \quad k \in \mathbb{N}.$$

We conclude that the odd-member subsequence $(r_1, r_3, r_5, \ldots)$ is strictly increasing and the even-member subsequence $(r_2, r_4, r_6, \ldots)$ is strictly decreasing. Finally, by completeness of $\mathbb{R}$, these two subsequences approach a **unique** real number τ, since, by the above, the even-odd differences approach zero (since the Fibonacci sequence is unbounded).

It is easy to find the value of the limit τ. Dividing through the defining equation $F_{n+1} = F_n + F_{n-1}$ by F_{n-1}, we have

$$\frac{F_{n+1}}{F_{n-1}} = \frac{F_{n+1}}{F_n} \cdot \frac{F_n}{F_{n-1}} = \frac{F_n}{F_{n-1}} + 1,$$

or equivalently,

$$r_n \cdot r_{n-1} = r_{n-1} + 1, \quad 2 \leq n \in \mathbb{N}.$$

Now, taking the limit on both sides as $n \to \infty$, and using Proposition 3.1.2 along with the fact that $\lim_{n\to\infty} r_{n-1} = \lim_{n\to\infty} r_n = \tau$, we obtain

$$\tau^2 = \tau + 1.$$

In other words, τ is the unique **positive** solution of the quadratic equation $x^2 - x - 1 = 0$. The Quadratic Formula[1] now gives

$$\tau = \frac{1 + \sqrt{5}}{2}.$$

This is the famous **golden number**[2] (or **golden ratio** or Euclid's **extreme and mean ratio**) of Greek antiquity.

[1] Here we use the well-known formula $(-b \pm \sqrt{b^2 - 4ac})/(2a)$ giving the two zeros of the quadratic polynomial $ax^2 + bx + c$. A full analysis of this is in Section 6.6.

[2] For a thorough discussion on the golden number, see the author's *Glimpses of Algebra and Geometry*, 2nd ed. Springer, New York, 2002.

History
The German mathematician and astronomer Johannes Kepler (1571–1630) noted that the ratio of consecutive Fibonacci numbers is " as 5 to 8 so is 8 to 13, practically, and as 8 is to 13, so is 13 to 21 almost," finally concluding that the ratios get closer and closer to the golden number. Note the vagueness of the concept of convergence predating the precise definition about two centuries.

The next example shows the interesting fact that, given an arithmetic sequence with integral difference, the square root of a natural number gets arbitrarily close to one of the members of the sequence.

Example 3.1.3 Let $0 < \epsilon \in \mathbb{R}$ and $d \in \mathbb{N}$. Show that

$$\epsilon < \left|\sqrt{m} - d \cdot n\right| < 2\epsilon$$

for some $m, n \in \mathbb{N}$.[3]

The proof is "ad hoc" and involves several careful choices. First, the expression within the absolute value above will be compared to a choice of a rational number $a/b \in \mathbb{Q}, a, b \in \mathbb{N}$ (comparable to $\epsilon/2$) as $2\epsilon < a/b < 4\epsilon$. We let $k \in \mathbb{N}$ (eventually large) and define $n = kb$ and $m = (dkb)^2 + dka$. With these choices, we need to estimate $0 < \sqrt{m} - dn = \sqrt{(dkb)^2 + dka} - dkb$ as $k \to \infty$. We now "rationalize the last radical expression" as

$$0 < \sqrt{m} - dn = \sqrt{(dkb)^2 + dka} - dkb = dkb\left(\sqrt{1 + \frac{a}{dkb^2}} - 1\right)$$

$$= dkb \frac{\frac{a}{dkb^2}}{\sqrt{1 + \frac{a}{dkb^2}} + 1} = \frac{a}{b} \frac{1}{\sqrt{1 + \frac{a}{dkb^2}} + 1} < \frac{a}{2b} < 2\epsilon.$$

The stated upper estimate follows. Since

$$\lim_{k \to \infty} \frac{1}{\sqrt{1 + \frac{a}{dkb^2}} + 1} = \frac{1}{2},$$

for k large enough, the lower estimate also follows.

We now return to the geometric sequence studied in Section 2.3. Let $|r| < 1$, $r \in \mathbb{R}$. We claim

$$\lim_{n \to \infty} \sum_{k=0}^{n} r^k = \lim_{n \to \infty} (1 + r + r^2 + \cdots + r^n) = \frac{1}{1 - r}.$$

[3]This was a problem in the Duke 2012 William Lowell Putnam Mathematical Competition Preparation.

To do this, we first state the **Finite Geometric Series Formula**:

$$\sum_{k=0}^{n} r^k = 1 + r + r^2 + \cdots + r^n = \frac{1 - r^{n+1}}{1 - r}, \quad r \neq 1.$$

There are several classical proofs of this; the quickest is by induction.

For $n = 0$, the formula is obvious. For the general induction step $n \Rightarrow n + 1$, we calculate

$$1 + r + r^2 + \cdots + r^n + r^{n+1} = \frac{1 - r^{n+1}}{1 - r} + r^{n+1} = \frac{1 - r^{n+2}}{1 - r}.$$

The formula follows.

According to our study of the geometric sequence in Section 2.3, for $|r| < 1$, we have $\lim_{n\to\infty} r^n = \lim_{n\to\infty} r^{n+1} = 0$. Using this in the Finite Geometric Series Formula, the stated limit relation above follows.

This limit is usually written in an infinite series form[4] called the **Infinite Geometric Series Formula**:

$$\sum_{n=0}^{\infty} r^n = 1 + r + r^2 + \cdots + r^n + \cdots = \frac{1}{1 - r}, \quad |r| < 1.$$

Remark The Finite Geometric Series Formula also shows that, if $r \geq 1$, then

$$\sum_{n=0}^{\infty} r^n = 1 + r + r^2 + \cdots + r^n + \cdots = \infty.$$

History

According to legends, a king gave the inventor of chess (possibly the ancient Indian Brahmin mathematician Sessa) the right to name his prize for the new game who then asked for an amount of grains of wheat (or rice) as follows. Place 1 grain of wheat on one square of a chess board, 2 on another, then 4, etc. each time the double amount of what has been placed on a previous square. Unaware of geometric progression, the king readily agreed. On an 8×8 chessboard, there are 64 squares so that the amount of grain requested by the inventor was

$$1 + 2 + 4 + 8 + \cdots + 2^{63} = 8^{64} - 1 = 18,446,744,073,709,551,615,$$

where we used the Finite Geometric Series Formula above. Taking 25 mg as the average weight of a grain of wheat, this amounts to approximately 461,168,601,842.73 metric tons of wheat. For comparison, this is almost 971 times the world rice production in 2014/2015 (approximately 475.04 million metric tons).

[4] If $(a_n)_{n\in\mathbb{N}_0}$ is a sequence, then we write $\sum_{n=0}^{\infty} a_n = \lim_{n\to\infty}(a_0 + a_1 + \cdots + a_n)$.

Borrowing a term in music, the next example should be considered as a "variation on the theme."

Example 3.1.4 Let $|r|, |s| < 1$, $r, s \in \mathbb{R}$. Calculate the limit

$$\lim_{n\to\infty} \frac{1 + r + r^2 + \cdots + r^n}{1 + s + s^2 + \cdots + s^n}.$$

Using the Finite Geometric Series above, we have

$$\lim_{n\to\infty} \frac{1 + r + r^2 + \cdots + r^n}{1 + s + s^2 + \cdots + s^n} = \lim_{n\to\infty} \frac{\frac{1-r^{n+1}}{1-r}}{\frac{1-s^{n+1}}{1-s}} = \frac{1-s}{1-r}.$$

We now briefly revisit the infinite decimal representation of a real number in Section 2.2. As noted there, a real number can be represented as an infinite sum:

$$a.d_1d_2d_3\ldots = a + \frac{d_1}{10} + \frac{d_2}{10^2} + \frac{d_3}{10^3} + \cdots,$$

where the decimal digits $d_1, d_2, d_3, \ldots$ range from 0 to 9.

We showed that a real number is rational if and only if its infinite decimal representation ends with a repeating pattern or if it terminates. As an application of the previous example, we now derive this in a less ad hoc manner. As in Section 2.2, we may start with the reduced repeating pattern

$$0.\overline{d_1d_2\ldots d_k} = d_1d_2\ldots d_k\left(\frac{1}{10^k} + \frac{1}{10^{2k}} + \cdots\right) = \frac{d_1d_2\ldots d_k}{10^k}\left(1 + \frac{1}{10^k} + \left(\frac{1}{10^k}\right)^2 + \cdots\right).$$

Since $1/10^k < 1$, the Infinite Geometric Series Formula applies. We obtain

$$0.\overline{d_1d_2\cdots d_k} = \frac{d_1d_2\cdots d_k}{10^k}\left(1 + \frac{1}{10^k} + \left(\frac{1}{10^k}\right)^2 + \cdots\right)$$

$$= \frac{d_1d_2\ldots d_k}{10^k}\frac{1}{1-\frac{1}{10^k}} = \frac{d_1d_2\ldots d_k}{10^k}\frac{10^k}{10^k-1} = \frac{d_1d_2\ldots d_k}{10^k-1}.$$

This is the formula that we arrived at in Section 2.2 using ad hoc methods.

Example 3.1.5 We have

$$\sum_{n=1}^{\infty} nr^n = \frac{r}{(1-r)^2}, \quad |r| < 1.$$

We calculate

$$\sum_{n=1}^{\infty} nr^n = \sum_{n=1}^{\infty} \overbrace{(r^n + \cdots + r^n)}^{n}$$

$$= r + (r^2 + r^2) + (r^3 + r^3 + r^3) + \cdots + (r^n + r^n + \cdots + r^n) + \cdots$$

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

Remark Once we know that the sum $\sum_{n=1}^{\infty} nr^n$, $|r| < 1$, is finite, there is a simpler way to determine its value. Letting $S(r)$ denote this sum, we calculate

$$S(r) = \sum_{n=1}^{\infty} nr^n = \sum_{n=1}^{\infty} r^n + \sum_{n=2}^{\infty} (n-1)r^n = \frac{r}{1-r} + r \sum_{n=2}^{\infty} (n-1)r^{n-1}$$

$$= \frac{r}{1-r} + r \sum_{n=1}^{\infty} nr^n = \frac{r}{1-r} + rS(r),$$

where we used the Finite Geometric Series Formula. Rearranging, we obtain $S(r) = r/(1-r)^2$.

Example 3.1.6 For $n \in \mathbb{N}$, we let

$$H_n = 1 + \frac{1}{2} + \frac{1}{3} + \cdots + \frac{1}{n}.$$

We claim

$$\sum_{n=1}^{\infty} \frac{1}{n} = \lim_{n \to \infty} H_n = \infty.$$

To show this, we first note that the sequence $(H_n)_{n \in \mathbb{N}}$ is strictly increasing so that it is either convergent to a finite limit or unbounded. For $n \in \mathbb{N}$, we calculate

$$H_{2^{n+1}} - H_{2^n} = \frac{1}{2^n + 1} + \frac{1}{2^n + 2} + \cdots + \frac{1}{2^{n+1}} > 2^n \frac{1}{2^{n+1}} = \frac{1}{2},$$

where we estimated the $2^{n+1} - 2^n = 2^n$ terms by the last (smallest) term $1/2^{n+1}$. Thus, for $n \in \mathbb{N}$, we have

$$H_{2^n} = (H_{2^n} - H_{2^{n-1}}) + (H_{2^{n-1}} - H_{2^{n-2}}) + \cdots + (H_2 - H_1) + H_1 \geq n \frac{1}{2} + 1.$$

By monotonicity of the limit, we obtain

$$\lim_{n\to\infty} H_n = \lim_{n\to\infty} H_{2^n} \geq \lim_{n\to\infty} \left(\frac{n}{2} + 1\right) = \infty.$$

The claim follows.

The infinite series

$$\sum_{n=1}^{\infty} \frac{1}{n} = \infty$$

in the example above is called the **harmonic series**.

Remark The partial sums of the harmonic series increase very slowly; for example, we have $H_{10^6} = 14.39272672\ldots$ and $H_{10^9} = 21.30048150\ldots$.

A variation on the theme is the following:

Example 3.1.7 For $2 \leq n \in \mathbb{N}$, find a formula for the sum

$$\frac{1}{[\sqrt{1}]} + \frac{1}{[\sqrt{2}]} + \frac{1}{[\sqrt{3}]} + \cdots + \frac{1}{[\sqrt{n^2-1}]},$$

in terms of H_n, $n \in \mathbb{N}$, where $[x]$, $x \in \mathbb{R}$, is the greatest integer $\leq x$.

As we showed in Section 2.1, for $k \in \mathbb{N}$, the square root $\sqrt{k}$ is a natural number if and only if $k = m^2$, the square of a natural number $m \in \mathbb{N}$. By the definition of the greatest integer function and monotonicity of the square root, the value $[\sqrt{k}]$ is the constant $m \in \mathbb{N}$ precisely when $m^2 \leq k \leq (m+1)^2 - 1$. This happens exactly $(m+1)^2 - 1 - m^2 + 1 = 2m + 1$ times, and therefore these terms contribute to the sum $(2m+1)/m$. Since $1 \leq m \leq n-1$, we obtain that the sum above is equal to

$$\frac{3}{1} + \frac{5}{2} + \frac{7}{3} + \cdots + \frac{2n-1}{n-1}.$$

We obtain

$$\frac{1}{[\sqrt{1}]} + \frac{1}{[\sqrt{2}]} + \frac{1}{[\sqrt{3}]} + \cdots + \frac{1}{[\sqrt{n^2-1}]} = 2(n-1) + H_{n-1}, \quad 2 \leq n \in \mathbb{N}.$$

Example 3.1.8 For each non-empty subset $A \subset \{1, 2, \ldots, n\}$, $n \in \mathbb{N}$, consider the ratio Σ_A/Π_A, where $\Sigma_A = \Sigma_{a\in A} a$, resp. $\Pi_A = \Pi_{a\in A} a$, is the sum, resp. product, of the elements in A.[5] Determine the following sums:

$$S_n = \sum_{\emptyset \neq A \subset \{1,\ldots,n\}} \frac{\Sigma_A}{\Pi_A} \quad \text{and} \quad P_n = \sum_{\emptyset \neq A \subset \{1,\ldots,n\}} \frac{1}{\Pi_A}$$

[5]The first part of this problem with explicit final formulas was in the USA Mathematics Olympiad, 1991.

$$S_n^{\wedge} = \sum_{\emptyset \neq A \subset \{1,\ldots,n\}} (-1)^{|A|} \frac{\Sigma_A}{\Pi_A} \quad \text{and} \quad P_n^{\wedge} = \sum_{\emptyset \neq A \subset \{1,\ldots,n\}} (-1)^{|A|} \frac{1}{\Pi_A}.$$

First, we determine P_n and $P_n^{\wedge}$, $n \in \mathbb{N}$. Clearly, we have $P_1 = 1$ and $P_1^{\wedge} = -1$, since the only non-empty subset of the set $\{1\}$ is the whole set itself.

We now notice that

$$P_n = \left(1 + \frac{1}{1}\right) \cdot \left(1 + \frac{1}{2}\right) \cdots \left(1 + \frac{1}{n}\right) - 1 = \frac{2}{1} \cdot \frac{3}{2} \cdots \frac{n+1}{n} - 1 = (n+1) - 1 = n.$$

This is because, expanding the parentheses, we obtain $1 = 1 \cdot 1 \cdots 1$ corresponding to the empty set $\emptyset$ (which is deducted) and products of the form

$$\frac{1}{i_1 \cdot i_2 \cdots i_k}, \quad 1 \leq i_1 < i_2 < \ldots < i_k \leq n, \ k = 1, 2, \ldots, n,$$

corresponding to the non-empty subset $A = \{i_1, i_2, \ldots, i_k\} \subset \{1, 2, \ldots, n\}$.

Similarly, for the alternating sum, we have

$$P_n^{\wedge} = \left(1 - \frac{1}{1}\right) \cdot \left(1 - \frac{1}{2}\right) \cdots \left(1 - \frac{1}{n}\right) - 1 = -1,$$

since, upon expanding, we obtain 1 and products of the form

$$\frac{(-1)^k}{i_1 \cdot i_2 \cdots i_k}, \quad 1 \leq i_1 < i_2 < \ldots < i_k \leq n, \ k = 1, 2, \ldots, n,$$

corresponding to the non-empty subset $A = \{i_1, i_2, \ldots, i_k\} \subset \{1, 2, \ldots, n\}$.

Second, we derive inductive formulas for S_n and $S_n^{\wedge}$, $n \in \mathbb{N}$. Clearly, we have $S_1 = 1$ and $S_1^{\wedge} = -1$, since the only non-empty subset of the set $\{1\}$ is the whole set itself. We claim

$$S_{n+1} = \left(1 + \frac{1}{n+1}\right) S_n + n + 1 \quad \text{and} \quad S_{n+1}^{\wedge} = \left(1 - \frac{1}{n+1}\right) S_n^{\wedge}, \quad n \in \mathbb{N}.$$

There are three types of non-empty subsets in $\{1, \ldots, n, n+1\}$, $n \in \mathbb{N}$. If the subset does not contain the element $n+1$, then it is a non-empty subset $A \subset \{1, \ldots, n\}$. If it contains the element $n+1$, then it can be of the form $A \cup \{n+1\}$, where $A \subset \{1, \ldots, n\}$ is non-empty, or it can be the singleton $A = \{n+1\}$.

The respective ratios of the first type of subsets add up to S_n. The ratio of the second type of subset $A \cup \{n+1\}$, $\emptyset \neq A \subset \{1, \ldots, n\}$, is calculated as

$$\frac{\Sigma_{A \cup \{n+1\}}}{\Pi_{A \cup \{n+1\}}} = \frac{\Sigma_A + (n+1)}{\Pi_A \cdot (n+1)} = \frac{\Sigma_A}{\Pi_A} \cdot \frac{1}{n+1} + \frac{1}{\Pi_A}.$$

The sum of these ratios gives $S_n/(n+1) + P_n$. Finally, the ratio corresponding to the third type of subset $A = \{n+1\}$ is equal to $(n+1)/(n+1) = 1$. Adding these, we obtain

$$S_{n+1} = S_n + \frac{S_n}{n+1} + n + 1, \quad n \in \mathbb{N},$$

where we used that $P_n = n$. The first inductive formula follows.

The proof of the second inductive formula is similar. For the second type of subset $A \cup \{n+1\}$, $\emptyset \neq A \subset \{1, \ldots, n\}$, we have

$$(-1)^{|A|+1} \frac{\Sigma_{A\cup\{n+1\}}}{\Pi_{A\cup\{n+1\}}} = (-1)^{|A|+1} \frac{\Sigma_A + (n+1)}{\Pi_A \cdot (n+1)} = -(-1)^{|A|} \frac{\Sigma_A}{\Pi_A} \cdot \frac{1}{n+1} - (-1)^{|A|} \frac{1}{\Pi_A}.$$

The sum of these ratios gives $-S_n^\wedge/(n+1) - P_n^\wedge$. Adding these, we obtain

$$S_{n+1}^\wedge = S_n^\wedge - \frac{S_n^\wedge}{n+1}, \quad n \in \mathbb{N},$$

where we used that $P_n^\wedge = -1$. The second inductive formula follows.

To solve the first inductive formula (for S_n, $n \in \mathbb{N}$), we claim

$$H_n = n + 1 - \frac{S_n + 1}{n+1}, \quad n \in \mathbb{N}.$$

Clearly, $H_1 = 2 - 2/2 = 1$. Moreover, we have

$$\begin{aligned} H_{n+1} &= n + 2 - \frac{S_{n+1} + 1}{n+2} = n + 2 - \frac{\frac{n+2}{n+1} S_n + n + 2}{n+2} = n + 1 - \frac{S_n}{n+1} \\ &= n + 1 - \frac{S_n + 1}{n+1} + \frac{1}{n+1} = H_n + \frac{1}{n+1}, \end{aligned}$$

and the claim now follows by simple induction. Playing this back to S_n, we finally obtain

$$S_n = n^2 + 2n - (n+1)H_n, \quad n \in \mathbb{N}.$$

The solution to the second recurrence formula is simpler. After rearranging, we obtain

$$(n+1)S_{n+1}^\wedge = nS_n^\wedge, \quad n \in \mathbb{N}.$$

This means that the expression on either side is constant and therefore equal to $S_1^\wedge = -1$. We get

$$S_n^{\wedge} = -\frac{1}{n}.$$

The example follows.

We now leave the harmonic series and consider the sum of **squares** of reciprocals of positive integers. In contrast to the harmonic series, we have the following:

Example 3.1.9 We have

$$\sum_{n=1}^{\infty} \frac{1}{n^2} = \lim_{n\to\infty} \left(1 + \frac{1}{2^2} + \frac{1}{3^2} + \cdots + \frac{1}{n^2}\right) \leq 2.$$

First, the sequence under the limit is strictly increasing, and so, like in the previous case, the limit is either finite or infinite. We claim that it is finite.

Indeed, we estimate the terms using the following:

$$\frac{1}{k^2} < \frac{1}{k(k-1)} = \frac{1}{k-1} - \frac{1}{k}, \quad 2 \leq k \in \mathbb{N}.$$

Using this for each term, we have

$$1 + \frac{1}{2^2} + \frac{1}{3^2} + \cdots + \frac{1}{n^2} < 1 + \left(1 - \frac{1}{2}\right) + \left(\frac{1}{2} - \frac{1}{3}\right) + \cdots + \left(\frac{1}{n-1} - \frac{1}{n}\right) = 2 - \frac{1}{n}.$$

Monotonicity of the limit now gives

$$\lim_{n\to\infty} \left(1 + \frac{1}{2^2} + \frac{1}{3^2} + \cdots + \frac{1}{n^2}\right) \leq \lim_{n\to\infty} \left(2 - \frac{1}{n}\right) = 2.$$

Remark 1 In Example 2.3.2, we noted that the defining condition for a Cauchy sequence cannot be replaced by the condition $\inf_{N\in\mathbb{N}_0} \sup_{n\geq N} |a_{n+1} - a_n| = 0$. On the other hand, if we impose the condition $|a_{n+1} - a_n| \leq 1/n^2$, $n \in \mathbb{N}$, then the sequence $(a_n)_{n\in\mathbb{N}}$ becomes convergent. This follows from the estimate in the example above. Indeed, for $m \geq n \geq 2$, $m, n \in \mathbb{N}$, we calculate

$$\begin{aligned} |a_{m+1} - a_n| &\leq |a_{m+1} - a_m| + |a_m - a_{m-1}| + \cdots + |a_{n+1} - a_n| \\ &\leq \frac{1}{m^2} + \frac{1}{(m-1)^2} + \cdots + \frac{1}{n^2} \\ &\leq \left(\frac{1}{m-1} - \frac{1}{m}\right) + \left(\frac{1}{m-2} - \frac{1}{m-1}\right) + \cdots + \left(\frac{1}{n-1} - \frac{1}{n}\right) = \frac{1}{n-1} - \frac{1}{m}. \end{aligned}$$

This shows that the sequence $(a_n)_{n\in\mathbb{N}}$ is Cauchy, thereby convergent.

Remark 2 In 1735, Euler announced[6] that

$$\sum_{n=1}^{\infty} \frac{1}{n^2} = \frac{\pi^2}{6}.$$

We will give an elementary proof of this in Section 11.7.

History
Calculating the exact value of this sum was first posed by Pietro Mengoli in 1644. It has baffled the leading mathematicians of the time, such as the Bernoulli family, for almost a century. It has brought international fame to the 28-year old Euler, and his solution was read in 1735 in the Saint Petersburg Academy of Sciences. Euler used some methods that were not justified at the time, but within six years he was able to provide a rigorous proof. This problem was subsequently named after his hometown (and that of the Bernoulli's) as the **Basel problem**.

As a straightforward generalization (in the use of the monotonicity of the limit), for $2 \leq p \in \mathbb{N}$, the infinite series

$$\sum_{n=1}^{\infty} \frac{1}{n^p} = \lim_{n\to\infty} \left(1 + \frac{1}{2^p} + \frac{1}{3^p} + \cdots + \frac{1}{n^p}\right)$$

converges. This is called the p-**series**.

In the next section, we will show that the p-series converges for any **real** $1 < p \in \mathbb{R}$. At present, analogously to the previous example, we show that the special case $p = 3/2$ can be done by simple estimates.

Example 3.1.10 We have

$$1 + \frac{1}{2\sqrt{2}} + \cdots + \frac{1}{n\sqrt{n}} \leq 3 - \frac{2}{\sqrt{n}}, \quad n \in \mathbb{N}.$$

As a consequence, we have

$$\sum_{n=1}^{\infty} \frac{1}{n\sqrt{n}} \leq 3.$$

To derive this inequality, we use Peano's Principle of Induction. For $n = 1$, the equality holds. For the general induction step $n \Rightarrow n + 1$, by the induction hypothesis, we have

$$1 + \frac{1}{2\sqrt{2}} + \cdots + \frac{1}{n\sqrt{n}} + \frac{1}{(n+1)\sqrt{n+1}} \leq 3 - \frac{2}{\sqrt{n}} + \frac{1}{(n+1)\sqrt{n+1}}$$

[6] Unlike the harmonic series, this series converges fast; for example, the first one thousand terms differ from $\pi^2/6$ by an error of 0.0009995001667.

Hence, it is enough to show that

$$3 - \frac{2}{\sqrt{n}} + \frac{1}{(n+1)\sqrt{n+1}} \leq 3 - \frac{2}{\sqrt{n+1}}.$$

Rearranging, we obtain

$$\frac{2n+3}{2n+2} \leq \sqrt{\frac{n+1}{n}}.$$

Squaring and eliminating the denominators, the inequality easily follows.

A (much simpler) variation on the theme is the following:

Example 3.1.11 Show that

$$\sum_{n=1}^{\infty} \frac{1}{(n+1)\sqrt{n} + n\sqrt{n+1}} = 1.$$

For $n \in \mathbb{N}$, we have

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

Hence the partial (finite) sums are telescopic, and we obtain

$$\sum_{n=1}^{N} \frac{1}{(n+1)\sqrt{n} + n\sqrt{n+1}} = 1 - \frac{1}{\sqrt{N+1}}, \quad N \in \mathbb{N}.$$

Letting $N \to \infty$, the example follows.

Exercises

3.1.1. Let $(a_n)_{n\in\mathbb{N}}$ and $(b_n)_{n\in\mathbb{N}}$ be real sequences such that $\lim_{n\to\infty}(a_n + b_n) = 2$ and $\lim_{n\to\infty}(a_n \cdot b_n) = 1$. Show that $\lim_{n\to\infty} a_n = \lim_{n\to\infty} b_n = 1$.

3.1.2. Let $(a_n)_{n\in\mathbb{N}}$ be a sequence with positive terms. Show that (a) $\sum_{n=1}^{\infty} a_n$ finite implies that $\sum_{n=1}^{\infty} \sqrt{a_n a_{n+1}}$ is also finite, but (b) the converse is false.

3.1.3. Let $(a_n)_{n\in\mathbb{N}_0}$ be an arithmetic sequence with difference $d \in \mathbb{R}$ and $(b_n)_{n\in\mathbb{N}_0}$ a geometric sequence with ratio $r \in \mathbb{R}$ such that $|r| < 1$. Show that

$$\sum_{n=0}^{\infty} a_n b_n = \frac{a_0 b_0}{1-r} + \frac{d b_0 r}{(1-r)^2}.$$

3.1.4. Let $a_0, a_1 \in \mathbb{R}$ and $0 < t < 1$, $t \in \mathbb{R}$, and define the real sequence $(a_n)_{n\in\mathbb{N}_0}$ inductively by $a_{n+1} = (1-t)a_n + ta_{n-1}$, $n \in \mathbb{N}$. Show that

$$\lim_{n\to\infty} a_n = \frac{ta_0 + a_1}{t+1}.$$

3.1.5. Let $n \in \mathbb{N}$. Determine the number of subsets of the set $\{1, 2, \ldots, 2n\}$ with the property that the sum of the smallest and largest elements is equal to $2n+1$.

3.1.6. Show that, for appropriate ranges $m, n \in \mathbb{N}$, the Fibonacci numbers (Example 3.1.2) satisfy the following identities:

i. $\displaystyle\sum_{k=1}^{n} F_k = F_{n+2} - 1$

ii. $\displaystyle\sum_{k=0}^{n-1} F_{2k+1} = F_{2n}$

iii. $\displaystyle\sum_{k=1}^{n} F_{2k} = F_{2n+1} - 1$

iv. $\displaystyle\sum_{k=1}^{n} F_k^2 = F_n F_{n+1}$

v. $F_n^2 - F_{n+m}F_{n-m} = (-1)^{n-m}F_m^2$

vi. $F_{3n} = F_{n+1}^3 + F_n^3 - F_{n-1}^3$

vii. $F_{3n+1} = F_{n+1}^3 + 3F_{n+1}F_n^2 - F_n^3$

viii. $F_{3n+2} = F_{n+1}^3 + 3F_{n+1}^2F_n + F_n^3$

3.1.7. Given $n \in \mathbb{N}$, show that the number of ways to split n as a sum of 1's and 2's is F_{n+1}. (For example, we have $1+1+1=3$, $1+2=3$, $2+1=3$ giving $F_4 = 3$.)

3.1.8. Show that F_n, $n \in \mathbb{N}$, is the number of n-digit binary integers[7] that have no consecutive zeros.

3.1.9. Derive Binet's formula

$$F_n = \frac{\tau^n - (-1)^n(1/\tau)^n}{\tau + 1/\tau},$$

where τ is the golden number (see Example 3.1.2).

[7] Sequences of n digits of 0's and 1's, starting with 1.

3.1.10. Derive the identity

$$F_{m+n+1} = F_{m+1}F_{n+1} + F_m F_n, \quad m, n \geq 0, \ m, n \in \mathbb{Z}.$$

As special cases, obtain the identities

$$F_{2n+1} = F_{n+1}^2 + F_n^2 \quad \text{and} \quad F_{2n} = F_{n+1}^2 - F_{n-1}^2.$$

3.1.11. Show that (a) $F_m | F_{mn}$, $m, n \in \mathbb{N}$, (b) $\gcd(F_n, F_{n+1}) = 1$, $n \in \mathbb{N}$, and (c) $\gcd(F_m, F_n) = F_{\gcd(m,n)}$, $m, n \in \mathbb{N}$.

3.1.12. Derive a closed formula for the sum $S_n = 1 + 11 + 111 + \cdots + \overbrace{11\ldots1}^{n}$ using the Finite Geometric Series Formula.

3.1.13. Show that H_n is not an integer for $2 \leq n \in \mathbb{N}$.

3.2 Roots, Rational and Real Exponents

Let $m \in \mathbb{N}$. In this section we will show that if $0 < a \in \mathbb{R}$ is a positive real number, then there exists a unique positive real number $0 < b \in \mathbb{R}$ such that $b^m = a$. In this case, b is called the m**th root** of a, and it is denoted by $b = \sqrt[m]{a}$. We call m the **degree** of the root. This concept clearly extends to zero; the mth root of zero is zero itself.

If m is **odd**, then $b^m = a$, $0 < a, b \in \mathbb{R}$, implies $(-b)^m = -a$. This shows that, for m odd, any real number $a \in \mathbb{R}$ has a unique mth root, and $\sqrt[m]{a} = -\sqrt[m]{-a}$ defines the mth root of a negative real number.

If m is **even** and $b^m = a$, $0 < a, b \in \mathbb{R}$, then $b^m \geq 0$ so that we must have $a \geq 0$. We see that negative real numbers do not have even degree roots. On the other hand, for m even, $b^m = a$, $0 < a, b \in \mathbb{R}$, implies $(-b)^m = a$ so that, besides $\sqrt[m]{a}$, we can also define the **negative** mth root $-\sqrt[m]{a}$. With this, for m even, the mth roots of $0 < a \in \mathbb{R}$ are $\pm\sqrt[m]{a}$.

For $0 < a \in \mathbb{R}$ and any m (regardless the parity), the mth root $\sqrt[m]{a} > 0$ is usually called the **principal** m**th root** of a.

For $a \geq 0$, $\sqrt{a}$ is referred to as the **square root** of a. (The degree 2 is not indicated explicitly.) For $a \in \mathbb{R}$, $\sqrt[3]{a}$ is called the **cube root** of a. For specific $n \geq 4$, the nth root is referred to by the respective ordinal number; for example, $\sqrt[4]{2}$ is the **fourth** root of 2, $\sqrt[5]{3}$ is the **fifth** root of 3, etc.

History

According to Eratosthenes of Cyrene (c. 276–190 BCE), the citizens of the island of Delos, stricken by the plague around 430 BCE, consulted the oracle of Apollo for aid. The oracle's answer was that the Delians should build an altar of Apollo of the same cubic shape as the original altar but twice the volume. The Delians later asked Plato to clarify the meaning of this, and his answer was that "the oracle meant, not that the god wanted an altar of double the size, but that he wished, in setting them the task, to shame the Greeks for their neglect of mathematics and their contempt

of geometry." This came down to us as the **Delian problem**, and it essentially asks to construct (using a straightedge (unmarked ruler) and a compass, or other means, see later) a line segment with length $\sqrt[3]{2}$. (See also the epitaph to Chapter 8 as well as a note to the solution of the Delian problem by conics.)
Bhāskara II treated roots extensively in his work *Bijaganita*; he was also the first to recognize that a positive number has two square roots.
$\surd$ or $\sqrt{\ }$ is the **radical sign**, and a is the **radicand**. (The radical sign always stretches over the radicand even if the latter is long; for example, we write $\sqrt{365}$ and not $\sqrt{3}65$.) A symbol for the square root was depicted as an intricate R by Regiomontanus (1436–1476). In Cardano's *Ars Magna* a variant of R was also used for Radix (base), a Latin word for "root," to indicate square roots. The symbol $\surd$ used today first appeared in print in 1525 by Cristoph Rudolff's book of computing entitled *Behend und hübsch Rechnung durch die kunstreichen regeln Algebre so gemeinicklich die Coss genent werden* (Nimble and beautiful calculation via the artful rules of algebra [which] are commonly called "coss"[8]). It is probable but not universally accepted that he invented this symbol to resemble the lowercase "r" for radix.

Remark The **shifting mth root algorithm** extracts the mth root of a positive real number digit-by-digit and thereby produces a real number in infinite decimal representation. The existence of the mth root of a real number also follows from Newton's Method (applied to the polynomial equation $x^m = a$), which, in this case, requires a minor modification of the Babylonian Method.

We now show that the mth root of a positive real number exists.

Proposition 3.2.1 *Let $0 < a \in \mathbb{R}$ and $m \in \mathbb{N}$. Then there exists a unique $0 < b \in \mathbb{R}$ such that $b^m = a$.*

Proof We first show existence. Let $0 < a \in \mathbb{R}$, and consider the set

$$A = \{0 < r \in \mathbb{R} \mid r^m > a\}.$$

By the Bernoulli inequality, we have

$$(1 + a)^m \geq 1 + ma > a,$$

so that $1 + a \in A$. In particular, A is non-empty.

Since 0 is an obvious lower bound for A, completeness of $\mathbb{R}$ implies that the infimum of A exists. We let $0 \leq b = \inf A$. We claim that b is the desired mth root of a, so that $b^m = a$ holds. This will also give $b > 0$ (as $b = 0$ cannot happen).

For every $n \in \mathbb{N}$, $b + 1/n$ is not a lower bound for A, and hence there exists $r_n \in A$ such that $(b \leq) r_n < b + 1/n$, $n \in \mathbb{N}$. By monotonicity of the limit, we obtain $\lim_{n\to\infty} r_n = b$. Raising both sides to the mth power, we have

$$\lim_{n\to\infty} r_n^m = \left(\lim_{n\to\infty} r_n\right)^m = b^m.$$

[8]Islamic mathematicians such as Muhammad ibn Mūsā Al-Khwārizmī used the word "shai" (thing) for the indeterminate/variable. This in Latin gave rise to the word "res," and in Italian "cosa" (thing). Algebra in Italy became "l'arte della cosa," in England "cossike arte" (the rule of coss), and in Germany "die Coss."

On the other hand, since $r_n \in A$, we have $a < r_n^m$, $n \in \mathbb{N}$, and again monotonicity of the limit gives $a \leq \lim_{n\to\infty} r_n^m$. Combining these, we obtain $a \leq b^m$. As a byproduct, we have $b > 0$.

To finish the proof of the existence, we claim that $a \geq b^m$ holds. Assume not. Since $0 < a/b^m < 1$, we can choose

$$0 < \delta < \frac{b}{m}\left(1 - \frac{a}{b^m}\right).$$

We calculate

$$(b-\delta)^m > \left(b - \frac{b}{m}\left(1 - \frac{a}{b^m}\right)\right)^m = b^m\left(1 - \frac{1}{m}\left(1 - \frac{a}{b^m}\right)\right)^m \geq b^m\left(1 - \left(1 - \frac{a}{b^m}\right)\right) = a,$$

where in the last estimate we used the Bernoulli inequality with $-(1/m)(1 - a/b^m) \geq -1$). This shows that $b - \delta \in A$, a contradiction, since $b = \inf A$. Thus, we have $a \geq b^m$. Existence follows.

For unicity, let $0 < b, c \in \mathbb{R}$ such that $b^m = c^m = a$. We may assume $b \leq c$ (since otherwise we would swap b and c). We have $b^m \leq c^m$ so that equalities must hold. Unicity holds.

Remark Unicity also follows from the identity[9]

$$x^m - y^m = (x - y)(x^{m-1} + x^{m-2}y + \cdots + xy^{m-2} + y^{m-1}), \quad x, y \in \mathbb{R}.$$

The mth root satisfies several identities, and they are simple consequences of unicity and the analogous identities for integral exponents. For $m, n, \in \mathbb{N}$, we have the following

$$\sqrt[m]{ab} = \sqrt[m]{a}\sqrt[m]{b}, \quad 0 \leq a, b \in \mathbb{R};$$

$$\sqrt[m]{\frac{a}{b}} = \frac{\sqrt[m]{a}}{\sqrt[m]{b}}, \quad 0 \leq a \in \mathbb{R},\ 0 < b \in \mathbb{R};$$

$$\sqrt[n]{\sqrt[m]{a}} = \sqrt[mn]{a}, \quad 0 \leq a \in \mathbb{R}.$$

The roots of real numbers can be nicely incorporated into our exponential notations. From now on we assume that the base is non-zero. Recall the identity $(b^m)^n = b^{m\cdot n}, \quad b \in \mathbb{R}$, where $m, n \in \mathbb{Z}$. Now, if b is an mth root of a, then $a = b^m$, and the identity above (for $m \cdot n = 1$) suggests that we should define $\sqrt[m]{a} = a^{1/m}$. Taking integral powers of both sides would give us exponentiation with rational exponents. We make this more precise as follows. We represent a positive rational number $0 < q \in \mathbb{Q}$ as a fraction $q = m/n$, $m, n \in \mathbb{N}$ and define

$$a^q = a^{m/n} = \sqrt[n]{a^m}, \quad 0 < a \in \mathbb{R}.$$

[9] Identities will be treated in Chapter 6.

We claim that this definition does not depend on the specific representation of the rational number as a **fraction**; that is, for $q = m/n = m'/n', m, m', n, n' \in \mathbb{N}$, we have $\sqrt[n]{a^m} = \sqrt[n']{a^{m'}}$. By unicity, raising both sides to the exponent $mn' = m'n$, we must have equality

$$(\sqrt[n]{a^m})^{m'n} = ((\sqrt[n]{a^m})^n)^{m'} = a^{mm'} = ((\sqrt[n']{a^{m'}})^{n'})^m = (\sqrt[n']{a^{m'}})^{mn'},$$

which is indeed the case. Thus, positive rational exponents are well-defined.

Extension to negative exponents is straightforward requiring

$$a^{-q} = \frac{1}{a^q}, \quad 0 < q \in \mathbb{Q}.$$

This, along with $a^0 = 1, 0 < a \in \mathbb{R}$, defines the extension to rational exponents. For $0 < a, b \in \mathbb{R}$ and $p, q \in \mathbb{Q}$, we have the following Identities:

$$a^{p+q} = a^p \cdot a^q, \quad a^{p-q} = \frac{a^p}{a^q}, \quad (a^p)^q = a^{pq}, \quad (ab)^p = a^p \cdot b^p.$$

These identities can be established in a straightforward manner using the analogous identities for integral exponents and the properties of the roots.

Rational exponents exhibit monotonicity properties that are useful in computations. For $1 < a \in \mathbb{R}$, the power a^q is strictly increasing in $q \in \mathbb{Q}$. Similarly, for $0 < a < 1$, the power a^q is strictly decreasing in $q \in \mathbb{Q}$. These follow directly from the first and second identities.

Example 3.2.1 Let $0 < a, b, c \in \mathbb{R}$, and define $u = ab$, $v = bc$, and $w = ca$.[10] Express a, b, c in terms of u, v, w.

We have $uvw = a^2b^2b^2$ so that $abc = \sqrt{uvw}$. With this, we have

$$a = \frac{abc}{bc} = \frac{\sqrt{uvw}}{v} = \sqrt{\frac{uw}{v}}.$$

Similarly, we obtain

$$b = \frac{abc}{ac} = \frac{\sqrt{uvw}}{w} = \sqrt{\frac{uv}{w}} \quad \text{and} \quad c = \frac{abc}{ab} = \frac{\sqrt{uvw}}{u} = \sqrt{\frac{vw}{u}}.$$

In Section 2.1, we showed that, for $n \in \mathbb{N}$, the square root $\sqrt{n}$ is a rational number if and only if n is a perfect square. Using similar technique, we can show that the mth root $\sqrt[m]{n}$ of a natural number $n \in \mathbb{N}$ is a rational number if and only if n is a power of m, that is, $n = a^m$ for some $a \in \mathbb{N}$.

[10] There are many ways to solve this problem. All involve fractional exponents and their respective identities.

Indeed, assume that $\sqrt[m]{n}$, $m, n \in \mathbb{N}$, is a rational number a/b, $a, b \in \mathbb{N}$, $\gcd(a, b) = 1$. By definition, we have $(a/b)^m = n$. This gives $a^m/b = nb^{m-1} \in \mathbb{N}$. We conclude that b divides a^m. Since $\gcd(a, b) = 1$, by the corollary of Proposition 1.3.1, b must divide a. This is a contradiction unless $b = 1$. Thus, we have $\sqrt[m]{n} = a$, and the claim follows.

A variation on the theme in Example 2.3.8 is the following:

Example 3.2.2 Let $n_1, n_2, n_3 \in \mathbb{N}$ be distinct, and assume that the linear relation

$$c_1\sqrt[3]{n_1} + c_2\sqrt[3]{n_2} + c_3\sqrt[3]{n_3} = 0$$

holds for some non-zero **rational** coefficients $0 \neq c_1, c_2, c_3 \in \mathbb{Q}$. Then the product $n_1n_2n_3$ must be a perfect cube.

As a special case, the linear relation above holds if $\sqrt[3]{n_1}$, $\sqrt[3]{n_2}$, $\sqrt[3]{n_3}$ are members of an arithmetic sequence; thereby, the same conclusion holds. In particular, the cube roots of three distinct primes cannot participate in an arithmetic sequence.[11]

By assumption, we have

$$c_1n_1^{1/3} + c_2n_2^{1/3} = -c_3n_3^{1/3},$$

where we used fractional exponents. We take the cube of both sides and use the well-known identity[12]

$$(x + y)^3 = x^3 + 3x^2y + 3xy^2 + y^3, \quad x, y \in \mathbb{R}.$$

We obtain

$$3c_1^2c_2n_1^{2/3}n_2^{1/3} + 3c_1c_2^2n_1^{1/3}n_2^{2/3} = -c_1^3n_1 - c_2^3n_2 - c_3^3n_3,$$

or equivalently,

$$3c_1c_2n_1^{1/3}n_2^{1/3}\left(c_1n_1^{1/3} + c_2n_2^{1/3}\right) = -c_1^3n_1 - c_2^3n_2 - c_3^3n_3.$$

Replacing the expression in parentheses by the original linear relation, we get

$$3c_1c_2c_3n_1^{1/3}n_2^{1/3}n_3^{1/3} = c_1^3n_1 + c_2^3n_2 + c_3^3n_3.$$

This gives

$$(n_1n_2n_3)^{1/3} = \frac{c_1^3n_1 + c_2^3n_2 + c_3^3n_3}{3c_1c_2c_3} \in \mathbb{Q},$$

a rational number. By the above, $n_1n_2n_3$ must be a perfect cube.

[11] This special case was a problem in the USA Mathematical Olympiad, 1972.

[12] As noted previously, identities will be treated in Chapter 6.

For the special case, if $\sqrt[3]{n_1}$, $\sqrt[3]{n_2}$, $\sqrt[3]{n_3}$ participate in an arithmetic sequence with difference d, then we have

$$\sqrt[3]{n_1} = \sqrt[3]{n_3} + a_1 d \quad \text{and} \quad \sqrt[3]{n_2} = \sqrt[3]{n_3} + a_2 d, \;\; a_1 \neq a_2, \; 0 \neq a_1, a_2 \in \mathbb{Z}.$$

Eliminating d, we obtain the linear relation

$$a_2 \sqrt[3]{n_1} - a_1 \sqrt[3]{n_2} + (a_1 - a_2)\sqrt[3]{n_3} = 0$$

with non-zero integer coefficients. The example follows.

Remark The corollary in the example above (concerning the cube roots of three distinct primes) is a special case of a well-known and general result concerning nth roots of primes. In 1940, Besicovitch proved the following theorem:[13]

Let $p_1, p_2, \ldots, p_l$, $2 \leq l \in \mathbb{N}$, be **distinct** primes, and $q_1, q_2, \ldots, q_l \in \mathbb{N}$ such that each q_i, $i = 1, 2, \ldots, l$, is relatively prime to the product $p_1 \cdot p_2 \cdots p_l$. Moreover, let $2 \leq m_1, m_2, \ldots, m_l \in \mathbb{N}$, and consider the positive roots $\sqrt[m_i]{p_i \cdot q_i}$, $i = 1, 2, \ldots, l$. Finally, let $p(x_1, x_2, \ldots, x_l)$ be a polynomial[14] in l indeterminates with **rational** coefficients such that, for $i = 1, 2, \ldots, l$, the degree of $p(x_1, x_2, \ldots, x_l)$ in x_i is $\leq m_i - 1$. Then

$$p(\sqrt[m_1]{p_1 q_1}, \sqrt[m_2]{p_2 q_2}, \ldots, \sqrt[m_l]{p_l q_l}) = 0$$

implies that $p(x_1, x_2, \ldots, x_l)$ is identically zero; that is, all the coefficients of p vanish.

Now, this result along with the proof of the second statement of Example 3.2.2 implies that no roots $\sqrt[m_1]{p_1}$, $\sqrt[m_2]{p_2}$, $\sqrt[m_3]{p_3}$, $2 \leq m_1, m_2, m_3 \in \mathbb{N}$, of distinct primes p_1, p_2, p_3 can participate in an arithmetic sequence.

Indeed, replacing the cube roots with the respective roots above, after eliminating the difference of the arithmetic sequence, we obtain

$$a_2 \sqrt[m_1]{p_1} - a_1 \sqrt[m_2]{p_2} + (a_1 - a_2) \sqrt[m_3]{p_3} = 0.$$

The theorem of Besicovitch above applied to the polynomial $p(x_1, x_2, x_3) = a_2 x_1 - a_1 x_2 + (a_1 - a_2)x_3$ (of degree 1 in each indeterminate) implies that $a_1 = a_2 = 0$. This is a contradiction.

The proof of the theorem of Besicovitch is elementary but complex, and it is beyond the scope of this book.

Example 3.2.3 Let $1 \leq a \in \mathbb{Q}$.[15] We have

$$\lim_{n \to \infty} \left(\sqrt[a]{a} \cdot \sqrt[a^2]{a^2} \cdots \sqrt[a^n]{a^n} \right) \leq a^{a/(a-1)^2}.$$

[13] Besicovitch, A.S., *On the linear independence of fractional powers of integers*, J. London Math. Soc. 15 (1940), 3-6.

[14] Polynomials will be treated in detail in Chapters 6–7.

[15] A special case ($a = 2$) was a problem in the Irish Mathematical Olympiad, 1997. Note also that actually equality holds by sequential continuity of the exponentiation, see Section 4.2.

Indeed, the general term in the parentheses can be written as a single exponent:

$$a^{1/a+2/a^2+\cdots+n/a^n}.$$

Using Example 3.1.5 (with $r = 1/a$), for the limit of the exponent as $n \to \infty$, we have

$$\sum_{n=1}^{\infty} \frac{n}{a^n} = \sum_{n=1}^{\infty} n \left(\frac{1}{a}\right)^n = \frac{1/a}{(1-1/a)^2} = \frac{a}{(a-1)^2}.$$

The limit relation follows since the terms of the series are positive.

Remark Exponentiation of the **zero** as a base with **positive** exponent is usually defined to be zero: $0^q = 0$ with $q > 0$. On the other hand, 0^0 is undefined. Exponentiation of a **negative** base and real exponents cannot be defined consistently. For example, we have $-1 = (-1)^1 = (-1)^{2/2} \neq \sqrt[2]{(-1)^2} = 1$. As another problem, for $k \in \mathbb{N}$, we have $(-1)^{\frac{2k}{2k+1}} = \sqrt[2k+1]{(-1)^{2k}} = \sqrt[2k+1]{1} = 1$. On the other hand, $\lim_{k\to\infty}(-1)^{\frac{2k}{2k+1}} = (-1)^1 = -1$, since $\lim_{k\to\infty} 2k/(2k+1) = \lim_{k\to\infty} 1/(1+1/(2k)) = 1$.

Example 3.2.4 Let $(a_n)_{n\in\mathbb{N}}$ be a real sequence such that $0 < a_n < 1$, $n \in \mathbb{N}$. Does this condition imply that $\lim_{n\to\infty} a_n^n = 0$?

The answer is "no:" Take, for example, $a_n = 1/\sqrt[n]{2}$, $n \in \mathbb{N}$.

Example 3.2.5 We have $\lim_{n\to\infty} \sqrt[n]{a} = 1, \quad 0 < a \in \mathbb{R}$.

To show this, first let $a > 1$. The sequence $(\sqrt[n]{a})_{n\in\mathbb{N}}$ is strictly decreasing: $\sqrt[n+1]{a} < \sqrt[n]{a}$, $n \in \mathbb{N}$. In addition, we have the lower bound $1 < \sqrt[n]{a}$, $n \in \mathbb{N}$. (By the Monotone Convergence Theorem, $\lim_{n\to\infty} \sqrt[n]{a} \geq 1$ exists, but we will not need this fact.)

We let $0 < b_n = \sqrt[n]{a} - 1 \in \mathbb{R}$, $n \in \mathbb{N}$. By the Bernoulli inequality, we have

$$a = (1+b_n)^n \geq 1 + nb_n, \quad n \in \mathbb{N}.$$

This gives

$$0 < \sqrt[n]{a} - 1 = b_n \leq \frac{a-1}{n} \quad n \in \mathbb{N}.$$

Using monotonicity of the limit, we obtain $\lim_{n\to\infty}(\sqrt[n]{a} - 1) = 0$. The limit follows in this case.

For $0 < a < 1$,[16] we have $\lim_{n\to\infty} 1/\sqrt[n]{a} = \lim_{n\to\infty} \sqrt[n]{1/a} = 1$ so that the limit follows again.

Remark If $\lim_{n\to\infty} \sqrt[n]{a} = L$ is assumed (as it follows from the Monotone Convergence Theorem), then, taking the subsequence of the even terms $(\sqrt[2m]{a})_{m\in\mathbb{N}}$

[16] The case $a = 1$ is obvious.

that also converges to L, we have

$$L^2 = \left(\lim_{n\to\infty} \sqrt[n]{a}\right)^2 = \left(\lim_{m\to\infty} \sqrt[2m]{a}\right)^2 = \lim_{m\to\infty} (\sqrt[2m]{a})^2 = \lim_{m\to\infty} \sqrt[m]{a} = L.$$

Since $1 \le L = L^2$, this gives $L = 1$.

A non-trivial variation on the theme is the following:

Example 3.2.6 We have

$$\lim_{n\to\infty} \sqrt[n]{a^n + b^n} = \max(a, b), \quad 0 < a, b \in \mathbb{R}.$$

First, if $a = b$, then $\sqrt[n]{a^n + b^n} = \sqrt[n]{2a^n} = \sqrt[n]{2} \cdot a$. By the previous example, $\lim_{n\to\infty} \sqrt[n]{2} = 1$. Thus, in this case, our limit formula follows.

If $a \neq b$, we may assume $a > b$ since otherwise we would switch a and b.

First Solution. We perform a reduction step. We write $\sqrt[n]{a^n + b^n} = b\sqrt[n]{1 + (a/b)^n}$. Letting $c = a/b > 1$, our limit formula reduces to the following:

$$\lim_{n\to\infty} \sqrt[n]{1 + c^n} = c, \quad 1 < c \in \mathbb{R}.$$

We define $a_n = \sqrt[n]{1 + c^n} - c > 0$, $n \in \mathbb{N}$. We need to show that $(a_n)_{n\in\mathbb{N}}$ is a null-sequence. We have

$$1 + c^n = (c + a_n)^n = c^n \left(1 + \frac{a_n}{c}\right)^n \ge c^n \left(1 + n\frac{a_n}{c}\right),$$

where in the last step we used the Bernoulli inequality. Dividing by c^n and simplifying, this gives $na_n/c \le 1/c^n$. Equivalently, we have

$$0 < a_n \le \frac{1}{nc^{n-1}}.$$

By monotonicity of the limit, we have $0 \le \lim_{n\to\infty} a_n \le \lim_{n\to\infty} 1/(nc^{n-1}) = 0$, where the last limit is zero because $c > 1$.

Second Solution. A much shorter proof can be obtained if we use Example 3.2.5. Assuming $a \le b$, we have

$$b = \lim_{n\to\infty} \sqrt[n]{b^n} \le \lim_{n\to\infty} \sqrt[n]{a^n + b^n} \le \lim_{n\to\infty} \sqrt[n]{2b^n} = \lim_{n\to\infty} \sqrt[n]{2} \lim_{n\to\infty} \sqrt[n]{b^n} = b.$$

Remark Example 3.2.6 can be generalized in two ways.

First, if $(a_n)_{n\in\mathbb{N}}$ and $(b_n)_{n\in\mathbb{N}}$ are sequences with positive members and

$$\lim_{n\to\infty} a_n = a > 0 \quad \text{and} \quad \lim_{n\to\infty} b_n = b > 0,$$

then we have

$$\lim_{n\to\infty} \sqrt[n]{a_n^n + b_n^n} = \max(a, b).$$

The proof is analogous to the one above replacing the constant c with $c_n = a_n/b_n$, $n \in \mathbb{N}$.

Second, if $a_1, \ldots, a_N$, $2 \leq N \in \mathbb{N}$, are positive real numbers, then we have

$$\lim_{n\to\infty} \sqrt[n]{a_1^n + \cdots + a_N^n} = \max(a_1, \ldots, a_N).$$

Indeed, this follows by Peano's Principle of Induction with respect to $2 \leq N \in \mathbb{N}$. For the general induction step $N \Rightarrow N + 1$, we use the first generalization above and calculate

$$\begin{aligned}\lim_{n\to\infty} \sqrt[n]{a_1^n + \cdots + a_N^n + a_{N+1}^n} &= \lim_{n\to\infty} \sqrt[n]{\left(\sqrt[n]{a_1^n + \cdots + a_N^n}\right)^n + a_{N+1}^n} \\ &= \max(\max(a_1, \ldots, a_N), a_{N+1}) = \max(a_1 \ldots, a_{N+1}).\end{aligned}$$

The following crucial proposition is a substantial generalization of Example 3.2.5.

Proposition 3.2.2 *For any rational null-sequence* $q : \mathbb{N}_0 \to \mathbb{Q}$, $q = (q_0, q_1, q_2, \ldots)$, *we have*

$$\lim_{n\to\infty} a^{q_n} = 1, \quad 0 < a \in \mathbb{R}.$$

Before getting to the proof, we generalize the Bernoulli inequality for rational exponents as follows.

Bernoulli Inequality (Rational Exponent). *For* $-1 < r \neq 0$, $r \in \mathbb{R}$, *we have*

$$(1 + r)^q > 1 + qr, \quad 1 < q \in \mathbb{Q}.$$

The Bernoulli inequality has an interesting symmetry. The **simultaneous** interchange of the indeterminates $q \leftrightarrow 1/q$ and $r \leftrightarrow qr$ (and raising both sides to the power q) transforms the inequality into itself with the inequality sign **reversed**:

$$(1 + r)^q < 1 + qr, \quad 0 < q < 1,\ q \in \mathbb{Q},\ -1 < r \neq 0,\ r \in \mathbb{R}.$$

We derive this second (equivalent) inequality.

It is convenient as well as instructive to reformulate this inequality in terms of a certain monotonicity property of the sequence[17]

[17] This sequence will be of paramount importance in Euler's treatment of the exponential function in Section 10.5.

$$e_n^*(s) = \left(1 + \frac{s}{n}\right)^n, \quad s \in \mathbb{R},\ n \in \mathbb{N}.$$

The monotonicity property in question is the following:

$$e_n^*(s) < e_{n+1}^*(s), \quad 0 \neq s > -n,\ n \in \mathbb{N},$$

and we claim that this is **equivalent** to the Bernoulli inequality for **rational exponents** above.[18]

First, substituting $q = n/(n+1)$ and $r = s/n$, $0 \neq s > -n$, into the Bernoulli inequality, the monotonicity property holds.

Second, assume that the monotonicity property holds. Let $q = m/n \in \mathbb{Q}$, $m < n$, $m, n \in \mathbb{N}$. By simple induction, for $-m < s \neq 0$, $s \in \mathbb{R}$, we have $e_m^*(s) < e_n^*(s)$. Substituting $s = mr$, $-1 < r \neq 0$, $r \in \mathbb{R}$, we obtain

$$(1+r)^m < \left(1 + \frac{m}{n}r\right)^n.$$

Taking the nth root of both sides, the Bernoulli inequality follows.

Finally, it remains to show that the monotonicity property above holds for $e_n^*(s)$.

We calculate

$$\begin{aligned}\frac{e_{n+1}^*(s)}{e_n^*(s)} &= \frac{\left(1 + \frac{s}{n+1}\right)^{n+1}}{\left(1 + \frac{s}{n}\right)^n} = \left(\frac{n}{n+s}\right)^{n+1} \left(1 + \frac{s}{n+1}\right)^{n+1} \left(1 + \frac{s}{n}\right) \\ &= \left(1 - \frac{s}{(n+s)(n+1)}\right)^{n+1} \left(1 + \frac{s}{n}\right) > \left(1 - \frac{s}{n+s}\right)\left(1 + \frac{s}{n}\right) = 1,\end{aligned}$$

where, in the last estimate, we used the Bernoulli inequality for natural exponents ($n + 1 \geq 2$). (Note that $s/((n+s)(n+1)) < 1$ since $s > -n$.)

Summarizing, we derived the Bernoulli inequality for rational exponents.

The simple substitution, $0 < a\ (= r + 1) \neq 1$, $a \in \mathbb{R}$, gives the equivalent form of the Bernoulli inequality

$$a^q < 1 + q(a-1), \quad 0 < q < 1,\ q \in \mathbb{Q}.$$

We need another version of this for negative exponents. Taking the reciprocals of both sides, we have

$$a^{-q} > \frac{1}{1 + q(a-1)} = \frac{1 - q(a-1)}{1 - q^2(a-1)^2} > 1 - q(a-1), \quad 0 < q < 1,\ q \in \mathbb{Q}.$$

[18] In this equivalence we assume that the Bernoulli inequality holds for **integral** exponents. This we have already shown by simple induction at the end of Section 2.1.

We now **assume** $1 < a \in \mathbb{R}$ and combine the two inequalities to obtain

$$1 - q(a-1) \le a^{-q} \le a^q \le 1 + q(a-1), \quad 0 \le q < 1.$$

Summarizing, for $1 < a \in \mathbb{R}$, we have

$$|a^q - 1| \le |q|(a-1), \quad |q| < 1,\ q \in \mathbb{Q}.$$

Proof of Proposition 3.2.2 First, assume $1 < a \in \mathbb{R}$. Let $q : \mathbb{N}_0 \to \mathbb{R}$, $q = (q_0, q_1, q_2, \ldots)$, be a rational null-sequence. By the inequality above and monotonicity of the limit superior, we have

$$0 \le \limsup_{n\to\infty} |a^{q_n} - 1| \le (a-1) \limsup_{n\to\infty} |q_n| = (a-1) \lim_{n\to\infty} |q_n| = 0.$$

The proposition follows in this case.

Second, assume $0 < a < 1$, $a \in \mathbb{R}$. (The case $a = 1$ is trivial.) By what we just proved, we have $\lim_{n\to\infty}(1/a)^{q_n} = 1$. Using Proposition 3.1.3, we have

$$\lim_{n\to\infty} a^{q_n} = \lim_{n\to\infty} \frac{1}{(1/a)^{q_n}} = \frac{1}{\lim_{n\to\infty}(1/a)^{q_n}} = 1.$$

The proposition follows.

Our first application of the Bernoulli inequality for rational exponents is the following:

Example 3.2.7 For $0 \le q \in \mathbb{Q}$ and $1 < a \in \mathbb{R}$, we have

$$\lim_{n\to\infty} \frac{n^q}{a^n} = 0.$$

Indeed, for $q + 1 < n \in \mathbb{N}$, we have

$$\frac{n^q}{a^n} = \frac{n^q}{\left(a^{n/(q+1)}\right)^{q+1}} < \frac{n^q}{\left(1 + \frac{n}{q+1}(a-1)\right)^{q+1}} < \frac{n^q}{\left(\frac{n}{q+1}(a-1)\right)^{q+1}} = \frac{1}{n}\left(\frac{q+1}{a-1}\right)^{q+1},$$

where we used the Bernoulli inequality for the rational exponent $1 < n/(q+1) \in \mathbb{Q}$. Using this, we have

$$0 \le \lim_{n\to\infty} \frac{n^q}{a^n} \le \left(\frac{q+1}{a-1}\right)^{q+1} \lim_{n\to\infty} \frac{1}{n} = 0.$$

The example follows.

Note the important consequence: For any $1 < a \in \mathbb{R}$ and $m \in \mathbb{N}_0$, there exists $N \in \mathbb{N}$ such that

$$a^n > n^m, \quad n \geq N.$$

Another simple application is the following stronger statement than the limit in Example 3.2.5.

Example 3.2.8 We have

$$\lim_{n\to\infty} \sqrt[n]{n} = 1.$$

We first claim that the sequence $(\sqrt[n]{n})_{n\in\mathbb{N}}$ is strictly decreasing for $3 \leq n \in \mathbb{N}$. Indeed, by Example 2.1.4, we have

$$n^{n+1} > (n+1)^n, \quad 3 \leq n \in \mathbb{N}.$$

Taking the $n(n+1)$th root of both sides, we obtain

$$\sqrt[n]{n} > \sqrt[n+1]{n+1}, \quad 3 \leq n \in \mathbb{N}.$$

The claim follows. Since 1 is an obvious lower bound of the sequence, the Monotone Convergence Theorem implies that $(\sqrt[n]{n})_{n\in\mathbb{N}}$ is convergent to a limit $L \geq 1$. To determine L, we take the subsequence $(\sqrt[2^n]{2^n})_{n\in\mathbb{N}}$ (which, necessarily, must converge to the same limit). We have

$$\lim_{n\to\infty} \sqrt[2^n]{2^n} = \lim_{n\to\infty} \left(2^n\right)^{1/2^n} = \lim_{n\to\infty} 2^{n/2^n}.$$

On the other hand, by the previous example, $(n/2^n)_{n\in\mathbb{N}}$ is a null-sequence. Applying Proposition 3.2.2, we obtain

$$\lim_{n\to\infty} 2^{n/2^n} = 1.$$

Thus, $L = 1$, and the example follows.

Remark The second part of Example 3.2.8 can also be completed (without the recourse to Proposition 3.2.2) as follows:

$$\begin{aligned} L^2 = \lim_{n\to\infty} (\sqrt[n]{n})^2 = \lim_{m\to\infty} (\sqrt[2m]{2m})^2 = \lim_{m\to\infty} \sqrt[m]{2m} \\ = \lim_{m\to\infty} (\sqrt[m]{2}\sqrt[m]{m}) = \lim_{m\to\infty} \sqrt[m]{2} \lim_{m\to\infty} \sqrt[m]{m} = L, \end{aligned}$$

where we used Example 3.2.5. Now $1 \leq L = L^2$ so that $L = 1$ follows.

We finish this cadre of examples by the following:

Example 3.2.9 We have $\lim_{n\to\infty} \sqrt[n]{n!} = \infty$.

We first claim that $(k+1)(n-k) \geq n$, $0 \leq k < n$. Indeed, we have

$$(k+1)(n-k)-n=(k+1)n-k(k+1)-n=kn-k(k+1)=k(n-k-1) \geq 0,\ \ 0 \leq k < n,$$

and the claim follows. Using this, we estimate

$$(n!)^2 = (1 \cdot n)(2 \cdot (n-1)) \cdot (3 \cdot (n-2)) \cdot \ldots \cdot (n \cdot 1) \geq n^n.$$

Taking the $(2n)$th root, we obtain $\sqrt[n]{n!} = \sqrt[2n]{(n!)^2} \geq \sqrt[2n]{n^n} = \sqrt{n}$. By monotonicity of the limit, we finally arrive at $\infty = \lim_{n\to\infty} \sqrt{n} \leq \lim_{n\to\infty} \sqrt[n]{n!}$. The example follows.

Remark 1 An alternative proof can be given as follows.

First, notice that the sequence $(\sqrt[n]{n!})_{n\in\mathbb{N}}$ is strictly increasing. To show this, let $n \in \mathbb{N}$. Multiplying both sides of the obvious inequality $n! < (n+1)^n$ by $(n!)^n$, we obtain $(n!)^{n+1} < ((n+1)n!)^n = (n+1)!^n$. Taking the $n(n+1)$th root of both sides, strict monotonicity follows.

Thus, $\lim_{n\to\infty} \sqrt[n]{n!}$ is either finite or infinite. It is enough to check this on a subsequence. Letting $n = 2m$ even, we have

$$(2m)! = m!(m+1)(m+2) \cdot \ldots \cdot (m+m) \geq m^m.$$

This gives $\sqrt[2m]{(2m)!} \geq \sqrt{m}$. By monotonicity of the limit again, we obtain $\lim_{n\to\infty} \sqrt[n]{n!} = \lim_{m\to\infty} \sqrt[2m]{(2m)!} \geq \lim_{m\to\infty} \sqrt{m} = \infty$.

Remark 2 A (2-step) refinement of Example 3.2.9 will yield the well-known Stirling formula; see the remark after Example 10.3.4.

Let $0 < a \in \mathbb{R}$. We define the power $a^r \in \mathbb{R}$ with **real** exponent $r \in \mathbb{R}$ as follows. Let $q : \mathbb{N}_0 \to \mathbb{Q}$, $q = (q_0, q_1, q_2, \ldots)$, be a rational sequence such that $\lim_{n\to\infty} q_n = r$. Then we define

$$a^r = \lim_{n\to\infty} a^{q_n}.$$

We need to show that the limit exists, and it does not depend on the rational sequence chosen for the exponent.

We first assume $1 < a \in \mathbb{R}$. We claim that $(a^{q_n})_{n\in\mathbb{N}_0}$ is a Cauchy sequence (thereby convergent by Proposition 2.3.3).

We start by observing that the **convergent** rational sequence q is bounded: $|q_n| \leq c$, $n \in \mathbb{N}_0$, for some $0 < c \in \mathbb{Q}$. Thus, by monotonicity for rational exponents, we have $|a^{q_n}| \leq a^c$, $n \in \mathbb{N}_0$. Moreover, since q is a (rational) Cauchy sequence, for (any) given $0 < \epsilon \in \mathbb{Q}$, there exists $N \in \mathbb{N}_0$ such that

$$|q_n - q_m| < \min\left(\frac{\epsilon}{a^c(a-1)}, 1\right), \quad m, n \geq N.$$

We now use the identities for rational exponentiation along with the Bernoulli estimate above (for $|q| = |q_n - q_m| < 1$). For $m, n \geq N$, we have

$$|a^{q_n} - a^{q_m}| = |a^{q_m}(a^{q_n - q_m} - 1)| = |a^{q_m}||a^{q_n - q_m} - 1| \leq a^c |q_n - q_m|(a - 1) < \epsilon.$$

The claim follows.

Second, if $0 < a < 1$, then, by what we just proved, for a rational convergent sequence $q : \mathbb{N}_0 \to \mathbb{Q}$, $q = (q_0, q_1, q_2, \ldots)$, the sequence $((1/a)^{q_n})_{n \in \mathbb{N}_0}$ is convergent. Using Proposition 3.1.3, the sequence $(a^{q_n})_{n \in \mathbb{N}_0}$ is also convergent.

Next, we claim that the real power a^r is well-defined; that is, it does not depend on the choice of the rational sequence $q : \mathbb{N}_0 \to \mathbb{Q}, q = (q_0, q_1, q_2, \ldots)$, convergent to $r \in \mathbb{R}$.

Indeed, let $q' : \mathbb{N}_0 \to \mathbb{Q}$, $q' = (q'_0, q'_1, q'_2, \ldots)$, be another rational sequence with limit r. Since q and q' have the same limit, $q - q'$ is a null-sequence. (In Cantor's construction of the real numbers discussed above, we have $q \sim q'$.) By Proposition 3.2.2, we have $\lim_{n \to \infty} a^{q_n - q'_n} = 1$. Therefore, using the identities for rational exponentiation, we obtain

$$\lim_{n \to \infty} a^{q_n} = \lim_{n \to \infty} \left(a^{q_n - q'_n} \cdot a^{q'_n} \right) = \lim_{n \to \infty} a^{q_n - q'_n} \lim_{n \to \infty} a^{q'_n} = \lim_{n \to \infty} a^{q'_n}.$$

The claim follows. (Instead of this proof of the second part, alternatively, we can construct the sequence $(q_0, q'_0, q_1, q'_1, \ldots)$ and appeal to the first part of the proof above.)

Exponentiation with positive base and real exponent satisfies the same identities as those with rational exponent. For $0 < a, b \in \mathbb{R}$ and $r, s \in \mathbb{R}$, we have the following Identities:

$$a^{r+s} = a^r \cdot a^s, \quad a^{r-s} = \frac{a^r}{a^s}, \quad (a^r)^s = a^{rs}, \quad (ab)^r = a^r \cdot b^r.$$

These identities can be established in a straightforward manner taking the limits of the analogous identities for rational exponents.

In addition, we also have the following monotonicity properties. For $1 < a \in \mathbb{R}$, the power a^r is strictly increasing in $r \in \mathbb{R}$. Similarly, for $0 < a < 1$, the power a^r is strictly decreasing in $r \in \mathbb{R}$.

Finally, to complete the circle, the Bernoulli inequality holds for real exponents (taking again the limit of the respective inequality for rational exponents).

Bernoulli Inequality (Real Exponent). *For $0 < a \neq 1$, $a \in \mathbb{R}$, we have*

$$a^r < 1 + r(a - 1), \ 0 < r < 1, \ r \in \mathbb{R} \quad \text{and} \quad a^r > 1 + r(a - 1), \ 1 < r \in \mathbb{R}.$$

As above, for $1 < a \in \mathbb{R}$, we can combine the two estimates and obtain

$$|a^r - 1| \leq |r| \cdot (a - 1), \quad |r| < 1, \ r \in \mathbb{R}.$$

Remark As a simple consequence, note that Example 3.2.7 holds for real exponent $0 \leq q \in \mathbb{R}$.

We now show a simple but important consequence of the Bernoulli inequality for real exponents.

Example (Young's Inequality) Let $0 < p, q \in \mathbb{R}$ such that $1/p + 1/q = 1$. Then, we have

$$xy \leq \frac{x^p}{p} + \frac{y^q}{q}, \quad 0 < x, y \in \mathbb{R},$$

with equality if and only if $x^p = y^q$.

If $x^p = y^q$, then the equality clearly holds. We assume $x^p \neq y^q$, substitute $u = x^p$ and $v = y^q$, and rewrite the (sharp) inequality in the equivalent form

$$u^{1/p} \cdot v^{1/q} < \frac{u}{p} + \frac{v}{q}, \quad u \neq v,\ 0 < u, v \in \mathbb{R}.$$

We "dehomogenize" by setting $a = u/v$, $0 < a \neq 1$, $a \in \mathbb{R}$, in yet another equivalent form

$$a^{1/p} < \frac{a}{p} + \frac{1}{q} = 1 + \frac{1}{p}(a - 1).$$

This, however, is the Bernoulli inequality for the exponent $0 < r = 1/p < 1$. The Young inequality follows.

Example 3.2.10 Determine the infimum $\inf_{0<r,s\in\mathbb{R}} (r^s + s^r)$.

For either $1 \leq r \in \mathbb{R}$ or $1 \leq s \in \mathbb{R}$, we have $r^s + s^r > 1$, and $\inf_{1\leq r,s\in\mathbb{R}} (r^s + s^r) = 1$. Thus, we may assume that $0 < r, s < 1$. The Bernoulli inequality then gives

$$\frac{r}{r^s} = r^{1-s} = (1 + (r - 1))^{1-s} \leq 1 + (r - 1)(1 - s) = r + s - rs,$$

or equivalently,

$$r^s \geq \frac{r}{r + s - rs}.$$

Swapping r and s and adding, we obtain

$$r^s + s^r \geq \frac{r}{r + s - rs} + \frac{s}{r + s - rs} = \frac{r + s}{r + s - rs} > 1.$$

Thus the value of the infimum is 1.

Now that we have the Bernoulli inequality for real exponents in place we return to the question of convergence for the p-series $\sum_{n=1}^{\infty} 1/n^p$ for $1 < p \in \mathbb{R}$.

First, we give an elementary approach and seek an upper bound for the partial sum

$$\sum_{k=1}^{n} \frac{1}{k^p} = 1 + \frac{1}{2^p} + \cdots + \frac{1}{n^p}.$$

In the previous section, using Peano's Principle of Induction, we showed that, for $p = 2$, an upper bound is $2 - 1/n$, and, for $p = 3/2$, an upper bound is $3 - 2/n^{1/2}$. As an easy generalization of this, we now claim

$$\sum_{k=1}^{n} \frac{1}{k^p} = 1 + \frac{1}{2^p} + \cdots + \frac{1}{n^p} \leq \frac{p}{p-1} - \frac{1}{p-1} \cdot \frac{1}{n^{p-1}}, \quad n \in \mathbb{N},\ 1 < p \in \mathbb{R}.$$

Note that this implies that the p-series $\sum_{n=1}^{\infty} 1/n^p$ converges for $1 < p \in \mathbb{R}$.

Remark The reader versed in elementary calculus would notice that the upper bound here also comes from the integral estimate

$$1 + \frac{1}{2^p} + \cdots + \frac{1}{n^p} < 1 + \int_1^n \frac{dt}{t^p} = 1 + \frac{1 - 1/n^{p-1}}{p-1}.$$

As noted above, we use induction with respect to $n \in \mathbb{N}$ to prove this claim. Throughout, we assume $1 < p \in \mathbb{R}$.

The initial case $n = 1$ is clear. For the general induction step $n \Rightarrow n + 1$, we assume that the inequality above holds. Using this as the induction hypothesis, we calculate

$$1 + \frac{1}{2^p} + \cdots + \frac{1}{n^p} + \frac{1}{(n+1)^p} \leq \frac{p}{p-1} - \frac{1}{p-1} \cdot \frac{1}{n^{p-1}} + \frac{1}{(n+1)^p}.$$

We need to show

$$\frac{p}{p-1} - \frac{1}{p-1} \cdot \frac{1}{n^{p-1}} + \frac{1}{(n+1)^p} \leq \frac{p}{p-1} - \frac{1}{p-1} \cdot \frac{1}{(n+1)^{p-1}},$$

or equivalently,

$$\frac{1}{(n+1)^p} \left(1 + \frac{n+1}{p-1}\right) \leq \frac{1}{p-1} \cdot \frac{1}{n^{p-1}}.$$

After simplification and elimination of the denominators, this becomes

$$(n+1)^p \geq (n+p)n^{p-1} = n^p + pn^{p-1}.$$

Dividing through by n^p, this becomes equivalent to

$$\left(1+\frac{1}{n}\right)^p \geq 1+\frac{p}{n}.$$

This, however, is the Bernoulli inequality for the real exponent $1 < p \in \mathbb{R}$. The claim follows.

Second, there is a much more powerful method to settle this and many other convergence and divergence questions. This is called the **Cauchy Condensation Test**, and it is very useful in the study of infinite series.

We begin with a **decreasing** sequence $(a_n)_{n\in\mathbb{N}}$ of infinite series of **non-negative** real numbers, $0 \leq a_{n+1} \leq a_n$, $n \in \mathbb{N}$, and form the infinite series $\sum_{k=1}^{\infty} a_k = a_1+a_2+\cdots+a_n+\cdots$. By definition, this series converges if the sequence of partial sums $(s_n)_{n\in\mathbb{N}}$, $s_n = a_1+a_2+\cdots+a_n$, $n \in \mathbb{N}$, has a (finite) limit. Since $a_n \geq 0$, $n \in \mathbb{N}$, the sequence $(s_n)_{n\in\mathbb{N}}$ is **increasing**. Therefore, by the Monotone Convergence Theorem, our original infinite series converges if and only if (any subsequence of) $(s_n)_{n\in\mathbb{N}}$ is **bounded**.

The crux is to compare our infinite series with the "condensed" series

$$a_1 + 2a_2 + 2^2a_{2^2} + \cdots + 2^n a_{2^n} + \cdots$$

Since this series also has non-negative terms, it is convergent if and only if its partial sums are bounded.

Now, the Cauchy Condensation Test states that the two series **equiconverge**; that is, one is convergent if and only if the other is convergent.

Remark An illustrative example to motivate "condensation" is the (divergent) harmonic series $\sum_{k=1}^{\infty} 1/k$ in Example 3.1.6 (along with the estimates there). Its condensed series is $\sum_{k=1}^{\infty} 2^k \cdot 1/2^k = \sum_{k=1}^{\infty} 1 = 1+1+1+\cdots = \infty$.

To derive the stated equiconvergence, we first compare the subsequence $(s_{2^n})_{n\in\mathbb{N}_0}$ of partial sums of our original series with those of the condensed series as follows. For $n \in \mathbb{N}_0$, we have

$$s_{2^{n+1}} - s_{2^n} = a_{2^n+1} + a_{2^n+2} + \cdots + a_{2^{n+1}} \geq 2^n \cdot a_{2^{n+1}} = \frac{2^{n+1}\cdot a_{2^{n+1}}}{2},$$

where we used the assumption that the sequence $(a_n)_{n\in\mathbb{N}}$ is decreasing. This gives

$$\begin{aligned} s_{2^n} &= (s_{2^n} - s_{2^{n-1}}) + (s_{2^{n-1}} - s_{2^{n-2}}) + \cdots + (s_2 - s_1) + a_1 \\ &\geq \frac{1}{2}\left(2^n a_{2^n} + 2^{n-1} a_{2^{n-1}} + \cdots + 2a_2 + a_1\right), \quad n \in \mathbb{N}. \end{aligned}$$

Thus, boundedness of the sequence of partial sums $(s_{2^n})_{n\in\mathbb{N}_0}$ of our original series implies boundedness of the partial sums of the condensed series.

For the converse, we compare the subsequence $(s_{2^n-1})_{n\in\mathbb{N}}$ of partial sums of our original series with those of the condensed series as follows. For $n \in \mathbb{N}$, we have

$$s_{2^{n+1}-1} - s_{2^n-1} = a_{2^n} + a_{2^n+1} + \cdots + a_{2^{n+1}-1} \leq 2^n \cdot a_{2^n},$$

where we used again the assumption that the sequence $(a_n)_{n\in\mathbb{N}}$ is decreasing.

This gives

$$\begin{aligned} s_{2^n-1} &= (s_{2^n-1} - s_{2^{n-1}-1}) + (s_{2^{n-1}-1} - s_{2^{n-2}-1}) + \cdots + (s_3 - s_1) + a_1 \\ &\leq 2^{n-1}a_{2^{n-1}} + 2^{n-2}a_{2^{n-2}} + \cdots + 2a_2 + a_1, \quad n \in \mathbb{N}. \end{aligned}$$

Thus, boundedness of the partial sums of the condensed series implies boundedness of the sequence of partial sums $(s_{2^n-1})_{n\in\mathbb{N}_0}$ of our original series.

The Cauchy Condensation Test follows. Note that we also obtained the following estimates for our infinite series:

$$\frac{1}{2}\left(a_1 + 2a_2 + 2^2a_{2^2} + \cdots\right) \leq a_1 + a_2 + a_3 + \cdots \leq a_1 + 2a_2 + 2^2a_{2^2} + \cdots$$

Example 3.2.11 Once again, consider the p-series

$$\sum_{k=1}^{\infty} \frac{1}{k^p} = 1 + \frac{1}{2^p} + \frac{1}{3^p} + \cdots + \frac{1}{n^p} + \cdots$$

for $0 < p \in \mathbb{R}$ **real**. We make use of the Cauchy Condensation Test. For $n \in \mathbb{N}$, we have

$$2^n \frac{1}{(2^n)^p} = \frac{1}{2^{n(p-1)}} = \left(\frac{1}{2^{p-1}}\right)^n.$$

Hence, the condensed series is geometric

$$1 + \frac{1}{2^{p-1}} + \left(\frac{1}{2^{p-1}}\right)^2 + \cdots + \left(\frac{1}{2^{p-1}}\right)^n + \cdots,$$

and this converges if $p > 1$ and diverges if $(0 <) p \leq 1$. The same therefore holds for the p-series. We recover our earlier result.

As the opposite case of the example above, for $p > 0$, it is natural to study the sequence with nth term[19]

$$s_p(n) = 1^p + 2^p + \cdots + (n-1)^p, \quad 2 \leq n \in \mathbb{N}.$$

[19]The shift in the base from n to $n - 1$ is a technical convenience.

For $p = 1, 2, 3$, the general term of the sequence has the following form:

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

$$1^2 + 2^2 + \cdots + (n-1)^2 = \frac{n(n-1)(2n-1)}{6} = \frac{1}{3}n^3 - \frac{1}{2}n^2 + \frac{1}{6}n$$

$$1^3 + 2^3 + \cdots + (n-1)^3 = \left(\frac{n(n-1)}{2}\right)^2 = \frac{1}{4}n^4 - \frac{1}{2}n^3 + \frac{1}{4}n^2$$

These can easily be shown by induction. As we will discuss later, the coefficients can be expressed in terms of the so-called **Bernoulli numbers**. At present, we are interested in the principal term, as the following example shows.

Example 3.2.12 For $0 < p \in \mathbb{R}$, we have

$$\frac{1}{p+1} - \frac{1}{n} < \frac{s_p(n)}{n^{p+1}} < \frac{1}{p+1}, \quad 2 \le n \in \mathbb{N}.$$

In particular, we have the limit

$$\lim_{n\to\infty} \frac{s_p(n)}{n^{p+1}} = \frac{1}{p+1}, \quad 0 < p \in \mathbb{R}.$$

To derive these inequalities, we employ the Bernoulli inequality

$$a^r > 1 + r(a-1), \quad 1 < r \in \mathbb{R},\ 0 < a \ne 1,\ a \in \mathbb{R}.$$

First, let $a = (k+1)/k$, $k = 1, \ldots, n-1$. We have

$$\left(\frac{k+1}{k}\right)^r > 1 + r\left(\frac{k+1}{k} - 1\right) = 1 + \frac{r}{k}, \quad k = 1, \ldots, n-1.$$

Simplifying, this gives $(k+1)^r - k^r > rk^{r-1}$, $k = 0, \ldots, n-1$. Summing up with respect to $k = 0, \ldots, n-1$, we obtain

$$n^r > r(1^{r-1} + 2^{r-1} + \cdots + (n-1)^{r-1}).$$

Substituting $0 < p = r - 1 \in \mathbb{R}$, we arrive at the following:

$$\frac{s_p(n)}{n^{p+1}} < \frac{1}{p+1}, \quad 2 \le n \in \mathbb{N}.$$

The upper estimate above follows.

Second, let $a = k/(k+1)$, $k = 0, \dots, n-1$. We have

$$\left(\frac{k}{k+1}\right)^r > 1 + r\left(\frac{k}{k+1} - 1\right) = 1 - \frac{r}{k+1}, \quad k = 0, \dots, n-1.$$

Simplifying, this gives $(k+1)^r - k^r < r(k+1)^{r-1}$, $k = 0, \dots, n-1$. Summing up with respect to $k = 0, \dots, n-1$, we obtain

$$n^r < r(1^{r-1} + 2^{r-1} + \cdots + n^{r-1}).$$

Substituting $0 < p = r - 1 \in \mathbb{R}$, we arrive at the following:

$$\frac{1}{p+1} < \frac{s_p(n) + n^p}{n^{p+1}}, \quad 2 \le n \in \mathbb{N}.$$

The lower estimate above follows. The proof is complete.

The previous example can be put in a more general framework that will be useful in the sequel.

Let $a < b$, $a, b \in \mathbb{R}$, and $f : [a, b] \to \mathbb{R}$ be a real function. For $n \in \mathbb{N}$, we subdivide the domain interval $[a, b]$ into n equal parts

$$a < a + \frac{b-a}{n} < a + 2\frac{b-a}{n} < \cdots < a + (n-1)\frac{b-a}{n} < a + n\frac{b-a}{n} = b$$

and define the **arithmetic mean**

$$\mathcal{A}_f(n, a, b) = \frac{1}{n}\sum_{k=1}^{n} f\left(a + k\frac{b-a}{n}\right).$$

Finally, we define the **mean** of f by the limit

$$\mathcal{A}_f(a, b) = \lim_{n\to\infty} \mathcal{A}_f(n, a, b),$$

where we tacitly assume that the limit exists.[20]

The mean is clearly linear, that is, for $f, g : [a, b] \to \mathbb{R}$ and $c \in \mathbb{R}$, we have

$$\mathcal{A}_{f+g} = \mathcal{A}_f + \mathcal{A}_g \quad \text{and} \quad \mathcal{A}_{c\cdot f} = c \cdot \mathcal{A}_f,$$

where we suppressed the dependence on the interval $[a, b]$.

The mean is also monotonic in the sense that if $f, g : [a, b] \to \mathbb{R}$ are real functions such that $f(x) \le g(x)$, $a \le x \le b$, then we have $\mathcal{A}_f \le \mathcal{A}_g$.

[20]The general theory of means is expounded in Hardy, G.H., Littlewood, J.E., and Pólya, G., *Inequalities*, 2nd ed. Cambridge University Press, 1988.

For $0 < p \in \mathbb{R}$, we define the **power function** $\mathfrak{p}_p$ by $\mathfrak{p}_p(x) = x^p$, $x \in \mathbb{R}$. For fixed $0 < x \in \mathbb{R}$, we now calculate the mean of $\mathfrak{p}_p$ over the interval $[0, x]$ as follows:

$$\mathcal{A}_{\mathfrak{p}_p}(n, x) = \frac{1}{n} \sum_{k=1}^{n} \left(k \frac{x}{n}\right)^p = x^p \cdot \frac{\sum_{k=1}^{n} k^p}{n^{p+1}} = x^p \cdot \frac{s_p(n+1)}{n^{p+1}},$$

where we suppressed 0, the initial end-point of the interval $[0, x]$. Taking the limit, we obtain

$$\mathcal{A}_{\mathfrak{p}_p}(x) = \lim_{n\to\infty} x^p \cdot \frac{s_p(n+1)}{n^{p+1}} = x^p \cdot \lim_{n\to\infty} \frac{s_p(n+1)}{(n+1)^{p+1}} \left(1 + \frac{1}{n}\right)^{p+1} = \frac{x^p}{p+1}.$$

Remark The reader versed in calculus will no doubt recognize the (right-)**Riemann sum**[21] and the **Riemann integral** as its limit as follows:

$$\int_0^x t^p dt = \lim_{n\to\infty} \left(\sum_{k=1}^{n} \left(k \frac{x}{n}\right)^p \cdot \frac{x}{n} \right) = \lim_{n\to\infty} \frac{s_p(n+1) x^{p+1}}{n^{p+1}} = \frac{x^{p+1}}{p+1}, \quad 0 < p \in \mathbb{R}.$$

Returning to the main line, we close this section by a simple observation on powers. In rare instances, an **irrational** number raised to an **irrational** exponent can be a rational number. A non-constructive proof is as follows.

Let $a \in \mathbb{N}$ be a natural number which is not a square. Then $\sqrt{a}$ is an irrational number. Consider this as the base of the real exponent $\sqrt{a}^{\sqrt{2}}$. Now, if this is a rational number, then we are done (since $\sqrt{2}$ is also irrational). If it is an irrational number, then we take this as a new base of the iterated exponent

$$r = \left(\sqrt{a}^{\sqrt{2}}\right)^{\sqrt{2}}.$$

Using an exponentiation identity, we calculate

$$r = \sqrt{a}^{\sqrt{2}\cdot\sqrt{2}} = \sqrt{a}^2 = a.$$

Since this is a natural number, the claim follows.

History

In 1900, the German mathematician David Hilbert (1862–1943) posed 23 main problems in mathematics. Part of the seventh problem is concerned with irrationality of rational numbers raised to exponents that are square roots of integers. (More precisely, he posed the problem whether an

[21] Let $a < b$, $a, b \in \mathbb{R}$, and $n \in \mathbb{N}$. Given a subdivision $a = x_0 < x_1 < \ldots < x_n = b$ of the closed interval $[a, b]$ and a function $f : [a, b] \to \mathbb{R}$, we define the left-Riemann sum of f (corresponding to this subdivision) by $\sum_{k=1}^{n} f(x_{k-1})(x_k - x_{k-1})$. The right-Riemann sum is defined by replacing $f(x_{k-1})$ by $f(x_k)$ in the sum.

algebraic number $\neq 0, 1$ raised to an algebraic irrational exponent is transcendental. Here a real number is algebraic if it is a root of a polynomial with rational coefficients; otherwise it is called transcendental.) As a special case he posed the problem of irrationality (transcendentality) of the real number

$$2^{\sqrt{2}} = 2.665144142690225188650297249 87\ldots$$

which was subsequently named after him. (Although this problem was positively resolved in 1930 by the Russian mathematician Rodion Kuzmin (1891–1949), this number is also called the Gelfond–Schneider number named after Aleksandr Gelfond (1906–1968) and Theodor Schneider (1911–1988), two major contributors to this problem and its generalizations.)

Exercises

3.2.1. Determine $\sqrt{16^{16x^2}}$, $x \in \mathbb{R}$.

3.2.2. Derive the inequalities

$$2\sqrt{n+1} - 2 < 1 + \frac{1}{\sqrt{2}} + \cdots + \frac{1}{\sqrt{n}} < 2\sqrt{n}, \quad 2 \leq n \in \mathbb{N}.$$

(Note the obvious consequence $\sum_{n=1}^{\infty} 1/\sqrt{n} = \infty$, the p-series for $p = 1/2$. More precisely, we have

$$2\frac{\sqrt{n+1}-1}{\sqrt{n}} < \frac{1 + \frac{1}{\sqrt{2}} + \cdots + \frac{1}{\sqrt{n}}}{\sqrt{n}} < 2,$$

which gives the limit

$$\lim_{n\to\infty} \frac{1 + \frac{1}{\sqrt{2}} + \cdots + \frac{1}{\sqrt{n}}}{\sqrt{n}} = 2,$$

as in Example 3.4.2.)

3.2.3. In this exercise, we outline a direct proof of the Bernoulli inequality for real exponents.[22] Let

$$A = \{q \in \mathbb{Q} \,|\, 0 < q < 1,\ (1+r)^q < 1 + qr,\ -1 < r \neq 0\}.$$

Show that A is dense in $(0, 1)$ using the following steps: (1) $1/2 \in A$, (2) $q \in A$ implies $1 - q \in A$, (3) $q, q' \in A$ implies $q \cdot q', (q + q')/2 \in A$, and (4) $\sum_{k=1}^{n} a_k 2^{-k} \in A$, $a_1, \ldots, a_n \in \{0, 1\}$. Finally, use density of A to show that $A = (0, 1)$.

[22] See Yuan-Chuan Li, Cheh-Chih Yeh, *Some Equivalent Forms of the Bernoulli's Inequality: A Survey*, Applied Mathematics, Vol. 4, No 7 (2013) 1070–1093; https://doi.org/10.4236/am.2013.47146.

3.3 Logarithms

Our starting point is the following fundamental result.

Proposition 3.3.1 *Let $1 < r \in \mathbb{R}$ and $0 < t \in \mathbb{R}$. Then there exists a unique $s \in \mathbb{R}$ such that $r^s = t$.*

The exponent $s \in \mathbb{R}$ in the proposition above is called the **logarithm of t to the base r** or the **base r-logarithm of t**, and it is denoted by $s = \log_r t$.

Proof Unicity follows directly from monotonicity of the exponentiation: For $1 < r \in \mathbb{R}$, if $s < s'$, then $r^s < r^{s'}$.

Turning to the proof of existence, for $1 < r \in \mathbb{R}$ and $0 < t \in \mathbb{R}$, we define

$$A = \{u \in \mathbb{R} \mid r^u < t\}.$$

Since $\lim_{n\to\infty} 1/r^n = 0$, we have $r^n > t$ for large $n \in \mathbb{N}$. Hence the set A is bounded above. We let $s = \sup A$. We claim that $r^s = t$ holds.

Assume $r^s < t$. We denote $1 < v = t/r^s \in \mathbb{R}$ and choose $2 \leq n \in \mathbb{N}$ such that $n > (r-1)/(v-1)$. The Bernoulli inequality gives

$$r^{1/n} < 1 + \frac{1}{n}(r-1) < v = t/r^s.$$

Using the exponential identities, this gives $r^{s+1/n} < t$. By the definition of A, this gives $s + 1/n \in A$. We obtain that s cannot be the supremum of A, a contradiction. We thus have $r^s \geq t$.

The argument to show $r^s \leq t$ is standard. For $n \in \mathbb{N}$, the number $s - 1/n$ cannot be an upper bound for A, and so there exists $u_n \in A$ such that $s - 1/n < u_n (\leq s)$. We choose a rational number $q_n \in \mathbb{Q}$ such that $s - 1/n < q_n < u_n \leq s$, $n \in \mathbb{N}$. By monotonicity of the limit, we have $\lim_{n\to\infty} q_n = s$. Since $r^{q_n} < r^{u_n} \leq t$, again by monotonicity of the limit and the definition of the real exponent, we have $r^s = \lim_{n\to\infty} r^{q_n} \leq t$. The proposition follows.

The logarithm defined by the proposition above can immediately be extended to bases $0 < r < 1$, $r \in \mathbb{R}$, by setting

$$\log_r s = -\log_{1/r} s.$$

With this, for $0 < r \in \mathbb{R}$, $r \neq 1$, and $0 < t \in \mathbb{R}$, we have

$$r^s = t \quad \text{if and only if} \quad s = \log_r t.$$

(Logarithm with base 1 is not defined.) From now on, the base is always understood to be a positive real number, not equal to one.

Clearly, we have $\log_r 1 = 0$ and $\log_r r = 1$. In addition, by the above, we have

$$r^{\log_r t} = t, \quad 0 < t \in \mathbb{R}.$$

The logarithm satisfies several identities that mirror those of the exponentiation.

For $0 < u, v \in \mathbb{R}$, we have

$$\log_r(uv) = \log_r(u) + \log_r(v), \quad \log_r\left(\frac{u}{v}\right) = \log_r(u) - \log_r(v), \quad \log_r(u^v) = v\log_r(u).$$

We first derive the last identity. For $0 < u, v \in \mathbb{R}$, $\log_r(u^v)$ is the unique real number s such that $r^s = u^v$. By the above, we have

$$u^v = \left(r^{\log_r u}\right)^v = r^{v\log_r u}.$$

Hence, $\log_r(u^v) = s = v\log_r(u)$, and the last identity follows.

For the first identity, for $0 < u, v \in \mathbb{R}$, we have

$$\log_r(uv) = \log_r(r^{\log_r u} r^{\log_r v}) = \log_r(r^{\log_r u + \log_r v}) = \log_r u + \log_r v.$$

The proof of the second identity is analogous.

Example 3.3.1 Which is bigger $5^{\log_7 3}$ or $3^{\log_7 5}$?

They are equal since

$$\log_7\left(5^{\log_7 3}\right) = \log_7 3 \cdot \log_7 5 = \log_7\left(3^{\log_7 5}\right).$$

Example 3.3.2 Solve the following system of equations:

$$2^u + 3^v = 5, \quad 8^u + 9^v = 17.$$

Clearly, $u = v = 1$ is a solution. To see if there are other solutions, we first set $x = 2^u$ and $y = 3^v$. The exponential identities give $8^u = 2^{3u} = x^3$ and $9^v = 3^{2v} = y^2$. In terms of x, y, the system of equations can be written as

$$x + y = 5, \quad x^3 + y^2 = 17.$$

Eliminating y, we obtain

$$x^3 + (5 - x)^2 - 17 = (x - 1)(x - 2)(x + 4) = 0.$$

For $x = 1$, we have $y = 4$, and these give $u = 0$ and $v = \log_3 4 = 2\log_3 2$. For $x = 2$, we have $y = 3$, and these give $u = 1$ and $v = 1$. Finally, $x = -4$ is not realized.

Returning to the main line, we have the **Change of Base Formula**

$$\log_r t = \log_r r' \cdot \log_{r'} t, \quad 0 < t \in \mathbb{R}.$$

This follows from the earlier identities as

$$r^{\log_r r' \cdot \log_{r'} t} = \left(r^{\log_r r'}\right)^{\log_{r'} t} = r'^{\log_{r'} t} = t = r^{\log_r t}.$$

Example 3.3.3 We have

$$\log_r t = \log_{r^2} t^2 = \log_{r^3} t^3 = \ldots, \quad 0 < t \in \mathbb{R}.$$

Indeed, using the Change of Base Formula and the logarithmic identities, for $n \in \mathbb{N}$, we have

$$\log_{r^n} t^n = \frac{\log_r t^n}{\log_r r^n} = \frac{n \log_r t}{n \log_r r} = \log_r t.$$

The idea in the previous example can be used in the following:

Example 3.3.4 Solve the system of equations[23]

$$\log_8(x) + \log_4(y^2) = 5, \quad \log_8(y) + \log_4(x^2) = 7.$$

Clearly, $0 < x, y \in \mathbb{R}$. We have

$$\log_8(x) + \log_4(y^2) = \log_{2^3}(\sqrt[3]{x})^3 + \log_{2^2}(y^2) = \log_2\left(\sqrt[3]{x}\right) + \log_2(y) = \log_2\left(\sqrt[3]{x}y\right) = 5.$$

This gives $\sqrt[3]{x}y = 2^5$. Similarly, we have $\log_8(y) + \log_4(x^2) = \log_2\left(x\sqrt[3]{y}\right) = 7$, or equivalently, $x\sqrt[3]{y} = 2^7$. The system of equations above therefore reduces to

$$xy^3 = 2^{15} \quad \text{and} \quad x^3 y = 2^{21}.$$

Eliminating y, we calculate $x^8 = (2^{21})^3/2^{15} = 2^{48}$. We obtain $x = 2^6 = 64$ and $y = 2^3 = 8$. The example follows.

Example 3.3.5 Determine

$$\left(\log_2 3\right)\left(\log_3 4\right)\cdots\left(\log_{2^n-1}(2^n)\right).$$

A simple induction in the use of the Change of Base formula shows that this expression is equal to $\log_2(2^n) = n \log_2 2 = n$.

Example 3.3.6 Write the following expression as a single logarithm:[24]

[23] The problem of calculating the product xy was in the American Invitational Mathematics Examination, 1984.

[24] The special case $n = 5$ was in the American High School Mathematics Examination, 1978.

$$\frac{1}{\log_2 x} + \frac{1}{\log_3 x} + \cdots + \frac{1}{\log_n x}, \quad 0 < x \neq 1, \ x \in \mathbb{R}.$$

A special case of the Change of Base Formula is the following:

$$\frac{1}{\log_b a} = \log_a b, \quad 0 < a, b \neq 1, \ a, b \in \mathbb{R}.$$

Using this, the expression above is rewritten as

$$\log_x 2 + \log_x 3 + \cdots + \log_x n = \log_x (n!).$$

The logarithm is **monotonic**. For $r > 1$, the logarithm $\log_r t$ is strictly increasing in $t \in \mathbb{R}$; that is, $0 < t < t', t, t' \in \mathbb{R}$, implies $\log_r t < \log_r t'$. For $0 < r < 1$, the logarithm $\log_r t$ is strictly decreasing in $t \in \mathbb{R}$; that is, $0 < t < t', t, t' \in \mathbb{R}$, implies $\log_r t > \log_r t'$.

It is enough to show the first statement. Let $r > 1$. If $0 < t < t', t, t' \in \mathbb{R}$, then we have

$$t = r^{\log_r t} < r^{\log_r t'} = t'.$$

By monotonicity of the exponentiation, this holds if and only if $\log_r t < \log_r t'$. The claim follows.

Example 3.3.7 Let $0 < r < 1$ be a real number chosen at random. What are the odds[25] that the integer $[\log_2 r]$ is even?

For $0 < r < 1$, the logarithm $\log_2 r$ is negative. We write an even negative integer in the form $-2n$, $n \in \mathbb{N}$. By the definition of the greatest integer, the condition $[\log_2 r] = -2n$ amounts to $-2n \leq \log_2 r < -2n + 1$, or equivalently, $2^{-2n} \leq r < 2^{-2n+1}$, $n \in \mathbb{N}$. The length of this interval is $2^{-2n+1} - 2^{-2n} = 2^{-2n}$. Summing up with respect to $n \in \mathbb{N}$, the probability that $[\log_2 r]$, $0 < r < 1$, is an even integer is $\sum_{n=1}^{\infty} 2^{-2n}$. This, however, is a geometric series, and the Infinite Geometric Series Formula gives

$$\sum_{n=1}^{\infty} 2^{-2n} = \sum_{n=1}^{\infty} \frac{1}{2^{2n}} = \sum_{n=1}^{\infty} \left(\frac{1}{4}\right)^n = \frac{1/4}{1 - 1/4} = \frac{1}{3}.$$

Returning to the main line, the Bernoulli inequality has a logarithmic counterpart. Recall that, for $0 < a \neq 1$, $a \in \mathbb{R}$, we have

$$a^r < 1 + r(a - 1), \ 0 < r < 1, \ r \in \mathbb{R} \quad \text{and} \quad a^r > 1 + r(a - 1), \ 1 < r \in \mathbb{R}.$$

[25] That is, "what is the probability…" The author could not help rewording this well-known contest preparation problem for the sake of making a pun.

Letting $a = 2$ and taking the base 2 logarithm of both sides, we obtain

$$r < \log_2(1+r),\ 0 < r < 1,\ r \in \mathbb{R} \quad \text{and} \quad r > \log_2(1+r),\ 1 < r \in \mathbb{R}.$$

Example 3.3.8 Let $1 < r, t \in \mathbb{R}$. Then $\log_r t$ is a rational number if and only if $r^m = t^n$ for some $m, n \in \mathbb{N}$. In particular, if $r, t > 1$ are integers, then rationality of $\log_r t$ implies that r and t must have the same prime divisors. Hence, $\log_2(3), \log_2(5), \ldots, \log_3(2)$, etc. are irrational numbers.

Letting $s = \log_r t > 0$, we have $r^s = t$. The number s is rational if and only if $s = m/n$ for some $m, n \in \mathbb{N}$. We thus have $\sqrt[n]{r^m} = t$, or equivalently, $r^m = t^n$.

Example 3.3.9 (Revisited) Recall, from Example 3.2.5 that, for $1 < a \in \mathbb{R}$ and $2 \leq n \in \mathbb{N}$, we have

$$0 < \sqrt[n]{a} - 1 < \frac{a-1}{n}.$$

We let $a = 2$ and rewrite this as

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

Taking the base 2 logarithm of both sides and simplifying, we obtain

$$\frac{1}{n} < \log_2\left(1 + \frac{1}{n}\right) = \log_2\left(\frac{n+1}{n}\right) = \log_2(n+1) - \log_2(n).$$

This gives

$$\frac{1}{n} - \log_2(n+1) < -\log_2(n), \quad 2 \leq n \in \mathbb{N}.$$

We now recall the partial sum of reciprocals (of the harmonic series):

$$H_n = 1 + \frac{1}{2} + \cdots + \frac{1}{n}, \quad n \in \mathbb{N}.$$

Adding H_{n-1}, $2 \leq n \in \mathbb{N}$, to both sides, we obtain

$$H_n - \log_2(n+1) < H_{n-1} - \log_2(n), \quad 2 \leq n \in \mathbb{N}.$$

We obtain that the sequence $(H_n - \log_2(n+1))_{n \in \mathbb{N}}$ is strictly decreasing. Since $H_1 - \log_2(2) = 0$, we arrive at the important inequality

$$H_n = 1 + \frac{1}{2} + \cdots + \frac{1}{n} < \log_2(n+1), \quad 2 \leq n \in \mathbb{N}.$$

(Note that equality holds for $n = 1$.)

Example 3.3.10 As a generalization of the p-series, for $0 < p, q \in \mathbb{R}$, we consider the infinite series

$$\sum_{n=2}^{\infty} \frac{1}{n^p \cdot (\log_r n)^q}.$$

By the Cauchy Condensation Test, this series equiconverges with the infinite series

$$\sum_{m=1}^{\infty} \frac{2^m}{(2^m)^p \cdot (\log_r 2^m)^q} = \frac{1}{(\log_r 2)^q} \sum_{m=1}^{\infty} \frac{(2^{1-p})^m}{m^q}.$$

If $p = 1$, then, up to the constant multiple in front of the summation, this becomes the q-series, which is convergent for $q > 1$ and divergent for $0 < q \leq 1$.

If $p > 1$ (and $q > 0$ is arbitrary), then we have

$$\sum_{m=1}^{\infty} \frac{(2^{1-p})^m}{m^q} = \sum_{m=1}^{\infty} \frac{1}{(2^{p-1})^m m^q} \leq \sum_{m=1}^{\infty} \frac{1}{(2^{p-1})^m}.$$

The last sum is geometric with ratio $0 < 1/2^{p-1} < 1$, and hence it is convergent.

If $p < 1$, then, letting $a = 2^{1-p} > 1$, we have

$$\lim_{m\to\infty} \frac{(2^{1-p})^m}{m^q} = \lim_{m\to\infty} \frac{a^m}{m^q} = \infty,$$

where, in the last equality, we used Example 3.2.7.

Exercises

3.3.1. Let $0 \neq a, b, c \in \mathbb{R}$, and $1 < x, y, z \in \mathbb{R}$ and $0 < w \in \mathbb{R}$ such that $\log_x w = a$, $\log_y w = b$, and $\log_{xyz} w = c$. Find[26] $\log_z w$ in terms of a, b, c.

3.3.2. Let $2 \leq n \in \mathbb{N}$. Solve $[\log_n(x)] = \log_n[x]$ for $1 \leq x \in \mathbb{R}$.

3.4 The Stolz–Cesàro Theorems

In this section, we discuss a powerful criterion for convergence of sequences due to Otto Stolz (published in 1885) and Ernesto Cesàro (1859–1906) (published in 1888).

[26] A special (numerical) case was a problem in the American Invitational Mathematics Examination, 1983.

Stolz–Cesàro Theorem. *Let $(a_n)_{n\in\mathbb{N}}$ and $(b_n)_{n\in\mathbb{N}}$ be real sequences such that $(b_n)_{n\in\mathbb{N}}$ is strictly increasing with $\lim_{n\to\infty} b_n = \infty$. Then, we have*

$$\liminf_{n\to\infty} \frac{a_n - a_{n-1}}{b_n - b_{n-1}} \le \liminf_{n\to\infty} \frac{a_n}{b_n} \le \limsup_{n\to\infty} \frac{a_n}{b_n} \le \limsup_{n\to\infty} \frac{a_n - a_{n-1}}{b_n - b_{n-1}}.$$

In particular,

$$\lim_{n\to\infty} \frac{a_n - a_{n-1}}{b_n - b_{n-1}} = \lim_{n\to\infty} \frac{a_n}{b_n},$$

provided that the limit on the left-hand side exists.

Proof It is enough to prove the inequality for the limit superior. Let $c \in \mathbb{R}$ such that

$$\limsup_{n\to\infty} \frac{a_n - a_{n-1}}{b_n - b_{n-1}} < c.$$

Then there exists $N \in \mathbb{N}_0$ such that

$$\frac{a_n - a_{n-1}}{b_n - b_{n-1}} < c, \quad n > N.$$

Thus, for $n > N$, we have

$$\begin{aligned} a_{N+1} - a_N &< c\,(b_{N+1} - b_N) \\ a_{N+2} - a_{N+1} &< c\,(b_{N+2} - b_{N+1}) \\ \cdots &\quad \cdots \\ a_{n-1} - a_{n-2} &< c\,(b_{n-1} - b_{n-2}) \\ a_n - a_{n-1} &< c\,(b_n - b_{n-1})\,. \end{aligned}$$

Adding, we obtain

$$a_n - a_N < c(b_n - b_N),$$

or equivalently

$$\frac{a_n}{b_n} < c + \frac{a_N - cb_N}{b_n}, \quad n > N.$$

Using this, we have

$$\limsup_{n\to\infty} \frac{a_n}{b_n} \le c + \limsup_{n\to\infty} \frac{a_N - cb_N}{b_n} = c,$$

where we used $\lim_{n\to\infty} b_n = \infty$.

The inequality and thereby the theorem follow.

Stolz–Cesàro Theorem (Equivalent Formulation). *Let $(a_n)_{n\in\mathbb{N}}$ and $(b_n)_{n\in\mathbb{N}}$ be real sequences such that $b_n > 0$, $n \in \mathbb{N}$, and $\lim_{n\to\infty} b_n = \infty$. Then, we have*

$$\liminf_{n\to\infty} \frac{a_n}{b_n} \le \liminf_{n\to\infty} \frac{a_1 + \cdots + a_n}{b_1 + \cdots + b_n} \le \limsup_{n\to\infty} \frac{a_1 + \cdots + a_n}{b_1 + \cdots + b_n} \le \limsup_{n\to\infty} \frac{a_n}{b_n}.$$

In particular,

$$\lim_{n\to\infty} \frac{a_n}{b_n} = \lim_{n\to\infty} \frac{a_1 + \cdots + a_n}{b_1 + \cdots + b_n},$$

provided that the limit on the left-hand side exists.

Proof This follows directly from the previous by the substitution $a_n \mapsto a_1 + \cdots + a_n$ and $b_n \mapsto b_1 + \cdots + b_n$, $n \in \mathbb{N}$.

Letting $b_n = n$ (or $b_n = 1$), $n \in \mathbb{N}$, we obtain the following special cases valid for **any** real sequence $(a_n)_{n\in\mathbb{N}}$:

$$\liminf_{n\to\infty}(a_n - a_{n-1}) \le \liminf_{n\to\infty} \frac{a_n}{n} \le \limsup_{n\to\infty} \frac{a_n}{n} \le \limsup_{n\to\infty}(a_n - a_{n-1}),$$

and

$$\liminf_{n\to\infty} a_n \le \liminf_{n\to\infty} \frac{a_1 + \cdots + a_n}{n} \le \limsup_{n\to\infty} \frac{a_1 + \cdots + a_n}{n} \le \limsup_{n\to\infty} a_n.$$

In particular,

$$\lim_{n\to\infty}(a_n - a_{n-1}) = \lim_{n\to\infty} \frac{a_n}{n},$$

and

$$\lim_{n\to\infty} a_n = \lim_{n\to\infty} \frac{a_1 + \cdots + a_n}{n},$$

provided that the limits on the left-hand sides exist. We call these the additive Stolz–Cesàro formulas.

Let $(a_n)_{n\in\mathbb{N}}$ be a real sequence with positive members. For $n \in \mathbb{N}$, let $b_n = \log_2(a_n)$, or equivalently, $a_n = 2^{b_n}$. Applying the exponential identities, we obtain

$$2^{\frac{b_1+\cdots+b_n}{n}} = \sqrt[n]{a_1 \cdots a_n}.$$

Using monotonicity of the exponentiation and the Stolz–Cesàro limit formulas above for the sequence $(b_n)_{n\in\mathbb{N}}$, we obtain

$$\liminf_{n\to\infty} a_n \le \liminf_{n\to\infty} \sqrt[n]{a_1 \cdots a_n} \le \limsup_{n\to\infty} \sqrt[n]{a_1 \cdots a_n} \le \limsup_{n\to\infty} a_n,$$

and

$$\liminf_{n\to\infty}\frac{a_n}{a_{n-1}}\le\liminf_{n\to\infty}\sqrt[n]{a_n}\le\limsup_{n\to\infty}\sqrt[n]{a_n}\le\limsup_{n\to\infty}\frac{a_n}{a_{n-1}}.$$

In particular,

$$\lim_{n\to\infty}a_n=\lim_{n\to\infty}\sqrt[n]{a_1\cdots a_n},$$

and

$$\lim_{n\to\infty}\frac{a_n}{a_{n-1}}=\lim_{n\to\infty}\sqrt[n]{a_n},$$

provided that the limits on the left-hand sides exist. We call these the multiplicative Stolz–Cesàro formulas.

Using the Stolz–Cesàro formulas, several of our earlier limits (derived using estimates with the Bernoulli inequality) can be obtained in a simple and direct way.

In particular, Examples 3.2.5, 3.2.8–3.2.9 follow using the multiplicative Stolz–Cesàro formula:

$$\lim_{n\to\infty}\sqrt[n]{a}=\lim_{n\to\infty}\frac{a}{a}=1,\quad a_n=a,\ 0<a\in\mathbb{R};$$

$$\lim_{n\to\infty}\sqrt[n]{n}=\lim_{n\to\infty}\frac{n}{n-1}=1,\quad a_n=n,\ 2\le n\in\mathbb{N};$$

$$\lim_{n\to\infty}\sqrt[n]{n!}=\lim_{n\to\infty}\frac{n!}{(n-1)!}=\lim_{n\to\infty}n=\infty.$$

Moreover, for Example 3.2.6, assuming $0<a<b$, $a,b\in\mathbb{R}$, we calculate

$$\lim_{n\to\infty}\sqrt[n]{a^n+b^n}=\lim_{n\to\infty}\frac{a^n+b^n}{a^{n-1}+b^{n-1}}=b\lim_{n\to\infty}\frac{1+(a/b)^n}{1+(a/b)^{n-1}}=b=\max(a,b),$$

where the geometric sequence $((a/b)^n)_{n\in\mathbb{N}}$ with ratio converges to zero since $0<a/b<1$. The two extensions of this limit can be treated in the same way.

Remark The **root test** and the **ratio test** are simple criteria for the convergence of an infinite series $\sum_{n=1}^{\infty}a_n$ with positive terms $0<a_n\in\mathbb{R}$, $n\in\mathbb{N}$. By the Monotone Convergence Theorem, $\sum_{n=1}^{\infty}a_n$ can have only two cases; it is either finite (convergence) or infinite.

The root test states that

$$\limsup_{n\to\infty}\sqrt[n]{a_n}<1$$

implies that $\sum_{n=1}^{\infty}a_n$ is finite.

Indeed, assume that $\limsup_{n\to\infty} \sqrt[n]{a_n} < r$ for some $0 < r < 1$, $r \in \mathbb{R}$. This means that there exists $N \in \mathbb{N}$ such that $\sqrt[n]{a_n} < r$ for $n \geq N$. Hence, $a_n < r^n$ for $n \geq N$. We obtain

$$\sum_{n=1}^{\infty} a_n = \sum_{n=1}^{N-1} a_n + \sum_{n=N}^{\infty} a_n < \sum_{n=1}^{N-1} a_n + \sum_{n=N}^{\infty} r^n = \sum_{n=1}^{N-1} a_n + \frac{r^N}{1-r},$$

where we used the Infinite Geometric Series Formula. (For $N = 1$, the finite sum is absent.) The root test follows.

The ratio test states that

$$\limsup_{n\to\infty} \frac{a_n}{a_{n-1}} < 1$$

implies that $\sum_{n=1}^{\infty} a_n$ is finite.

By the multiplicative Stolz–Cesàro Theorem above, the latter limit superior does not exceed the former, so that the ratio test is a direct consequence of the root test.[27]

In a similar vein, if

$$\liminf_{n\to\infty} \sqrt[n]{a_n} > 1 \quad \text{or} \quad \liminf_{n\to\infty} \frac{a_n}{a_{n-1}} > 1,$$

then $\sum_{n=1}^{\infty} a_n = \infty$.

We now return to the main line and give new applications of the Stolz–Cesàro Theorems.

Example 3.4.1 Let $(a_n)_{n\in\mathbb{N}}$ be a real sequence. We have

$$\lim_{n\to\infty} \frac{a_1 + \cdots + a_n}{n^{p+1}} = \frac{1}{p+1} \lim_{n\to\infty} \frac{a_n}{n^p}, \quad -1 < p \in \mathbb{R}.$$

Indeed, by the first Stolz–Cesàro formula, we have

$$\lim_{n\to\infty} \frac{a_1 + \cdots + a_n}{n^{p+1}} = \lim_{n\to\infty} \frac{a_n}{n^{p+1} - (n-1)^{p+1}} = \lim_{n\to\infty} \frac{a_n}{n^p} \cdot \lim_{n\to\infty} \frac{n^p}{n^{p+1} - (n-1)^{p+1}}.$$

We calculate the reciprocal of the last limit as

$$\lim_{n\to\infty} \frac{n^{p+1} - (n-1)^{p+1}}{n^p} = \lim_{n\to\infty} \frac{n^p + n^{p-1}(n-1) + \cdots + n(n-1)^{p-1} + (n-1)^p}{n^p}$$
$$= \lim_{n\to\infty} \left(1 + \left(1 - \frac{1}{n}\right) + \cdots + \left(1 - \frac{1}{n}\right)^{p-1} + \left(1 - \frac{1}{n}\right)^p\right) = p + 1,$$

[27] It is a bit of irony that the root test implies the ratio test, yet, in specific examples, the ratio test is far more useful.

where we used the identity

$$u^{p+1} - v^{p+1} = (u - v)\left(u^p + u^{p-1}v + \cdots + uv^{p-1} + v^p\right).$$

The example follows.

As a special case, letting $a_n = n^p$, $n \in \mathbb{N}$, we obtain

$$\lim_{n\to\infty} \frac{s_p(n)}{n^{p+1}} = \lim_{n\to\infty} \frac{s_p(n+1)}{n^{p+1}} = \frac{1}{p+1}, \quad -1 < p \in \mathbb{R},$$

where

$$s_p(n) = 1^p + 2^p + \cdots + (n-1)^p, \quad 2 \le n \in \mathbb{N}.$$

We recover the limit in Example 3.2.12 (Note the slightly extended range $-1 < p \in \mathbb{R}$.)

Example 3.4.2 Show that

$$\lim_{n\to\infty} \frac{1 + \frac{1}{\sqrt{2}} + \cdots + \frac{1}{\sqrt{n}}}{\sqrt{n}} = 2.$$

(Note that Exercise 3.2.2 at the end of Section 3.2 gives precise lower and upper bounds for $1 + \frac{1}{\sqrt{2}} + \cdots + \frac{1}{\sqrt{n}}$ and thereby provides an alternative derivation of the limit above.)

Letting $a_n = 1 + \frac{1}{\sqrt{2}} + \cdots + \frac{1}{\sqrt{n}}$ and $b_n = \sqrt{n}$, we use the first Stolz–Cesàro limit relation and calculate

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

Example 3.4.3 Show that

$$\lim_{n\to\infty} \sqrt[n]{H_n} = \lim_{n\to\infty} \sqrt[n]{1 + \frac{1}{2} + \cdots + \frac{1}{n}} = 1.$$

Recall from Example 3.1.6 that $\lim_{n\to\infty} H_n = \infty$. We now use the multiplicative Stolz–Cesàro formula to obtain

$$\lim_{n\to\infty} \sqrt[n]{H_n} = \lim_{n\to\infty} \frac{H_n}{H_{n-1}} = \lim_{n\to\infty} \left(1 + \frac{1/n}{1 + 1/2 + \cdots + 1/(n-1)}\right) = 1.$$

Example 3.4.4 Let $(a_n)_{n\in\mathbb{N}}$ be a real sequence with limit $\lim_{n\to\infty} a_n = a$. Show that

$$\lim_{n\to\infty} \frac{na_1 + (n-1)a_2 + \cdots + 2a_{n-1} + a_n}{n^2} = \frac{a}{2}.$$

We use the first Stolz–Cesàro limit relation **twice** with obvious roles as follows:

$$\lim_{n\to\infty} \frac{na_1 + (n-1)a_2 + \cdots + 2a_{n-1} + a_n}{n^2} = \lim_{n\to\infty} \frac{a_1 + \cdots + a_n}{n^2 - (n-1)^2}$$
$$= \lim_{n\to\infty} \frac{a_1 + \cdots + a_n}{2n-1} = \lim_{n\to\infty} \frac{a_n}{2} = \frac{a}{2}.$$

Exercises

3.4.1. Find the limit

$$\lim_{n\to\infty} \frac{\log_2(n!)}{n \log_2(n)}.$$

3.4.2. Let $(a_n)_{n\in\mathbb{N}}$ be a real sequence with positive terms such that $\lim_{n\to\infty} a_n/n = \infty$. Show that

$$\lim_{n\to\infty} \frac{1}{\sqrt{n}} \sum_{k=1}^{n} \frac{1}{\sqrt{a_k}} = 0.$$

Chapter 4
Limits of Real Functions

"Nothing takes place in the world
whose meaning is not that of
some maximum or minimum.
Leonhard Euler (1707–1783)

The principal aim of this chapter is to give a short introduction to the limit inferior and limit superior and (thereby) the limit for functions. Many (arithmetic and analytic) properties of these functional limits can be derived by establishing their link with sequential limits. In our largely classical approach, continuity and differentiability of real functions are also introduced and treated here as special limits (stopping short of fully developed advanced differential calculus) mainly because the derivative as a limit is an indispensable tool for later developments. For future purposes, we also give quick proofs of the Extreme Values Theorem, the Intermediate Value Theorem, and the Fermat Principle.

4.1 Limit Inferior and Limit Superior

Let $\mathcal{D}$ be a set and $f : \mathcal{D} \to \mathbb{R}$ a function with domain $\mathcal{D}$. Given a subset $\mathcal{C} \subset \mathcal{D}$, we define the **supremum** of the function f on $\mathcal{C}$ by

$$\sup_{\mathcal{C}} f = \sup_{x \in \mathcal{C}} f(x) = \sup\{f(x) \,|\, x \in \mathcal{C}\},$$

where the first equality is notation and the last is the definition. If the supremum exists, then we say that the function f is **bounded above** on $\mathcal{C}$ and write $\sup_{\mathcal{C}} f < \infty$. If the supremum does not exist, then we say that the function f is **unbounded above** on $\mathcal{C}$ and write $\sup_{\mathcal{C}} f = \infty$.

G. Toth, *Elements of Mathematics*, Undergraduate Texts in Mathematics,
https://doi.org/10.1007/978-3-030-75051-0_4

Similarly, the **infimum** of f over $\mathcal{C}$ is

$$\inf_{\mathcal{C}} f = \inf_{x\in\mathcal{C}} f(x) = \inf\{f(x) \mid x \in \mathcal{C}\}.$$

If the infimum exists, then the function f is **bounded below** on $\mathcal{C}$, $\inf_{\mathcal{C}} f > -\infty$; if it does not exist, then f is **unbounded below**, $\inf_{\mathcal{C}} f = -\infty$.
Finally, f is **bounded** if it is bounded above and below, or equivalently, we have $\sup_{\mathcal{C}} |f| < \infty$.

Remark We have $|\sup_{\mathcal{C}} f| \le \sup_{\mathcal{C}} |f|$; in particular, boundedness of f on $\mathcal{C}$ implies $|\sup_{\mathcal{C}} f| < \infty$. The converse, however, fails, that is, $|\sup_{\mathcal{C}} f| < \infty$ does not imply boundedness of f on $\mathcal{C}$. (Let $f : (-\infty, 0) \to \mathbb{R}$ be defined by $f(x) = 1/x$, $x < 0$. Then, we have $\sup_{(-\infty,0)} f = 0$, but f is not bounded on $(-\infty, 0)$.)

Let $f : \mathcal{D} \to \mathbb{R}$ be a **real function**, that is, the domain of definition $\mathcal{D} \subset \mathbb{R}$ is a set of real numbers. Let $c \in \mathbb{R}$, and assume that, for some $0 < d \in \mathbb{R}$, the function f is defined on the **deleted open interval**

$$(c-d, c+d)^\circ = (c-d, c+d) \setminus \{c\} = (c-d, c) \cup (c, c+d) \subset \mathcal{D}.$$

For $0 < \delta \le d$, we consider

$$\bar{S}(\delta) = \sup_{(c-\delta,c+\delta)\setminus\{c\}} f = \sup_{0<|x-c|<\delta} f(x).$$

The function $\bar{S} : \delta \in (0, d] \to \mathbb{R}$ (depending on f and c) is **increasing**; for $0 < \delta'' \le \delta' \le d$, we have $\bar{S}(\delta'') \le \bar{S}(\delta')$ (since $(c-\delta'', c+\delta'') \subset (c-\delta', c+\delta')$).
With this, we define the **limit superior** of f at c as the infimum of $\bar{S}$ over $(0, d]$, that is, we set

$$\limsup_{x\to c} f(x) = \inf_{0<\delta\le d} \bar{S}(\delta) = \inf_{0<\delta\le d} \sup_{0<|x-c|<\delta} f(x).$$

Similarly, to define the **limit inferior** of f at c, for $0 < \delta \le d$, we consider

$$\underline{S}(\delta) = \inf_{(c-\delta,c+\delta)\setminus\{c\}} f = \inf_{0<|x-c|<\delta} f(x).$$

The function $\underline{S} : \delta \in (0, d] \to \mathbb{R}$ (depending on f and c) is **decreasing**; for $0 < \delta'' \le \delta' \le d$, we have $\underline{S}(\delta'') \ge \underline{S}(\delta')$.
With this, we define the **limit inferior** of f at c as the supremum of $\underline{S}$ over $(0, d]$, that is, we set

$$\liminf_{x\to c} f(x) = \sup_{0<\delta\le d} \underline{S}(\delta) = \sup_{0<\delta\le d} \inf_{0<|x-c|<\delta} f(x).$$

Remark The limit superior and limit inferior are often indicated by overline and underline:

$$\overline{\lim_{x\to c}}\, f(x) = \limsup_{x\to c} f(x) \quad \text{and} \quad \underline{\lim_{x\to c}}\, f(x) = \liminf_{x\to c} f(x).$$

Since multiplying both sides of an inequality by a negative number reverses the inequality sign, we have

$$\limsup_{x\to c} f(x) = -\liminf_{x\to c}(-f(x)).$$

The connection with the concept of limit superior and limit inferior of sequences is as follows:

Proposition 4.1.1 *Let $c \in \mathbb{R}$ and $f : \mathcal{D} \to \mathbb{R}$, $(c-d, c+d)^\circ \subset \mathcal{D}$, $0 < d \in \mathbb{R}$, a bounded real function. Then, for any convergent real sequence $(x_n)_{n\in\mathbb{N}}$, $0 < |x_n - c| < d$, $n \in \mathbb{N}$, with limit $\lim_{n\to\infty} x_n = c$, we have*

$$\liminf_{x\to c} f(x) \le \liminf_{n\to\infty} f(x_n) \le \limsup_{n\to\infty} f(x_n) \le \limsup_{x\to c} f(x).$$

Moreover, there exist convergent real sequences $(\bar{x}_n)_{n\in\mathbb{N}}$, $0 < |\bar{x}_n - c| < d$, $n \in \mathbb{N}$, and $(\underline{x}_n)_{n\in\mathbb{N}}$, $0 < |\underline{x}_n - c| < d$, $n \in \mathbb{N}$, with limit

$$\lim_{n\to\infty} \underline{x}_n = \lim_{n\to\infty} \bar{x}_n = c$$

such that

$$\liminf_{x\to c} f(x) = \lim_{n\to\infty} f(\underline{x}_n) \le \lim_{n\to\infty} f(\bar{x}_n) = \limsup_{x\to c} f(x).$$

Proof Since taking opposites interchanges the limit superior and limit inferior for both sequences and functions, it is enough to prove the proposition for the limit superior.

Let $(x_n)_{n\in\mathbb{N}}$, $0 < |x_n - c| < d$, $n \in \mathbb{N}$, be a convergent real sequence with limit $\lim_{n\to\infty} x_n = c$. By convergence, for any $0 < \delta \le d$, there exists $N \in \mathbb{N}$, such that $|x_n - c| < \delta$ for $n \ge N$. This gives

$$\limsup_{n\to\infty} f(x_n) = \inf_{N\in\mathbb{N}} \sup_{n\ge N} f(x_n) \le \inf_{0<\delta\le d} \sup_{0<|x-c|<\delta} f(x) = \limsup_{x\to c} f(x).$$

The first statement of the proposition follows.

Once again, it is enough to prove the second statement for the limit superior. Let

$$\bar{L} = \limsup_{x\to c} f(x) = \inf_{0<\delta\le d} \sup_{0<|x-c|<\delta} f(x).$$

Let $n \in \mathbb{N}$. The definition of the limit superior implies that there exists $0 < \delta_n \leq d$ such that for all $0 < \delta \leq \delta_n$, we have

$$\bar{L} - \frac{1}{n} < \sup_{0<|x-c|<\delta} f(x) < \bar{L} + \frac{1}{n}.$$

We may choose δ_n, $n \in \mathbb{N}$, such that $\lim_{n\to\infty} \delta_n = 0$. By the estimate above, for every $n \in \mathbb{N}$, there exists $\bar{x}_n$ such that $|\bar{x}_n - c| < \delta_n$, and

$$\bar{L} - \frac{1}{n} < f(\bar{x}_n) < \bar{L} + \frac{1}{n}.$$

This gives $\lim_{n\to\infty} f(\bar{x}_n) = \bar{L}$. The second statement follows.

As a simple corollary, we have

$$\liminf_{x\to c} f(x) \leq \limsup_{x\to c} f(x).$$

We define the **limit** $\lim_{x\to c} f(x)$ if equality holds, and, in this case, the limit is equal to the common value of the limit superior and the limit inferior.

As an immediate corollary to the proposition above, we obtain the following:

Corollary *Let $c \in \mathbb{R}$ and $f : \mathcal{D} \to \mathbb{R}$, $(c-d, c+d)^\circ \subset \mathcal{D}$, $0 < d \in \mathbb{R}$, be a real function. Then $\lim_{x\to c} f(x) = L$ if and only if, for any convergent real sequence $(x_n)_{n\in\mathbb{N}}$, $0 < |x_n - c| < d$, $n \in \mathbb{N}$, with limit $\lim_{n\to\infty} x_n = c$, we have $lim_{n\to\infty} f(x_n) = L$.*

Remark As for sequences, for a real function $f : \mathcal{D} \to \mathbb{R}$, $(c-d, c+d)^\circ \subset \mathcal{D}$, $c \in \mathbb{R}$, $0 < d \in \mathbb{R}$, we have $\lim_{x\to c} f(x) = L$ if and only if

$$\inf_{0<\delta\leq d} \sup_{0<|x-c|<\delta} |f(x) - L| = 0.$$

This is a compact reformulation of the usual definition of the limit $\lim_{x\to c} f(x) = L$.

For every $0 < \epsilon$, there exists $0 < \delta \leq d$ such that $0 < |x - c| < \delta$ implies $|f(x) - L| < \epsilon$.

History

This so-called ϵ-δ definition of the limit goes back to Bolzano in 1817, but, as noted previously, it was published posthumously. The modern formulation and notation above is due to Weierstrass.

Given a real function $f : \mathcal{D} \to \mathbb{R}$, $(c-d, c+d)^\circ \subset \mathcal{D}$, $0 < d \in \mathbb{R}$, we define the **infinite limit** $\lim_{x\to c} f(x) = \infty$ as

$$\liminf_{x\to c} f(x) = \sup_{0<\delta\leq d} \inf_{0<|x-c|<\delta} f(x) = \infty.$$

Once again, this is a compact reformulation of the customary definition of the infinite limit $\lim_{x\to c} f(x) = \infty$.

For every $0 < M \in \mathbb{R}$, there exists $0 < \delta \leq d$ such that $0 < |x - c| < \delta$ implies $M \leq f(x)$.

In a similar vein, we define the **infinite limit** $\lim_{x\to c} f(x) = -\infty$ if

$$\limsup_{x\to c} f(x) = \inf_{0<\delta\leq d} \sup_{0<|x-c|<\delta} f(x) = -\infty,$$

or equivalently:

For every $0 < M \in \mathbb{R}$, there exists $0 < \delta \leq d$ such that $0 < |x - c| < \delta$ implies $f(x) \leq -M$.

Note that the corollary above holds with L replaced by $\pm\infty$.

Returning to the main line, the proposition above also allows to transplant our previous results on the limit superior and limit inferior of sequences to those of functions. For arithmetic properties of the limit superior and limit inferior, we have

$$\begin{aligned}\liminf_{x\to c} f(x) + \liminf_{x\to c} g(x) &\leq \liminf_{x\to c}(f(x) + g(x)) \\ &\leq \limsup_{x\to c}(f(x) + g(x)) \leq \limsup_{x\to c} f(x) + \limsup_{x\to c} g(x).\end{aligned}$$

Proposition 4.1.1 combined with our earlier results on sequences in Section 3.1 has several consequences.

First, Proposition 3.1.1 gives the following:

Proposition 4.1.2 *Let $f, g : \mathcal{D} \to \mathbb{R}$, $(c - d, c + d)^\circ \subset \mathcal{D}$, $0 < d \in \mathbb{R}$, be real functions, and assume that f is bounded and $\lim_{x\to c} g(x)$ exists. Then, we have*

$$\begin{aligned}\limsup_{x\to c}(f(x) + g(x)) &= \limsup_{x\to c} f(x) + \lim_{x\to c} g(x) \\ \liminf_{x\to c}(f(x) + g(x)) &= \liminf_{x\to c} f(x) + \lim_{x\to c} g(x).\end{aligned}$$

In particular, if $\lim_{x\to c} f(x)$ and $\lim_{x\to c} g(x)$ both exist, then so does $\lim_{x\to c}(f(x) + g(x))$, and we have

$$\lim_{x\to c}(f(x) + g(x)) = \lim_{x\to c} f(x) + \lim_{x\to c} g(x).$$

Second, Proposition 3.1.2 gives the following:

Proposition 4.1.3 *Let $f, g : \mathcal{D} \to \mathbb{R}$, $(c - d, c + d)^\circ \subset \mathcal{D}$, $0 < d \in \mathbb{R}$, be real functions, and assume that f is bounded, and $\lim_{x\to c} g(x)$ exists and non-negative. Then we have*

$$\limsup_{x\to c}(f(x) \cdot g(x)) = \limsup_{x\to c} f(x) \cdot \lim_{x\to c} g(x)$$

$$\liminf_{x\to c}(f(x)\cdot g(x)) = \liminf_{x\to c} f(x)\cdot \lim_{x\to c} g(x).$$

In particular, if $\lim_{x\to c} f(x)$ *and* $\lim_{x\to c} g(x)$ *both exist, then so does* $\lim_{x\to c}(f(x)\cdot g(x))$*, and we have*

$$\lim_{x\to c}(f(x)\cdot g(x)) = \lim_{x\to c} f(x)\cdot \lim_{x\to c} g(x).$$

By a simple induction, we have

$$\lim_{x\to c} f(x)^m = \Big(\lim_{x\to c} f(x)\Big)^m, \quad m\in\mathbb{N},$$

provided that $\lim_{x\to c} f(x)$ exists.

Third, Proposition 3.1.3 gives the following:

Proposition 4.1.4 *Let* $f, g : \mathcal{D}\to\mathbb{R}$*,* $(c-d, c+d)^\circ\subset\mathcal{D}$*,* $0<d\in\mathbb{R}$*, be real functions. Assume that* $\lim_{x\to c} f(x)$ *exists,* $g(x)$ *is nowhere zero, and* $\lim_{x\to c} g(x)$ *exists and is non-zero. Then, we have*

$$\lim_{x\to c}\frac{f(x)}{g(x)} = \frac{\lim_{x\to c} f(x)}{\lim_{x\to c} g(x)}.$$

Sometimes a function is only defined or considered on an interval $(c, c+d)$, $0<d\in\mathbb{R}$, and we wish to know the limiting properties of f as $x\in(c, c+d)$ approaches c. Replacing the deleted neighborhood $(c-\delta, c+\delta)^\circ$ with the interval $(c, c+\delta)$, $0<\delta\le d$, we arrive at the concept of the **right-sided limit superior and inferior**:

$$\limsup_{x\to c^+} f(x) = \inf_{0<\delta\le d}\sup_{0<x-c<\delta} f(x) \quad\text{and}\quad \liminf_{x\to c^+} f(x) = \sup_{0<\delta\le d}\inf_{0<x-c<\delta} f(x),$$

and the **right-sided limit** $\lim_{x\to c^+} f(x)$ when $\liminf_{x\to c^+} f(x) = \limsup_{x\to c^+} f(x)$ with the limit being equal to this common value.

In a similar vein, replacing the deleted neighborhood $(c-\delta, c+\delta)^\circ$ with the interval $(c-\delta, c)$, $0<\delta\le d$, we have the **left-sided limit superior and limit inferior**

$$\limsup_{x\to c^-} f(x) = \inf_{0<\delta\le d}\sup_{0<c-x<\delta} f(x) \quad\text{and}\quad \liminf_{x\to c^-} f(x) = \sup_{0<\delta\le d}\inf_{0<c-x<\delta} f(x),$$

and the **left-sided limit** $\lim_{x\to c^-} f(x)$ when $\liminf_{x\to c^-} f(x) = \limsup_{x\to c^-} f(x)$ with the limit being equal to this common value.

All the previous statements hold for one-sided limits with appropriate modifications.

Remark We note that, for a function $f : \mathcal{D} \to \mathbb{R}$, $(c, c+d) \subset \mathcal{D}$, $c \in \mathbb{R}$, $0 < d \in \mathbb{R}$, we have $\lim_{x \to c^+} f(x) = L$ if and only if, for every $0 < \epsilon \in \mathbb{R}$, there exists $0 < \delta \leq d$ such that $0 < x - c < \delta$ implies $|f(x) - L| < \epsilon$.

Similarly, for a function $f : \mathcal{D} \to \mathbb{R}$, $(c - d, c) \subset \mathcal{D}$, $c \in \mathbb{R}$, $0 < d \in \mathbb{R}$, we have $\lim_{x \to c^-} f(x) = L$ if and only if, for every $0 < \epsilon \in \mathbb{R}$, there exists $0 < \delta \leq \delta_0$ such that $0 < c - x < \delta$ implies $|f(x) - L| < \epsilon$.

One-sided limits are often used to evaluate regular (two-sided) limits. This is based on the obvious statement that $\lim_{x \to c} f(x)$ exists if and only if $\lim_{x \to c^+} f(x) = \lim_{x \to c^-} f(x)$, and, in this case, the limit is equal to this common value.

Next, we define the **limit at infinity**. Let $0 < K_0 \in \mathbb{R}$ and $f : (K_0, \infty) \to \mathbb{R}$ be a real function. We define the limit superior and limit inferior at infinity of f by

$$\limsup_{x \to \infty} f(x) = \limsup_{u \to 0^+} f(1/u) \quad \text{and} \quad \liminf_{x \to \infty} f(x) = \liminf_{u \to 0^+} f(1/u).$$

The limit at infinity $\lim_{x \to \infty} f(x)$ exists if $\limsup_{x \to \infty} f(x) = \liminf_{x \to \infty} f(x)$, and, in this case, the limit is equal to this common value. The limit relation $\lim_{x \to \infty} f(x) = L$ means that, for every $0 < \epsilon \in \mathbb{R}$, there exists $K_0 \leq K \in \mathbb{R}$ such that $K \leq x$ implies $|f(x) - L| < \epsilon$.

Finally, we define $\lim_{x \to \infty} f(x) = \infty$ by $\liminf_{x \to \infty} f(x) = \infty$. This means that, for every $0 < M \in \mathbb{R}$, there exists $K_0 \leq K \in \mathbb{R}$ such that $K \leq x$ implies $M \leq f(x)$.

The limit superior and limit inferior at negative infinity are defined by taking opposites in an obvious way.

Exercise

4.1.1. Let $f : \mathcal{D} \to \mathbb{R}$, $(c-d, c+d)^\circ \subset \mathcal{D}$, $0 < d \in \mathbb{R}$, be a positive real function. Show that

$$\liminf_{x \to c} \frac{1}{f(x)} = \frac{1}{\limsup_{x \to c} f(x)}.$$

4.2 Continuity

Let $c \in \mathbb{R}$ and $0 < d \in \mathbb{R}$. A real function $f : \mathcal{D} \to \mathbb{R}$, $(c - d, c + d) \subset \mathcal{D}$, is said to be **continuous** at c if

$$\lim_{x \to c} f(x) = f(c).$$

We call $f : \mathcal{D} \to \mathbb{R}$, $[c, c+d) \subset \mathcal{D}$, **right-continuous** at c if $\lim_{x\to c^+} f(x) = f(c)$. Similarly, $f : \mathcal{D} \to \mathbb{R}$, $(c-d, c] \subset \mathcal{D}$, is **left-continuous** at c if $\lim_{x\to c^-} f(x) = f(c)$. Clearly, $f : \mathcal{D} \to \mathbb{R}$, $(c-d, c+d) \subset \mathcal{D}$, is continuous at c if it is right-continuous and left-continuous at c.

Let $a < b$, $a, b \in \mathbb{R}$. A real function $f : (a, b) \to \mathbb{R}$ is **continuous on** (a, b) if it is continuous at any $c \in (a, b)$. If f is defined on the half-closed interval $[a, b)$, resp. $(a, b]$, then, for continuity on $[a, b)$, resp. $(a, b]$, we require continuity on (a, b) **and** right-continuity of f at a, resp. left-continuity at b. Finally, $f : [a, b] \to \mathbb{R}$ is continuous on $[a, b]$ if it is continuous on (a, b), right-continuous at a, and left-continuous at b.

By the proposition of the previous section, a real function $f : (c-d, c+d) \to \mathbb{R}$, $c \in \mathbb{R}$, $0 < d \in \mathbb{R}$, is continuous at c if and only if for any convergent real sequence $(x_n)_{n\in\mathbb{N}}$, $n \in \mathbb{N}$, with limit $\lim_{n\to\infty} x_n = c$, we have $\lim_{n\to\infty} f(x_n) = f(c)$.[1] Similar statements hold for right- and left-continuity by restricting the sequence to the respective sides.

Extreme Values Theorem. *Let $a < b$, $a, b \in \mathbb{R}$, and $f : [a, b] \to \mathbb{R}$ be a continuous function. Then $\sup_{x\in[a,b]} f(x)$ and $\inf_{x\in[a,b]} f(x)$ are finite and attained; that is, we have $f(c) = \sup_{x\in[a,b]} f(x)$ and $f(d) = \inf_{x\in[a,b]} f(x)$ for some $c, d \in [a, b]$.*

Proof It is enough to treat the supremum. Let $\sup_{x\in[a,b]} f(x) = L \leq \infty$. By the definition of the supremum, there exists a real sequence $(x_n)_{n\in\mathbb{N}}$, $x_n \in [a, b]$, $n \in \mathbb{N}$, such that $\lim_{n\to\infty} f(x_n) = L$. By the Bolzano–Weierstrass Theorem, $(x_n)_{n\in\mathbb{N}}$ has a convergent subsequence $(x_{n_k})_{k\in\mathbb{N}}$ with limit $\lim_{k\to\infty} x_{n_k} = c \in [a, b]$, say. Clearly, we have $\lim_{k\to\infty} f(x_{n_k}) = L$. By Corollary to Proposition 4.1.1 and the definition of continuity, $L = f(c)$. Hence L is finite and it is attained.

The theorem follows.

A direct consequence of Propositions 4.1.2–4.1.3 of the previous section is the following:

Proposition 4.2.1 *Let $c \in \mathbb{R}$ and $0 < d \in \mathbb{R}$. Let $f, g : (c-d, c+d) \to \mathbb{R}$ be real functions, and assume that f and g are continuous at c. Then the functions $f + g$ and $f \cdot g$ are continuous at c.*

An obvious consequence of continuity is the following: If $f, g : \mathcal{D} \to \mathbb{R}$, $(c-d, c+d) \subset \mathcal{D}$, $c \in \mathbb{R}$, $0 < d \in \mathbb{R}$, are continuous functions at c such that $f(c) < g(c)$, then there exists $0 < \delta \leq d$ such that $f(x) < g(x)$ for $|x - c| < \delta$.

To show this, we first note that $f - g$ is continuous at c.[2] Let $0 < \epsilon = (g(c) - f(c))/2$, and choose $0 < \delta \leq d$ such that

$$|x - c| < \delta \quad \Rightarrow \quad |(g(x) - f(x)) - (g(c) - f(c))| < \epsilon = \frac{g(c) - f(c)}{2}.$$

[1] This property is termed **sequential continuity**. In our case of single-variable (and also multivariate) real functions, this is equivalent to continuity.

[2] By Proposition 4.2.1 since constant functions (such as -1) are obviously continuous.

The last inequality implies

$$0 < \frac{g(c) - f(c)}{2} < g(x) - f(x), \quad |x - c| < \delta.$$

The statement follows.

What we just proved clearly holds for right- or left-continuity with appropriate modifications.

Proposition 4.1.4 gives the following:

Proposition 4.2.2 *Let $c \in \mathbb{R}$ and $0 < d \in \mathbb{R}$. Let $f, g : (c - d, c + d) \to \mathbb{R}$ be real functions, and assume that f and g are continuous at c. If $g(c) \neq 0$, then f/g is also continuous*[3] *at c.*

Since $\lim_{x\to c} 1 = 1$ and $\lim_{x\to c} x = c$, $c \in \mathbb{R}$, are (near) tautologies, Proposition 4.1.1 along with a simple induction implies that $\lim_{x\to c} x^n = c^n$, $n \in \mathbb{N}$. As we will see in Section 6.1, the integral powers x^n, $n \in \mathbb{N}_0$, are the basic building blocks of polynomials and rational functions. More precisely, a polynomial is a finite sum of powers x^n, $n \in \mathbb{N}_0$, multiplied by real numbers, and rational functions are quotients of polynomials. It follows that every polynomial is continuous everywhere, and every rational function is continuous on its domain.

Intermediate Value Theorem. *Let $a < b$, $a, b \in \mathbb{R}$, and $f : [a, b] \to \mathbb{R}$ be a continuous function. Let $M \in \mathbb{R}$ be between $f(a)$ and $f(b)$, that is, $\min(f(a), f(b)) < M < \max(f(a), f(b))$. Then, we have $f(c) = M$ for some $c \in (a, b)$.*

Proof We may assume $f(a) < f(b)$, so that $f(a) < M < f(b)$. Let

$$A = \{x \in [a, b] \mid f(x) \leq M\}.$$

Clearly, A is non-empty (since $a \in A$). Let $c = \sup A \in [a, b]$. We claim that $c \in (a, b)$. Indeed, since f is right-continuous at a and $f(a) < M$, we have $a < c$. Since f is left-continuous at b and $M < f(b)$, we have $c < b$. These give $c \in (a, b)$.

Let $0 < \epsilon \in \mathbb{R}$. By continuity of f at c, there exists $0 < \delta \leq d$, $d = \min(b - c, c - a)$, such that $|x - c| < \delta$ implies $f(x) - \epsilon < f(c) < f(x) + \epsilon$.

By the definition of the supremum, there exists $c' \in (c - \delta, c]$ such that $c' \in A$; that is, $f(c') \leq M$. This and the continuity above give $f(c) < f(c') + \epsilon \leq M + \epsilon$.

Again by the definition of the supremum, there exists $c'' \in (c, c + \delta)$ such that $c'' \notin A$; that is, $f(c'') > M$. This and the continuity above give $M - \epsilon < f(c'') - \epsilon < f(c)$.

Combining these, we obtain $M - \epsilon < f(c) < M + \epsilon$. Since $0 < \epsilon \in \mathbb{R}$ was arbitrary, $f(c) = M$ follows.

[3] Since $g(c) \neq 0$, we also have $g(x) \neq 0$ for $|x - c| < \delta$ with $0 < \delta \in \mathbb{R}$ small enough.

Example 4.2.1 Let $f : [0, 1] \to [0, 1]$ be a continuous function. Then there exists $c \in [0, 1]$ such that[4] $f(c) = c$. Indeed, we may assume that $f(0) \neq 0$ and $f(1) \neq 1$ (since otherwise there is nothing to prove). Consider the function $g : [0, 1] \to \mathbb{R}$ defined by $g(x) = f(x) - x$, $x \in [0, 1]$. We have $g(0) = f(0) > 0$ and $g(1) = f(1) - 1 < 0$. By the Intermediate Value Theorem, we have $g(c) = 0$ for some $c \in [0, 1]$. This gives $f(c) = c$ as claimed.

Corollary *Let $f : \mathcal{D} \to \mathbb{R}$ be a real function. If f is continuous and injective on an interval $\mathcal{I} \subset \mathcal{D}$, then it is strictly monotonic on $\mathcal{I}$.*

Proof Let $a < b$, $a, b \in \mathcal{I}$. Since f is injective on $\mathcal{I}$, we have $f(a) \neq f(b)$. We may assume that $f(a) < f(b)$.

We claim that f is strictly increasing on the interval $[a, b]$. Assume not. There exist $x < x'$, $x, x' \in [a, b]$, such that $f(x') < f(x)$. ($f(x) \neq f(x')$ by injectivity again.)

We first claim that $f(a) \leq f(x')$. Indeed, otherwise we have $f(x') < f(a) < f(b)$ with $a < x' < b$ ($x' = b$ cannot happen by injectivity), and, by the Intermediate Value Theorem, we have $f(c) = f(a)$ for some $c \in [x', b]$, contradicting injectivity.

Second, we have $f(x) \leq f(b)$, since otherwise $f(a) < f(b) < f(x)$ with $a < x < b$. By the Intermediate Value Theorem again, we have $f(c) = f(b)$, $c \in [a, x]$, contradicting injectivity again.

Summarizing, we have $a < x < x' < b$ and $f(a) < f(x') < f(x) < f(b)$ (with strict inequalities throughout). By the Intermediate Value Theorem again, there exists $c \in [x', b]$ such that $f(c) = f(x)$, once again contradicting to injectivity. The corollary follows.

Remark The assumption on continuity in the corollary above is essential; the function in Example 0.3.5 is injective but neither monotonic nor continuous (except at 0).

We have seen that arithmetic operations of functions preserve continuity. The next proposition states that continuity is also preserved by composition of functions. It is a direct consequence of (sequential) continuity of the participating functions.

Proposition 4.2.3 *Let $f : \mathcal{D} \to \mathbb{R}$, $(c - d, c + d) \subset \mathcal{D}$, $c \in \mathbb{R}$, $0 < d \in \mathbb{R}$, and $g : \mathcal{E} \to \mathbb{R}$, $(f(c) - e, f(c) + e) \subset \mathcal{E}$, $0 < e \in \mathbb{R}$, be real functions. Assume that f is continuous at c and g is continuous at $f(c)$. Then the composition $g \circ f$ is continuous at c.*

Example 4.2.2 The converse of the previous proposition is obviously false; for example, let $f : \mathbb{R} \to \mathbb{R}$ be **any** real function and $g : \mathbb{R} \to \mathbb{R}$ the identically zero function. For a less trivial example, let $f, g : \mathbb{R} \to \mathbb{R}$ be defined by $f(x) = x^2$ and

[4] This statement also holds with $[0, 1]$ replaced by an arbitrary closed interval $[a, b]$. In this form, it is often termed as the 1-dimensional **Brouwer fixed point theorem**, even though the latter (in dimensions ≥ 2) is more subtle.

$$g(x) = \begin{cases} 1, & \text{if} \quad x \geq 0 \\ 0, & \text{if} \quad x < 0. \end{cases}$$

Then, f is continuous everywhere, and $g \circ f$, being the constant function 1, is also continuous everywhere, but g is discontinuous at 0.

We now extend the definition of the **power function** $\mathfrak{p}_r : \mathcal{D} \to \mathbb{R}$, $\mathfrak{p}_r(x) = x^r$, $x \in \mathcal{D} \subset \mathbb{R}$ (Section 3.2) to **any** real exponent $r \in \mathbb{R}$ as follows:

For **zero exponent** $r = 0$, the domain $\mathcal{D}$ of the power function $\mathfrak{p}_0$ is $\mathcal{D} = \mathbb{R} \setminus \{0\}$, and we have $\mathfrak{p}_0(x) = x^0 = 1$, $0 \neq x \in \mathbb{R}$. (Recall that 0^0 is undefined.)[5]

For a **positive rational exponent** $r = m/n \in \mathbb{Q}$, $m, n \in \mathbb{N}$, the domain $\mathcal{D}$ of the power function $\mathfrak{p}_r$ is $\mathcal{D} = \{0 \leq x \in \mathbb{R}\}$ if n is even and $\mathcal{D} = \mathbb{R}$ if n is odd.

For a **negative rational exponent** $r = -m/n \in \mathbb{Q}$, $m, n \in \mathbb{N}$, the domain $\mathcal{D}$ of the power function $\mathfrak{p}_r$ is $\mathcal{D} = \{0 < x \in \mathbb{R}\}$ if n is even and $\mathcal{D} = \mathbb{R} \setminus \{0\}$ if n is odd.

For a **positive irrational exponent** $0 < r \in \mathbb{R} \setminus \mathbb{Q}$, the domain $\mathcal{D}$ of the power function $\mathfrak{p}_r$ is $\mathcal{D} = \{0 \leq x \in \mathbb{R}\}$.

For a **negative irrational exponent** $0 < r \in \mathbb{R} \setminus \mathbb{Q}$, the domain $\mathcal{D}$ of the power function $\mathfrak{p}_r$ is $\mathcal{D} = \{0 < x \in \mathbb{R}\}$.

We now proceed to show that the power function $\mathfrak{p}_r : \mathcal{D} \to \mathbb{R}$ is continuous on its domain $\mathcal{D}$. Since the set of positive real numbers is included in $\mathcal{D}$ in all cases, we first show continuity of $\mathfrak{p}_r$ at $0 < c \in \mathbb{R}$.

We claim that

$$\lim_{x \to c} x^r = c^r, \quad 0 < c \in \mathbb{R}, \ r \in \mathbb{R}.$$

Replacing the variable x by x/c, the limit above reduces to the following:

$$\lim_{x \to 1} x^r = 1, \quad r \in \mathbb{R}.$$

First assume that $|r| < 1$. By the combined Bernoulli inequality, we have

$$|x^r - 1| \leq |r| \cdot |x - 1|, \quad 1 < x \in \mathbb{R}.$$

By monotonicity of the right-limit, we obtain

$$0 \leq \lim_{x \to 1^+} |x^r - 1| \leq |r| \lim_{x \to 1^+} |x - 1| = |r| \lim_{x \to 1} |x - 1| = 0.$$

For the left-limit, we calculate

$$\lim_{x \to 1^-} x^r = \lim_{x \to 1^-} \frac{1}{1/x^r} = \lim_{x \to 1^-} \frac{1}{(1/x)^r} = \lim_{x \to 1^+} \frac{1}{x^r} = \frac{1}{\lim_{x \to 1^+} x^r} = 1,$$

[5] Clearly, $\lim_{x \to 0} x^0 = 1$. This is one of the reasons why sometimes 0^0 is defined as 1.

where we changed the variable $0 < x < 1$ to $1 < 1/x$. The limit relation above and hence continuity follow in this case.

For $1 < |r|$, let $n \in \mathbb{N}$ such that $|r| < n$. Then we have $|r/n| < 1$, and the limit relation above holds for the exponent n/r. We have

$$\lim_{x\to 1} x^r = \lim_{x\to 1} \left(x^{r/n}\right)^n = \left(\lim_{x\to 1} x^{r/n}\right)^n = 1^n = 1.$$

The limit relation and thereby continuity at $0 < c \in \mathbb{R}$ follow in general.

The case of rational exponents $r = \pm m/n \in \mathbb{Q}$, $m, n \in \mathbb{N}$, can be reduced to exponents that are reciprocals of (non-zero) integers since $\mathfrak{p}_{m/n}(x) = x^{m/n} = (x^{1/n})^m = (\mathfrak{p}_{1/n}(x))^m$, $x \in \mathcal{D}$. Indeed, for $n \in \mathbb{N}$ odd and $0 < c \in \mathbb{R}$, we have

$$\lim_{x\to -c} x^{1/n} = \lim_{x\to c} \sqrt[n]{-x} = -\lim_{x\to c} \sqrt[n]{x} = -c^{1/n} = -\sqrt[n]{c} = (-c)^{1/n},$$

where we used continuity of the power function at $0 < c \in \mathbb{R}$.

Finally, the (possibly only right-)continuity at $c = 0$ follows from simple applications of the exponential identities.

Remark Let $2 \leq n \in \mathbb{N}$ and $0 < c \in \mathbb{R}$. Choose $m \in \mathbb{N}$ such that $c < m^n$. Consider the power function $\mathfrak{p}_n(x) = x^n$, restricted to $x \in [0, m]$. We have $\mathfrak{p}(0) = 0$ and $\mathfrak{p}(m) = m^n$. Since $\mathfrak{p}_n$ is continuous and $0 < c < m^n$, the Intermediate Value Theorem implies that there exists $0 < a < m$, $a \in \mathbb{R}$, such that $\mathfrak{p}_n(a) = a^n = c$. This establishes the existence of the nth root $a = \sqrt[n]{c}$. This was treated in Section 3.2 using different methods.

Exercise

4.2.1. Define the real function $f : \mathbb{R} \to \mathbb{R}$ as follows. For $0 \neq x \in \mathbb{Q}$ rational, let $f(x) = 1/b$, where $x = a/b$, $\gcd(a, b) = 1$, $a \in \mathbb{Z}$, $b \in \mathbb{N}$; $f(0) = 1$, and, for $x \in \mathbb{R} \setminus \mathbb{Q}$ irrational, let $f(x) = 0$. Show that f is continuous at every irrational point and discontinuous at every rational point.

4.3 Differentiability

Of particular importance is the **difference quotient** of a function. Given $c \in \mathbb{R}$, assume that the domain of a single-variable real function f contains the interval $(c - d, c + d)$, where $0 < d \in \mathbb{R}$. Then the difference quotient of f at c is defined by

$$\mathfrak{m}_f(x, c) = \frac{f(x) - f(c)}{x - c}, \quad 0 < |x - c| < d.$$

Of paramount importance in differential calculus is the limit

$$\lim_{x\to c} \mathfrak{m}_f(x,c) = \lim_{x\to c} \frac{f(x)-f(c)}{x-c}.$$

We call f **differentiable** at c if this limit exists, and the actual value of the limit, called the **derivative** of f at c, will be denoted by $f'(c)$.

History
There is compelling evidence that some of the basic properties of the derivative and therefore those of differential calculus were discovered by Bhāskara II, predating Newton and Leibniz about 500 years. He used these properties for astronomical calculations. Finally, note that the notation $f'(c)$ for the derivative of a function f at c (although sometimes erroneously attributed to Newton) is due to Joseph-Louis Lagrange (1736–1813).

The importance of this limit is easily understood by the following interpretation of the derivative. We consider all the linear functions that take the same value at c as the function f and select the one whose values "best approximate" the values of f. A linear function that takes the same value as f at c has the general equation $y = f(c) + m(x-c)$ with $m \in \mathbb{R}$ as an indeterminate. Best approximation is interpreted as the infimum

$$\inf_{m\in\mathbb{R}} \lim_{x\to c} \left| \frac{f(x)-(f(c)+m(x-c))}{x-c} \right|.$$

This, however, can be written as

$$\inf_{m\in\mathbb{R}} \lim_{x\to c} \left| \frac{f(x)-f(c)}{x-c} - m \right|,$$

and the zero infimum is clearly attained by $m = f'(c)$, the derivative (assuming that it exists).

Let $c \in \mathbb{R}$ and $f : \mathcal{D} \to \mathbb{R}$, $(c-d, c+d) \subset \mathcal{D}$, $0 < d \in \mathbb{R}$, be a real function, and assume that f is differentiable at c. The line given by the equation $y = f(c) + f'(c)(x-c)$ is called the **tangent line** to the graph $G(f)$ of the function f at c.

Taking the right- and left-limits in the definition of the difference quotient, we arrive at the concept of **right-** and **left-derivatives**. More precisely, for a real function $f : \mathcal{D} \to \mathbb{R}$, $[c, c+d) \subset \mathcal{D}$, resp., $(c-d, c] \subset \mathcal{D}$, $c \in \mathbb{R}$, $0 < d \in \mathbb{R}$, we define the right-, resp., left-derivatives at $c \in \mathbb{R}$ by

$$f'_+(c) = \lim_{x\to c^+} \mathfrak{m}_f(x,c) = \lim_{x\to c^+} \frac{f(x)-f(c)}{x-c},$$

$$f'_-(c) = \lim_{x\to c^-} \mathfrak{m}_f(x,c) = \lim_{x\to c^-} \frac{f(x)-f(c)}{x-c}.$$

As for limits, the derivative $f'(c)$ exists if and only if the right- and left-derivatives $f'_+(c)$ and $f'_-(c)$ exist and they are equal.

There are natural instances when one of the one-sided derivatives or both exist.

Example 4.3.1 Recall that a real function $f : (-d, d) \to \mathbb{R}$, $0 < d \leq \infty$, is called **even**, resp., **odd**, if $f(-x) = f(x)$, resp., $f(-x) = -f(x)$, $|x| < d$.
Assuming that f is even and that the one-sided derivatives exist, we calculate

$$f'_+(c) = \lim_{x\to 0^+} \frac{f(x)-f(0)}{x} = \lim_{x\to 0^-} \frac{f(-x)-f(0)}{-x} = -\lim_{x\to 0^-} \frac{f(x)-f(0)}{x} = -f'_-(c).$$

This shows that, if the derivative of an even function at 0 exists, then it must be zero. For example, for the absolute value function $f(x) = |x|$, $x \in \mathbb{R}$, the left- and right-derivatives are $f'_\pm = \pm 1$, and the derivative at 0 does not exist.

Assuming that f is odd and that the one-sided derivatives exist, then clearly $f(0) = 0$, and we have

$$f'_-(c) = \lim_{x\to 0^+} \frac{f(x)}{x} = \lim_{x\to 0^-} \frac{f(-x)}{-x} = \lim_{x\to 0^-} \frac{f(x)}{x} = f'_-(c).$$

This shows that, if the left- and right-derivatives at 0 of an odd function exist, then they must be equal, and therefore the (two-sided) derivative at 0 also exists.

Let $f : \mathcal{D} \to \mathbb{R}$, $(c - d, c + d) \subset \mathcal{D}$, $0 < d \in \mathbb{R}$, be a real function. We call c a **critical point** of f if either f is **not** differentiable at c or $f'(c) = 0$.

The importance of critical points lies in the **Fermat Principle**: If $f : \mathcal{D} \to \mathbb{R}$, $(c - d, c + d) \subset \mathcal{D}$, $0 < d \in \mathbb{R}$, assumes its supremum or infimum at c, then c is a critical point of f.

Indeed, assume that f assumes its supremum at c; that is, we have $f(x) \leq f(c)$ for all $|x - c| < d$. If f is not differentiable at c, then we are done. Assume that $f'(c)$ exists. For $0 < x - c < d$, we have

$$\mathfrak{m}_f(x, c) = \frac{f(x) - f(c)}{x - c} \leq 0.$$

Therefore, for the right-derivative, we have $f'_+(c) = \lim_{x\to c^+} \mathfrak{m}_f(x, c) \leq 0$. Similarly, For $0 < c - x < d$, we have

$$\mathfrak{m}_f(x, c) = \frac{f(x) - f(c)}{x - c} \geq 0.$$

Therefore, for the left-derivative, we have $f'_-(c) = \lim_{x\to c^-} \mathfrak{m}_f(x, c) \geq 0$. Since we assume that $f'(c)$ exists, the right- and left-limits must coincide. We obtain $f'(c) = 0$. The Fermat Principle follows.

An important consequence of the Fermat Principle is the following: Assume that $f : \mathcal{I} \to \mathbb{R}$ is a continuous function on an interval $\mathcal{I} \subset \mathbb{R}$. If f has no critical point on $\mathcal{I}$, then f is strictly monotonic on $\mathcal{I}$.

Since f is continuous, injectivity implies strict monotonicity. (See the corollary to the Intermediate Value Theorem above.) Assume, on the contrary, that f fails to be injective on $\mathcal{I}$. This means that there exist $x' < x''$, $x', x'' \in \mathcal{I}$, such that

$$f(x') = f(x'').$$

We restrict f to the closed interval $[x', x''] \subset \mathcal{I}$. Since f is continuous, by the Extreme Values Theorem, it assumes its supremum or infimum at a point c of the open interval (x', x''). By the Fermat Principle, c is a critical point of f, a contradiction.

We now note the simple fact that differentiability implies continuity. Indeed, assume that the real function $f : (c - d, c + d) \to \mathbb{R}$, $0 < d$, $c, d \in \mathbb{R}$, is differentiable at c. Using the formula

$$f(x) = f(c) + \mathfrak{m}_f(x, c) \cdot (x - c), \quad |x - c| < d,$$

we obtain

$$\lim_{x \to c} f(x) = f(c) + \lim_{x \to c} \big(\mathfrak{m}_f(x, c) \cdot (x - c)\big) = f(c) + f'(c) \cdot 0 = f(c).$$

Continuity follows.

Example 4.3.2 Define the function $f : \mathbb{R} \to \mathbb{R}$ by

$$f(x) = \begin{cases} x^2 & \text{if } x \in \mathbb{Q} \\ -x^2 & \text{if } x \in \mathbb{R} \setminus \mathbb{Q}. \end{cases}$$

Where is f differentiable?

For $c \neq 0$, f is not continuous at c. To show this, let $(x_n)_{n \in \mathbb{N}}$ be a rational sequence and $(x'_n)_{n \in \mathbb{N}}$ an irrational sequence (a sequence whose members are irrational numbers) such that $\lim_{n \to \infty} x_n = \lim_{n \to \infty} x'_n = c$. We have $\lim_{n \to \infty} f(x_n) = c^2$ and $\lim_{n \to \infty} f(x'_n) = -c^2$. Since $c \neq 0$, f is not continuous at c. As such, f is not differentiable at c either.

For $c = 0$, we have

$$\lim_{x \to 0} \mathfrak{m}_f(x, 0) = \lim_{x \to 0} \frac{f(x)}{x} = \lim_{x \to 0} \frac{\pm x^2}{x} = \lim_{x \to 0} (\pm x) = 0.$$

Hence f is differentiable (only) at 0.

The difference quotient has important arithmetic properties. Letting $f, g : (c - d, c + d) \to \mathbb{R}$, $c \in \mathbb{R}$, $0 < d \in \mathbb{R}$, be real functions, for $|x - c| < d$, we have

$$\mathfrak{m}_{f+g}(x,c) = \mathfrak{m}_f(x,c) + \mathfrak{m}_g(x,c)$$

$$\mathfrak{m}_{f\cdot g}(x,c) = \mathfrak{m}_f(x,c)\cdot g(c) + f(c)\cdot \mathfrak{m}_g(x,c) + \mathfrak{m}_f(x,c)\cdot \mathfrak{m}_g(x,c)\cdot (x-c)$$

$$\mathfrak{m}_{f/g}(x,c) = \frac{\mathfrak{m}_f(x,c)\cdot g(c) - f(c)\cdot \mathfrak{m}_g(x,c)}{g(c)^2 + g(c)\cdot \mathfrak{m}_g(x,c)\cdot (x-c)},$$

where in the last formula we assume $g(x) \neq 0$ for $|x-c| < d$.

Indeed, we calculate

$$\begin{aligned}\mathfrak{m}_{f+g}(x,c) &= \frac{(f(x)+g(x)) - (f(c)+g(c))}{x-c}\\ &= \frac{f(x)-f(c)}{x-c} + \frac{g(x)-g(c)}{x-c} = \mathfrak{m}_f(x,c) + \mathfrak{m}_g(x,c);\end{aligned}$$

$$\begin{aligned}\mathfrak{m}_{f\cdot g}(x,c) &= \frac{f(x)\cdot g(x) - f(c)\cdot g(c)}{x-c}\\ &= \frac{f(x)-f(c)}{x-c}\cdot g(c) + f(c)\cdot \frac{g(x)-g(c)}{x-c} + \frac{f(x)-f(c)}{x-c}\cdot \frac{g(x)-g(c)}{x-c}\cdot (x-c)\\ &= \mathfrak{m}_f(x,c)\cdot g(c) + f(c)\cdot \mathfrak{m}_g(x,c) + \mathfrak{m}_f(x,c)\cdot \mathfrak{m}_g(x,c)\cdot (x-c);\end{aligned}$$

$$\begin{aligned}\mathfrak{m}_{f/g}(x,c) &= \frac{f(x)/g(x) - f(c)/g(c)}{x-c} = \frac{f(x)\cdot g(c) - f(c)\cdot g(x)}{g(c)g(x)(x-c)}\\ &= \frac{\big(\mathfrak{m}_f(x,c)(x-c) + f(c)\big)\, g(c) - f(c)\,\big(\mathfrak{m}_g(x,c)(x-c) + g(c)\big)}{g(c)\,\big(g(c) + \mathfrak{m}_g(x,c)(x-c)\big)\,(x-c)}\\ &= \frac{\mathfrak{m}_f(x,c)\cdot g(c) - f(c)\cdot \mathfrak{m}_g(x,c)}{g(c)^2 + g(c)\cdot \mathfrak{m}_g(x,c)\cdot (x-c)}.\end{aligned}$$

Assuming that f and g are differentiable at c, taking the limits as $x \to c$, we obtain the following differentiation formulas:

$$\begin{aligned}(f+g)' &= f' + g'\\ (f\cdot g)' &= f'\cdot g + f\cdot g'\\ \left(\frac{f}{g}\right)' &= \frac{f'\cdot g - f\cdot g'}{g^2},\end{aligned}$$

where the dependence on c has been suppressed.

As a generalization of the second (product) formula, a simple induction gives

$$(f^n)' = n(f^{n-1})\cdot f', \quad n \in \mathbb{N}.$$

To close this section, we claim that the **derivative of the power function** $\mathfrak{p}_r : \mathcal{D} \to \mathbb{R}$, $r \in \mathbb{R}$, $\mathfrak{p}_r(x) = x^r$, $x \in \mathcal{D}$, is the following:

$$\mathfrak{p}_r'(c) = \lim_{x \to c} \frac{x^r - c^r}{x - c} = rc^{r-1}, \quad 0 < c \in \mathbb{R}.$$

Dividing by c^{r-1}, this limit simplifies to

$$\lim_{x \to 1} \frac{x^r - 1}{x - 1} = r, \quad r \in \mathbb{R}.$$

To derive this limit, we first make two reduction steps. First, the limit clearly holds for $r = 0$. Since

$$\frac{x^{-r} - 1}{x - 1} = -\frac{1}{x^r} \cdot \frac{x^r - 1}{x - 1}, \quad 1 \neq x \in \mathbb{R},$$

it is enough to prove the limit above for $0 < r \in \mathbb{R}$. Second, for $0 < x \neq 1$, $x \in \mathbb{R}$, we have

$$\frac{x^r - 1}{x - 1} = \frac{\frac{1}{1/x^r} - 1}{\frac{1}{1/x} - 1} = \left(\frac{1}{x}\right)^{1-r} \cdot \frac{(1/x)^r - 1}{(1/x) - 1}.$$

This shows that it is enough to derive the **right-limit**

$$\lim_{x \to 1^+} \frac{x^r - 1}{x - 1} = r, \quad 0 < r \in \mathbb{R}.$$

After these reductions, we first consider the case when the exponent $r \in \mathbb{Q}$ is **rational**; $r = m/n$, $m, n \in \mathbb{N}$. We have

$$\lim_{x \to 1} \frac{x^{m/n} - 1}{x - 1} = \lim_{x \to 1} \frac{x^m - 1}{x^n - 1},$$

where we changed the variable from x to $x^{1/n}$ and used $\lim_{x \to 1} x^{1/n} = 1$. We now use the Finite Geometric Series Formula twice

$$\begin{aligned} \lim_{x \to 1} \frac{x^m - 1}{x^n - 1} &= \lim_{x \to 1} \frac{(x-1)(x^{m-1} + x^{m-2} + \cdots + x + 1)}{(x-1)(x^{n-1} + x^{n-2} + \cdots + x + 1)} \\ &= \lim_{x \to 1} \frac{x^{m-1} + x^{m-2} + \cdots + x + 1}{x^{n-1} + x^{n-2} + \cdots + x + 1} = \frac{m}{n}. \end{aligned}$$

The limit relation above follows for rational exponents.

We now turn to the case of general exponent $0 < r \in \mathbb{R}$. Let $(q_n)_{n \in \mathbb{N}}$ be a rational sequence with limit $\lim_{n \to \infty} q_n = r$. We may assume that $|r - q_n| < 1$, $n \in \mathbb{N}$. For

$0 < x \neq 1, x \in \mathbb{R}$, we calculate

$$\left|\frac{x^r - 1}{x - 1} - r\right| = \left|x^{q_n} \cdot \frac{x^{r-q_n} - 1}{x - 1} + \left(\frac{x^{q_n} - 1}{x - 1} - q_n\right) - (r - q_n)\right|$$
$$\leq \left|x^{q_n}\right| \cdot \left|\frac{x^{r-q_n} - 1}{x - 1}\right| + \left|\frac{x^{q_n} - 1}{x - 1} - q_n\right| + |r - q_n|.$$

We will calculate the right-limit of this, so that, from now on, we may assume $1 < x \in \mathbb{R}$. Since $|r - q_n| < 1$, $n \in \mathbb{N}$, the combined Bernoulli inequality gives

$$\left|\frac{x^{r-q_n} - 1}{x - 1}\right| \leq |r - q_n|.$$

Using this, our estimate above reduces to

$$\left|\frac{x^r - 1}{x - 1} - r\right| \leq \left(\left|x^{q_n}\right| + 1\right) \cdot |r - q_n| + \left|\frac{x^{q_n} - 1}{x - 1} - q_n\right|.$$

Again by the Bernoulli inequality, we have

$$\left|x^{q_n}\right| \leq 1 + |q_n||x - 1| \leq 1 + (1 + r)(x - 1), \quad 1 < x,$$

since

$$|q_n| - |r| \leq |r - q_n| < 1.$$

Putting everything together, we arrive at the estimate

$$\left|\frac{x^r - 1}{x - 1} - r\right| \leq (2 + (1 + r)(x - 1)) \cdot |r - q_n| + \left|\frac{x^{q_n} - 1}{x - 1} - q_n\right|.$$

Let $0 < \epsilon \in \mathbb{R}$. Since $\lim_{n\to\infty} q_n = r$, we can choose $N \in \mathbb{N}$ such that

$$|r - q_n| < \frac{\epsilon}{2(3 + r)}, \quad n \geq N.$$

Fix $n \geq N$. By the case of rational exponents above, there exists $0 < \delta < 1$ such that, for $0 < |x - 1| < \delta$, we have

$$\left|\frac{x^{q_n} - 1}{x - 1} - q_n\right| < \epsilon/2.$$

With these choices, we have

$$\left|\frac{x^r-1}{x-1}-r\right| \le (2+(1+r)(x-1))\cdot|r-q_n| + \left|\frac{x^{q_n}-1}{x-1}-q_n\right|$$
$$< (2+(1+r))\frac{\epsilon}{2(3+r)} + \frac{\epsilon}{2} = \epsilon.$$

The limit relation above and hence the claimed differentiation formula for the power function follow.

Remark We have seen in the previous section that, for rational exponents $r = \pm m/n$, $m, n \in \mathbb{N}$, with n odd, the domain of the power function $\mathfrak{p}_{m/n}$ includes all negative real numbers. For n odd, we have $\mathfrak{p}_{m/n}(-x) = (-1)^m \mathfrak{p}_{m/n}(x)$, $x \in \mathbb{R}$. Using this, a simple computation shows that, for $0 \neq c \in \mathbb{R}$, the differentiation formula for the power function still holds.
It remains to consider the case $c = 0$. For positive (rational or irrational) exponents, the power function $\mathfrak{p}_r$ is defined (at least) for non-negative real numbers. For the right-derivative, we have

$$(\mathfrak{p}_r)'_+(0) = \lim_{x\to 0^+}\frac{x^r}{x} = \lim_{x\to 0^+} x^{r-1} = \begin{cases} 0 & \text{if } 1 < r \\ 1 & \text{if } r = 1 \\ \infty & \text{if } 0 < r < 1. \end{cases}$$

The left-derivative is defined only for positive rational exponents $r = m/n$, $m, n \in \mathbb{N}$, with n odd. In this case, we have

$$(\mathfrak{p}_{m/n})'_-(0) = \lim_{x\to 0^-} x^{(m-n)/n} = (-1)^{m+1} \lim_{x\to 0^+} x^{m/n-1} = \begin{cases} 0 & \text{if } 1 < m/n \\ 1 & \text{if } m/n = 1 \\ \pm\infty & \text{if } 0 < m/n < 1. \end{cases}$$

As a byproduct, we see that the (two-sided) derivative exists if and only if $r = m/n \ge 1$, in particular, if $r = m \in \mathbb{N}$ $(n = 1)$.

We now return to the main line. Using the derivative of the power function above, for natural exponents $n \in \mathbb{N}$, we have

$$\mathfrak{p}'_n = n\mathfrak{p}_{n-1}, \quad n \in \mathbb{N}.$$

By the differentiation formulas above, we recover the earlier result to the effect that every polynomial is differentiable everywhere, and every rational function, that is, the quotient of two polynomials, is differentiable on its domain. This is not the case for algebraic functions, however.[6] For example, the cube root function $\mathfrak{p}_{1/3}$ is defined everywhere, but its derivative does not exist at 0.

[6] Algebraic functions will be treated in detail in Section 9.

Example 4.3.3 Show that

$$\lim_{x\to 1} \frac{(1-x)(1-\sqrt{x})(1-\sqrt[3]{x})\cdots(1-\sqrt[n]{x})}{(1-x)^n} = \frac{1}{n!}.$$

We write the expression in the limit as the product of n factors

$$\frac{1-\sqrt[k]{x}}{1-x}, \quad k = 1, \ldots, n.$$

The limit of the kth factor is calculated as

$$\lim_{x\to 1} \frac{1-\sqrt[k]{x}}{1-x} = \lim_{x\to 1} \frac{x^{1/k}-1}{x-1} = \frac{1}{k}.$$

Taking the product for $k = 1, \ldots, n$, the example follows.

Exercises

4.3.1. Define the sequence of functions $f_n : [-1, 1] \to \mathbb{R}$, where $n \in \mathbb{N}_0$, inductively as $f_0(x) = |x|$, $f_n(x) = |f_{n-1}(x) - 1/2^n|$, $n \in \mathbb{N}$. Determine the set of points where f_n is not differentiable.

4.3.2. Give an example of a function $g : \mathbb{R} \to \mathbb{R}$, which is discontinuous everywhere except at 0 where it is differentiable and $g'(0) = 1$.

Chapter 5
Real Analytic Plane Geometry

"Let it have been postulated
1. To draw a straight-line from any point to any point.
2. And to produce a finite straight-line continuously in a straight-line. 3. And to draw a circle with any center and radius. 4. And that all right-angles are equal to one another. 5. And that if a straight-line falling across two (other) straight-lines makes internal angles on the same side (of itself whose sum is) less than two right-angles, then the two (other) straight-lines, being produced to infinity, meet on that side (of the original straight-line) that the (sum of the internal angles) is less than two right-angles (and do not meet on the other side)."
The five postulates in Euclid's *Elements*, translated by Richard Fitzpatrick.

Among the few choices of systems of axioms to construct a geometric model of the plane (for example, via Euclid or Hilbert), we take the least strenuous path; and, in making use of the real number system already in place, we develop real analytic plane geometry using Birkhoff's axioms of metric geometry. One of the main purposes of this chapter is to explain what is classically known as the Cantor–Dedekind Axiom: The real number system is order isomorphic to the linear continuum of geometry. This is the root of one of the faults of Euclid's axioms (as the ancient Greeks had no way of knowing the real number system), and this is resolved by the Birkhoff Postulate of Line Measure. But, unlike the original approaches of Hilbert and Birkhoff, we are working here in a concrete model, $\mathbb{R}^2$, built from the real number system $\mathbb{R}$ of Chapter 2. Verifying that the Birkhoff postulates hold in our concrete model is much less demanding than the synthetic (purely axiomatic) approach. Nevertheless, our model-oriented exposition still encounters some struggle, as in Sections 5.6–5.7, where the existence and properties of the circular arc length are shown using purely metric tools and paving the way to trigonometry (Chapter 11). This also gives a precise answer to the question: "What is π?" Once again, this relies on the Least Upper Bound Property

G. Toth, *Elements of Mathematics*, Undergraduate Texts in Mathematics,
https://doi.org/10.1007/978-3-030-75051-0_5

of the real number system, the main common thread with the first two chapters. A natural offspring of this technical passage is concluded with an optional section on the (often neglected) Principle of Shortest Distance, given here in full detail. This can be skipped (at least at the first reading), since it is used only for deriving the reflection properties of some of the conics in Chapter 8.

To ease up the complexity of the material, we make frequent side tours to develop metric properties of many geometric configurations. We determine all Pythagorean triples not by elementary number theory but via analytic geometry: the method of rational slopes. We introduce here additional important tools that will play pivotal roles in the sequel: the Cauchy–Schwarz inequality, the AM-GM inequality, and their offsprings. Finally, still in this chapter, we present Archimedes' duplication method to approximate π, once again with a view to algebraic formulas for many special angles given subsequently in trigonometry in Chapter 11.

5.1 The Birkhoff Metric Geometry

Recall that an axiomatic system contains a set of **primitives** or, more pointedly, **undefined terms** and **basic assumptions** or **axioms**.

Once the set of primitives and the set of axioms are given, any subsequent statements, called **propositions**, **lemmas**, or **theorems**, must be logical consequences of the axioms and previously proved theorems. In an axiomatic system, there are also **definitions**, which baptize previously undefined entities that are (usually) combinations of primitives and previously defined terms.

A **model** is an interpretation of the primitives in which the axioms become true statements.

Euclidean geometry, the geometry of the plane, has been axiomatized in the *Elements* (Books I-IV and VI) by Euclid.

History
Book I of the *Elements* begins with 23 definitions; a few are as follows:[1]

"1. A **point** is that of which there is no part.[2]
2. And a **line** is a length without breadth.
3. And the extremities of a line are points.
4. A **straight-line** is (any) one which lies evenly with points on itself. ...
8. And a plane **angle** is the inclination of the lines to one another, when two lines in a plane meet one another, and are not lying in a straight-line.
9. And when the lines containing the angle are straight then the angle is called **rectilinear**.
10. And when a straight-line stood upon (another) straight-line makes adjacent angles (which are) equal to one another, each of the equal angles is a right-angle, and the former straight-line is called a **perpendicular** to that upon which it stands.

[1]The excerpts quoted here are from the English edition and translation by Richard Fitzpatrick of the Greek text of J.L. Heiberg from *Euclidis Elementa, edidit et Latine interpretatus est J.L. Heiberg, in aedibus B.G. Teubneri, 1883–1885*.

[2]The numbering follows the original translation.

11. An **obtuse angle** is one greater than a right-angle.
12. And an **acute angle** (is) one less than a right-angle. ...
15. A **circle** is a plane figure contained by a single line [which is called a circumference], (such that) all of the straight-lines radiating toward [the circumference] from one point among those lying inside the figure are equal to one another.
16. And the point is called the **center** of the circle.
17. And a **diameter** of the circle is any straight-line, being drawn through the center, and terminated in each direction by the circumference of the circle. (And) any such (straight-line) also cuts the circle in half.
18. And a **semi-circle** is the figure contained by the diameter and the circumference cuts off by it. And the center of the semi-circle is the same (point) as (the center of) the circle.. ..
20. And of the trilateral figures: an **equilateral triangle** is that having three equal sides, an **isosceles (triangle)** that having only two equal sides, and a **scalene (triangle)** that having three unequal sides.
21. And further of the trilateral figures: a **right-angled triangle** is that having a right-angle, an **obtuse-angled (triangle)** that having an obtuse angle, and an **acute-angled (triangle)** that having three acute angles.
22. And of the quadrilateral figures: a **square** is that which is right-angled and equilateral, a **rectangle** that which is right-angled but not equilateral, a **rhombus** that which is equilateral but not right-angled, and a **rhomboid** that having opposite sides and angles equal to one another which is neither right-angled nor equilateral. And let quadrilateral figures besides these be called **trapezia**.
23. **Parallel lines** are straight-lines which, being in the same plane, and being produced to infinity in each direction, meet with one another in neither (of these directions)."

Euclid divided the set of basic assumptions into **postulates** and **common notions**. The postulates are related to geometry and the common notions referred to logic (or common sense).
The 5 postulates are as in the epithet for this chapter above.
There are 5 common notions as follows:

"1. Things equal to the same thing are also equal to one another.
2. And if equal things are added to equal things then the wholes are equal.
3. And if equal things are subtracted from equal things then the remainders are equal.
4. And things coinciding with one another are equal to one another.
5. And the whole [is] greater than the part."

Euclid's axioms have subtle faults.[3] The first, and most obvious, is that he did not recognize the need of undefined terms or primitives; instead, he tried to define them. (See, for example, Definitions 1 and 2 above.) The second, and more serious, is that he relied on unpostulated preconceptions that he thought to be too obvious to justify. As an illustration to this, we consider the very first statement, Proposition 1 in Book I, where he proves the existence of an equilateral triangle with a given side (and therefore with given two end-points) by constructing the third vertex as an intersection point of two circles. There is no axiom that guarantees that these two circles intersect at all. This needs to be remedied by either adding this as a "circle-circle" axiom or adding axioms from which this would follow as a "circle-circle" proposition. Moreover, once this problem is fixed, there is, once again, no guarantee that this third intersection point is non-collinear with the first two (the end-points of

[3]For a somewhat overly critical account on Euclid, see Russell, B., *The Teaching of Euclid*, The Mathematical Gazette, 2 (33) (1902) 165–167.

the given line segment). If it is collinear, then one of the points would be between the other two, and this violates the fifth common notion as above.

After critical examination of Euclid's axioms, David Hilbert in his book *Grundlagen der Geometrie* (published in 1899) set forth a more complex and more comprehensive system of axioms. For plane geometry, in Hilbert's system, the primitives are point and line and three primitive relations: (1) incidence (containment), two binary relations linking points and lines, (2) order (betweenness), a ternary relation between points, and (3) congruence, two binary relations, one linking segments, and another linking angles. Hilbert barely mentions circles, but, for example, the circle-circle statement above follows from his Axiom of Continuity (the latter mimicking the Dedekind cuts). We will not need a detailed discussion on this as we will work with yet another system of axioms.

In 1932, George D. Birkhoff (1884–1944) introduced a new set of four postulates for plane Euclidean geometry, often referred to as the **Birkhoff axioms**. The Birkhoff system created what is called **metric geometry**. Metric geometry has axioms for distance and angle measure. Betweenness and congruence are defined in terms of distance and angle measure. The Birkhoff postulates are based on the use of the scale and the protractor. Since this system is built upon the ordered field of real numbers $\mathbb{R}$, it is particularly well suited to us.

The **primitives** in the Birkhoff system are (1) **point**, (2) **line**, a set of points, (3) **distance**, a real number $d(A, B) \in \mathbb{R}$ associated with any two points A and B, and (4) **angle**, formed by any three ordered points A, O, B, $A \neq O \neq B$, denoted by $\angle AOB$ (with O being the vertex of the angle), possessing an **angle measure** $\mu(\angle AOB) \in \mathbb{R}$, a real number determined mod 2π, that is, up to (addition of) an integer multiple of 2π.

The set of all points is called the **plane**, and it is denoted by $\mathbb{P}$. We tacitly assume that $\mathbb{P}$ has at least two points.[4]

An initial set of **definitions** in the Birkhoff system are as follows:

Parallel Lines: Two lines ℓ' and ℓ'' are parallel if, as sets of points, they are **equal**, $\ell = \ell'$, or **disjoint**, $\ell' \cap \ell'' = \emptyset$.

Betweenness:[5] If A, B, C are three distinct points, then we say that C is **between** A and B, written as $A * C * B$, if $d(A, C) + d(C, B) = d(A, B)$.

Line Segment: Given two points A and B, the **line segment** $[A, B]$ is the set of points C such that $A * C * B$ together with the **end-points** A and B.

Half-line or Ray, End-Point: The half-line ℓ' with end-point O is defined by two distinct points O and A in a line ℓ as the set of points B of ℓ such that O is **not** between A and B.

Triangle: If A, B, C are three distinct points, the line segments $[A, B]$, $[B, C]$, $[C, A]$ are said to form a triangle $\triangle[A, B, C]$ with sides as these line segments and

[4]Strictly speaking, the concept of plane is a definition, and the assumption that it has at least two points is a postulate.

[5]This corresponds to Hilbert's order relation.

vertices A, B, C. If A, B, C are collinear, then we say that the triangle $\triangle[A, B, C]$ is degenerate.

The Birkhoff postulates are as follows:

I. Point-Line Postulate: For any two distinct points A and B, there is a unique line ℓ such that $A, B \in \ell$.

II. Postulate of Line Measure: For every line ℓ, there is a one-to-one correspondence $c_\ell : \ell \to \mathbb{R}$, called a **metric coordinate function** of ℓ, such that, for every $A, B \in \ell$, we have $|c_\ell(A) - c_\ell(B)| = d(A, B)$.

Remark The first two postulates have many implications. Since $\mathbb{P}$ contains at least two points, it also contains a line (through them), and by the Postulate of Line Measure, it must contain infinitely many points (corresponding to all real numbers and therefore of cardinality of $\mathbb{R}$). In addition, the distance must be non-negative and symmetric since, for any two points A and B in a line ℓ, we have $d(A, B) = |c_\ell(A) - c_\ell(B)| = |c_\ell(B) - c_\ell(A)| = d(B, A) \geq 0$. Moreover, since c_ℓ is one-to-one, $d(A, B) > 0$ if and only if $A \neq B$. (If $A = B$, then ℓ can be chosen to be any line containing this point and another point C distinct from this.) For a distance, it is also usually required that it satisfies the **Triangle Inequality**; that is, for any three points A, B, C, we have $d(A, C) \leq d(A, B) + d(B, C)$. This, however, follows from the additional postulates below. Finally, note that symmetry of the distance implies that, for any three distinct points A, B, C, we have $A * C * B$ if and only if $B * C * A$. In addition, among three distinct points A, B, C there is at most one that is between the other two.[6]

III. Postulate of Angle Measure: For every point O, there is a one-to-one correspondence α_O between the set of all half-lines with end-point O and the set of real numbers $\mathbb{R} \pmod{2\pi}$ such that, for every two half-lines ℓ' and ℓ'' with end-point O, we have[7] $\alpha_O(\ell'') - \alpha_O(\ell') = \mu(\angle A'OA'')$, where $A' \in \ell'$ and $A'' \in \ell''$.

Remark Note that this postulate implies that $\mu(\angle AOB) = \mu(\angle A'OB')$ if A, A' and B, B' are one the same half-lines with end-point O.

IV. Postulate of Similarity: Given two triangles $\triangle[A, B, C]$ and $\triangle[A', B', C']$ and $0 < k \in \mathbb{R}$ such that $d(C', A') = kd(C, A)$, $d(C', B') = kd(C, B)$, and $\mu(\angle A'C'B') = \mu(\angle ACB)$, then $d(A', B') = kd(A, B)$, $\mu(\angle B'A'C') = \mu(\angle BAC)$, and $\mu(\angle C'B'A') = \mu(\angle CBA)$.

Remark The triangles $\triangle[A, B, C]$ and $\triangle[A', B', C']$ in the Postulate of Similarity above are called **similar** and **congruent** if $k = 1$.

Instead of pursuing the axiomatic approach,[8] in the next section, we follow a more rapid course by creating a model for the Birkhoff plane $\mathbb{P}$, called the **Cartesian plane**.

---

[6] These are axioms of the Hilbert system.

[7] As sets of real numbers $\pmod{2\pi}$.

[8] See Birkhoff, G.D., *A Set of Postulates for Plane Geometry, Based on Scale and Protractor*, Annals of Mathematics, Second Series, Vol. 33, No. 2 (1932) 329–345.

Exercise

5.1.1. The following is Lemma XXIII in Book I of Newton's "Principia":[9] "If two given right lines, as AC, BD, terminating in given points A, B, are in a given ratio one to the other, and the right line CD, by which the indeterminate points C, D are joined is cut in K in a given ratio: I say, that the point K will be placed in a given right line." Using modern language, we let ℓ_0 and ℓ_1 be two non-collinear half-lines with common end-point O, say, and with two points $A \in \ell_0$ and $B \in \ell_1$, $A \neq O \neq B$. Given $0 < r, s \in \mathbb{R}$, we want to find the set of points $K \in [C, D]$, $A \neq C \in \ell_0$, $B \neq D \in \ell_1$, such that $A \in [O, C]$ and $B \in [O, D]$ and $d(B, D)/d(A, C) = r$ and $d(C, K) = d(D, K) = s$.[10]

5.2 The Cartesian Model of the Birkhoff Plane

We **define** the Cartesian plane as the Cartesian product $\mathbb{R}^2 = \mathbb{R} \times \mathbb{R}$ and its elements as the **points**. Each point P is represented by a pair $(x, y) \in \mathbb{R}^2$ of real numbers; x is the first and y is the second coordinate of P.

Whenever convenient, we will use the additive structure[11] in $\mathbb{R}^2$ and write $P + P' = (x + x', y + y')$ for the sum of points $P = (x, y)$ and $P' = (x', y')$, and also write $cP = (cx, cy)$ for the constant multiple, $c \in \mathbb{R}$, of the point $P = (x, y)$.

Note also that in $\mathbb{R}^2$, the two axes divide the plane into four (closed) **quadrants**:

$$I = \{(x, y) \in \mathbb{R}^2 \mid x \geq 0 \text{ and } y \geq 0\}$$

$$II = \{(x, y) \in \mathbb{R}^2 \mid x \leq 0 \text{ and } y \geq 0\}$$

$$III = \{(x, y) \in \mathbb{R}^2 \mid x \leq 0 \text{ and } y \leq 0\}$$

$$IV = \{(x, y) \in \mathbb{R}^2 \mid x \geq 0 \text{ and } y \leq 0\}.$$

We **define** a line ℓ in $\mathbb{R}^2$ as a set of points given by a **linear equation** $ax - by = c$, where the coefficients are real numbers, $a, b, c \in \mathbb{R}$. We tacitly assume that the coefficients a and b of the linear terms do not vanish simultaneously, that is, we have $a^2 + b^2 > 0$. Thus, a **line** ℓ is defined as

[9] The quote is Florian Cajori's edition of Andrew Motte's English translation in 1729 of Sir Isaac Newton's *Philosophiae Naturalis Principia Mathematica*, published in 1687.

[10] Newton used this lemma to show that "if two points proceed with a uniform motion in right lines, and their distance be divided in a given ratio, the dividing point will be either at rest or proceed uniformly in a right line."

[11] We will not use the vector space structure of $\mathbb{R}^2$, nor the usual geometric concepts such as the dot product, etc.

$$\ell = \{(x, y) \in \mathbb{R}^2 \mid ax - by = c\}, \quad a^2 + b^2 > 0, \quad a, b, c \in \mathbb{R}.$$

It follows from the definition that any line has infinitely many points.

If the coefficients $a, b, c \in \mathbb{R}$ define ℓ, then, for any $0 \neq t \in \mathbb{R}$, the coefficients $ta, tb, tc \in \mathbb{R}$ obviously define the same line ℓ.

Remark If $b \neq 0$, it is customary to call the ratio $m = a/b$ the **slope** (steepness) of the line ℓ given by the equation $ax - by = c$. If, in addition, $P = (x_0, y_0)$ is a point on ℓ, then we have $ax - by = ax_0 - by_0$. Dividing through b, we obtain the so-called **point-slope form** of the equation of the line $y - y_0 = m(x - x_0)$.

Recall that two lines ℓ and ℓ' are called **parallel** if they are **equal** or if they are **disjoint**. We will now derive algebraic criteria for these in terms of the coefficients of the equations that define the lines.

Let ℓ be given by the equation $ax - by = c$ $(a^2 + b^2 > 0)$ and ℓ' given by $a'x - b'y = c'$ $(a'^2 + b'^2 > 0)$. We put these together to form a system of equations

$$ax - by = c \quad \text{and} \quad a'x - b'y = c'.$$

Eliminating the indeterminates y and x gives the following reduced system:

$$(ab' - a'b)x = b'c - bc' \quad \text{and} \quad (ab' - a'b)y = a'c - ac'.$$

I. First, we claim that if ℓ and ℓ' contain **at least two** distinct common points, then

$$a' = ta, \quad b' = tb, \quad c' = tc,$$

for some $0 \neq t \in \mathbb{R}$; and consequently, the two lines ℓ and ℓ' are equal.

Indeed, since the reduced system above has at least two solutions, we must have $ab' - a'b = 0$ (since otherwise we would have a unique solution). This implies $b'c - bc' = 0$ and $a'c - ac' = 0$. We now put these together as a system

$$ab' = a'b, \quad b'c = bc', \quad a'c = ac'.$$

If a, b, c are all non-zero, then we have

$$\frac{a'}{a} = \frac{b'}{b} = \frac{c'}{c}.$$

Setting this equal to $t \in \mathbb{R}$, the claim follows.

If $a = 0$, then $b \neq 0$ (since $a^2 + b^2 > 0$), so that $a' = 0$ and $b' \neq 0$ (since, again, $a'^2 + b'^2 > 0$). If, in addition, $c = 0$, then $c' = 0$, and $b'/b = t \in \mathbb{R}$, and the claim follows again. If $c \neq 0$, then $c' \neq 0$, and we have $b'/b = c'/c = t \in \mathbb{R}$, and the claim follows again. The remaining cases are similar.

II. Second, if the distinct parallel lines ℓ and ℓ' are given by the linear equations above, the corresponding reduced system of equations has no solution, and we must have

$$ab' = a'b, \quad b'c \neq bc', \quad a'c \neq ac'.$$

With a reasoning similar to the one above, we obtain that two distinct lines ℓ and ℓ' are parallel if and only if, for some $0 \neq t \in \mathbb{R}$, we have

$$a' = ta, \quad b' = tb, \quad c' \neq tc.$$

Combining these two cases, as a byproduct, we see that the relation of being "parallel" is an equivalence relation on the set of all lines. We call an equivalence class a **pencil of parallel lines**. Thus, a pencil consists of all lines that are parallel to one another. The discussion above also yields that a pencil of parallel lines is given by the equations $ax - by = c$, $a^2 + b^2 > 0$, where $c \in \mathbb{R}$ varies through **all** real numbers.

As another application, we now show that the **Axiom of Parallelism** or **Playfair Axiom** (equivalent to Postulate 5 of Euclid as in the epitaph of this chapter) holds: Given a line ℓ and a point P_0 **not** on the line, there exists a **unique** line ℓ' that contains the point P_0 and is parallel to ℓ.

History

The Greek philosopher Proclus Lycaeus (412–485), in his commentary about Euclid's Proposition 31 in Book 1, states what is now named after the Scottish mathematician John Playfair (1748–1819), the Playfair Axiom. The critical part of the axiom is unicity. In his *Elements of Geometry* (published in 1795), Playfair himself stated this part of the axiom as "Two intersecting straight lines cannot be both parallel to the same straight line." Playfair acknowledged that he borrowed this from the same statement made ten years earlier by the English mathematician and clergyman William Ludlam (1680–1728).

Let ℓ be given by $ax - by = c$ and $P_0 = (x_0, y_0)$. Since $P_0 \notin \ell$, we have $ax_0 - by_0 \neq c$. We define the line ℓ' by $ax - by = ax_0 - by_0$. By construction, $P_0 \in \ell'$, and, by the above, ℓ and ℓ' are parallel. Existence follows.

For unicity, let ℓ' and ℓ'' be two lines parallel to ℓ and containing P_0. Since being parallel is an equivalence relation, ℓ' and ℓ'' are parallel. Since they have the common point P_0, they must be equal. Unicity follows.

The **Point-Line Postulate** of Birkhoff asserts the existence and uniqueness of a line containing two distinct points. We now show that this postulate holds in our model $\mathbb{R}^2$.

Given two distinct points $P_0 = (x_0, y_0)$ and $P_1 = (x_1, y_1)$, an equation of a line containing these points is given by

$$(y_1 - y_0)x - (x_1 - x_0)y = x_0y_1 - x_1y_0.$$

Indeed, simple substitution shows that the coordinates of P_0 and P_1 both satisfy this equation. In addition, this equation is clearly linear with $a = y_1 - y_0$ and $b = x_1 - x_0$ and $c = x_0y_1 - x_1y_0$, and we also have $a^2 + b^2 = (x_1 - x_0)^2 + (y_1 - y_0)^2 > 0$ (as $P_0 \neq P_1$). We conclude that this is an equation of a line containing the given points P_0 and P_1.

Unicity is clear since we have proved that two lines that have at least two common points must coincide.

The Point-Line Postulate follows.

Remark The equation above for the line containing two points $P_0 = (x_0, y_0)$ and $P_1 = (x_1, y_1)$ can be written in the following compact form:

$$(x_0 - x)(y - y_1) - (x - x_1)(y_0 - y) = 0.$$

Although we will not need it in the future, we note that, with the variable point $Q = (x, y)$, the left-hand side is the (signed) area of the parallelogram with vertices at the origin, $P_0 - Q$, $Q - P_1$, and $P_0 - P_1$. It expresses the fact that Q is on the line containing the points P_0 and P_1 if and only if the parallelogram is degenerate (has area zero), that is, its four vertices are collinear.

To derive the **Postulate of Line Measure**, we introduce the **affine** (or **convex**) parametrization for a line.

Assume that the line ℓ contains two distinct points $P_0 = (x_0, y_0)$ and $P_1 = (x_1, y_1)$. For $t \in \mathbb{R}$, we define the point

$$P_t = (1 - t)P_0 + tP_1 = ((1 - t)x_0 + tx_1, (1 - t)y_0 + ty_1).$$

We claim that $\ell = \{P_t \mid t \in \mathbb{R}\}$. The indeterminate $t \in \mathbb{R}$ is called an **affine parameter** of the line ℓ.

First, for $t \in \mathbb{R}$, we have $P_t \in \ell$ since

$$\begin{aligned}
&(y_1 - y_0)\left((1 - t)x_0 + tx_1\right) - (x_1 - x_0)\left((1 - t)y_0 + ty_1\right) \\
&\quad = (1 - t)\left((y_1 - y_0)x_0 - (x_1 - x_0)y_0\right) + t\left((y_1 - y_0)x_1 - (x_1 - x_0)y_1\right) \\
&\quad = (1 - t)(x_0y_1 - x_1y_0) + t(x_0y_1 - x_1y_0) = x_0y_1 - x_1y_0.
\end{aligned}$$

We need to show the converse. If $x_0 \neq x_1$, then we let $t = (x - x_0)/(x_1 - x_0)$, or equivalently, $x = (1 - t)x_0 + tx_1$. Substituting this into the equation of the line, a simple computation gives $y = (1 - t)y_0 + ty_1$. If $y_0 \neq y_1$, then we let $t = (y - y_0)/(y_1 - y_0)$, or equivalently, $y = (1 - t)y_0 + ty_1$. Substituting this into the equation of the line again, we obtain $x = (1 - t)x_0 + tx_1$. The converse follows.

In the Birkhoff plane, the concept of **betweenness** and the derived concepts of a line segment and half-line are defined in terms of the distance (yet to be introduced here). We now adopt a different definition for betweenness and will show later that this definition coincides with Birkhoff's definition (in terms of the distance).

Given three points A, B, C, we say that C is **between** A and B, written as $A * C * B$, if, setting $A = P_0$ and $B = P_1$, we have $C = P_t$ for some $0 < t < 1$.

With this, we define the **line segment** with end-points A and B by

$$[A, B] = \{P_t \mid 0 \leq t \leq 1\}, \quad A = P_0, \ B = P_1.$$

In particular, the line segment $[A, B]$ is part of the line ℓ that contains the points A and B.

Finally, we claim that the **half-line** ℓ' defined by two distinct points O (the end-point of ℓ') and $A \in \ell'$ is given by

$$\ell' = \{P_t \mid t \geq 0\}, \quad O = P_0, \ A = P_1.$$

Arguing by contradiction, we need to see for what points $B = P_t$, $t \in \mathbb{R}$, (on the line ℓ containing ℓ') is the point O between A and B. This condition holds if, for some $0 < s < 1$, $s \in \mathbb{R}$, we have

$$\begin{aligned} O &= (1-s)A + sB = (1-s)P_1 + sP_t \\ &= (1-s)P_1 + s((1-t)P_0 + tP_1) \\ &= s(1-t)P_0 + (1-s(1-t))P_1. \end{aligned}$$

Since $O = P_0$, this gives $s(1-t) = 1$, or equivalently, $t = 1 - 1/s$. This shows that $0 < s < 1$ if and only if $t < 0$. The claim follows.

Exercises

5.2.1. A triangular array of points is given by $T = \{(a, b) \in \mathbb{N}_0 \times \mathbb{N}_0 \mid 0 \leq b \leq a, a+b \leq 6, a+b \text{ even}\}$. How many non-degenerate triangles can be formed with vertices chosen as points of T?

5.2.2. Show that if $a, b : \mathbb{N}_0 \to \mathbb{R}$ are arithmetic sequences with differences d and e, then the points $(a_n, b_n) \in \mathbb{R}^2$, $n \in \mathbb{N}_0$, are on the same line in $\mathbb{R}^2$.

5.2.3. Given two parallel lines by the equations $y = mx + b_1$ and $y = mx + b_2$, show that the distance (the length of a perpendicular line segment with end-points on each) is equal to

$$\frac{|b_1 - b_2|}{\sqrt{m^2+1}}.$$

5.2.4. Let $A = \bigcup_{r \in [0,1]} [(r, 0), (0, 1-r)]$, the union of all line segments (in the first quadrant $\mathbf{I}$ of $\mathbb{R}^2$) with end-points $(r, 0)$ and $(0, 1-r)$, $r \in [0, 1]$. Show that A is given by the inequality $\sqrt{x} + \sqrt{y} \leq 1$, $(x, y) \in \mathbf{I}$.

5.3 The Cartesian Distance

We now introduce the **Cartesian distance** $d : \mathbb{R}^2 \to \mathbb{R}$ as follows: Given two points $P_0 = (x_0, y_0)$ and $P_1 = (x_1, y_1)$, we define

$$d(P_0, P_1) = \sqrt{(x_0 - x_1)^2 + (y_0 - y_1)^2}.$$

Example 5.3.1 As a simple application of the Cartesian distance formula, we ask the following question: What kind of numbers can arise as the distance between two points A and B in $\mathbb{R}^2$ whose coordinates are **integers**?

Letting $P_0 = (x_0, y_0)$ and $P_1 = (x_1, y_1)$, our assumption is $x_0, x_1, y_0, y_1 \in \mathbb{Z}$. In particular, $a = x_0 - x_1$ and $b = y_0 - y_1$ are also integers. Using the Cartesian distance formula, the problem can be reformulated as follows: Given two integers $a, b \in \mathbb{Z}$, what kind of number is $\sqrt{a^2 + b^2}$?

Since $a^2 + b^2$ is a natural number (discarding the case when $a = b = 0$, that is, when the two points P_0 and P_1 coincide), we know from our earlier study that $\sqrt{a^2 + b^2}$ is an irrational if and only if $a^2 + b^2$ is not a square, that is, there does not exist $c \in \mathbb{N}$ satisfying $a^2 + b^2 = c^2$. Thus, we see that $\sqrt{a^2 + b^2}$ is either irrational or a non-negative integer c satisfying $a^2 + b^2 = c^2$. A triple (a, b, c), $a, b, c \in \mathbb{N}$, satisfying $a^2 + b^2 = c^2$ is called **Pythagorean triple**, and we will study them in Section 5.7. Note, in particular, the interesting consequence that a (genuine) positive fraction (with non-zero denominator) cannot be the distance between two points with integral coordinates.

Before getting into the detailed study of the distance, we show that the Postulate of Line Measure holds in our model $\mathbb{R}^2$.

We let a line ℓ be given by two of its (distinct) points $P_0 = (x_0, y_0)$ and $P_1 = (x_1, y_1)$, and let $c : \ell \to \mathbb{R}$, $c(P_t) = t, t \in \mathbb{R}$, be the corresponding affine coordinate function. For $s, t \in \mathbb{R}$, we calculate the distance $d(P_s, P_t)$ as follows:

$$d(P_s, P_t) = \sqrt{((t - s)(x_0 - x_1))^2 + ((t - s)(y_0 - y_1))^2} = |t - s| d(P_0, P_1).$$

For $s = 0$, this gives $d(P_0, P_t) = |t| d(P_0, P_1)$, $t \in \mathbb{R}$. We now let $t_0 = d(P_0, P_1) > 0$, discard the old P_1, and replace it with the new $\bar{P}_1 = P_{1/t_0}$ to obtain a new affine parametrization $\bar{c}_\ell$ of the line ℓ with $\bar{P}_0 = P_0$ and the new $\bar{P}_1 (= P_{1/t_0})$. With respect to this new parametrization, we have $d(\bar{P}_0, \bar{P}_1) = d(P_0, P_{1/t_0}) = d(P_0, P_1)/t_0 = 1$. With this, by the computation above, we have $d(\bar{P}_s, \bar{P}_t) = |t - s|$, $s, t \in \mathbb{R}$. Finally, we set $\bar{c}_\ell(\bar{P}_t) = t$, $t \in \mathbb{R}$. Clearly, $\bar{c}_\ell$ is a metric coordinate function since $|\bar{c}_\ell(\bar{P}_t) - \bar{c}_\ell(\bar{P}_s)| = |t - s| = d(\bar{P}_s, \bar{P}_t)$, $s, t \in \mathbb{R}$. The Postulate of Line Measure follows.

We now turn to the properties of the Cartesian distance d:

1. **Non-negativity**: $d(P_0, P_1) \geq 0$ for all $P_0, P_1 \in \mathbb{R}^2$, and $d(P_0, P_1) = 0$ if and only if $P_0 = P_1$.
2. **Symmetry**: $d(P_0, P_1) = d(P_1, P_0)$ for all $P_0, P_1 \in \mathbb{R}^2$.
3. **(Strict) Triangle Inequality**: $d(P_0, P_1) \leq d(P_0, Q) + d(Q, P_1)$ for all $P_0, P_1, Q \in \mathbb{R}^2$. The triangle inequality is **strict** in the sense that equality holds if and only if $Q \in [P_0, P_1]$.

Remark Strictness of the triangle inequality above (the second statement in 3) shows that our definition of betweenness is equivalent to Birkhoff's.

Non-negativity and symmetry follow from the Postulate of Line Measure (as noted above). We only need to show the triangle inequality.

We let $P_0 = (x_0, y_0)$, $P_1 = (x_1, y_1)$, and $Q = (x_2, y_2)$ and denote $a = x_0 - x_2$, $b = x_2 - x_1$, and $c = y_0 - y_2$, $d = y_2 - y_1$, so that, we have $a + b = x_0 - x_1$ and $c + d = y_0 - y_1$.

For the triangle inequality, we need to show

$$\sqrt{(a+b)^2 + (c+d)^2} \leq \sqrt{a^2 + c^2} + \sqrt{b^2 + d^2}.$$

Squaring both sides, we have

$$(a+b)^2 + (c+d)^2 \leq a^2 + b^2 + c^2 + d^2 + 2\sqrt{a^2 + c^2}\sqrt{b^2 + d^2}.$$

Expanding and simplifying, we obtain

$$ab + cd \leq \sqrt{a^2 + c^2}\sqrt{b^2 + d^2}.$$

Squaring both sides again, we arrive at the **Cauchy–Schwarz inequality**:

$$(ab + cd)^2 \leq (a^2 + c^2)(b^2 + d^2).$$

Since the steps that we made are reversible, we obtain that the triangle inequality is equivalent to the Cauchy–Schwarz inequality above.

The latter, however, is a direct consequence of the identity

$$(ab + cd)^2 + (ad - bc)^2 = (a^2 + c^2)(b^2 + d^2),$$

which can be verified by expanding all parentheses. (On the left-hand side, the "hybrid terms" $abcd$ cancel, and the "biquadratic terms" a^2b^2, etc. on both sides are the same.)

Thus, the Cauchy–Schwarz inequality and thereby the triangle inequality follow.

Remark The identity above is a special case of Brahmagupta's identity ($d = -1$) discussed in Section 2.1. Note also that, for $a, b, c, d \in \mathbb{N}$, this identity gives the following interesting fact: If $m, n \in \mathbb{N}$ are sums of squares of integers, then so is the product $m \cdot n$.

Finally, we now turn to the proof of strictness of the triangle inequality: For $P_0, P_1, Q \in \mathbb{R}$, we have $Q \in [P_0, P_1]$ if and only if $d(P_0, P_1) = d(P_0, Q) + d(Q, P_1)$.

We use the notations as above: $P_0 = (x_0, y_0)$, $P_1 = (x_1, y_1)$, and $Q = (x_2, y_2)$. For the "if" part, assuming that equality holds in the triangle inequality, and thereby in the Cauchy–Schwarz inequality, the identity above implies

$$ad - bc = (x_0 - x_2)(y_2 - y_1) - (x_2 - x_1)(y_0 - y_2) = 0.$$

As in the remark (for $x = x_2$ and $y = y_2$) before the proof of the Postulate of Line Measure (Section 5.2), this means that Q is on the line ℓ containing the points P_0 and P_1. Let $c_\ell : \ell \to \mathbb{R}$ be the affine coordinate function associated with P_0 and P_1. Since $Q \in \ell$, we have $Q = P_t$ for some $t \in \mathbb{R}$. With this, we have $d(P_0, Q) = d(P_0, P_t) = |t|d(P_0, P_1)$ and $d(Q, P_1) = d(P_t, P_1) = |1 - t|d(P_0, P_1)$. Hence $|t| + |1 - t| = 1$ holds. This means that $0 < t < 1$ so that $Q = P_t \in [P_0, P_1]$. The claim follows.

The "only if" part is obvious since $Q \in [P_0, P_1]$ implies $Q = P_t$ for some $0 \le t \le 1$, and thus $d(P_0, Q) + d(Q, P_1) = d(P_0, P_t) + d(P_t, P_1) = td(P_0, P_1) + (1 - t)d(P_0, P_1) = d(P_0, P_1)$.

Example 5.3.2 Given $A, B \in \mathbb{R}^2$, the **midpoint** between A and B is a point $M \in \mathbb{R}^2$ such that $d(A, M) = d(B, M) = d(A, B)/2$. By the above, the midpoint is unique, and it is given by $M = (1/2)A + (1/2)B$. In terms of an affine parameter with $A = P_0$ and $B = P_1$, we have $M = P_{1/2}$.

The considerations above lead to the important concept of **orientation** in our model $\mathbb{R}^2$. We have seen above that if $P_0 = (x_0, y_0)$, $P_1 = (x_1, y_1)$, and $P_2 = (x_2, y_2)(= Q)$ are three **non-collinear** points, then[12]

$$\omega(P_0, P_1, P_2) = (x_2 - x_0)(y_2 - y_1) - (x_2 - x_1)(y_2 - y_0) \neq 0.$$

If $\omega(P_0, P_1, P_2) > 0$, then we say that the **ordered** triple (P_0, P_1, P_2) is **positively oriented**; otherwise ($\omega(P_0, P_1, P_2) < 0$), we say that (P_0, P_1, P_2) is **negatively oriented**. Clearly, if (P_0, P_1, P_2) is positively oriented, then so are the triples (P_1, P_2, P_0) and (P_2, P_0, P_1); and any other triples, such as (P_0, P_2, P_1), are negatively oriented.

The origin of our coordinate system in $\mathbb{R}^2$, a point in the positive first axis, and a point in the positive second axis (in this order) form a positively oriented triple.

Remark We usually list the vertices of a non-degenerate triangle $\triangle[A, B, C]$ such that (A, B, C) is positively oriented, $\omega(A, B, C) > 0$ (that is, they correspond to the uppercase letters of the English alphabet in increasing order).

Exercise

5.3.1. Let P_n, $3 \le n \in \mathbb{N}$, be the perimeter of a regular n-sided polygon such that its sides are tangent to a given circle. Show that the sequence $(P_{2^n})_{2 \le n \in \mathbb{N}}$ is strictly decreasing.

[12] We changed the sign to match with the customary positive orientation of $\mathbb{R}^2$.

5.4 The Triangle Inequality

The triangle inequality gets its name from its application to the side lengths of a triangle $\triangle[A, B, C]$. We denote the side lengths as follows: $a = d(B, C)$, $b = d(C, A)$, and $c = d(A, B)$. Then, for a non-degenerate triangle, the triangle inequality states the following three inequalities:

$$a < b + c, \quad b < c + a, \quad c < a + b.$$

These inequalities are difficult to work with. Introducing, however, the new indeterminates

$$u = \frac{b + c - a}{2}, \quad v = \frac{c + a - b}{2}, \quad w = \frac{a + b - c}{2},$$

the triangle inequalities simply translate into $u, v, w > 0$. The system above can easily be inverted to obtain $a = v + w$, $b = w + u$, $c = u + v$. We will give a simple geometric interpretation of this in the next section.

Remark This substitution is often termed as the "Ravi Substitution." It is an old problem solving strategy.[13]

In the applications, we need a simple but fundamental inequality as follows.

Example 5.4.1 For all $x, y \in \mathbb{R}$, we have $4xy \leq (x + y)^2$. Moreover, equality holds if and only if $x = y$.

Indeed, expanding and simplifying, the inequality gives $0 \leq x^2 - 2xy + y^2$, or equivalently, $0 \leq (x - y)^2$. Since the steps are reversible, the stated inequality follows. Note that equality holds if and only if $x = y$.

The **geometric mean** of two non-negative numbers $0 \leq x, y \in \mathbb{R}$ is defined as $\sqrt{xy}$, while the **arithmetic mean** is $(x + y)/2$. For $x, y \geq 0$, taking the square root on both sides of the inequality in the example above, we obtain

$$\sqrt{xy} \leq \frac{x + y}{2}.$$

This asserts that the geometric mean is always less than or equal to the arithmetic mean. It is usually called the **AM-GM inequality**. This, and its extensions to several variables, will play a paramount importance later.

Remark Let $x, y > 0$, and assume that they appear (anywhere) in a **geometric sequence** with a middle term between them. Then this middle term is equal to the geometric mean $\sqrt{xy}$. Indeed, in the geometric sequence x, z, y, the consecutive ratios are $z/x = y/z$. Thus, we have $z^2 = xy$, so that $z = \sqrt{xy}$.

[13] See, for example, Engel, A., *Problem solving strategies*, Springer, Berlin, 1997.

There are literally hundreds of problems in mathematical contests that reduce to a version of the AM-GM inequality. Herewith we give a few.

Example 5.4.2 For any $0 < u, v, w \in \mathbb{R}$, we have

$$(u+v)(v+w)(w+u) \geq 8uvw.$$

Indeed, by the AM-GM inequality, we have $u + v \geq 2\sqrt{uv}$. Applying this to all pairs in u, v, w, we obtain

$$(u+v)(v+w)(w+u) \geq 8\sqrt{uv}\sqrt{vw}\sqrt{wu} = 8uvw.$$

The inequality follows.

The inequality just derived implies that, for any $0 < a, b, c \in \mathbb{R}$, we have

$$abc \geq (a+b-c)(b+c-a)(c+a-b).$$

To show this, first note that only at most one of the factors on the right-hand side can be negative or zero. Indeed, if $a + b - c \leq 0$ and $b + c - a \leq 0$, say, then, adding, we obtain $2b \leq 0$, a contradiction. In addition, if exactly one of the factors on the right-hand side is negative or zero, then we are done since the left-hand side is positive. Thus, we may assume that all factors in the right-had side are positive. This means that a, b, c can be thought of as the side lengths of a triangle. Applying the substitution above, our inequality is transformed into the inequality of Example 5.4.2.

Example 5.4.3 For $0 < u, v, w \in \mathbb{R}$, we have

$$\sqrt{2}(\sqrt{u} + \sqrt{v} + \sqrt{w}) \leq \sqrt{u+v} + \sqrt{v+w} + \sqrt{w+u}, \quad u, v, w > 0.$$

Indeed, squaring both sides and simplifying, this inequality reduces to

$$2(\sqrt{uv} + \sqrt{vw} + \sqrt{wu}) \leq \sqrt{(u+v)(v+w)} + \sqrt{(v+w)(w+u)} + \sqrt{(w+u)(u+v)}.$$

We now claim that

$$\sqrt{uv} + \sqrt{vw} \leq \sqrt{(u+v)(v+w)}.$$

Once this is proved, performing the cyclic permutation $u \mapsto v \mapsto w \mapsto u$ twice, and adding the corresponding inequalities, our inequality follows. Thus, it remains to show this last inequality. Squaring again and simplifying, we have $2\sqrt{uv^2w} \leq uw + v^2$. But this is just another form of the AM-GM inequality.

Note that, by the substitution above, if a, b, c are the side lengths of a triangle, the inequality of Example 5.4.3 above gives[14]

$$\sqrt{a+b-c}+\sqrt{b+c-a}+\sqrt{c+a-b}\le\sqrt{a}+\sqrt{b}+\sqrt{c}.$$

Example 5.4.4 Show that, for $0\le x_1,\dots,x_n\in\mathbb{R}$, $2\le n\in\mathbb{N}$, we have

$$\sum_{1\le i<j\le n}x_ix_j\cdot\sum_{k=1}^{n}x_k^2\le\frac{1}{8}\left(\sum_{k=1}^{n}x_k\right)^4.$$

Indeed, using the AM-GM inequality as in Example 5.4.1, we calculate

$$\begin{aligned}\left(\sum_{k=1}^{n}x_k\right)^4&=\left(\sum_{k=1}^{n}x_k^2+2\sum_{1\le i<j\le n}x_ix_j\right)^2\\&\ge 4\left(\sum_{k=1}^{n}x_k^2\right)\left(2\sum_{1\le i<j\le n}x_ix_j\right)=8\sum_{1\le i<j\le n}x_ix_j\cdot\sum_{k=1}^{n}x_k^2\end{aligned}$$

The inequality now follows.

We complete this cadre of examples by one that shows how the AM-GM inequality can sometimes be used to solve a system of non-linear equations.

Example 5.4.5 Solve the following system of equations for $x, y, z\in\mathbb{R}$:

$$x+y=2,\quad xy-z^2=1.$$

First, $xy=1+z^2\ge 1$, so that, by the first equality, we obviously have $x, y>0$. Now, the AM-GM inequality gives $2=x+y\ge 2\sqrt{xy}$. Hence $xy\le 1$. Combining this with the previous inequality, we obtain $xy=1$, and consequently $x=y=1$. Finally, the second equality gives $z=0$.

We finish this section by another application of the AM-GM inequality: The **Babylonian Method** on how to approximate the square root of a natural number $a\in\mathbb{N}$ by rational numbers.

We assume that $a\in\mathbb{N}$ is not a perfect square. We let $0<q_0\in\mathbb{Q}$ and define the sequence $(q_n)_{n\in\mathbb{N}_0}$ inductively by

$$q_{n+1}=\frac{1}{2}\left(q_n+\frac{a}{q_n}\right),\quad n\ge 0.$$

[14] This was a problem in the Asian Pacific Mathematical Competition, 1996.

Remark We will give a geometric interpretation of this formula in Section 8.2. This can be substantially generalized to what is known as **Newton's Method**, and it goes far beyond our rational approximations of the square root of a natural number. This particular case was also known to the ancient Babylonians, and this is why it carries the name "Babylonian Method."

Clearly, $(q_n)_{n\in\mathbb{N}_0}$ is a sequence of rational numbers. We claim that it is decreasing from the **first** term q_1 onward, and $\lim_{n\to\infty} q_n = \sqrt{a}$.

By our initial choice, we have $q_0 > 0$. Assuming $q_n > 0$, $n \in \mathbb{N}_0$, a quick look at the inductive formula above gives $q_{n+1} > 0$. It now follows from Peano's Principle of Induction that $q_n > 0$ for all $n \in \mathbb{N}_0$.
We can actually say more. The AM-GM inequality gives

$$0 < a = q_n \cdot \frac{a}{q_n} < \frac{1}{4}\left(q_n + \frac{a}{q_n}\right)^2 = q_{n+1}^2.$$

Note the sharp inequality in the middle as $q_n \neq a/q_n$ (since otherwise we would have $q_n^2 = a$, and a would be a perfect square). We thus obtain $q_{n+1}^2 > a$ for all $n \in \mathbb{N}_0$.

Rearranging this last inequality, we have

$$q_{n+1} > \frac{1}{2}\left(q_{n+1} + \frac{a}{q_{n+1}}\right) = q_{n+2}, \quad n \in \mathbb{N}_0.$$

We see that the sequence $(q_n)_{n\in\mathbb{N}}$ is strictly decreasing and bounded from below. By (Cauchy) completeness of $\mathbb{R}$, we have $\lim_{n\to\infty} q_n = r \in \mathbb{R}$, and, in addition, we also have $r^2 \geq a$. Letting $n \to \infty$ in the inductive definition of the sequence $(q_n)_{n\in\mathbb{N}}$ above, we obtain

$$r = \frac{1}{2}\left(r + \frac{a}{r}\right).$$

Rearranging, we obtain $r^2 = a$. This finally gives[15] $r = \sqrt{a}$.

To find the rate of convergence, we introduce the **relative error**

$$\delta_n = \frac{q_n}{\sqrt{a}} - 1 > -1, \quad n \in \mathbb{N}_0.$$

Clearly, $\delta_n \neq 0$ since $\sqrt{a}$ is irrational. We rewrite this as $q_n = \sqrt{a}(\delta_n + 1)$, $n \in \mathbb{N}_0$. We now claim that the following inductive relation holds for the relative error:

$$\delta_{n+1} = \frac{\delta_n^2}{2(\delta_n + 1)}, \quad n \in \mathbb{N}_0.$$

[15] Note that this can also serve as a definition of $\sqrt{a}$ as the equivalence class of the rational Cauchy sequence $(q_n)_{n\in\mathbb{N}}$.

Note that this implies $\delta_n > 0$ for $n \in \mathbb{N}$. (The initial δ_0 may be negative.) We now calculate

$$\begin{aligned}\delta_{n+1} &= \frac{q_{n+1}}{\sqrt{a}} - 1 = \frac{q_n + a/q_n}{2\sqrt{a}} - 1 = \frac{q_n^2 + a}{2\sqrt{a}q_n} - 1 \\ &= \frac{a(1+\delta_n)^2 + a}{2\sqrt{a}\sqrt{a}(1+\delta_n^2)} - 1 = \frac{(1+\delta_n)^2 + 1}{2(\delta_n + 1)} - 1 = \frac{\delta_n^2}{2(\delta_n + 1)}.\end{aligned}$$

The inductive formula for the relative error above follows.

Since $\delta_n > 0$ for $n \in \mathbb{N}$, this implies

$$\delta_{n+1} = \frac{\delta_n}{\delta_n + 1} \cdot \frac{\delta_n}{2} \leq \frac{\delta_n}{2} \quad \text{and} \quad \delta_{n+1} = \frac{1}{\delta_n + 1} \cdot \frac{\delta_n^2}{2} \leq \frac{\delta_n^2}{2}.$$

(The first estimate is better than the second for $\delta_n > 1$, which may happen for some initial values of the indices $n \in \mathbb{N}$.) Putting these together, we obtain

$$0 < \delta_{n+1} \leq \min\left(\frac{\delta_n}{2}, \frac{\delta_n^2}{2}\right), \quad n \in \mathbb{N}.$$

The first estimate implies $\delta_2 \leq \delta_1/2$ $(n = 1)$, $\delta_3 \leq \delta_2/2 \leq \delta_1/2^2$ $(n = 2)$, $\delta_4 \leq \delta_3/2 \leq \delta_1/2^3$ $(n = 3)$, etc. In general, we have $0 < \delta_{n+1} \leq \delta_1/2^n$, $n \in \mathbb{N}$. In particular, we obtain $\lim_{n\to\infty} \delta_n = 0$. The Babylonian approximation method is established.

The following table depicts the first 5 iterates of the Babylonian Method for $\sqrt{2}$ starting with $q_0 = 1$:

n	q_n	δ_n
0	1	$-2.92893218 \cdot 10^{-1}$
1	$\frac{3}{2} = 1.5$	$6.06601778 \cdot 10^{-2}$
2	$\frac{17}{12} = 1.41\overline{6}\ldots$	$1.73460668 \cdot 10^{-3}$
3	$\frac{577}{408} = 1.414\overline{2156862745098039}\ldots$	$1.50182509 \cdot 10^{-6}$
4	$\frac{665857}{470832} \approx 1.41421356237468991062629 6$	$1.12773761 \cdot 10^{-12}$
5	$\frac{886731088897}{627013566048} \approx 1.41421356237309504880 2$	$6.35896059 \cdot 10^{-25}$

Exercise

5.4.1. A triangle with side lengths that form three consecutive terms in a geometric sequence exists if and only if the ratio q of the geometric sequence satisfies $1/\tau < q < \tau$, where τ is the golden number.

5.5 Lines and Circles

We now return to our Cartesian distance d and study its invariance properties under some transformations (self-maps) of the plane $\mathbb{R}^2$.

For $W \in \mathbb{R}^2$, we define the **translation** by W as the map $T_W : \mathbb{R}^2 \to \mathbb{R}^2$ given by $T_W(P) = P + W$, $P \in \mathbb{R}^2$. In coordinates, if $W = (u, v) \in \mathbb{R}^2$, then we have

$$T_W(P) = P + W = (x + u, y + v), \quad P = (x, y) \in \mathbb{R}^2.$$

The Cartesian distance d is invariant (unchanged) under translations; that is, for $W \in \mathbb{R}^2$, we have

$$d(T_W(P_0), T_W(P_1)) = d(P_0, P_1), \quad P_0, P_1 \in \mathbb{R}^2.$$

Indeed, this is obvious since, in the definition of the Cartesian distance, we take differences of the respective coordinates of P_0 and P_1, and thus the coordinates u and v of W cancel.

Another type of transformation of the plane $\mathbb{R}^2$ that we will utilize is the (positive) **quarter-turn**. We define the (positive) quarter-turn $S_0 : \mathbb{R}^2 \to \mathbb{R}^2$ about the origin by $S_0(P) = S_0(x, y) = (-y, x)$, $P = (x, y) \in \mathbb{R}^2$. Its square $S_0^2 = S_0 \circ S_0 : \mathbb{R}^2 \to \mathbb{R}^2$ (by composition) is the negative of the identity $-\mathrm{id}_{\mathbb{R}^2}$, the **half-turn** about the origin, given by $S_0^2(x, y) = (-x, -y)$, $P = (x, y) \in \mathbb{R}^2$. The (positive) quarter-turn about any point $O \in \mathbb{R}^2$ is defined by the composition $S_O = T_O \circ S_0 \circ T_{-O}$. Once again, the square S_O^2 is the half-turn about the point O. We call O the **center** of S_O.

Once again, it follows easily that the Cartesian distance d is invariant under a quarter-turn about any center.

These transformations are **affine** in the sense that they map lines to lines preserving the respective affine coordinate functions. (This follows immediately from the strict triangle inequality or by direct computation.) In particular, for $P_0, P_1 \in \mathbb{R}^2$, we have $T_W([P_0, P_1]) = [T_W(P_0), T_W(P_1)]$, $W \in \mathbb{R}^2$, and $S_O([P_0, P_1]) = [S_O(P_0), S_O(P_1)]$, $O \in \mathbb{R}^2$. In addition, these transformations preserve the relation being parallel, that is, they send pencils of parallel lines to pencils of parallel lines.

A translation sends a line to a **parallel** line. Indeed, if a line ℓ is given by the equation $ax - by = c$, $a^2 + b^2 > 0$, then a translation T_W with $W = (u, v) \in \mathbb{R}^2$ sends ℓ into a line with equation $a(x - u) - b(y - v) = c$, that is, $ax - by = c + au - bv$.

A half-turn sends a line to a **parallel** line. Indeed, since translations do the same (by the above), it is enough to show this for the half-turn about the origin, the negative of the identity map. Now, if a line ℓ is given by the equation $ax - by = c$, $a^2 + b^2 > 0$, then the half-turn about the origin sends ℓ to the line with equation $-ax + by = c$, that is, $ax - by = -c$.

Finally note that these transformations preserve ω, and therefore they are **orientation preserving** in the sense that if (P_0, P_1, P_2) is a positively oriented triple, then so are the corresponding transformed triples.

Remark The adjective "positive" for the quarter-turn S_O about a point $O \in \mathbb{R}^2$ is due to the fact that, for any $P \in \mathbb{R}^2$, $P \neq O$, the triple $(O, P, S_O(P))$ is **positively** oriented. Indeed, since translations are orientation preserving, it is enough to show this for the quarter-turn about the origin. Now, if $P = (x, y)$, $x^2 + y^2 > 0$, then we have $\omega((0, 0), (x, y), (-y, x)) = x^2 + y^2 > 0$.

We say that two lines are **perpendicular** if one is obtained from the other by a quarter-turn. As simple computation shows, if a line ℓ is given by the equation $ax - by = c, a^2 + b^2 > 0$, then, for $O = (u, v) \in \mathbb{R}^2$, the transformed perpendicular line $S_O(\ell)$ is given by the equation $bx + ay = c + a(v - u) + b(v + u)$.

On the other hand, as noted above, a pencil of parallel lines is given by the equation $ax - by = c$, where the constant $c \in \mathbb{R}$ varies over all real numbers. It follows that the relation of being perpendicular depends only on the pencils of parallel lines that each of the two perpendicular lines is participating in. In other words, if two lines are perpendicular, then so are any two lines in the respective pencils of parallel lines.

Since every pencil of parallel lines contains a unique representative through the origin, and $S_0^2 = -\mathrm{id}$, it also follows that the relation being perpendicular is symmetric.

Finally, if two lines ℓ and ℓ' are intersected by another line perpendicular to both, then ℓ and ℓ' are parallel.

Example 5.5.1 (Perpendicular Bisector) Determine the set of points that are **equidistant** from two given distinct points P_0 and P_1 on the plane $\mathbb{R}^2$.

I. Algebraic Solution. Let $P_0 = (x_0, y_0)$ and $P_1 = (x_1, y_1)$, $P_0 \neq P_1$. A variable point $P = (x, y)$ is equidistant from P_0 and P_1 if and only if $d(P, P_0) = d(P, P_1)$. Using the distance formula, after squaring, we calculate

$$\begin{aligned}(x - x_0)^2 + (y - y_0)^2 &= (x - x_1)^2 + (y - y_1)^2\\ 2(x_0 - x_1)x + 2(y_0 - y_1)y &= x_0^2 - x_1^2 + y_0^2 - y_1^2\\ 2(x_0 - x_1)x + 2(y_0 - y_1)y &= (x_0 - x_1)(x_0 + x_1) + (y_0 - y_1)(y_0 + y_1).\end{aligned}$$

A final rearrangement and grouping the multiples of $(x_0 - x_1)$ and $(y_0 - y_1)$ give the symmetric form

$$(x_1 - x_0)\left(x - \frac{x_0 + x_1}{2}\right) + (y_1 - y_0)\left(y - \frac{y_0 + y_1}{2}\right) = 0.$$

First, this equation is linear (since $(x_0 - x_1)^2 + (y_0 - y_1)^2 > 0$), and hence it must represent a line. Second, it is also clear (by substituting $x = (x_0 + x_1)/2$ and $y = (y_0 + y_1)/2$) that this line contains the **midpoint** M of the line segment $[P_0, P_1]$ (Example 5.3.2) since

$$M = \left(\frac{x_0 + x_1}{2}, \frac{y_0 + y_1}{2} \right).$$

Finally, as shown above, an equation of the line containing the points $P_0 = (x_0, y_0)$ and $P_1 = (x_1, y_1)$ is given by

$$(y_1 - y_0)x - (x_1 - x_0)y = x_0 y_1 - x_1 y_0.$$

Comparing the equations of these lines, we see that they are perpendicular.

We conclude that the set of points equidistant from two distinct points P_0 and P_1 is the **perpendicular bisector** of the line segment $[P_0, P_1]$ that contains their **midpoint** M. As a byproduct, we also recover the **midpoint formula** for M above (Example 5.3.2 again).

II. Geometric Solution After Euclid.[16] Let P be a point such that $d(P, P_0) = d(P, P_1)$. Then the triangle $\triangle[P, P_0, P_1]$ is isosceles. Thus, by the Theorem on Isosceles Triangles in Euclid's *Elements* (Book I, Proposition 5), the angles at P_0 and P_1 are congruent. Consider the midpoint M of the line segment $[P_0, P_1]$. Then the sub-triangles $\triangle[P, P_0, M]$ and $\triangle[P, P_1, M]$ are congruent.[17] Thus, these sub-triangles must have right angle at M. We obtain that P is on the perpendicular bisector of the line segment $[P_0, P_1]$.

History

In the discussion above we used the **Theorem on Isosceles Triangles**: The angles opposite to the equal sides of an isosceles triangle are equal. It is also called the **pons asinorum**, "bridge of donkeys" in Latin. It is either a somewhat derogatory phrase pointing to and challenging the reader to tackle this first non-trivial proposition in the *Elements* and pass this bridge to get to harder ones, or the isosceles triangle depicts an actual pointed bridge that only a brave and sure-footed donkey can pass.

We now introduce another fundamental concept of Greek geometry, the concept of a circle. Given a point $O \in \mathbb{R}^2$ and a positive real number $0 < r \in \mathbb{R}$, we define the **circle** of radius r and center at O as the set[18]

$$\mathbb{S}_{O,r} = \{P \in \mathbb{R}^2 \mid d(P, O) = r\}.$$

Letting $O = (u, v)$ and $P = (x, y)$, the Cartesian distance formula gives the equation of the circle $\mathbb{S}_{O,r}$ as

$$(x - u)^2 + (y - v)^2 = r^2,$$

[16] This geometric solution can be reworded to become a simple consequence of Birkhoff's Postulates of Angle Measure and Similarity. The validity of these postulates in our model will be proved in Section 5.7. Hence, for a change, we give here a proof based on Euclid's *Elements*.

[17] Warning: Congruence of the triangles $\triangle[P, P_0, M]$ and $\triangle[P, P_1, M]$ also follows from the observation that the lengths of the three pairs of sides of these triangles are equal, but in the *Elements*, this occurs after Proposition 5 of Book I.

[18] Compare this with primitive #15 in the *Elements* as stated at the beginning of this chapter.

where we squared both sides to eliminate the square root. If the center of the circle is the origin 0, then we write $\mathbb{S}_r = \mathbb{S}_{0,r}$. Similarly, if the circle has a unit radius ($r = 1$), then we write $\mathbb{S}_O = \mathbb{S}_{O,1}$. Finally, $\mathbb{S}$ denotes the unit radius circle with center at the origin.

As a simple application of the perpendicular bisector and the concept of a circle, consider three distinct points $A, B, C \in \mathbb{R}^2$ on the plane. We ask the following question: Is there any point on the plane which is equidistant from all these three points?

If A, B, C are **collinear**, then the answer is no. Indeed, the set of points equidistant from A and B is the perpendicular bisector of the line segment $[A, B]$, and the set of points equidistant from B and C is the perpendicular bisector of $[B, C]$. These bisectors are distinct and perpendicular to the line that passes through A, B, C. Thus, they are parallel and have no common intersection.

Assume now that A, B, C are **not collinear**. Consider the perpendicular bisectors ℓ_a, ℓ_b, ℓ_c of the line segments $[B, C], [C, A], [A, B]$. By the previous step, each pair of bisectors intersects in a point. We claim that these (three) points are the same. Indeed if O is the common intersection of ℓ_a and ℓ_b, then $d(O, B) = d(O, C)$ and $d(O, C) = d(O, A)$. Therefore, $d(O, A) = d(O, B)$ so that O is equidistant from A and B, and hence it is on the bisector ℓ_c. We obtain that if A, B, C are not collinear, then there is a **unique** point equidistant from all these three points.

This conclusion can be put into a familiar framework if we consider the points A, B, C, the vertices of the triangle $\triangle[A, B, C]$. Since $d(O, A) = d(O, B) = d(O, C) = R$, say, we see that the circle with center O and radius R contains the points A, B, C. This is the unique circle **circumscribed** about the triangle. We call this the **circumcircle** and its radius the **circumradius** R of the triangle.

We now return to the main line and study the possible configurations of a circle and a line. We claim that there are three possibilities; namely, the circle and the line may be disjoint, meet at one point, or meet at two points. When a circle meets a line at exactly one point, we say that the line is **tangent** to the circle. A line is **secant** to a circle if they intersect at exactly two points.[19]

Discarding the case when the circle and the line are disjoint, we assume that the circle $\mathbb{S}_{O,r}$ and the line ℓ intersect at least in one point $P_0 \in \mathbb{S}_{O,r} \cap \ell$. For simplicity, we translate the entire configuration such that the center of the circle is at the origin. The equation of the circle $\mathbb{S}_r$ above reduces to

$$x^2 + y^2 = r^2.$$

We let $P_0 = (x_0, y_0)$, so that $x_0^2 + y_0^2 = r^2$. As usual, we write the equation of the line ℓ through P_0 as

$$ax - by = ax_0 - by_0, \quad a^2 + b^2 > 0.$$

[19]The words "tangent" and "secant" are derived from "tangere" and "secare," respectively, which in Latin mean "to touch" and "to cut."

After simplification and factoring, we obtain that any intersection point $P = (x, y) \in \mathbb{S}_r \cap \ell$ satisfies the system of equations

$$(x - x_0)(x + x_0) + (y - y_0)(y + y_0) = 0$$
$$a(x - x_0) = b(y - y_0).$$

Assuming $P \neq P_0$, that is, discarding the solution $x = x_0$ and $y = y_0$, we obtain $b(x + x_0) + a(y + y_0) = 0$. A simple computation now gives

$$x = x_0 - 2b\frac{bx_0 + ay_0}{a^2 + b^2} \quad \text{and} \quad y = y_0 - 2a\frac{bx_0 + ay_0}{a^2 + b^2}.$$

This is a solution different from $x = x_0$ and $y = y_0$ (which has been discarded) if and only if $bx_0 + ay_0 \neq 0$.

Turning the question around, we see that the circle $\mathbb{S}_r$ and the line ℓ have a **unique** intersection point $P_0 = (x_0, y_0)$ if and only if $bx_0 + ay_0 = 0$. By the discussion at the beginning of this section, $bx + ay = 0$ is an equation of a line **perpendicular** to the tangent line ℓ, and the latter has the equation $ax - by = ax_0 - by_0$.

As a final note, we claim that the entire tangent line ℓ with the exception of the point of tangency P_0 lies in the **exterior** of the circle $\mathbb{S}_r$.

Indeed, letting $P_t = (x_0 + tb, y_0 + at)$, $t \in \mathbb{R}$, the equation of the tangent line ℓ in the form $a(x - x_0) = b(y - y_0)$ above clearly shows that $\ell = \{P_t \mid t \in \mathbb{R}\}$. We now calculate the distance as

$$d(P_t, P_t)^2 = (x_0 + tb)^2 + (y_0 + ta)^2 = x_0^2 + y_0^2 + 2t(bx_0 + ay_0) + t^2(a^2 + b^2)$$
$$= r^2 + t^2(a^2 + b^2) \geq r^2$$

with equality if and only if $t = 0$, that is, at the point of tangency P_0. The claim follows.

Summarizing, we obtain that through any point P_0 of a circle $\mathbb{S}_{O,r}$ (with center O), there is a unique line ℓ that is **tangent** to $\mathbb{S}_{O,r}$, and it is characterized by the property that it is perpendicular to the radial line containing P_0 and the center O of the circle. Any other line through P_0 is a secant to $\mathbb{S}_{O,r}$, that is, it intersects the circle in two distinct points. Finally, the entire tangent line lies in the exterior of the circle $\mathbb{S}_{O,r}$, except the point of tangency P_0.

History

The ancient Greek mathematicians elevated the study of geometric configurations to the discipline of **Geometry**. These include lines, polygons, circles, parabolas, ellipses, hyperbolas, their metric properties and mutual relationships, such as tangents, secants, intersections, etc. As noted previously, the ancient Greeks also created **Geometric Algebra**, which associated algebraic terms with geometric objects, such as length, perimeter, area, etc. **Algebra** as a discipline separate from Geometry (and Arithmetic) was established by Muhammad ibn Mūsā al-Khwārizmī. To a large extent it was an early theory of equations that studied solutions of linear and quadratic equations. This theory had different developments by Diophantus of Alexandria and the Indian mathematician Brahmagupta. Throughout the Middle Ages Arabic scholars raised this discipline

to new heights. The modern symbolic representation of variables and constants, introduced by the French mathematician François Viète (Latin Vieta) (1540–1603) and subsequently brought to perfection by Descartes, put algebra on a solid foundation. But it was the introduction and systematic use of coordinate systems that revolutionized mathematics by establishing a bridge between geometry and algebra.

As an application, we introduce the concept of **distance of a point from a line**. Let O be a point and ℓ a line. By definition, the distance of O from ℓ is

$$d(O, \ell) = \inf\{d(O, P) \mid P \in \ell\}.$$

If $O \in \ell$, then $d(O, \ell) = 0$. We may therefore assume that $O \notin \ell$.

We claim that there is a unique circle with center at O and with ℓ as a tangent line to the circle.

Indeed, by what we proved above, this circle is obtained by taking the line ℓ' through O perpendicular to ℓ, and the radius of the circle is the distance of the intersection point $P_0 = \ell \cap \ell'$ from the center O.

Since ℓ is tangent to the circle, all the points on ℓ except P_0 are in the exterior of the circle. Thus, the radius of the circle realizes the infimum above, and therefore it is the distance $d(O, \ell)$.

To obtain an explicit formula, let ℓ be given by the equation $ax - by = c$, $a^2 + b^2 > 0$, and let $O = (u, v) \in \mathbb{R}^2$. The equation of ℓ' through O and perpendicular to ℓ is given by $bx + ay = bu + av$. Putting these two equations together, a short computation gives the intersection point $P_0 = \ell \cap \ell'$ as

$$P_0 = \left(\frac{b(bu + av) + ac}{a^2 + b^2}, \frac{a(bu + av) - bc}{a^2 + b^2} \right).$$

By a short computation, we arrive at the distance of P_0 from $O = (u, v)$ as

$$d(O, \ell) = \frac{|au - bv - c|}{\sqrt{a^2 + b^2}}.$$

Example 5.5.2 (Angular Bisector) Determine the set of points that are **equidistant** from two given distinct lines ℓ_0 and ℓ_1 on the plane $\mathbb{R}^2$.

If ℓ_0 and ℓ_1 are parallel, then the set of points equidistant from both lines is the parallel line midway between ℓ_0 and ℓ_1.

Assume now that ℓ_0 and ℓ_1 intersect in a point C. The intersecting lines ℓ_0 and ℓ_1 split the plane into four angular sectors. Let P be a point such that $d(P, \ell_0) = d(P, \ell_1)$. We may assume that $P \neq C$ (since C is clearly equidistant from both lines with zero distance) and also that P is not on any of these two lines. Thus, P is contained in one of the open angular sectors. Let $P_0 \in \ell_0$ and $P_1 \in \ell_1$ be such that $d(P, P_0) = d(P, \ell_0) = d(P, \ell_1) = d(P, P_1)$. The two **right** triangles $\triangle[P, P_0, C]$ and $\triangle[P, P_1, C]$ are congruent since they have two equal sides, and they also have a

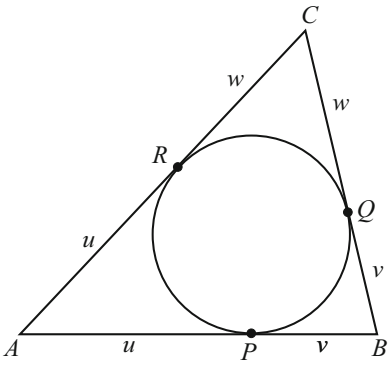


Fig. 5.1 The incircle of a triangle.

common side.[20] Thus their angles at C are equal. We obtain that P is on the bisector of the angular sector that contains P. We conclude that the set of points equidistant from two intersecting lines is the pair of perpendicular angular bisectors.

As an application, consider three lines ℓ_0, ℓ_1, and ℓ_2 that are the extensions of the sides of a (non-degenerate) triangle $\triangle[P_0, P_1, P_2]$. This triangle is the intersection of three angular sectors, one from each vertex (see Figure 5.1). The bisectors corresponding to each angular sector intersect in a point C. This is the unique point equidistant from all the three lines. We obtain that C is the center of the unique **inscribed** circle touching each line at the points where the distance of C and the lines are realized. We call this the **incircle** and its radius the **inradius** of the triangle.

Remark If $\triangle[A, B, C]$ is a (non-degenerate) triangle, then its incircle touches each side of the triangle at a specific point, $P \in [A, B]$, $Q \in [B, C]$, $R \in [C, A]$, say. Clearly, we have $d(A, P) = d(A, R)$, $d(B, P) = d(B, Q)$, and $d(C, Q) = d(C, R)$. Denoting these distances by u, v, and w, we obtain the substitution

$$a = v + w, \quad b = w + u, \quad c = u + v.$$

This gives the geometric interpretation of the substitution at the beginning of the previous section.

We now turn to study **secant** lines. Let $P_0, P_1 \in \mathbb{S}_{O,r}$ be two distinct points on the circle. Once again we translate the entire configuration so that the center of the circle $\mathbb{S}_r$ is at the origin, and thereby it has the equation $x^2 + y^2 = r^2$. By the strict triangle inequality, we have $d(P_0, P_1) \leq d(0, P_0) + d(0, P_1) = 2r$ with equality if and only if $0 \in [P_0, P_1]$.

[20] In this example we assume Birkhoff's Postulates of Angle Measure and Similarity and, consequently, the Pythagorean Theorem, whose validity, in our model, will be proved in Section 5.7. This is pedagogically justified since this example is a perfect fit for our present line of argument. Alternatively, one can also refer here to Euclid's *Elements*.

Letting $P_0 = (x_0, y_0)$, $x_0^2 + y_0^2 = r^2$, and $P_1 = (x_1, y_1)$, $x_1^2 + y_1^2 = r^2$, we calculate

$$\begin{aligned} d(P_0, P_1)^2 = (x_0 - x_1)^2 + (y_0 - y_1)^2 &= x_0^2 + y_0^2 + x_1^2 + y_1^2 - 2x_0x_1 - 2y_0y_1 \\ &= 2r^2 - 2(x_0x_1 + y_0y_1). \end{aligned}$$

Now recall the affine coordinate function that parametrizes the line through P_0 and P_1 under which the point $P_t = (1-t)P_0 + tP_1$ corresponds to the parameter $t \in \mathbb{R}$.

Using the result of the previous computation, we have

$$\begin{aligned} d(0, P_t)^2 &= ((1-t)x_0 + tx_1)^2 + ((1-t)y_0 + ty_1)^2 \\ &= (1-t)^2r^2 + t^2r^2 + 2t(1-t)(x_0x_1 - y_0y_1) \\ &= (1-t)^2r^2 + t^2r^2 + t(1-t)(2r^2 - d(P_0, P_1)^2) \\ &= r^2 - t(1-t)d(P_0, P_1)^2. \end{aligned}$$

In particular, we see that, for $t \in [0, 1]$, we have $d(0, P_t) \leq r$ with equality if and only if $t = 0, 1$.

Summarizing, given two distinct points P_0 and P_1 on a circle $\mathbb{S}_{O,r}$, for $t \in [0, 1]$, we have $d(P_t, O) \leq r$ (with equality if and only if $t = 0, 1$), that is, the line segment $[P_0, P_1]$ is contained in the interior of the circle $\mathbb{S}_{O,r}$. Similarly, for $t \notin [0, 1]$, we have $d(P_t, O) > r$.

Exercises

5.5.1. Calculate the length of the hypotenuse of the right triangle $\triangle[A, B, C]$ with right angle at C, where $d(A, C) = d(A, M) = 1$ and M is the midpoint of the hypotenuse $[A, B]$.

5.5.2. Use the pons asinorum to prove **Thales' Theorem**: If A, B, C are distinct points on a circle $\mathbb{S}_{O,r}$ and $O \in [A, B]$ (that is, $[A, B]$ is a diameter), then $\angle ACB$ is a right angle. Generalize this to the case when $[A, B]$ is a chord of the circle, $O \notin [A, B]$; the **Central Angle Theorem**: If C is on the longer circular arc of $\mathbb{S}_{O,r}$ with end-points A, B, then $\mu(\angle AOB) = 2\mu(\angle ACB)$; and if C is on the shorter circular arc of $\mathbb{S}_{O,r}$ with end-points A, B, then $\mu(\angle AOB) = 2(\pi - \mu(\angle BCA))$.

5.5.3. The **power** $\mathfrak{p}_S(P)$ of a point $P \in \mathbb{R}^2$ with respect to a circle $S = \mathbb{S}_{O,r}$ is defined as $\mathfrak{p}_S(P) = d(P, O)^2 - r^2$. (Note that the power is zero for points on the circle, negative for points inside the circle, and positive for points outside the circle.) (a) Prove the **Intersecting Chords Theorem**: Let P be outside S. Show that any line through P that meets S in the points $A, B \in S$;

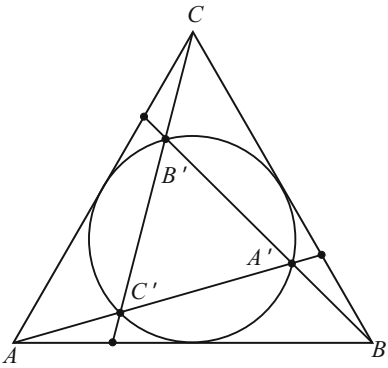


Fig. 5.2 An occurrence of the golden number τ.

we have $\mathfrak{p}_S(P) = d(P, A) \cdot d(P, B)$. (Note that, for a line tangent to S, the point of tangency is $A = B$ so that we have $\mathfrak{p}_S(P) = d(P, A)^2$, and the circle with center at P and radius $d(P, A)$ is orthogonal to S.) (b) Extend (a) to the case when P is inside S. (c) Let $S_1 = \mathbb{S}_{O_1,r_1}$ and $S_2 = \mathbb{S}_{O_2,r_2}$ be two disjoint circles. Show that the set of points in $P \in \mathbb{R}^2$ that have the same power with respect to S_1 and S_2 is a line, the so-called **radical line**. (d) Generalize the radical line to the case of intersecting circles or two circles with one inside the other. (e) Prove **Monge's Theorem**: Given three disjoint circles (with non-parallel radical axes), there is a circle orthogonal to all three.

5.5.4. Let $\triangle[A, B, C]$ be an equilateral triangle and S the incircle. Let $\triangle[A', B', C']$ be an equilateral triangle inscribed in S, that is, $A', B', C' \in S$, such that the line extensions of the sides of this triangle pass through the vertices of $\triangle[A, B, C]$; that is, $A' \in [B, B']$, $B' \in [C, C']$, $C' \in [A, A']$. Show that the ratio $d(A', C')/d(A, C')$ is the golden number τ (see Figure 5.2).

5.5.5. Given a line segment $[A, B]$ on the plane $\mathbb{R}^2$, determine the set of points $C \in \mathbb{R}^2$ such that the non-degenerate triangle $\triangle[A, B, C]$ has obtuse angle at C.

5.5.6. Let A and B be two points on the plane unit distance apart, and $0 < q < 1$, $q \in \mathbb{R}$. Show that the set $\{P \in \mathbb{R}^2 \mid d(P, A) = q \cdot d(P, B)\}$ is a circle, and determine its center and radius (in terms of q).

5.5.7. In a triangle $\triangle[A, B, C]$, let the angular bisector of the (interior) angle at the vertex A intersect the opposite side at the point $D \in [B, C]$. Prove the **Angle Bisector Theorem** $d(A, B)/d(A, C) = d(B, D)/d(C, D)$.

5.5.8. Consider two parallel chords of a circle S with lengths a and b which are d distance apart.[21] Let a third parallel chord of length c be in the midway of the first two. Express c in terms of a, b, d.

[21] Generalization of a problem in the American High School Mathematics Examination, 1995.

5.5.9. Three circles of common radius $0 < r \in \mathbb{R}$ are mutually and externally tangent. (a) If they are also internally tangent to a larger circle of radius $0 < R \in \mathbb{R}$, then find R/r. (b) Find the perimeter of the triangle whose sides are tangent to each pair of the three circles.

5.5.10. The vertices of a square of side length 2 are the centers of four circles of radius 1. Find the radius of the smaller circle externally tangent to these four circles whose center is the center of the square.

5.5.11. Two tangents are drawn from a point A to a circle of radius $0 < r \in \mathbb{R}$ and center O. Another tangent to the circle meets the two tangent lines at points B and C such that the triangle $\triangle[A, B, C]$ is disjoint from the circle. Find the perimeter of the triangle in terms of $d(A, O)$ and r.

5.5.12. A circle touches all four sides of an isosceles trapezoid. Find the radius r of the circle in terms of the parallel side lengths (bases) a and c of the trapezoid.

5.6 Arc Length on the Unit Circle

In this section we make a detailed and rigorous study of the **arc length** of circular arcs of a **unit** radius circle $\mathbb{S}_O$ with center $O \in \mathbb{R}^2$. The material presented here is technically demanding, and the readers who have only marginal interest in axiomatic developments may skip it. The results of this section will only be used for the establishment of the Birkhoff angle measure in the next section and for the proof of the Law of Cosines in trigonometry discussed in the last chapter of this book.

Let $P_0, P_1 \in \mathbb{S}_O$ be distinct points. The secant line through P_0 and P_1 divides the circle $\mathbb{S}_O$ into two **circular arcs** with end-points P_0 and P_1. These two circular arcs can be obtained from the equation of the line through the points $P_0 = (x_0, y_0)$ and $P_1 = (x_1, y_1)$ by intersecting the circle $\mathbb{S}_O$ with the two half-planes

$$(y_1 - y_0)x - (x_1 - x_0)y \gtreqless x_0y_1 - x_1y_0$$

defined by the equation of the (common boundary) line.

If $d(P_0, P_1) = 2$, then, by the strict triangle inequality, $O \in [P_0, P_1]$, and the two circular arcs are called **semi-circles**.[22] They are congruent via the half-turn about the center O. A line through O and perpendicular to the line extension of $[P_0, P_1]$ splits these two semi-circles into four **quarter-circles**. These quarter-circles are permuted (cyclically) by the quarter-turn S_O.

Assume now that $d(P_0, P_1) < 2$ so that, by the strict triangle inequality, $O \notin [P_0, P_1]$. Unless stated otherwise, we will always denote by $\mathcal{C} \subset \mathbb{S}_O$ the circular arc,[23] which is in the opposite side of the line extension of $[P_0, P_1]$ to the center O.

[22] Compare this with definition #18 in the *Elements* at the beginning of this chapter.

[23] As we will see below, $\mathcal{C}$ is the **shorter** (arc length) circular arc with end-points P_0 and P_1.

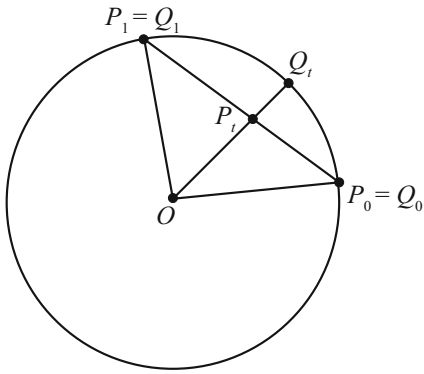


Fig. 5.3 Parametrization of a circular arc.

The other circular arc[24] will be denoted by $\mathcal{C}^c$ and will be called the **complement** of $\mathcal{C}$. For uniformity, for $d(P_0, P_1) = 2$, $\mathcal{C}$ (and $\mathcal{C}^c$) will denote either of the semi-circles with end-points P_0 and P_1.

Once again, recall the affine coordinate function that parametrizes the line through P_0 and P_1 under which the point $P_t = (1-t)P_0 + tP_1$ corresponds to the parameter $t \in \mathbb{R}$.

By the computations at the end of the previous section, we have

$$d(O, P_t)^2 = 1 - t(1-t)d(P_0, P_1)^2.$$

For $t \in [0, 1]$, we define[25]

$$Q_t = \frac{1}{d(O, P_t)} P_t + \left(1 - \frac{1}{d(O, P_t)}\right) O, \quad Q_0 = P_0, \ Q_1 = P_1.$$

Clearly, we have $d(O, Q_t) = d(O, P_t)/d(O, P_t) = 1$, or equivalently, $Q_t \in \mathbb{S}_O$, $t \in [0, 1]$ (see Figure 5.3). By the first formula defining the half-planes above, we see that the points Q_t, $t \in [0, 1]$, are in the half-plane that does not contain the point O. This gives

$$\mathcal{C} = \{Q_t \,|\, 0 \le t \le 1\}.$$

(If $Q \in \mathcal{C}$, then $[O, Q]$ and $[P_0, P_1]$ must intersect in a point P_t, $t \in [0, 1]$, say, so that $Q = Q_t$ holds.)

With this preparation, we now define the arc length of $\mathcal{C}$.

A **partition** of the interval $[0, 1]$ is a finite strictly increasing sequence:

$$(t_0, t_1, \ldots, t_{n-1}, t_n), \quad 0 = t_0 < t_1 < \cdots < t_{n-1} < t_n = 1, \quad n \in \mathbb{N}.$$

[24] The circular arc $\mathcal{C}^c$ is **not** the set-theoretic complement of $\mathcal{C}$ with respect to the whole circle $\mathbb{S}_O$ because $\mathcal{C}$ and $\mathcal{C}^c$ overlap in the two common end-points P_0 and P_1.

[25] Geometrically, the point $Q_t \in \mathcal{C}$ is obtained from P_t by radial projection from the center O.

We denote by Π the set of all partitions of $[0, 1]$ (for all $n \in \mathbb{N}$). A partition $(t_0, t_1, \ldots, t_{n-1}, t_n) \in \Pi$ of $[0, 1]$ defines an open **polygonal path** of $\mathcal{C}$ connecting $Q_0 (= P_0)$ and $Q_1 (= P_1)$ with consecutive vertices $Q_0 = Q_{t_0}, Q_{t_1}, \ldots, Q_{t_{n-1}}, Q_{t_n} = Q_1 \in \mathcal{C}$ with union $\bigcup_{i=1}^{n} [Q_{t_{i-1}}, Q_{t_i}]$. The **length** of a polygonal path is $\sum_{i=1}^{n} d(Q_{t_{i-1}}, Q_{t_i})$, the sum of the lengths of the participating line segments.

We now define the arc length of the circular arc $\mathcal{C} \subset \mathbb{S}_O$ by

$$\mathcal{L}_{\mathcal{C}} = \sup \left\{ \sum_{i=1}^{n} d(Q_{t_{i-1}}, Q_{t_i}) \,\middle|\, (t_0, t_1, \ldots, t_{n-1}, t_n) \in \Pi \right\}.$$

We need to show that the arc length $\mathcal{L}_{\mathcal{C}}$ is a (finite) real number; that is, the set of lengths of all polygonal paths of $\mathcal{C}$ is bounded above.

Let ℓ_0, resp. ℓ_1, denote the half-line with end-point O and containing P_0, resp. P_1 (For the notations introduced here and below, refer to Figure 5.4.). Let m_0, resp. m_1, be the tangent line to $\mathcal{C}$ through the point $Q_0 = P_0$, resp. $Q_1 = P_1$. By the results of the previous section, m_0, resp. m_1, is perpendicular to the line extension of ℓ_0, resp. ℓ_1. The lines m_0 and m_1 cannot be parallel since $O \notin [P_0, P_1]$. Let M be the intersection point of m_0 and m_1.

We claim that

$$\mathcal{L}_{\mathcal{C}} \leq d(Q_0, M) + d(Q_1, M).$$

Consider a polygonal path $\bigcup_{i=1}^{n} [Q_{t_{i-1}}, Q_{t_i}]$ with vertices $Q_0 = Q_{t_0}, Q_{t_1}, \ldots, Q_{t_{n-1}}, Q_{t_n} = Q_1 \in \mathcal{C}$ corresponding to a partition $(t_0, t_1, \ldots, t_{n-1}, t_n) \in \Pi$ of $[0, 1]$ (see again Figure 5.4). For $i = 0, 1, 2, \ldots, n$, let h_i, resp. k_i, be the line through Q_{t_i} and parallel to m_1, resp. m_0. Let h_i and m_0 meet at the point R_i, $i = 0, \ldots, n$; in particular, $R_0 = Q_0$ and $R_n = M$. Similarly, let k_i and m_1 meet at the point S_i, $i = 0, \ldots, n$; in particular, $S_0 = M$ and $S_n = Q_1$. Finally, for $i = 1, \ldots, n$, let h_i and k_{i-1} meet at the point T_i.

With these notations, by the triangle inequality, we have

$$d(Q_{t_{i-1}}, Q_{t_i}) \leq d(Q_{t_{i-1}}, T_i) + d(Q_{t_i}, T_i), \quad i = 1, \ldots, n.$$

On the other hand, $d(Q_{t_{i-1}}, T_i) = d(R_{i-1}, R_i)$ and $d(Q_{t_i}, T_i) = d(S_{i-1}, S_i)$, $i = 1, \ldots, n$, as the respective points are vertices of parallelograms.[26]

The next lemma will imply that $R_{i-1} * R_i * R_{i+1}$ and $S_{i-1} * S_i * S_{i+1}$, $i = 1, \ldots, n-1$. Once this is proved, it will follow that $\sum_{i=1}^{n} d(R_{i-1}, R_i) = d(R_0, R_n) = d(Q_0, M)$ and $\sum_{i=1}^{n} d(S_{i-1}, S_i) = d(S_0, S_n) = d(Q_1, M)$, so that the stated upper bound above holds, and the supremum defining the arc length is finite.

[26]The opposite sides of a parallelogram have equal lengths. This follows from translation invariance of the distance as shown at the beginning of Section 5.5.

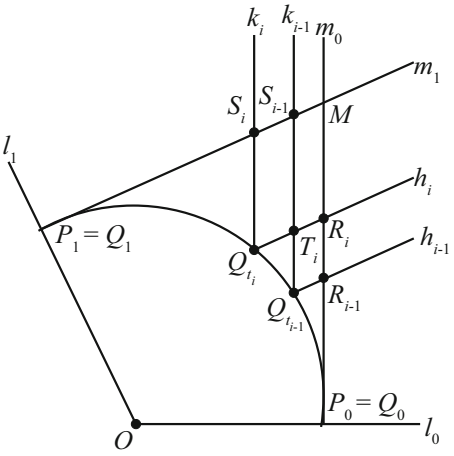


Fig. 5.4 Upper bound for the arc length.

In the lemma, without loss of generality, we assume that the center of the ambient circle is at the origin.

Lemma *Let* $\mathbb{S}$ *be the unit circle with center at the origin and* $P_0, P_1 \in \mathbb{S}$ *two distinct points with* $d(P_0, P_1) < 2$. *For* $t \in [0, 1]$, *let* $Q_t = P_t/d(0, P_t) \in \mathbb{S}$, $P_t = (1-t)P_0 + tP_1$; *and let* sP_0, $s \in \mathbb{R}$, *be the intersection of the radial line extension of* $[0, P_0]$ *and the line through* Q_t *perpendicular to this radial line. Then we have*

$$s = \frac{2 - td^2}{2\sqrt{t^2d^2 - td^2 + 1}}, \quad d = d(P_0, P_1).$$

In particular, s, *as a function of* $t \in [0, 1]$, *is strictly decreasing for* $0 \le t \le \min(2/d^2, 1)$ *and strictly increasing*[27] *for* $\min(2/d^2, 1) \le t \le 1$ *(see Figure 5.5).*

Proof We let $P_0 = (x_0, y_0)$, $x_0^2 + y_0^2 = 1$, and $P_1 = (x_1, y_1)$, $x_1^2 + y_1^2 = 1$. Setting $d = d(P_0, P_1)$, the computations at the end of the previous section give $d^2 = 2 - 2(x_0x_1 - y_0y_1)$ and $d(0, P_t)^2 = t^2d^2 - td^2 + 1$.

Since $d < 2$, for $t \in [0, 1]$, we have

$$d(0, P_t)^2 = t^2d^2 - td^2 + 1 = (t - 1/2)^2d^2 + 1 - d^2/4 > 0.$$

This gives

$$d(0, P_t) = \sqrt{t^2d^2 - td^2 + 1}, \quad t \in [0, 1].$$

[27] Monotonicity changes only if $d^2 > 2$, and then it does across $s = 0$, that is, when the sign of s changes from positive to negative; $s = 0$ corresponds to Q_{2/d^2} and P_{2/d^2} being perpendicular to P_0.

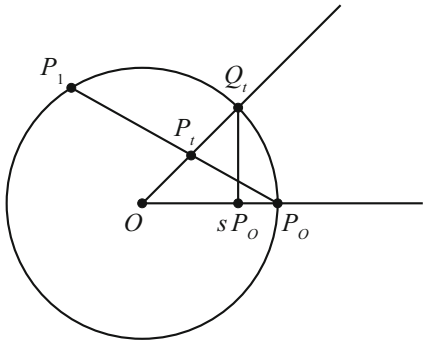


Fig. 5.5 Illustration for the monotonicity lemma.

The equation of the line extension of the radial line segment $[0, P_0]$ is of the form $y_0x - x_0y = 0$. The pencil of parallel lines perpendicular to this are described by the equations $x_0x + y_0y = c$, where $c \in \mathbb{R}$. Now, the value of the constant c is determined by the constraint that the perpendicular line must pass through the point $Q_t = P_t/d(0, P_t)$. This gives

$$c = \frac{x_0((1-t)x_0 + tx_1) + y_0((1-t)y_0 + ty_1)}{\sqrt{t^2d^2 - td^2 + 1}} = \frac{1 - t + t(x_0x_1 + y_0y_1)}{\sqrt{t^2d^2 - td^2 + 1}}$$
$$= \frac{1 - t + t(1 - d^2/2)}{\sqrt{t^2d^2 - td^2 + 1}} = \frac{2 - td^2}{2\sqrt{t^2d^2 - td^2 + 1}}.$$

On the other hand, by definition, the point $sP_0 = (sx_0, sy_0)$ must be contained in this perpendicular line. This gives $s = s(x_0^2 + y_0^2) = c$. Putting everything together, we obtain

$$s = \frac{2 - td^2}{2\sqrt{t^2d^2 - td^2 + 1}}.$$

It remains to show the last statement of the lemma, the monotonicity properties of s with respect to t.

First, let $0 \le t \le \min(2/d^2, 1)$. Since in this range $s \ge 0$, it is enough to show that s^2 is strictly decreasing. Now, a simple computation gives

$$s^2 = \frac{d^2}{4} + \left(1 - \frac{d^2}{4}\right) \frac{1 - td^2}{t^2d^2 - td^2 + 1}.$$

Since $d < 2$, we need to show that, for $0 \le t < t' \le \min(2/d^2, 1)$, we have

$$\frac{1 - td^2}{t^2d^2 - td^2 + 1} > \frac{1 - t'd^2}{t'^2d^2 - t'd^2 + 1}.$$

Eliminating the denominators, simplifying and factoring, this becomes

$$(t' - t)(td^2 + t'd^2 - (td^2)(t'd^2)) > 0.$$

This, however, clearly follows[28] since $0 \leq td^2 < t'd^2 \leq 2$. The claimed monotonicity is proved.

Second, for $\min(2/d^2, 1) \leq t \leq 1$, we have $s \leq 0$. Squaring again, the same argument gives the opposite monotonicity property. The lemma follows.

To establish the existence of (an upper bound of) the arc length for a circular arc $\mathcal{C} \subset \mathbb{S}_O$, we followed a **geometric method**. We now describe another essentially **analytic method** to obtain the same result.

First, we derive a geometric formula that will be used several times in the future. Consider a (non-degenerate) triangle $\triangle[A, B, C]$ with (non-collinear) vertices A, B, C. We let

$$A_0 = \frac{1}{d(C, A)} A + \left(1 - \frac{1}{d(C, A)}\right) C$$
$$B_0 = \frac{1}{d(C, B)} B + \left(1 - \frac{1}{d(C, B)}\right) C$$

In other words, A_0, and respectively B_0, is the point at unit distance from C on the half-line with end-point C and containing A, and respectively B. We then have the following important formula:[29]

$$d(A_0, B_0)^2 = 2 + \frac{d(A, B)^2 - d(C, A)^2 - d(C, B)^2}{d(C, A)d(C, B)}.$$

Indeed, using $A_0 - C = (A - C)/d(C, A)$ and $B_0 - C = (B - C)/d(C, B)$, and translation invariance of the distance multiple times, we calculate[30]

$$d(A_0, B_0)^2 = d(A_0 - C, B_0 - C)^2 = d\left(\frac{A - C}{d(A, C)}, \frac{B - C}{d(B, C)}\right)^2$$
$$= d\left(\frac{A - C}{d(A, C)} - \frac{B - C}{d(B, C)}, 0\right)^2 = d\left(\frac{A - C}{d(A, C)}, 0\right)^2 + \left(\frac{B - C}{d(B, C)}, 0\right)^2$$

[28] $0 < a, b \leq 2$ implies $(a + b)/(ab) = 1/a + 1/b \geq 1/2 + 1/2 = 1$.

[29] In different (non-axiomatic) developments, this formula is equivalent to the so-called Law of Cosines.

[30] It is customary to set $|P| = d(P, 0)$, the distance of a point P from the origin. Algebraically, $|P|^2$ is then the sum of squares of the two coordinates of P. Translation invariance then gives $d(P, Q)^2 = d(P - Q, 0)^2 = |P - Q|^2$. Using the fact that this is a quadratic form in the coordinates of the points P and Q, the computations above become more familiar.

$$+\frac{1}{d(A,C)d(B,C)}\left(d(A-B,0)^2-d(A-C,0)^2-d(B-C,0)^2\right)$$
$$=d(C,A_0)^2+d(C,B_0)^2+\frac{d(A,B)^2-d(C,A)^2-d(C,B)^2}{d(C,A)d(C,B)}.$$

Since $d(C,A_0)=d(C,B_0)=1$, the formula follows.

Returning to the main line, recall that we use the affine parametrization $P_t=(1-t)P_0+tP_1$, $t\in[0,1]$, for the line segment $[P_0,P_1]$, and the parametrization Q_t, $t\in[0,1]$, for the circular arc $\mathcal{C}$. For $t,t'\in[0,1]$, letting $P_t=A$, $P_{t'}=B$, $O=C$, so that $Q_t=A_0$, $Q_{t'}=B_0$, the formula above is rewritten as

$$d(Q_t,Q_{t'})^2=2+\frac{d(P_t,P_{t'})^2-d(O,P_t)^2-d(O,P_{t'})^2}{d(O,P_t)d(O,P_{t'})}$$

For future reference, we include here a useful equivalent form of this as

$$d(Q_t,Q_{t'})^2=2-\frac{2-(t+t'-2tt')d^2}{\sqrt{t^2d^2-td^2+1}\sqrt{t'^2d^2-t'd^2+1}},\quad d=d(P_0,P_1),$$

where we used $d(P_t,P_{t'})^2=(t-t')^2d^2$, $d(O,P_t)=\sqrt{t^2d^2-td^2+1}$ and $d(O,P_{t'})=\sqrt{t'^2d^2-t'd^2+1}$.

We rewrite the original formula and estimate

$$d(Q_t,Q_{t'})^2=\frac{d(P_t,P_{t'})^2-(d(O,P_t)-d(O,P_{t'}))^2}{d(O,P_t)d(O,P_{t'})}\le\frac{d(P_t,P_{t'})^2}{d(O,P_t)d(O,P_{t'})}$$

Setting, as usual, $d=d(P_0,P_1)<2$, for the denominator, we have

$$d(O,P_t)^2=t^2d^2-td^2+1=\left(t-\frac{1}{2}\right)^2d^2+1-\frac{d^2}{4}\ge1-\frac{d^2}{4}.$$

With this, we arrive at

$$d(Q_t,Q_{t'})\le\frac{d}{\sqrt{1-d^2/4}}|t-t'|,\quad t,t'\in[0,1],\ d=d(P_0,P_1)<2.$$

We express this by saying that the map $Q:[0,1]\to\mathbb{R}^2$, $Q(t)=Q_t$, $t\in[0,1]$, satisfies the **Lipschitz condition** with **Lipschitz constant** $d/\sqrt{1-d^2/4}$.

Applying this to a partition $(t_0,t_1,\dots,t_{n-1},t_n)\in\Pi$ of $[0,1]$, we see that the length of the corresponding polygonal path

$$\sum_{i=1}^{n}d(Q_{t_{i-1}},Q_{t_i})\le\frac{d}{\sqrt{1-d^2/4}}\sum_{i=1}^{n}(t_i-t_{i-1})=\frac{d}{\sqrt{1-d^2/4}}.$$

Taking the supremum for all polygonal paths, we finally obtain

$$\mathcal{L}_{\mathcal{C}} \leq \frac{d}{\sqrt{1-d^2/4}}, \quad d = d(P_0, P_1) < 2.$$

(Note that, as a simple computation shows, this upper estimate is the same as the one we obtained above by our geometric method).

We now explore the properties of the arc length of circular arcs in $\mathbb{S}_O$. We say that **additivity** holds in a circular arc if whenever it is split into two circular arcs $\mathcal{C}_1$ and $\mathcal{C}_2$ (by a common end-point), then we have

$$\mathcal{L}_{\mathcal{C}_1 \cup \mathcal{C}_2} = \mathcal{L}_{\mathcal{C}_1} + \mathcal{L}_{\mathcal{C}_2}.$$

We first claim that additivity holds in a circular arc $\mathcal{C} \subset \mathbb{S}_O$ with end-points P_0 and P_1 satisfying $d(P_0, P_1) < 2$ (and $\mathcal{C}$ and O are in opposite sides of the line extension of $[P_0, P_1]$).

Let $Q \in \mathcal{C}$, $P_0 \neq Q \neq P_1$. Let $\mathcal{C}_1 \subset \mathcal{C}$, resp. $\mathcal{C}_2 \subset \mathcal{C}$, be the circular arc with end-points P_0 and Q, resp. P_1 and Q. We have $\mathcal{C} = \mathcal{C}_1 \cup \mathcal{C}_2$. We claim that

$$\mathcal{L}_{\mathcal{C}} = \mathcal{L}_{\mathcal{C}_1} + \mathcal{L}_{\mathcal{C}_2}.$$

Indeed, since a partition of $[P_0, Q]$ and a partition of $[P_1, Q]$ can be united to define a partition of $[P_0, P_1]$, taking suprema, we see that $\mathcal{L}_{\mathcal{C}} \geq \mathcal{L}_{\mathcal{C}_1} + \mathcal{L}_{\mathcal{C}_2}$. On the other hand, letting $0 < \epsilon \in \mathbb{R}$, we can choose a polygonal path $\bigcup_{i=1}^{n} [Q_{t_{i-1}}, Q_{t_i}]$ such that

$$\mathcal{L}_{\mathcal{C}} - \epsilon \leq \sum_{i=1}^{n} d(Q_{t_{i-1}}, Q_{t_i}).$$

By the triangle inequality, adding $Q \in \mathcal{C}$ to a polygonal path for $\mathcal{C}$ increases its length. If Q participates in a polygonal path for $\mathcal{C}$, then this path can be split into the union of two polygonal paths, one for $\mathcal{C}_1$ and the other for $\mathcal{C}_2$. Once again, taking suprema, we obtain $\mathcal{L}_{\mathcal{C}} - \epsilon \leq \mathcal{L}_{\mathcal{C}_1} + \mathcal{L}_{\mathcal{C}_2}$. Since this is true for all $0 < \epsilon \in \mathbb{R}$, we obtain $\mathcal{L}_{\mathcal{C}} \leq \mathcal{L}_{\mathcal{C}_1} + \mathcal{L}_{\mathcal{C}_2}$. Additivity in the circular arc $\mathcal{C}$ follows.

Remark The arc length $\mathcal{L}_{\mathcal{C}}$ of a circular arc $\mathcal{C}$ with end-points P_0 and P_1 depends only on the distance $d(P_0, P_1) < 2$. This means that, if $\mathcal{C}'$ is another circular arc with end-points P_0' and P_1' such that $d(P_0', P_1') = d(P_0, P_1) < 2$, then we have $\mathcal{L}_{\mathcal{C}} = \mathcal{L}_{\mathcal{C}'}$. Indeed, a partition of the interval $[0, 1]$ induces a partition of $[P_0, P_1]$ and $[P_0', P_1']$, and the associated polygonal paths have the same lengths because, by the formula above, the distance $d(Q_t, Q_{t'})$ depends only on $d(P_0', P_1') = d(P_0, P_1)$ and the parameters $t, t' \in [0, 1]$. These equal lengths contribute the same amount to the suprema that define the arc lengths $\mathcal{L}_{\mathcal{C}}$ and $\mathcal{L}_{\mathcal{C}'}$, which thereby must be equal.

Finally, note an important special case. Let $\mathcal{C}$ be a circular arc with the end-points P_0 and P_1. We claim that $d(P_0, P_1) = \sqrt{2}$ if and only if the line extensions of the half-lines ℓ_0 and ℓ_1 (from the center O to the respective end-points) are perpendicular.

Indeed, assuming (for simplicity) that the center is at the origin, the equation of the line extension of the half-line ℓ_0 through P_0 is $y_0x - x_0y = 0$. The quarter-turn S_0 sends this line to the perpendicular line with the equation $x_0x + y_0y = 0$. This line contains the point $P_1 = (x_1, y_1)$ if and only if $x_0x_1 + y_0y_1 = 0$. Since $d(P_0, P_1)^2 = 2 - 2(x_0x_1 + y_0y_1)$ (see the end of the previous section), this is equivalent to $d(P_0, P_1)^2 = 2$. The claim follows.

With this, we can introduce the positive real number $\pi \in \mathbb{R}$ such that, for $d(P_0, P_1) = \sqrt{2}$, we have $\mathcal{L}_{\mathcal{C}} = \pi/2$. By the remark above, the arc length of a circular arc depends only on the distance between its end-points, so that π is well-defined.

We now **extend** the definition of the arc length to any circular arc of $\mathbb{S}_O$ using additivity.

If $d(P_0, P_1) < 2$, then the arc length of the circular arc $\mathcal{C}$ with end-points P_0 and P_1 has been defined above as the supremum of the lengths of its polygonal paths. In particular, for any quarter-circle $\mathcal{C}$ ($d(P_0, P_1) = \sqrt{2}$), we have $\mathcal{L}_{\mathcal{C}} = \pi/2$.

If $d(P_0, P_1) = 2$, then, by the sharp triangle inequality, the points P_0 and P_1 are collinear, and the center O is at the midpoint of the line segment $[P_0, P_1]$. We define $\mathcal{L}_{\mathcal{C}} = \mathcal{L}_{\mathcal{C}^c} = \pi$ for either of the semi-circles $\mathcal{C}$ or $\mathcal{C}^c$ with end-points P_0 and P_1. (Note that they are congruent via the half-turn S_O^2.)

If $d(P_0, P_1) < 2$, then we define the arc length of the circular arc $\mathcal{C}^c$ **complementary** to $\mathcal{C}$ by $\mathcal{L}_{\mathcal{C}^c} = 2\pi - \mathcal{L}_{\mathcal{C}}$.

Finally, we define the arc length of the entire circle $\mathbb{S}_O$ to be 2π.

We note that the arc length, being defined in terms of the Cartesian distance, is preserved under translations and half-turns.

We now claim that **additivity** holds in any circular arc of $\mathbb{S}_O$.

First, we show additivity in a semi-circle $\mathcal{C}$. Let $\mathcal{C}$ have end-points P_0 and P_1, and $Q \in \mathcal{C}$ with $P_0 \neq Q \neq P_1$. Then Q splits $\mathcal{C}$ into two circular arcs: $\mathcal{C}_1$ with end-points P_0 and Q and $\mathcal{C}_2$ with end-points P_1 and Q. We claim that $\mathcal{L}_{\mathcal{C}} = \mathcal{L}_{\mathcal{C}_1} + \mathcal{L}_{\mathcal{C}_2} = \pi$.

Indeed, either $\mathcal{C}_1$ or $\mathcal{C}_2$ contains a quarter-circle. Assume, without loss of generality, that the first does. If $\mathcal{C}_1$ is itself a quarter-circle, then so is $\mathcal{C}_2$ and the statement holds. Otherwise, split $\mathcal{C}_1$ into a quarter circle with one end-point at P_0 and a circular arc $\mathcal{C}_1'$ with one end-point at Q. Then, by additivity in $\mathcal{C}_1$ already shown, we have $\mathcal{L}_{\mathcal{C}_1} = \pi/2 + \mathcal{L}_{\mathcal{C}_1'}$. Since $\mathcal{C}_1'$ and $\mathcal{C}_2$ join to form another quarter-circle, again by additivity in quarter-circles already shown, we also have $\mathcal{L}_{\mathcal{C}_1'} + \mathcal{L}_{\mathcal{C}_2} = \pi/2$. Putting these together, we obtain $\mathcal{L}_{\mathcal{C}_1} + \mathcal{L}_{\mathcal{C}_2} = (\pi/2 + \mathcal{L}_{\mathcal{C}_1'}) + (\pi/2 - \mathcal{L}_{\mathcal{C}_1'}) = \pi = \mathcal{L}_{\mathcal{C}}$. Additivity in semi-circles follows.

Next, we need to show additivity in a circular arc of the form $\mathcal{C}^c$ complementary to the circular arc $\mathcal{C}$ with end-points P_0 and P_1 such that $d(P_0, P_1) < 2$. Let $Q \in \mathcal{C}^c$ distinct from P_0 and P_1. Then Q splits $\mathcal{C}^c$ into two circular arcs: $\mathcal{C}_1$ with one end-point at P_0 and another circular arc $\mathcal{C}_2$ with one end-point at P_1. We need to show $\mathcal{L}_{\mathcal{C}^c} = \mathcal{L}_{\mathcal{C}_1} + \mathcal{L}_{\mathcal{C}_2}$.

First, assume that either $\mathcal{C}_1$ or $\mathcal{C}_2$ contains a half-circle. Without loss of generality, we may assume that $\mathcal{C}_1$ does. We have $\mathcal{L}_{\mathcal{C}_1} = \mathcal{L}_{(\mathcal{C}\cup\mathcal{C}_2)^c} = 2\pi - \mathcal{L}_{\mathcal{C}\cup\mathcal{C}_2} = 2\pi - (\mathcal{L}_{\mathcal{C}} + \mathcal{L}_{\mathcal{C}_2})$, where we used additivity in $\mathcal{C} \cup \mathcal{C}_2$ (including the case when $\mathcal{C} \cup \mathcal{C}_2$ is a semi-circle). On the other hand, we have $\mathcal{L}_{\mathcal{C}^c} = 2\pi - \mathcal{L}_{\mathcal{C}} = 2\pi - (2\pi - (\mathcal{L}_{\mathcal{C}_1} + \mathcal{L}_{\mathcal{C}_2})) = \mathcal{L}_{\mathcal{C}_1} + \mathcal{L}_{\mathcal{C}_2}$. Additivity follows in this case.

Second, assume neither $\mathcal{C}_1$ nor $\mathcal{C}_2$ contains a semi-circle. Let $P_2 \in \mathbb{S}_O$ be the opposite to P_0 with respect to O. Since neither $\mathcal{C}$ nor $\mathcal{C}_1$ contains semi-circles, we have $P_2 \in \mathcal{C}_2$, $P_1 \neq P_2 \neq Q$. Then P_2 splits $\mathcal{C}_2$ into two circular arcs: $\mathcal{C}_2'$ with one end-point at P_1 and another $\mathcal{C}_2''$ with one end-point at Q. Clearly, $\mathcal{C}_1 \cup \mathcal{C}_2''$ is a semi-circle, so that, by the previous case, we have $\mathcal{L}_{\mathcal{C}^c} = \mathcal{L}_{\mathcal{C}_1\cup\mathcal{C}_2''} + \mathcal{L}_{\mathcal{C}_2'}$. On the other hand, by additivity in semi-circles, we also have $\mathcal{L}_{\mathcal{C}_1\cup\mathcal{C}_2''} = \mathcal{L}_{\mathcal{C}_1} + \mathcal{L}_{\mathcal{C}_1''}$. Putting these together, we obtain $\mathcal{L}_{\mathcal{C}^c} = \mathcal{L}_{\mathcal{C}_1\cup\mathcal{C}_2''} + \mathcal{L}_{\mathcal{C}_2'} = \mathcal{L}_{\mathcal{C}_1} + \mathcal{L}_{\mathcal{C}_1''} + \mathcal{L}_{\mathcal{C}_2'} = \mathcal{L}_{\mathcal{C}_1} + \mathcal{L}_{\mathcal{C}_2}$. This finishes the second case. Additivity in complementary circular arcs follows.

Finally, note that additivity in the entire circle $\mathbb{S}_O$ follows from the definitions as any split consists of complementary pairs of circular arcs.

The proof of the additivity of the arc length in general is now complete.

Remark In view of the additivity and the forthcoming discussion, it is convenient to define the arc length of a single point on $\mathbb{S}_O$ to be zero.

As the final task in this section, we claim that any given number $0 \leq r \leq 2\pi$ arises as the arc length of a circular arc $\mathcal{C} \subset \mathbb{S}_O$, that is, we have $\mathcal{L}_{\mathcal{C}} = r$. By additivity, it is enough to show this for $0 < r < \pi/2$. Let ℓ_0 and ℓ_1 be perpendicular half-lines with common end-point O such that, for $P_0 \in \ell_0$ and $P_1 \in \ell_1$, $d(O, P_0) = d(O, P_1) = 1$, and hence $d(P_0, P_1) = \sqrt{2}$. The unit interval $[0, 1]$ parametrizes the line segment $[P_0, P_1]$ by the affine coordinate function $P_t = (1 - t)P_0 + tP_1$, $t \in [0, 1]$, and the quarter-circle with end-points P_0 and P_1 by $Q_t = P_t/d(O, P_t) \in \mathbb{S}_O$, $t \in [0, 1]$.

For $t \in [0, 1]$, we denote by $\mathcal{C}_t \subset \mathbb{S}_O$ the circular arc with end-points $Q_0 (= P_0)$ and Q_t. The quarter-circle itself is then equal to $\mathcal{C}_1$, and we have

$$\mathcal{L}_{\mathcal{C}_0} = 0 \quad \text{and} \quad \mathcal{L}_{\mathcal{C}_1} = \pi/2.$$

We first study the properties of the arc length $\mathcal{L}_{\mathcal{C}_t}$ as a function of $t \in [0, 1]$. Since $d = d(P_0, P_1) = \sqrt{2}$, we have $d/\sqrt{1 - d^2/4} = 2$, so that the previous Lipschitz estimate gives

$$d(Q_t, Q_{t'}) \leq 2|t - t'|, \quad t, t' \in [0, 1].$$

On the other hand, the explicit formula for this distance specializes to

$$\begin{aligned} d(Q_t, Q_{t'})^2 &= 2 - \frac{2 - 2(t + t' - 2tt')}{\sqrt{2t^2 - 2t + 1}\sqrt{2t'^2 - 2t' + 1}} \\ &= 2 - \frac{2(1 - t)(1 - t')}{\sqrt{2t^2 - 2t + 1}\sqrt{2t'^2 - 2t' + 1}}, \quad t, t' \in [0, 1]. \end{aligned}$$

This gives

$$\max_{t,t' \in [0,1]} d(Q_t, Q_{t'}) = \sqrt{2}.$$

We now let $t, t' \in [0, 1]$ and consider the circular arc $\mathcal{C}_{t,t'} \subset \mathbb{S}_O$ with end-points Q_t and $Q_{t'}$. The general upper estimate derived earlier for the arc length gives

$$\mathcal{L}_{\mathcal{C}_{t,t'}} \leq \frac{d(Q_t, Q_{t'})}{\sqrt{1 - d(Q_t, Q_{t'})^2/4}} \leq 2\sqrt{2}|t - t'|, \quad t, t' \in [0, 1],$$

where we used the Lipschitz estimate and the maximum above. Finally, by additivity, the arc length of the path $\mathcal{C}_{t,t'}$ is equal to $\mathcal{L}_{\mathcal{C}_{t,t'}} = |\mathcal{L}_{\mathcal{C}_t} - \mathcal{L}_{\mathcal{C}_{t'}}|$. Putting these together, we obtain

$$|\mathcal{L}_{\mathcal{C}_t} - \mathcal{L}_{\mathcal{C}_{t'}}| \leq 2\sqrt{2}|t - t'|, \quad t, t' \in [0, 1].$$

This means that the arc length $\mathcal{L}_{\mathcal{C}_t}$ as a function of $t \in [0, 1]$ satisfies the Lipschitz property with Lipschitz constant $2\sqrt{2}$.

We now note the simple fact that the Lipschitz property above implies **continuity** of the function $t \mapsto \mathcal{L}_{\mathcal{C}_t}, t \in [0, 1]$. (Indeed, for any $0 < \epsilon \in \mathbb{R}$, we can choose $\delta = \epsilon/(2\sqrt{2})$, universally.) Hence, by the Intermediate Value Theorem (Section 4.2), for any given $0 \leq r \leq 2\pi$, there exists $t \in [0, 1]$ such that $\mathcal{L}_{\mathcal{C}_t} = r$. The claim follows.

Remark The arc length $\mathcal{L}_{\mathcal{C}}$ of a circular arc $\mathcal{C}$ with end-points P_0 and P_1 depends only on the distance $d(P_0, P_1) < 2$. We now make the additional claim that the correspondence that associates with the arc length $\mathcal{L}_{\mathcal{C}}$ of a circular arc $\mathcal{C}$ with end-points P_0 and P_1 and the distance $d(P_0, P_1)$ is **strictly increasing** in the sense that if $\mathcal{C}'$ and $\mathcal{C}''$ are circular arcs, then $\mathcal{L}_{\mathcal{C}'} < \mathcal{L}_{\mathcal{C}''}$ if and only if $d(P_0', P_1') < d(P_0'', P_1'')$ for the corresponding end-points.

Indeed, let $\mathcal{C}$ be a circular arc with end-points P_0 and P_1, $d(P_0, P_1) < 2$, such that $\max(\mathcal{L}_{\mathcal{C}'}, \mathcal{L}_{\mathcal{C}''}) < \mathcal{L}_{\mathcal{C}}$. Then, by the above, there exist $\mathcal{C}_0' \subset \mathcal{C}$ with one end-point at P_0 and congruent to $\mathcal{C}'$ and $\mathcal{C}_0'' \subset \mathcal{C}$ with one end-point at P_0 and congruent to $\mathcal{C}''$. Now the claim is equivalent to monotonicity of the distance $d(Q_t, Q_0)$ in $t \in [0, 1]$. As for this, first note that, as a special case of the general formula derived earlier, we have

$$d(Q_t, Q_0) = 2 - \frac{2 - td^2}{\sqrt{t^2d^2 + td^2 + 1}}, \quad d = d(P_0, P_1).$$

Now, monotonicity follows from the proof of the lemma above.

Exercise

5.6.1. Let $\triangle[A, B, C]$ be a triangle with vertices $A, B, C \in \mathbb{R}^2$ and side lengths $a, b, c \in \mathbb{R}$. Let $0 < r \in \mathbb{R}$ such that $2r < \min(a, b, c)$. Consider the configuration of three circles with centers A, B, C and radius r. What is the shortest length of a band that stretches around the outside of the three circles?

5.7 The Birkhoff Angle Measure

We now turn to the concept of **angle measure** μ for angles in the **Birkhoff Postulate of Angle Measure** (Section 5.1) in our model $\mathbb{R}^2$.

We first define the **angle measure** for angles in our model $\mathbb{R}^2$. Let $\angle AOB$ be an angle formed by the ordered triple (O, A, B), $A \neq O \neq B$. Let $\mathbb{S}_O$ be the unit radius circle with center O.

If (O, A, B) is a positively oriented triple, then we define the angle measure as the arc length $\mu(\angle AOB) = \mathcal{L}_\mathcal{C} \pmod{2\pi}$, where $\mathcal{C} \subset \mathbb{S}_O$ is the circular arc with end-points A_0 and B_0, the points at unit distance from O on the half-line with end-point O and containing A and B.[31]

If A, O, B are collinear, then we define $\mu(\angle AOB) = 0 \pmod{2\pi}$ if O is not between A and B, and $\mu(\angle AOB) = \pi \pmod{2\pi}$ if O is between A and B.

If (O, A, B) is a negatively oriented triple, then we define the angle measure as $\mu(\angle AOB) = -\mathcal{L}_\mathcal{C} \pmod{2\pi}$, where $\mathcal{C} \subset \mathbb{S}_O$ is the circular arc with end-points A_0 and B_0 as above.

Note that the angle measure $\mu(\angle AOB)$ depends only on the half-lines ℓ_0 and ℓ_1 with end-point O containing A and B, respectively. Thus, we can write $\mu(\angle \ell_0 O \ell_1) = \mu(\angle AOB)$.

To derive the Birkhoff Postulate of Angle Measure, for $O \in \mathbb{R}^2$, we need to define the function α_O on the set of all half-lines with end-point O to the real numbers modulo 2π. For O at the origin 0, we let $\alpha_0(\ell) = \mu(\angle \ell_+ 0 \ell)$, where ℓ is any half-line with end-point 0 and ℓ_+ is the positive first axis. With this, we define α_O by translating all the geometric entities from O to the origin 0.

[31] Note that the triple (O, A_0, B_0) is also positively oriented. Recall also that, according to our conventions, $\mathcal{C}$ is the shorter arc length circular arc with end-points A_0 and B_0.

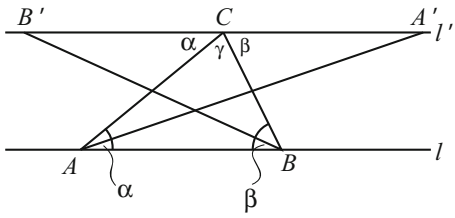


Fig. 5.6 The sum of angles in a triangle is π.

First, according to our last result in the previous section, all real numbers in $[0, 2\pi]$ arise as arc lengths of circular arcs in $\mathbb{S}_O$. This implies that α_O is a **surjective** map onto all real numbers mod 2π.

To show that α_O is an **injective** map, it is enough to derive the characteristic property of the angle measure: For every two half-lines ℓ_0 and ℓ_1 with common end-point O, we have

$$\alpha_O(\ell_1) - \alpha_O(\ell_0) = \mu(\angle \ell_0 O \ell_1) \pmod{2\pi}.$$

This, however, is a direct consequence of the **additivity** of the arc length derived in the last section. The Birkhoff Postulate of Angle Measure follows.

The angle measure being in place, we now derive some basic metric properties of triangles.

First, we claim that the sum of the measures of the interior angles in a triangle $\triangle[A, B, C]$ is equal to π. We introduce the customary notation for triangles as follows. The vertices A, B, C of a triangle are arranged such that the triple (A, B, C) is positively oriented. We denote the side lengths as follows: $a = d(B, C)$, $b = d(C, A)$, $c = d(A, B)$. For the angle measures of the three interior angles corresponding to the vertices A, B, C, we use the first three letters of the Greek alphabet: $\alpha = \mu(\angle BAC)$, $\beta = \mu(\angle CBA)$, and $\gamma = \mu(\angle ACB)$ (see Figure 5.6).

We now claim that

$$\alpha + \beta + \gamma = \pi.$$

To show this, let ℓ be the line through A and B. Let ℓ' be the image of ℓ under the half-turn about the midpoint of the line segment $[A, C]$. Then ℓ' is a line parallel to ℓ through the vertex C. Let $B' \in \ell'$ be the image of B under this half-turn. Then the angles $\angle BAC$ and $\angle B'CA$ are congruent under this half-turn and the side $[A, C]$ is shared by one of the half-lines in both angles. Hence, we have $\alpha = \mu(\angle BAC) = \mu(\angle B'CA)$. Similarly, the half-turn about the midpoint of $[B, C]$ brings ℓ to a line ℓ'' through C parallel to ℓ. By unicity of parallel lines through the same point, we obtain $\ell' = \ell''$. Let A' be the image of A under this second half-turn. As before, the angles $\angle CBA$ and $\angle BCA'$ are congruent and the line segment $[B, C]$ is shared by one of the half-lines in both angles. Hence $\beta = \mu(\angle CBA) = \mu(\angle BCA')$. The three angles $\angle B'CA$, $\angle ACB$, $\angle BCA'$ can be joined (by deleting the shared rays) to

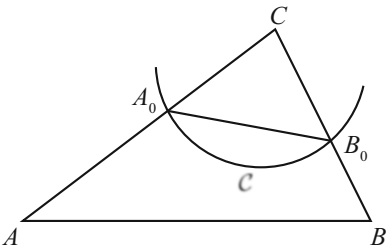


Fig. 5.7 Proof of Birkhoff's Postulate of Similarity.

form a straight angle $\angle B'CA'$ with angle measure π. Finally, since $\alpha = \mu(\angle B'CA)$, $\gamma = \mu(\angle ACB)$, $\beta = \mu(\angle BCA')$, the claim follows.

We are now ready to derive the **Birkhoff Postulate of Similarity**: Given two triangles $\triangle[A, B, C]$ and $\triangle[A', B', C']$ and $0 < k \in \mathbb{R}$ such that $d(C', A') = kd(C, A)$, $d(C', B') = kd(C, B)$, and $\mu(\angle A'C'B') = \mu(\angle ACB)$, then $d(A', B') = kd(A, B)$, $\mu(\angle B'A'C') = \mu(\angle BAC)$, and $\mu(\angle C'B'A') = \mu(\angle CBA)$.

We start with a (non-degenerate) triangle $\triangle[A, B, C]$ with (non-collinear) vertices A, B, C. Recall the fundamental formula

$$d(A_0, B_0)^2 = 2 + \frac{d(A, B)^2 - d(C, A)^2 - d(C, B)^2}{d(C, A)d(C, B)}$$

derived in the previous section (Figure 5.7). Here the point A_0 is on the half-line with end-point C and containing A such that $d(A_0, C) = 1$. Similarly, the point B_0 is on the half-line with end-point C and containing B such that $d(B_0, C) = 1$. Therefore, we have $A_0, B_0 \in \mathbb{S}_C$, the unit circle with center at C.

Recall that the arc length of the circular arc $\mathcal{C} \in \mathbb{S}_C$ with end-points A_0 and B_0 uniquely determines and is uniquely determined by the distance $d(A_0, B_0)$ between its end-points. Since this arc length is, by definition, the angle measure $\mu(\angle ACB)$, the same holds for the angle measure $\mu(\angle ACB)$ and the distance $d(A_0, B_0)$.

Now let $\triangle[A', B', C']$ be another triangle, and assume that, for some $0 < k \in \mathbb{R}$, we have $d(C', A') = kd(C, A)$, $d(C', B') = kd(C, B)$, and $\mu(\angle A'C'B') = \mu(\angle ACB)$. Applying the formula above for $\triangle[A', B', C']$, we have

$$\begin{aligned} d(A'_0, B'_0)^2 &= 2 + \frac{d(A', B')^2 - d(C', A')^2 - d(C', B')^2}{d(C', A')d(C', B')} \\ &= 2 + \frac{d(A', B')^2 - k^2 d(C, A)^2 - k^2 d(C, B)^2}{k^2 d(C, A)d(C, B)}. \end{aligned}$$

By what we said above, the assumption $\mu(\angle A'C'B') = \mu(\angle ACB)$ implies $d(A'_0, B'_0) = d(A_0, B_0)$. Comparing the two formulas above, we obtain $d(A', B') = kd(A, B)$.

Now that all the three respective side lengths of the two triangles $\triangle[A, B, C]$ and $\triangle[A', B', C']$ are a constant ($k > 0$) multiple of each other, we can write down the formula above with the vertices permuted cyclically $(A, B, C) \mapsto (B, C, A) \mapsto (C, A, B)$. The right-hand sides of these formulas are the same for the corresponding triangles. Using the same reasoning as above, we see that the left-hand sides give $\mu(\angle B'A'C') = \mu(\angle BAC)$ and $\mu(\angle C'B'A') = \mu(\angle CBA)$.

The Birkhoff Postulate of Similarity follows.

Remark The following version of Birkhoff's Postulate of Similarity easily follows from the original postulate: Given two triangles $\triangle[A, B, C]$ and $\triangle[A', B', C']$ such that $\mu(\angle A'C'B') = \mu(\angle ACB)$ and $\mu(\angle C'B'A') = \mu(\angle CBA)$ (and consequently $\mu(\angle B'A'C') = \mu(\angle BAC)$), there is $0 < k \in \mathbb{R}$ such that $d(C', A') = kd(C, A)$, $d(C', B') = kd(C, B)$, and $d(A', B') = kd(A, B)$.

The **Pythagorean Theorem** is a direct consequence of the fundamental formula above.

We denote the side lengths of our (non-degenerate) triangle $\triangle[A, B, C]$ as follows: $a = d(B, C)$, $b = d(C, A)$, $c = d(A, B)$. We also let ℓ_0 be the line extension of the side $[C, A]$ and ℓ_1 the line extension of the side $[C, B]$.

The Pythagorean Theorem states that $a^2 + b^2 = c^2$, if and only if ℓ_0 and ℓ_1 are perpendicular.

Let A_0 and B_0 be as in the proof above. The formula above gives

$$d(A_0, B_0)^2 = 2 + \frac{c^2 - a^2 - b^2}{ab}.$$

On the other hand, we showed that ℓ_0 and ℓ_1 are perpendicular if and only if $d(A_0, B_0) = \sqrt{2}$. The Pythagorean Theorem follows.

Remark Clearly, the postulated Cartesian distance formula is actually equivalent to the Pythagorean Theorem.

As an application, we finish this section by solving the classical problem of determining all **right** triangles with **integral** side lengths.

A triple (a, b, c) consisting of natural numbers $a, b, c \in \mathbb{N}$ is called **Pythagorean** if it satisfies the equation

$$a^2 + b^2 = c^2.$$

The name comes from the Pythagorean Theorem as discussed above: If a right triangle has integral side lengths a, b, and c (the hypotenuse), then the triple (a, b, c) is Pythagorean. We will now derive the complete list of Pythagorean triples.

History

All Pythagorean triples have been known since antiquity. The Babylonian clay tablet, Plimpton 322 [32] (c. 1900–1600 BCE, about 1000 years before Pythagoras) contains a list of Pythagorean

[32] The numeral refers to the G.A. Plimpton Collection in Columbia University.

triples which includes (4961,6480,8161). More about the trigonometric interpretation of this tablet will be given in Section 11.2.

It is widely held that the ancients used ropes with equally spaced knots bent over a triangle to survey land and to construct temples. In ancient Egypt the "stretching the cord" ceremony (with invoking Seshat, the goddess of wisdom, measurement, and writing) marked the inauguration of a temple project (see, for example, the middle of the third register of the Palermo Stone, 5th Dynasty, c. 2392–2283 BCE). Note that this method of forming a right angle is still used today in architecture.

For the rope to form a right triangle using a Pythagorean triple (a, b, c), there had to be $a+b+c-1$ knots. Interestingly, an often overlooked fact is that this construction of a right triangle uses the **converse** of the Pythagorean Theorem: If the triple (a, b, c) satisfies the Pythagorean equation above, then the triangle with side lengths a, b, and c has a right angle (opposite to the side of length c).

The first infinite sequence of Pythagorean triples was discovered by the Pythagoreans: $(a, b, c) = (n, (n^2-1)/2, (n^2+1)/2)$, where $3 \leq n \in \mathbb{N}$ is odd. (Note that b and c are consecutive numbers.) Plato discovered another sequence $(a, b, c) = (4n, 4n^2-1, 4n^2+1)$ with $n \in \mathbb{N}$. Finally, Euclid in Book X of the *Elements* derived the full list of Pythagorean triples but attempted no proof that the list was complete. The list of all Pythagorean triples is also expounded in the third century work *Arithmetica* by Diophantus.

The Pythagorean theorem and Pythagorean triples were known in India in the Vedic period. The *Sulba Sutras* (c. 800–c. 500 BCE) contain an elaborate list of rules to construct altars for fire sacrifice, and this involves the Pythagorean theorem. The *Baudhayana Sulba Sutra* has the following sequence of Pythagorean triples: (3, 4, 5), (5, 12, 13), (8, 15, 17), (7, 24, 25), (12, 35, 37).

If (a, b, c) is Pythagorean, then so is (ka, kb, kc) for any natural number $k \in \mathbb{N}$ (since the Pythagorean equation above can be multiplied through by k^2). Conversely, if a, b, and c have a common divisor k, then we can divide through the Pythagorean equation by k^2 and conclude that $(a/k, b/k, c/k)$ is also Pythagorean. The integers in this last triple have no common divisor.

A Pythagorean triple is called **primitive** if the numbers a, b, and c are **relatively prime**; that is, if the only natural number that divides all three of them is 1. We now claim that this is the case if and only if any of the three pairs (a, b), (b, c), or (a, c) is relatively prime. Indeed, if, for example, a and b have a common divisor $k > 1$, then they also have a common **prime** divisor p. Since p divides both a and b, it also divides $a^2 + b^2 = c^2$. Being a prime, p then divides c. Thus, p is a common divisor of a, b, and c. Based on this, from now on, we may restrict ourselves to finding all primitive Pythagorean triples.

Dividing both sides of the Pythagorean equation above by c^2, we obtain

$$\left(\frac{a}{c}\right)^2 + \left(\frac{b}{c}\right)^2 = 1.$$

Equivalently, the positive rational numbers $x = a/c$ and $y = b/c$ satisfy the equation $x^2 + y^2 = 1$. Notice that the pairs (a, c) and (b, c) are relatively prime and this property is equivalent to having **irreducible** fractions a/c and b/c in which all common factors are canceled. In other words, the positive fractions a/c and b/c

satisfying the equation above represent the Pythagorean triple (a, b, c) along with all the multiples (ka, kb, kc) with $k \in \mathbb{N}$.

The equation $x^2 + y^2 = 1$ is the equation of the unit circle $\mathbb{S}$. We call a point $P = (x, y)$ on the plane $\mathbb{R}^2$ **rational** if both x and y are rational numbers. We see that, for a Pythagorean triple (a, b, c), the point $(a/c, b/c)$ is a rational point on the unit circle $\mathbb{S}$. Note that, by construction, both a/c and b/c are positive so that the point $(a/c, b/c)$ is in the interior of the **first quadrant** I of $\mathbb{R}^2$ (that is, the boundary points on the positive first and second axes are excluded).

We now turn the question around and seek to describe all rational points on the open **quarter unit circle** connecting the points $(1, 0)$ and $(0, 1)$ (where openness means that the end-points $(1, 0)$ and $(0, 1)$ are excluded).

We consider a point (rational or not) on this quarter-circle as the **second** intersection point of $\mathbb{S}$ with a secant line that contains $(0, 1)$ (as the first intersection point). The general equation of a line[33] $ax - by = c$ through $(0, 1)$ reduces to $ax - by = -b$.

In Section 5.5, we determined the second intersection point of a secant line and the unit circle $\mathbb{S}$ with first common point $P_0 = (x_0, y_0)$ in general. In our case ($x_0 = 0$ and $y_0 = 1$), this second intersection point specializes to

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

This point is contained in the open first quadrant if and only if $ab < 0$ and $a^2 - b^2 < 0$. In terms of the slope $m = a/b$, the equation of the line can be rewritten as $y = mx + 1$. The second intersection point is

$$\left(-\frac{2m}{1 + m^2}, \frac{1 - m^2}{1 + m^2}\right),$$

where the slope m is contained in the open interval $(-1, 0)$. (Zero slope corresponds to the horizontal line across $(0, 1)$ tangent to $\mathbb{S}$, and the line with slope -1 intersects $\mathbb{S}$ at $(1, 0)$.)

Now the crux is that this point is a rational point if and only if the slope m is rational. Thus, for all values $m \in (-1, 0) \cap \mathbb{Q}$, we obtain all rational points on the open quarter-circle. Letting $m = -a/b$, $a, b \in \mathbb{N}$, we obtain all Pythagorean triples as

$$\left(2ab, b^2 - a^2, a^2 + b^2\right), \quad a < b, \ a, b \in \mathbb{N}.$$

The following table shows a few values:

[33] We use here the letters a, b, and c for the coefficients in the typical equation of a line, not to be confused with the same letters occurring in the Pythagorean triples above.

a	b	$2ab$	b^2-a^2	b^2+a^2
1	2	4	3	5
1	3	6	8	10
1	4	8	15	17
1	5	10	24	26
1	6	12	35	37
2	3	12	5	13
2	4	16	12	20
2	5	20	21	29
3	4	24	7	25
3	5	30	16	34
4	5	40	9	41
...	...	...	...	...
40	81	6480	4961	8161

We included the five triplets from *Baudhayana Sulba Sutra*. Note also the last line from the Babylonian tablet.

Example 5.7.1 Find all $n < 200$, $n \in \mathbb{N}$, such that $n^2+(n+1)^2$ is a perfect square.[34]

The problem is equivalent to finding all Pythagorean triples $(n, n+1, m)$, where $n < 200$ and $m \in \mathbb{N}$.

By the above, we have two cases:

I. $n = 2ab, n+1 = b^2-a^2, m = a^2+b^2, a < b, a, b \in \mathbb{N}$.
II. $n = b^2-a^2, n+1 = 2ab, m = a^2+b^2, a < b, a, b \in \mathbb{N}$.

In Case I, we have $2ab+1 = b^2-a^2$, or equivalently, $b^2-2ab-(a^2+1) = 0$. Solving this as a quadratic equation in b in terms of a, we obtain $b = a \pm \sqrt{2a^2+1}$. Only the positive square root is realized. In addition, $2a^2+1$ must be a perfect square. Since $a^2 < ab < 100$, we have $a < 10$. The cases $a = 1, 2, \ldots, 9$ give only $a = 2$ as a solution. Hence $b = 5$, and $n = 2ab = 20$. This gives the Pythagorean triple $(20, 21, 29)$.

Case II is analogous. We have $b^2 - 2ab - (a^2 - 1) = 0$ so that $b = a \pm \sqrt{2a^2-1}$. This gives $a = 1, 5$. The corresponding Pythagorean triples are $(3, 4, 5)$ and $(5, 12, 13)$.

Exercises

5.7.1. Let R be a rectangle with vertices A, B, C, D with the right angle at the vertex A **trisected** by two half-lines ℓ' and ℓ''. Assume that these half-lines meet the opposite sides at interior points: $E' = \ell' \cap [B, C]$ and $E'' =$

[34] This was a problem in the Nordic Mathematical Contests, 1998. Note, however, that the general solution without the upper bound is contained in Sierpiński, W., *Elementary Theory of Numbers*, 2nd ed. North Holland, 1985.

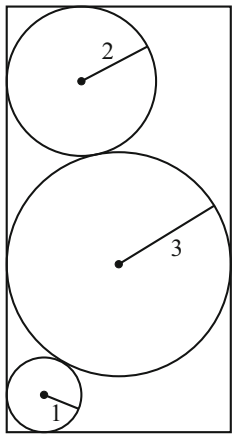


Fig. 5.8 Illustration to Exercise 5.7.3.

$\ell'' \cap \in [C, D]$. Express the side lengths $d(A, B)$ and $d(B, C)$ in terms of the distances $d(B, E')$ and $d(C, E'')$.

5.7.2. Two congruent (but distinct) rectangles overlap and share the longer (common) diagonal. If the side lengths of the rectangles are $0 < b < a \in \mathbb{R}$, then show that the overlap is a rhombus, and find its side length.

5.7.3. Three circles stacked up fit snugly in a rectangle (see Figure 5.8). The bottom, middle, and top circles have radii 1, 3 and 2, respectively. Each circle touches the left-side of the rectangle. The middle (largest) circle touches both vertical sides of the rectangle. Calculate the height of the rectangle.

5.7.4. For $m \in \mathbb{N}_0$, let $S_m = \mathbb{S}_{(0,2m),1}$. (The unit circles S_m, $m \in \mathbb{N}_0$, are lined up along the first axis.) Fix $2 \leq n \in \mathbb{N}$, and let ℓ be the line through the origin $(0, 0)$ and tangent to S_n. Let A_n, $B_n \in S_1$ be the intersection points of ℓ secant to S_1. Calculate $d(A_n, B_n)$.

5.7.5. The Fibonacci numbers can be used to construct Pythagorean triples. Show that, for $n \geq 3$, the triple $(2F_n F_{n-1}, F_n^2 - F_{n-1}^2, F_{2n-1})$ is Pythagorean.

5.7.6. Show that two angles with perpendicular sides are either equal or supplementary (that is, they together make a straight angle).

5.7.7. Show that a right triangle has side lengths that form three consecutive terms in an arithmetic sequence if and only if the side lengths are $3d$, $4d$, and $5d$, where d is the difference of the arithmetic sequence.

5.7.8. A right triangle has the property that the length of the hypothenuse is twice the length of the altitude from the vertex corresponding to the right angle. Show that the triangle is isosceles.

5.8 The Principle of Shortest Distance*

Since we can measure the distance between points in our model of the Cartesian plane $\mathbb{R}^2$, it is natural to ask the question: What is the **shortest** path between two (distinct) points P_0 and P_1 on the plane?

First, a **polygonal path** connecting P_0 and P_1 is an (open) polygon with end-points P_0 and P_1; that is, a union $\bigcup_{i=1}^{n}[Q_{i-1}, Q_i]$ such that $Q_0 = P_0$ and $Q_n = P_1$. The length of a polygonal path is defined as $\sum_{i=1}^{n} d(Q_{i-1}, Q_i)$, the sum of the lengths of the participating line segments in the polygonal path.

We now claim that the **shortest** polygonal path connecting P_0 and P_1 is the straight-line segment $[P_0, P_1]$ between P_0 and P_1, that is, in which $Q_i \in [P_0, P_1]$ for all $i = 0, 1, \ldots, n$. We show this by Peano's Principle of Induction with respect to $n \in \mathbb{N}$.

For $n = 1$, there is nothing to prove. Assume that the statement holds for $n \in \mathbb{N}$, and let $\sum_{i=1}^{n+1} d(Q_{i-1}, Q_i)$ be a polygonal path with end-points P_0 and P_1 consisting of $n + 1$ line segments. By the strict triangle inequality, for (any) given $i = 1, \ldots, n$, we have $d(Q_{i-1}, Q_{i+1}) \le d(Q_{i-1}, Q_i) + d(Q_i, Q_{i+1})$ with equality if and only if $Q_i \in [Q_{i-1}, Q_{i+1}]$. Since our path is the shortest, it follows that $Q_i \in [Q_{i-1}, Q_{i+1}]$. We now replace the two line segments $[Q_{i-1}, Q_i]$ and $[Q_i, Q_{i+1}]$ by the single line segment $[Q_{i-1}, Q_{i+1}]$ without altering the overall length of the polygonal path. The new polygonal path consists of n line segments, so that the induction hypothesis applies. We obtain $Q_1, \ldots, Q_{i-1}, Q_{i+1}, \ldots, Q_n \in [P_0, P_1]$. Since $Q_i \in [Q_{i-1}, Q_{i+1}]$, we also have $Q_i \in [P_0, P_1]$. The claim follows.

We now extend this to more general paths. We say that a subset $\mathcal{C} \subset \mathbb{R}^2$ with two specified points $P_0, P_1 \in \mathcal{C}$, $P_0 \ne P_1$, is a (simple) **rectifiable curve** if there exists a one-to-one[35] **Lipschitz** map $Q : [0, 1] \to \mathbb{R}^2$ such that the range of Q is $\mathcal{C}$ and $Q(0) = P_0$ and $Q(1) = P_1$. The map Q is usually called a **parametrization** of $\mathcal{C}$. The Lipschitz property means that, for some (Lipschitz) constant $L \in \mathbb{R}$, we have

$$d(Q(t), Q(t')) \le L|t - t'|, \quad t, t' \in [0, 1].$$

With this, we define the arc length of $\mathcal{C}$ by

$$\mathcal{L}_{\mathcal{C}} = \sup \left\{ \sum_{i=1}^{n} d(Q(t_{i-1}), Q(t_i)) \,\middle|\, (t_0, t_1, \ldots, t_{n-1}, t_n) \in \Pi \right\},$$

where the supremum is over the set Π of all partitions

$$(t_0, t_1, \ldots, t_{n-1}, t_n) \in \Pi, \quad 0 = t_0 < t_1 < \cdots < t_{n-1} < t_n = 1, \quad n \in \mathbb{N}.$$

We need to show that the arc length $\mathcal{L}_{\mathcal{C}}$ is a finite (real) number, or equivalently, the lengths of polygonal paths of $\mathcal{C}$ (in the supremum above) induced by all partitions of [0, 1] are bounded above. This is guaranteed by the Lipschitz property, since, for any partition $(t_0, t_1, \ldots, t_{n-1}, t_n) \in \Pi$, we have

[35] The property of being simple, that is, one-to-one, excludes "self-intersections." As we consider here only open curves and minimize the arc length, imposing this does not restrict the generality.

$$\sum_{i=1}^{n} d(Q(t_{i-1}), Q(t_i)) \le L \sum_{i=1}^{n} (t_i - t_{i-1}) = L.$$

Thus, the arc length $\mathcal{L}_{\mathcal{C}}$ exists.

The next example shows that some parametrization of a rectifiable curve may not be Lipschitz.[36]

Example 5.8.1 Consider the set $\mathcal{C} = \{(x, \sqrt{x}) \,|\, x \in [0, 1]\} \subset \mathbb{R}^2$ with specified points $P_0 = (0, 0)$ and $P_1 = (1, 1)$.

First, we let $Q : [0, 1] \to \mathbb{R}^2$ be the map defined by $Q(t) = (t, \sqrt{t})$, $t \in [0, 1]$. Clearly, the range of Q is $\mathcal{C}$, and $Q(0) = P_0$ and $Q(1) = P_1$.

We claim that Q does not have the Lipschitz property.

Assuming the contrary, there exists a Lipschitz constant $L \in \mathbb{R}$ such that

$$d(Q(t), Q(t')) = \sqrt{(t - t')^2 + (\sqrt{t} - \sqrt{t'})^2} \le L|t - t'|, \quad t, t' \in [0, 1].$$

Squaring, we see that $L \ge 1$, and we have $|\sqrt{t} - \sqrt{t'}| \le \sqrt{L^2 - 1}|t - t'|$, $t, t' \in [0, 1]$. Setting $t' = 0$, this gives $\sqrt{t} \le Mt$, $t \in [0, 1]$, where $M = \sqrt{L^2 - 1}$. This, however, is a contradiction since $1 \le M\sqrt{t}$ cannot hold for $0 < t < 1$, $t \in \mathbb{R}$, small enough.

Second, we let $Q' : [0, 1] \to \mathbb{R}^2$ be the map defined by $Q'(t) = (t^2, t)$, $t \in [0, 1]$. As before, the range of Q' is $\mathcal{C}$, and $Q'(0) = P_0$ and $Q'(1) = P_1$.

We claim that Q' is a Lipschitz map with Lipshitz constant $L = \sqrt{5}$; that is, we have

$$d(Q'(t), Q'(t')) = \sqrt{(t^2 - t'^2)^2 + (t - t')^2} \le \sqrt{5}|t - t'|, \quad t, t' \in [0, 1].$$

Squaring, and simplifying, we obtain $(t^2 - t'^2)^2 \le 4(t - t')^2$, $t, t' \in [0, 1]$. Factoring and simplifying again, this reduces to $|t + t'| \le 2$, $t, t' \in [0, 1]$. This obviously holds. The Lipschitz property holds as claimed.

Returning to the main line, we need to show unicity of the arc length; that is, the definition of the arc length $\mathcal{L}_{\mathcal{C}}$ of a rectifiable curve $\mathcal{C}$ does not depend on the parametrization (as long as it is Lipschitz). Let $\mathcal{C}$ be a rectifiable curve with specified points $P_0, P_1 \in \mathcal{C}$. If $Q, Q' : [0, 1] \to \mathbb{R}^2$ are both one-to-one Lipschitz maps with common range $\mathcal{C}$ and $Q(0) = Q'(0) = P_0$ and $Q'(1) = Q'(1) = P_1$, then we claim that the arc lengths defined by Q and Q' are equal.

First, for $t \in [0, 1]$, we let $s(t) \in [0, 1]$ be the unique real number such that $Q(t) = Q'(s(t))$. This defines a function $s : [0, 1] \to [0, 1]$, $s(0) = 0$, $s(1) = 1$, which is clearly bijective, that is, one-to-one and onto.

[36] In somewhat more generality, a curve on the plane is called rectifiable if it has bounded variation, that is, if the supremum above is finite. It can be shown that for a curve of bounded variation, there is always a Lipschitz parametrization as above.

Lemma *The function s is strictly increasing.*

Proof Since s is one-to-one, it is enough to show that it is continuous. (See Corollary to the Intermediate Value Theorem in Section 4.2.) We show sequential continuity of s. Let $(t_n)_{n\in\mathbb{N}}$ be a sequence in $[0, 1]$ such that $\lim_{n\to\infty} t_n = t_0$. We need to prove that the corresponding sequence $(s(t_n))_{n\in\mathbb{N}}$ is convergent and has limit $s(t_0)$. First, let $\liminf_{n\to\infty} s(t_n) = \underline{L}$. Choose a convergent subsequence $(s(t_{n_k}))_{k\in\mathbb{N}}$ such that $\lim_{k\to\infty} s(t_{n_k}) = \underline{L}$. (The existence of this subsequence follows easily from the definition of the limit inferior.) Since Q' is continuous (as it is Lipschitz), we have $\lim_{k\to\infty} Q'(s(t_{n_k})) = Q'(\underline{L})$. Since $Q(t) = Q'(s(t))$, $t \in [0, 1]$, this gives $\lim_{k\to\infty} Q(t_{n_k}) = Q'(\underline{L})$. On the other hand, by continuity of Q, this limit is $Q(t_0)$. Thus, we have $Q'(\underline{L}) = Q(t_0)$. Second, let $\limsup_{n\to\infty} s(t_n) = \overline{L}$. Repeating the previous argument (almost verbatim), we obtain $Q'(\overline{L}) = Q(t_0)$. Hence $Q'(\underline{L}) = Q'(\overline{L})$. Since Q' is one-to-one, we obtain $\underline{L} = \overline{L} (= L$, say), and we conclude that the sequence $(s(t_n))_{n\in\mathbb{N}}$ is convergent to this common value L. Finally, we have $Q'(L) = Q(t_0) = Q'(s(t_0))$, and, once again since Q' is one-to-one, we arrive at $\lim_{n\to\infty} s(t_n) = L = s(t_0)$. The lemma follows.

We now return to our rectifiable curve $\mathcal{C} \subset \mathbb{R}^2$. Let $\mathcal{L}_{\mathcal{C}}$ and $\mathcal{L}'_{\mathcal{C}}$ denote the arc length of $\mathcal{C}$ with respect to Q and Q', respectively. We claim that $\mathcal{L}_{\mathcal{C}} = \mathcal{L}'_{\mathcal{C}}$.

Let $0 < \epsilon \in \mathbb{R}$. The interval $[0, 1]$ has a partition $(t_0, t_1, \ldots, t_n) \in \Pi$, $t_0 = 0$ and $t_n = 1$, such that, for the associated polygonal path $\bigcup_{i=1}^{n} [Q(t_{i-1}), Q(t_i)]$ with $Q(0) = P_0$ and $Q(1) = P_1$, we have

$$\mathcal{L}_{\mathcal{C}} - \epsilon < \sum_{i=1}^{n} d(Q(t_{i-1}), Q(t_i)).$$

For $i = 1, \ldots, n$, we let $s_i = s(t_i) \in [0, 1]$, so that $Q(t_i) = Q'(s_i)$. Now the crux is that, according to the lemma above, the finite sequence $(s_0, s_1, \ldots, s_n)$ is monotonic with $s_0 = 0$ and $s_n = 1$, and thereby it forms a partition of $[0, 1]$. Hence, we have

$$\sum_{i=1}^{n} d(Q'(s_{i-1}), Q'(s_i)) \le \mathcal{L}'_{\mathcal{C}}.$$

Putting these together, we obtain $\mathcal{L}_{\mathcal{C}} - \epsilon < \mathcal{L}'_{\mathcal{C}}$. Since ϵ was arbitrary, we arrive at $\mathcal{L}_{\mathcal{C}} \le \mathcal{L}'_{\mathcal{C}}$. Reversing the roles of Q and Q', we obtain $\mathcal{L}'_{\mathcal{C}} = \mathcal{L}_{\mathcal{C}}$ as claimed. Independence of the arc length from parametrization follows.

We are now able to show that the shortest path between two points is the straight-line segment. Let $\mathcal{C}$ be a rectifiable path with specified points $P_0, P_1 \in \mathcal{C}$, and let $Q : [0, 1] \to \mathbb{R}^2$ be a Lipschitz map with range $\mathcal{C}$ and $Q(0) = P_0$ and $Q(1) = P_1$. Given any polygonal path $\bigcup_{i=1}^{n} [Q_{i-1}, Q_i]$, $Q_0 = P_0$ and $Q_n = P_1$, by the discussion above, we have

$$d(P_0, P_1) \leq \sum_{i=1}^{n} d(Q(t_{i-1}), Q(t_i)) \leq \mathcal{L}_{\mathcal{C}}.$$

If $\mathcal{L}_{\mathcal{C}}$ is minimal, then equalities hold. It follows that the polygonal path is the line segment $[P_0, P_1]$. Since this holds for all polygonal paths contributing to the supremum that defines the arc length $\mathcal{L}_{\mathcal{C}}$, we obtain that $\mathcal{C} = [P_0, P_1]$. Thus, the shortest path between two points is the straight-line segment.

As an application, we derive the **Principle of Least Distance**: A light ray reflected in a mirror has the same **angle of incidence** as the **angle of reflection**. (The angle of incidence is the angle that a ray makes with the perpendicular line to the surface at the point of incidence, and the angle of reflection is the angle made by a reflected ray with the same perpendicular line.)

The connection of this to the discussion on arc length above is physics: The light ray always travels along a path of shortest length.

To be specific, let A and B be two points on the **same** side of a line ℓ in the plane $\mathbb{R}^2$. The line ℓ represents the mirror; the light ray is emitted at A, reflected in ℓ, and detected at B. From A, the ray reaches ℓ in the shortest possible path, a straight-line segment, and after bouncing off from ℓ at a point C, once again, it reaches B in a straight-line segment. Thus, we can now ask the more precise question:

At what point C of ℓ is the sum of distances $d(A, C) + d(C, B)$ minimal?[37]

In what follows, we will describe a simple solution that employs the concept of **reflection** in a line. Given a line ℓ in $\mathbb{R}^2$, we define the reflection $\rho_\ell : \mathbb{R}^2 \to \mathbb{R}^2$ in ℓ as follows: Let $P \in \mathbb{R}^2$, and consider the line ℓ' through P perpendicular to ℓ. Let $Q \in \ell \cap \ell'$ be the intersection point of these two lines. Now, let $P' \in \ell'$ be the unique point such that Q is the midpoint of P and P'. We define $\rho_\ell(P) = P'$. (Note that $P = P'$ if and only if $P \in \ell$. In other words, the points on the line ℓ are the fixed points of ρ_ℓ.)

We claim that ρ_ℓ is distance preserving; that is, we have $d(\rho_\ell(A), \rho_\ell(B)) = d(A, B)$ for all $A, B \in \mathbb{R}^2$. For simplicity, we let $A' = \rho_\ell(A)$ and $B' = \rho_\ell(B)$. We also let $Q = (1/2)(A + A') \in \ell$ and $R = (1/2)(B + B') \in \ell$. We may assume that $A \notin \ell$ and $B \notin \ell$ since otherwise the proof is much simpler.

The triangles $\triangle[A, Q, R]$ and $\triangle[A', R, Q]$ are congruent since they have a common side $[Q, R]$, right angles at the vertex Q, and congruent sides $[Q, A]$ and $[Q, A']$, that is, we have $d(Q, A) = d(Q, A')$. By the Birkhoff Postulate of Similarity, we have $d(R, A) = d(R, A')$, and the angles $\angle ARQ$ and $\angle QRA'$ at the common vertex R are congruent. Now, consider the triangles $\triangle[A, R, B]$ and $\triangle[A', B', R]$. Their angles $\angle BRA$ and $\angle A'RB'$ at the common vertex R are congruent since $\mu(\angle ARQ) + \mu(\angle BRA) = \pi/2$ and $\mu(\angle A'RB') + \mu(\angle QRA') = \pi/2$. In addition, by the definition of ρ_ℓ, we have $d(R, B) = d(R, B')$, and, as noted above, $d(R, A) = d(R, A')$. Thus, by the Birkhoff Postulate of Similarity, we

[37] The Principle of Least Distance asserts that at C, the angle of incidence and the angle of reflection are equal. This determines the point C uniquely. This principle is usually proved in calculus using a minimization technique. In reality, it is much simpler.

obtain $d(A, B) = d(A', B')$. Thus, ρ_ℓ preserves distances. Note that, ρ_ℓ changes the angle measure to the opposite sign.

Now, we return to the Principle of Least Distance. Let $B' = \rho_\ell(B)$. Since reflection preserves distances, we have $d(C, B) = d(C, B')$ so that $d(A, C) + d(C, B) = d(A, C) + d(C, B')$. As we have shown above, the shortest path between the points A and B' is the straight-line segment. Hence the light ray bounces off at the point C, the intersection of ℓ and the line segment $[A, B']$. At C the opposite angles between the line perpendicular to ℓ and the line segment connecting A and B' are equal. One of the angles is the angle of incidence of the light ray. The other angle is equal to the angle of reflection of the light ray since reflection in a line preserves angles. The Principle of Least Distance follows.

Exercises

5.8.1. Let $\angle \ell' O \ell''$ be an angle in $\mathbb{R}^2$ formed by two half-lines ℓ' and ℓ'' meeting at O, and assume that it is acute; that is, the angle measure $\mu(\angle \ell' O \ell'') \in (0, \pi/2)$. Let A be a point in the corresponding (open) acute angular sector. Find $B \in \ell'$ and $C \in \ell''$ such that the (possibly degenerate) triangle $\triangle[A, B, C]$ has the least perimeter.

5.8.2. Use the proof of the Lemma following Example 5.8.1 to show that the inverse of a continuous bijection $f : I \to J$ between closed intervals I and J is continuous.

5.9 π According to Archimedes*

Attempts to approximate π, the ratio of the circumference and the diameter of a circle, can be found in virtually all ancient societies.[38] Archimedes devised the first rigorous (inductive) procedure to obtain rational approximations of π.

His method started with two regular hexagons, one **inscribed** and the other **circumscribed** about the unit circle $\mathbb{S}_O$ with center at a point O. The induction consists of systematically doubling the sides (while keeping the resulting polygons inscribed and circumscribed). Archimedes stopped at the 96-sided polygons. Approximating at each stage the various radical expressions by ingeniously chosen fractions, he finally arrived at the estimate

$$3\frac{10}{71} < \pi < 3\frac{1}{7}.$$

[38] For a short history of π, see the author's *Glimpses of Algebra and Geometry*, 2nd ed. Springer, New York, 2002.

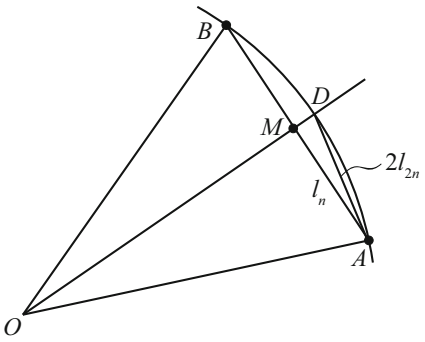


Fig. 5.9 Archimedes' duplication; inscribed polygon.

In this section we discuss Archimedes' method focusing on the relevant radical expressions rather than their approximating fractions.

Let $\mathcal{P}_n$ and $\mathcal{Q}_n$, $n \geq 3$, be regular n-sided polygon inscribed and, respectively, circumscribed about $\mathbb{S}_O$. The vertices of $\mathcal{P}_n$ lie (equidistantly spaced) on $\mathbb{S}_O$, and the midpoints of the sides of $\mathcal{Q}_n$ are (equidistantly spaced) points of tangency of the sides to $\mathbb{S}_O$. Let l_n and L_n denote **half** of the side length of $\mathcal{P}_n$and $\mathcal{Q}_n$, respectively.

Since $\mathcal{P}_n$ and $\mathcal{Q}_n$ have n sides, we have $nl_n < \pi < nL_n$, $n \geq 3$.

Archimedes established the following inductive formulas:

$$l_{2n} = \sqrt{\frac{1-\sqrt{1-l_n^2}}{2}} \quad \text{and} \quad L_{2n} = \sqrt{\frac{1}{L_n^2}+1} - \frac{1}{L_n}.$$

To show the first, let $[A, B]$ be a side of $\mathcal{P}_n$ and consider the triangle $\triangle[A, B, O]$, where O is the center of $\mathbb{S}_O$ (see Figure 5.9). Let the bisector of the angle $\angle AOB$ of angle measure $2\pi/n$ intersect $[A, B]$ at the midpoint M and further the unit circle at the point D.

Since $\triangle[O, A, M]$ is a right triangle with right angle at M and $d(A, M) = l_n$, the Pythagorean Theorem gives $d(O, M) = \sqrt{1-l_n^2}$. Since $1 = d(O, D) = d(O, M) + d(M, D)$, the triangle $\triangle[A, D, M]$ is also a right triangle with right angle at M, and $d(A, D) = 2l_{2n}$, the Pythagorean Theorem once again gives

$$\left(1-\sqrt{1-l_n^2}\right)^2 + l_n^2 = 4l_{2n}^2.$$

Expanding and simplifying, we obtain $1 - 2l_{2n}^2 = \sqrt{1-l_n^2}$. This gives

$$l_{2n}^2 = \frac{1-\sqrt{1-l_n^2}}{2}.$$

Taking square roots on both sides, our first formula follows.

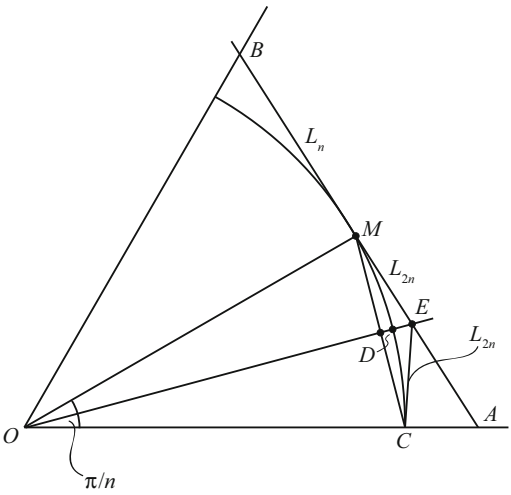


Fig. 5.10 Archimedes' duplication; circumscribed polygon.

For the second relation, let $[A, B]$ be a side of $\mathcal{Q}_n$ with midpoint M, the point of tangency of this side with $\mathbb{S}_O$ (for the notations here and below, see Figure 5.10). Let C be the intersection of the radial line segment $[O, A]$ with $\mathbb{S}_O$, and D, resp. E, the intersections of the angular bisector of the angle $\angle AOM$ of measure π/n with the circle, resp. the line segment $[A, M]$. We have $d(A, M) = d(B, M) = L_n$ and $d(C, E) = d(M, E) = L_{2n}$.

The hypotenuse $[O, A]$ of the right triangle $\triangle[O, A, M]$ has length $\sqrt{1 + L_n^2}$ so that the length of $[A, C]$ is $\sqrt{1 + L_n^2} - 1$. Finally, the triangles $\triangle[O, A, M]$ and $\triangle[A, C, E]$ are similar, so that we have

$$L_n = \frac{\sqrt{1 + L_n^2} - 1}{L_{2n}}.$$

Rearranging, our second formula follows.

Since the hexagon is made up by six equilateral triangles, a simple geometric consideration gives $2l_6 = 1$ and $L_6 = 1/\sqrt{3}$. We now iterate the relations above (starting with $n = 6$). It is somewhat easier to iterate the first on the doubles

$$2l_{2n} = \sqrt{2 - \sqrt{4 - (2l_n)^2}}.$$

For half of the perimeters (nl_n), starting with $6l_6 = 3$, a simple computation gives

$$12l_{12} = 6\sqrt{2 - \sqrt{3}} \approx 3.1058285412$$

$$24l_{24} = 12\sqrt{2 - \sqrt{2 + \sqrt{3}}} \approx 3.1326286132$$

$$48l_{48} = 24\sqrt{2 - \sqrt{2 + \sqrt{2 + \sqrt{3}}}} \approx 3.1393502030$$

$$96l_{96} = 48\sqrt{2 - \sqrt{2 + \sqrt{2 + \sqrt{2 + \sqrt{3}}}}} \approx 3.1410319508$$

These are lower bounds for π with increasing accuracy.

For the circumscribed polygons, using repeated elimination of the square roots in the denominators (by the difference of squares identity), starting with $L_6 = 1/\sqrt{3}$, we have

$$L_{12} = \sqrt{3+1} - \sqrt{3} = 2 - \sqrt{3}$$

$$L_{24} = \sqrt{\frac{1}{(2-\sqrt{3})^2} + 1} - \frac{1}{2-\sqrt{3}} = \sqrt{(2+\sqrt{3})^2 + 1} - (2+\sqrt{3})$$

$$= 2\sqrt{2+\sqrt{3}} - (2+\sqrt{3}) = (\sqrt{6}+\sqrt{2}) - (2+\sqrt{3}) = (\sqrt{3}-\sqrt{2})(\sqrt{2}-1)$$

$$L_{48} = \sqrt{\frac{1}{(\sqrt{3}-\sqrt{2})^2(\sqrt{2}-1)^2} + 1} - \frac{1}{(\sqrt{3}-\sqrt{2})(\sqrt{2}-1)}$$

$$= \sqrt{(\sqrt{3}+\sqrt{2})^2(\sqrt{2}+1)^2 + 1} - (\sqrt{3}+\sqrt{2})(\sqrt{2}+1)$$

$$L_{96} = \sqrt{\left(\sqrt{(\sqrt{3}+\sqrt{2})^2(\sqrt{2}+1)^2 + 1} + (\sqrt{3}+\sqrt{2})(\sqrt{2}+1)\right)^2 + 1}$$

$$- \left(\sqrt{(\sqrt{3}+\sqrt{2})^2(\sqrt{2}+1)^2 + 1} + (\sqrt{3}+\sqrt{2})(\sqrt{2}+1)\right).$$

For half of the perimeters (nL_n), starting with $6L_6 = 2\sqrt{3} \approx 3.4641016151$, we obtain

$$12L_{12} = 12(2 - \sqrt{3}) \approx 3.2153903091$$

$$24L_{24} = 24(\sqrt{3} - \sqrt{2})(\sqrt{2} - 1) \approx 3.1596599420$$

$$48l_{48} \approx 3.1460862151$$

$$96l_{96} \approx 3.1427145996.$$

These are upper bounds for π with increasing accuracy.

Using the inductive formulas above, we see that l_n and L_n can be written as nested square roots; therefore, as lengths of line segments, they are constructible by straightedge and compass. The sequence $(6l_6, 12l_{12}, 24l_{24}, 48l_{48}, \ldots)$ is strictly increasing and the sequence $(6L_6, 12L_{12}, 24L_{24}, 48L_{48}, \ldots)$ is strictly decreasing. Moreover, the (positive) differences $6(L_6 - l_6), 12(L_{12} - l_{12}), 24(L_{24} - l_{24}), 48(L_{48} - l_{48}), \ldots$ decrease to zero. By the Monotone Convergence Theorem, there is a unique real number between these two sequences. This is the number π.

Example 5.9.1 Is $5\pi^2 - 31\pi + 48$ positive or negative?

We have $5\pi^2 - 31\pi + 48 = (\pi - 3)(5\pi - 16) = 5(\pi - 3)(\pi - 32/10) < 0$.

Exercise

5.9.1. Using a straightedge and a compass, construct a regular octagon (8 sides) and a regular dodecagon (12 sides).

Chapter 6
Polynomial Expressions

"Of course I had progressed far beyond Vulgar Fractions and the Decimal System. We were arrived in an 'Alice-in-Wonderland' world, at the portals of which stood 'A Quadratic Equation.' This with a strange grimace pointed the way to the Theory of Indices, which again handed on the intruder to the full rigors of the Binomial Theorem."
in My Early Life by Sir Winston Churchill (1874–1965)

In this chapter we begin our study of the simplest mathematical expressions, the **polynomials**. We start with the simplest case: The binomial formula. It is presented here with full arithmetic and historical details, with many identities, and along with its principal, mostly combinatorial, applications including Bernoulli's derangements. The Division Algorithm for Integers discussed in Section 1.3 leads directly to its polynomial analogue, the Division Algorithm for Polynomials, or polynomial long division, and its offspring, the synthetic division. They reveal a great deal of information about the behavior of polynomials. We accompany these with many examples of (sometimes highly technical) polynomial factorizations. These exhibit beautiful interplays with divisibility properties of integers. Turning to a somewhat more advanced level, we derive the fundamental theorem on symmetric polynomials (leading to a very simple but non-standard derivation of the quadratic formula), the Viète relations, and the Newton–Girard formulas for power sums. Amongst the many applications of the Viète relations, we give an arithmetic proof of the allegedly most challenging problem ever posted on the International Mathematical Olympiad, in 1988. Finally, we briefly return to the Cauchy–Schwarz inequality, introduced in Section 5.3, in a multivariate setting accompanied by the Chebyshev sum inequality.

G. Toth, *Elements of Mathematics*, Undergraduate Texts in Mathematics,
https://doi.org/10.1007/978-3-030-75051-0_6

6.1 Polynomials

A **polynomial** is constructed from an indeterminate (variable, parameter, etc.) x (or t, u, etc.) and (real) numbers under the operations of **addition** and **multiplication**. The indeterminate x follows the usual rules of arithmetic, including exponentiation:

$$x^n = \overbrace{x \cdot x \cdot x \cdots x}^{n \text{ factors}}, \quad n \in \mathbb{N}.$$

Exponentiation is defined inductively by setting $x^0 = 1$, and $x^{n+1} = x \cdot x^n, n \in \mathbb{N}_0$.

A polynomial with x as an indeterminate is usually denoted by $p(x)$.

Examples for polynomials (in the indeterminate x) are

$$ax^2+bx+c,\ a, b, c \in \mathbb{R}; \quad \left(1+\frac{x}{365}\right)^{365}; \quad 1+x+\frac{x^2}{2!}+\frac{x^3}{3!}+\cdots+\frac{x^n}{n!},\ n \in \mathbb{N}.$$

History

In his work *La géometrie*, Descartes made a widespread use of letters to denote numbers (from the beginning of the alphabet such as a, b, c, etc.), and indeterminates (from the end of the alphabet such as x, y, z, t, u, v, etc). He used first superscripts to denote exponents.

More generally, when the role of the indeterminate[1] is played by a mathematical **entity** E (such as another expression, function, etc.) then we arrive at the concept of **polynomial expression**. Emphasizing the role of the entity, it is also called a **polynomial in** E.

Examples for polynomial expressions are

$$\sqrt{2}^5 + \sqrt{2} + 1; \quad \frac{d_1 d_2 \ldots d_k}{10^k}\left(1 + \frac{1}{10^k} + \left(\frac{1}{10^k}\right)^2 + \cdots + \left(\frac{1}{10^k}\right)^n\right).$$

The first is a polynomial expression in $\sqrt{2}$, and the second is a polynomial expression in $1/10^k$.

These definitions can be naturally extended to polynomials in several indeterminates $x, y, z \ldots$, and $x_1, x_2, x_3, \ldots, x_n, n \in \mathbb{N}$, etc., and to polynomial expressions in finitely many entities $E_1, E_2, \ldots, E_n$. In these cases the respective polynomials are usually denoted by $p(x, y)$, $p(x, y, z)$, $p(x_1, x_2, \ldots, x_n)$, etc.

Polynomials can be **evaluated** on numbers by substitution; that is, by performing the operations that the polynomial is made up on numbers instead of indeterminates. A polynomial $p(x)$ evaluated on a specific number c is denoted by $p(c)$, a polynomial $p(x, y)$ evaluated on (a, b) is denoted by $p(a, b)$, etc.

[1] According to modern terminology, the unknown quantity or quantities within a polynomial (regarded as an expression) are called **indeterminates**, and they are called **variables** only when the polynomial is considered as a function. It is, however, widespread to retain the classical terminology and use the word "variable" in both expressions and functions.

History
In the ancient Near East, during the so-called cradle of civilization (4th millennium BCE), people used a soft and malleable metal, copper, to make tools, weapons, and armor. One "day" (the time varies according to regions) they discovered that even a small amount of arsenic and later tin added to liquid copper not only makes the alloy better in casting, but it also makes the final product much stronger. The Bronze Age began. Although "historical bronzes" show a great variety in composition (which largely depended on availability), a typical bronze consists of 88% copper and 12% tin. Ancient bronze-smiths were well aware of this. Using B, C, and T for the amount of bronze, copper, and tin in a metal alloy, we can express this as $B = C+T$, $C = 0.88 \cdot B$, $T = 0.12 \cdot B$. However simple, these are one of the oldest equations people seem to have used, at least empirically. The right-hand sides are polynomial (linear) expressions in the indeterminates B, C, and T.

Using arithmetic operations (applied to both numbers and indeterminates), a polynomial can be brought to a finite sum of **monomials**. A monomial is the product of a (real) number and indeterminates raised to **integral** powers. The number in the monomial is called its **coefficient**, and the sum of the (integral) exponents is the degree of the monomial.

The **degree** of a polynomial $p(x)$, $p(x, y)$, etc., denoted by $\deg p(x)$, $\deg p(x, y)$, etc., is the **maximum** of the degrees of the monomials contained in $p(x)$, $p(x, y)$, etc. In case of several indeterminates, the degree may be attained by several monomials within the polynomial. Oftentimes a monomial expression is referred to as a **term**, and **like terms** are monomials with the same indeterminates raised to the same natural exponents. For example, x^2y and xy^2 are unlike terms, whereas $\sqrt{2}x^2y^2$ and $\sqrt{3}x^2y^2$ are like terms.

A **binomial** is a polynomial expression which can be written as the sum of two monomials. In a similar vein, a **trinomial** is the sum of three monomials. We will discuss binomials and trinomials in the forthcoming sections.

Remark The identically zero expression is considered as a polynomial with no degree. Unless stated explicitly, we always tacitly assume that the polynomials in our study are non-zero.

For a given polynomial $p(x)$, the equation $p(x) = 0$ is called a **polynomial equation**. Any solution of a polynomial equation is called a **root** of the polynomial. Finding a root (or roots) of a polynomial is one of the oldest problems in mathematics.

Remark The term **root** is traditional. It refers to the fact that low degree polynomial equations are usually solved by extraction of roots of certain expressions in the coefficients.[2]

Turning to polynomials of several indeterminates, the sets

$$\{(x, y) \in \mathbb{R}^2 \mid p(x, y) = 0\}$$
$$\{(x, y, z) \in \mathbb{R}^3 \mid p(x, y, z) = 0\}$$
$$\{(x_1, x_2, \ldots, x_n) \in \mathbb{R}^n \mid p(x_1, x_2, \ldots, x_n) = 0\}$$

[2]The modern terminology applied to the much wider class of functions calls a solution of the functional equation $f(x) = 0$ the **zero** of the function f.

etc. are called the **zero-sets** of the respective polynomials. The cases of two and three variables are especially important as they offer visual images in $\mathbb{R}^2$ and $\mathbb{R}^3$.

Examples for polynomials in the indeterminates x, y are

$$ax - by - c;^{3} \quad \frac{x^2}{a^2} + \frac{y^2}{b^2} - 1;^{4} \quad x^2 - d \cdot y^2 - 1,\ d \in \mathbb{N};^{5} \quad (x+y)^n,\ n \in \mathbb{N};^{6}$$

and examples for polynomials in the indeterminates x, y, z are

$$x^3 + y^3 + z^3 - 3xyz;^{7} \quad x^n + y^n - z^n,\ n \in \mathbb{N}.^{8}$$

An example for a polynomial expression in the entities $\sqrt{2}, \sqrt{3}$ is $(\sqrt{2} + \sqrt{3})^6$.

Although conceptually different, a polynomial can be transformed into a polynomial expression by replacing its indeterminates by entities, and vice versa, a polynomial expression can be reduced to a polynomial in the reverse way.

For example, the polynomial expression $\sqrt{2}^5 + \sqrt{2} + 1$ can be turned into the polynomial $x^5 + x + 1$, and the polynomial $ax^2 + bx + c, a, b, c \in \mathbb{R}$, above can be turned into the polynomial expression $a\sqrt{3}^2 + b\sqrt{3} + c$ in the entity $\sqrt{3}$.

As far as the general theory is concerned, it is therefore sufficient to consider polynomials only.

On the other hand, polynomial expressions arise naturally in various branches of mathematics; for example, in trigonometry, polynomial expressions in trigonometric functions, the so-called **trigonometric polynomials**, play significant roles. (See Section 11.3.) Similarly, in linear algebra, polynomial expressions in matrix entities, the so-called matrix polynomials, are objects of primary interest.

The Point-Line Postulate of Birkhoff's Geometry[9] says that, for any two distinct points, there is a unique line passing through them. Since lines are given by linear equations (Section 5.2), this implies that, given any two distinct real numbers $x_1, x_2 \in \mathbb{R}$, $x_1 \neq x_2$, and $y_1, y_2 \in \mathbb{R}$, there exists a linear (degree ≤ 1) polynomial $p(x)$ such that $p(x_1) = y_1$ and $p(x_2) = y_2$. The concept of **Lagrange (interpolation) polynomial** generalizes this observation as follows.

Example 6.1.1 Let $x_1, x_2, \ldots, x_n \in \mathbb{R}, 2 \leq n \in \mathbb{N}$, be distinct, and $y_1, y_2, \ldots, y_n \in \mathbb{R}$. Then there exists a unique polynomial $\ell(x)$ of degree $< n$ such that $\ell(x_i) = y_i$, $i = 1, 2, \ldots, n$.

---

[3] The zero-set $ax - by - c = 0$ is the generic equation of a line discussed in Section 5.2.

[4] The zero-set is the ellipse in normal form to be discussed in Section 8.3.

[5] $x^2 - d \cdot y^2 - 1 = 0$ is Pell's equation discussed in Section 2.1.

[6] The expansion of this is the Binomial Formula to be discussed in Section 6.3.

[7] This polynomial is related to the AM-GM inequality in three indeterminates.

[8] The zero-set of this polynomial is the so-called Fermat curve related to Fermat's Last Theorem.

[9] The first postulate of Euclid's *Elements*; see Section 5.1.

The existence is given by the so-called **Lagrange form**

$$\ell(x) = \sum_{i=1}^{n} y_i \ell_i(x),$$

where

$$\ell_i(x) = \prod_{\substack{j=1 \\ j \neq i}}^{n} \frac{x - x_j}{x_i - x_j}, \quad i = 1, 2, \ldots, n.$$

Clearly, for $i = 1, 2, \ldots, n$, we have $\ell_i(x_i) = 1$, and $\ell_i(x_j) = 0$ if $j \neq i$, $j = 1, 2, \ldots, n$. It is also clear that the degree of $\ell(x)$ is less than n.

Finally, note that unicity is a direct consequence of the Factor Theorem (to be discussed in Section 6.5) since a non-zero polynomial of degree $< n$ cannot have n distinct roots.

History

The concept of Lagrange polynomial was discovered by the British mathematician Edward Waring (1736–1798). It must have been known to Euler (as it is a direct consequence of one of his formulas published a few years later). In 1795 Lagrange published the formula above, and it was subsequently named after him.

We now discuss some famous examples for evaluating polynomials on integers.

Example 6.1.2 The polynomial $p(x) = x^2 + x + 41$ evaluated at 40 gives

$$p(40) = 40^2 + 40 + 41 = 40^2 + 2 \cdot 40 + 1 = 41^2 = 1681,$$

a square, in particular, a composite number. On the other hand, it is an amazing fact, discovered by Euler in 1772, that the values of $p(x)$ on **all** the first 40 integers starting with 0 are **prime numbers.** They are

$$p(0) = 41,\ p(1) = 43,\ p(2) = 47,\ p(3) = 53,\ p(4) = 61,\ p(5) = 71,\ p(6) = 83,$$

$$p(7) = 97,\ p(8) = 113,\ p(9) = 131,\ p(10) = 151,\ p(11) = 173,\ p(12) = 197,$$

$$p(13) = 223,\ p(14) = 251,\ p(15) = 281,\ p(16) = 313,\ p(17) = 347,\ p(18) = 383,$$

$$p(19) = 421,\ p(20) = 461,\ p(21) = 503,\ p(22) = 547,\ p(23) = 593,\ p(24) = 641,$$

$$p(25) = 691,\ p(26) = 743,\ p(27) = 797,\ p(28) = 853,\ p(29) = 911,\ p(30) = 971,$$

$$p(31) = 1033,\ p(32) = 1097,\ p(33) = 1163,\ p(34) = 1231,\ p(35) = 1301,$$

$$p(36) = 1373,\ p(37) = 1447,\ p(38) = 1523,\ p(39) = 1601.$$

A similar example is provided by the polynomial

$$q(x) = x^2 - 79x + 1601 = (x - 40)^2 + (x - 40) + 41,$$

for which $q(n)$ is prime for $n = 1, 2, 3, \ldots, 79$ (with each prime repeated twice).

In contrast, an "opposite example" is given by the following:

Example 6.1.3 For the polynomial $p(x) = x^6 + 1091$, the values $p(n)$ with $n = 1, 2, \ldots, 3905$ are **composite** numbers but

$$p(3906) = 3,551,349,655,007,944,406,147$$

is a **prime**.

This example needs a computer algebra system. First, for $n \in \mathbb{N}$ odd, $p(n)$ is clearly even, so that we need to calculate $p(n)$ only when $n = 2m$, $m \in \mathbb{N}$, is even. The first ten values are

m	$(2m)^6 + 1091$	prime factorization
1	1155	$3 \cdot 5 \cdot 7 \cdot 11$
2	5187	$3 \cdot 7 \cdot 13 \cdot 19$
3	47747	$7 \cdot 19 \cdot 359$
4	263235	$3 \cdot 5 \cdot 7 \cdot 23 \cdot 109$
5	1001091	$3 \cdot 7 \cdot 13 \cdot 19 \cdot 193$
6	2987075	$5^2 \cdot 7 \cdot 13^2 \cdot 101$
7	7530627	$3 \cdot 13 \cdot 193093$
8	1677830	$3 \cdot 7 \cdot 13 \cdot 41 \cdot 1499$
9	34013315	$5 \cdot 7 \cdot 353 \cdot 2753$
10	64001091	$3 \cdot 7 \cdot 11 \cdot 461 \cdot 601$

For the last composite number, we have

$$p(3905) = 2^2 \cdot 3 \cdot 7^2 \cdot 19 \cdot 1133850409 \cdot 279923617.$$

Turning to the next example, you may have wondered what was the role of the number of (non-leap) years 365 in the polynomial $(1 + x/365)^{365}$ noted at the beginning of this section. The next example is to clarify this.

Example 6.1.4 Suppose we have an initial deposit P in a checking account in a bank that gives x interest **compounded daily**. How much will our principal and interest be after one year?

For a moment, we keep the number n of compounding periods within a year an indeterminate. After the first period the bank adds P times x/n amount to our principal, and we end up with the amount

$$P + P \cdot \frac{x}{n} = P\left(1 + \frac{x}{n}\right).$$

This is our new principal at the beginning of the second period. Thus, after the second period, we have

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

Assume now that we wait t years. Since there are $n \cdot t$ compounding periods in t years, we arrive at the **Compound Interest Formula** giving the compounded amount (our principal plus interest after t years), the so-called **Future Value**, as

$$P\left(1+\frac{x}{n}\right)^{n\cdot t}.$$

This is a polynomial in the indeterminate x of degree $n \cdot t$ assuming that the latter is an integer. (Observe that t can be any (rational) number.)

Finally, the polynomial $(1 + x/365)^{365}$ gives the future value of a deposit of $P = \$1$ after one year, $t = 1$, with daily compounding $n = 365$.

History

In studying compounded interest, it was Jacob Bernoulli who first considered $(1 + 1/n)^n$ for large n. (This is the idealized situation with principal \$1 and 100% interest.) We will see later that, as n increases indefinitely, this expression approaches the number e.

Returning to the main line, applying the laws of arithmetic, a polynomial $p(x)$ can be brought to the form

$$p(x) = a_n x^n + a_{n-1}x^{n-1} + \cdots + a_1 x + a_0$$

as a finite sum of monomials in **descending order**. Here a_n, the **leading coefficient**, is tacitly assumed to be non-zero so that the degree of $p(x)$ is n. A polynomial $p(x)$ with leading coefficient 1 is called **monic**.

As we will see later, the large-scale behavior of a polynomial $p(x)$ is determined by its leading coefficient. The descending order above is to emphasize this.

Low degree polynomials have specific names and notation.

Polynomials of degree ≤ 1 are called **linear**,[10] and they can be brought to the point-slope form

$$p(x) = y_0 + m(x - x_0).$$

A degree two polynomial is called **quadratic**, and it is usually written as a **trinomial**

$$p(x) = ax^2 + bx + c.$$

Polynomials of degree 3, 4, 5, 6, etc. are called **cubic**, **quartic**, **quintic**, **sextic**, etc.

Remark 1 The expanded form of a polynomial is not always the most convenient to reveal its structure; see, for example, $(1 + x/365)^{365}$ as above.

[10] Also including constant polynomials.

Remark 2 At times it is more convenient to write $p(x)$ in **ascending order**

$$p(x) = a_0 + a_1 x + \cdots + a_{n-1} x^{n-1} + a_n x^n.$$

This is the preferred form for a general expression of the product of two polynomials (as we will see shortly), and also when the polynomial is a finite portion of an infinite power series.

Exercises

6.1.1. Let $p(x) = x^3 - 3x - 2$. Determine the polynomial whose roots are those of $p(x)$ plus 1.

6.1.2. Find all integer solutions of the equation $x^3 = 2y^3 + 4z^3$.

6.1.3. Let $a, b \in \mathbb{Z}$ with $a \neq 0$ such that a does not divide b. Show that the quadratic polynomial $ax^2 + bx + b - a$ has no root amongst the natural numbers.

6.2 Arithmetic Operations on Polynomials

Arithmetic operations such as addition, subtraction, multiplication, and division can be applied to polynomials.

In this section we discuss the first three of these operations. (Division of polynomials is more complex, and it is deferred to Section 6.5.) As before, we will treat polynomials of a single indeterminate in detail with occasional examples of polynomials in several indeterminates.

Since indeterminates of polynomials obey the same laws of arithmetic as (real) numbers, addition, subtraction, and multiplication of polynomials are defined naturally.

The **sum** of two polynomials is obtained by adding up all the monomials in each of the polynomials. When adding two polynomials of the same degree, the degree of the sum is less than or equal to the degree of the polynomials. If the two polynomials have different degrees, then the degree of the sum is the larger of the degrees of the participating polynomials.

Subtraction of a polynomial from another is the same as addition of the negative (in which all monomials changed to their negatives).

Multiplying polynomials follows the distributive law applied repeatedly. The **product** of two polynomials is the sum of all possible products of pairs of monomials that participate in their respective polynomials. The degree of the product is the sum of the degrees of the participating polynomials.

More specifically, in the single indeterminate case, consider two polynomials

$$p(x) = a_0 + a_1x + a_2x^2 + \cdots + a_nx^n$$

and

$$q(x) = b_0 + b_1x + b_2x^2 + \cdots + b_mx^m$$

of degrees n and m, where we used ascending order for convenience.

The sum $p(x)+q(x)$ is the polynomial of degree less than or equal to $\max(m, n)$ whose monomials are the sums of all monomials in $p(x)$ and $q(x)$. The coefficient c_k of the monomial c_kx^k, $0 \le k \le \max(m, n)$, in the sum is equal to $a_k + b_k$, where we tacitly assume that undefined coefficients are set to be zero.

To form the product $p(x)q(x)$ of two polynomials $p(x)$ and $q(x)$ as above, each monomial in $p(x)$ has to be multiplied with each monomial in $q(x)$, and then these products have to be added. The product $p(x)q(x)$ is a polynomial of degree $n + m$ written as

$$p(x)q(x) = c_0 + c_1x + c_2x^2 + \cdots + c_{n+m}x^{n+m}.$$

Forming the coefficients c_k, $0 \le k \le n+m$, follows the so-called **Cauchy Product Rule** (named after Augustin–Louise Cauchy (1789–1857)): The k-th coefficient c_k is the sum of the terms a_ib_j with $i + j = k$ and $0 \le i \le n$, $0 \le j \le m$. (This is because the corresponding product of monomials is $(a_ix^i)(b_jx^j) = a_ib_jx^{i+j} = a_ib_jx^k$.)

Thus, we have

$$p(x)q(x) = a_0b_0 + (a_0b_1 + a_1b_0)x + (a_0b_2 + a_1b_1 + a_2b_0)x^2 + \cdots + a_nb_mx^{n+m}.$$

Example 6.2.1 In multiplying polynomials, does the product $p(x)q(x)$ contain at least as many monomials as $p(x)$ or $q(x)$?

The answer is **no**. For example, consider the product $(x^2-\sqrt{2}x+1)(x^2+\sqrt{2}x+1)$. Using the Cauchy Product Rule, we multiply each monomial and obtain

$$\begin{aligned}(1 - \sqrt{2}x + x^2)(1 + \sqrt{2}x + x^2) = 1 + (\sqrt{2} - \sqrt{2})x + (1 - \sqrt{2}^2 + 1)x^2 \\ + (\sqrt{2} - \sqrt{2})x^3 + x^4 = 1 + x^4.\end{aligned}$$

We see that the product contains fewer monomials than each of the factors.

Example 6.2.2 Show that, for $n \in \mathbb{N}$, we have

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

This is a simple induction with respect to $n \in \mathbb{N}$. For $n = 1$, we have

$$(1 + x)(1 + x^2) = 1 + x + x^2 + x^3 = 1 + x + x^2 + x^{2^2-1}.$$

For the general induction step $n \Rightarrow n+1$ we use the induction hypothesis and calculate

$$\begin{aligned}
&(1+x)(1+x^2)(1+x^4)\cdots(1+x^{2^n})(1+x^{2^{n+1}}) \\
&\quad = (1+x+x^2+x^3+\cdots+x^{2^{n+1}-1})(1+x^{2^{n+1}}) \\
&\quad = (1+x+x^2+x^3+\cdots+x^{2^{n+1}-1}) + x^{2^{n+1}}(1+x+x^2+x^3+\cdots+x^{2^{n+1}-1}) \\
&\quad = 1+x+x^2+x^3+\cdots+x^{2^{n+1}-1}+x^{2^{n+1}}+x^{2^{n+1}+1}+\cdots+x^{2^{n+2}-1}.
\end{aligned}$$

The identity follows.

We continue with examples of polynomials in several indeterminates.

Example 6.2.3 Derive the identity

$$(-x+y+z)^2+(x-y+z)^2+(x+y-z)^2+(x+y+z)^2 = 4(x^2+y^2+z^2).$$

By the Cauchy Product Rule, we have

$$(x+y+z)^2 = x^2+y^2+z^2+2xy+2yz+2zx,$$

and the other three terms can be obtained by replacing x, y, z by their negatives. The crux is that in the sum of the four terms above all hybrid terms xy, yz, zx cancel. Hence, counting the pure quadratic terms, we obtain $4(x^2+y^2+z^2)$. The example follows.

For the next step, recall from Section 3.1 the Finite Geometric Series Formula

$$1-r^n = (1-r)(1+r+r^2+\cdots+r^{n-1}),$$

where we multiplied out with $1-r$ and shifted the exponent n down to $n-1$.

We now **homogenize** this formula by substituting $r = y/x$ and multiplying out by x^n. We obtain the following important identity

$$x^n - y^n = (x-y)(x^{n-1}+x^{n-2}y+\cdots+xy^{n-2}+y^{n-1}), \quad 2 \le n \in \mathbb{N}.$$

The frequently occurring cases $n = 2, 3$ are

$$\begin{aligned}
x^2 - y^2 &= (x-y)(x+y) \\
x^3 - y^3 &= (x-y)(x^2+xy+y^2).
\end{aligned}$$

We call these the **difference of squares**, and the **difference of cubes** identities.

For $n \in \mathbb{N}$ **odd**, replacing y by its negative, we obtain

$$x^n+y^n=(x+y)(x^{n-1}-x^{n-2}y+\cdots-xy^{n-2}+y^{n-1}), \ 3 \le n \in \mathbb{N}, \ n \text{ odd}.$$

In particular, for $n = 3$, we have

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

Remark In Example 6.2.1 above we may have proceeded as

$$(x^2-\sqrt{2}x+1)(x^2+\sqrt{2}x+1)=(x^2+1-\sqrt{2}x)(x^2+1+\sqrt{2}x)=(x^2+1)^2-(\sqrt{2}x)^2=x^4+1.$$

An important consequence of the identities above is the following: For any polynomial $p(x)$ with **integer** coefficients, and $a, b \in \mathbb{Z}$ distinct, we have

$$a - b \mid p(a) - p(b).$$

This follows from the identity above. Indeed, letting

$$p(x) = c_n x^n + c_{n-1}x^{n-1} + \cdots + c_1 x + c_0, \quad c_0, c_1, \ldots, c_{n-1}, c_n \in \mathbb{Z},$$

we have

$$p(a) - p(b) = c_n(a^n - b^n) + c_{n-1}(a^{n-1} - b^{n-1}) + \cdots + c_1(a - b).$$

On the other hand, by the identity above, for each $k = 1, \ldots, n$, we have

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

In particular, we have $a - b | a^k - b^k$, and thus $a - b | p(a) - p(b)$.

Example 6.2.4 Let $p(x)$ be a polynomial with integer coefficients.[11] If, for n integers $a_1, a_2, a_3, \ldots, a_n \in \mathbb{Z}$, we have $p(a_1) = a_2$, $p(a_2) = a_3, \ldots, p(a_{n-1}) = a_n$, $p(a_n) = a_1$ then $|a_1 - a_2| = |a_2 - a_3| = \ldots = |a_{n-1} - a_n| = |a_n - a_1|$.

By the discussion before this example, the conditions on $p(x)$ give

$$\begin{aligned} a_1 - a_2 | p(a_1) - p(a_2) &= a_2 - a_3 | p(a_2) - p(a_3) = \ldots \\ &= a_{n-1} - a_n | p(a_{n-1}) - p(a_n) = a_n - a_1 | p(a_n) - p(a_1) = a_1 - a_2. \end{aligned}$$

The statement follows.

Exercises

6.2.1. Consider the polynomial

$$p(x) = \frac{x^5}{5} + \frac{x^3}{3} + \frac{7x}{15}.$$

Show that $p(n) \in \mathbb{Z}$ for $n \in \mathbb{Z}$.

[11] A special case ($n = 3$) was part of a problem in the USA Mathematical Olympiad, 1974.

6.2.2. Determine 999^3.
6.2.3. Solve $(1 + x^2)(1 + x^4) = 4x^3$.
6.2.4. Let $a \in \mathbb{R}$. Solve $(x + y)^2 = (x - a)(y + a)$ for x, y.

6.3 The Binomial Formula

In this section we develop a general binomial formula for the expansion of the power $(x + y)^n$ for any natural exponent $n \in \mathbb{N}$.

The special cases of quadratic and cubic binomial formulas ($n = 2, 3$)

$$(x + y)^2 = x^2 + 2xy + y^2 \quad \text{and} \quad (x + y)^3 = x^3 + 3x^2y + 3xy^2 + y^3$$

are well-known and can easily be derived.

Example 6.3.1 Factor the polynomial $x^2 + 2xy + y^2 - z^2$.

We recognize that the first three terms match with the quadratic binomial formula above. Using this, we calculate

$$x^2 + 2xy + y^2 - z^2 = (x + y)^2 - z^2 = (x + y - z)(x + y + z),$$

where, in the last step, we used the difference of squares identity.

Example 6.3.2 Show that, for every $m \in \mathbb{N}$, there exists $n \in \mathbb{N}$ such that $m + n + 1$ is a perfect square, and $mn + 1$ is a perfect cube.

From the cubic binomial formula above, the second condition is easily satisfied with $n = m^2 + 3m + 3$ since $mn + 1 = m^3 + 3m^2 + 3m + 1 = (m + 1)^3$. This also works for the first condition since $m + n + 1 = m^2 + 4m + 4 = (m + 2)^2$.

To begin with the study of the binomial formula, we take a closer look at how the quadratic and cubic binomial formulas are derived.

In the quadratic case, we have

$$(x + y)^2 = (x + y)(x + y) = xx + xy + yx + yy.$$

Combining the middle like terms, we obtain $x^2 + 2xy + y^2$.

In the cubic case we have a similar pattern

$$(x+y)^3 = (x+y)(x+y)(x+y) = xxx+xxy+xyx+yxx+xyy+yxy+yyx+yyy.$$

Combining, we arrive at $x^3 + 3x^2y + 3xy^2 + y^3$.

One common feature of these expansions is that all the terms have the same degree. Thus, expanding $(x + y)^n$, all terms have to be of the form $x^{n-k}y^k$, $k = 0, \ldots, n$. Therefore the possible terms are

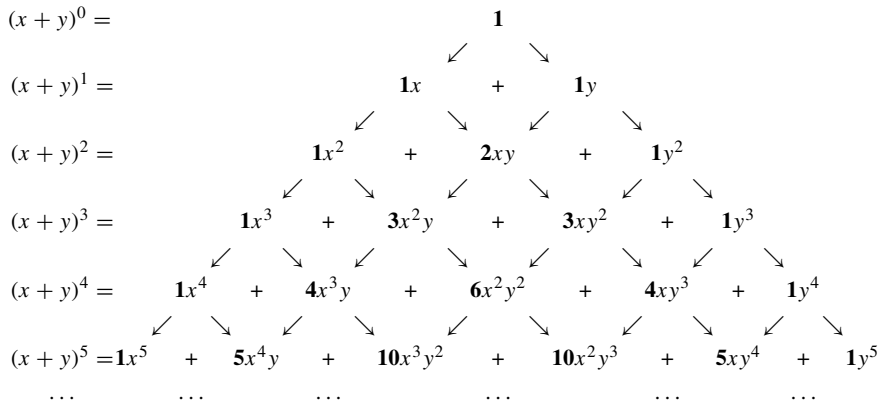
$$x^n, x^{n-1}y, x^{n-2}y^2, \ldots, x^2y^{n-2}, xy^{n-1}, y^n.$$

Another common feature of these expansions is that each term comes with unit coefficient, so that, after combining, the coefficient of each monomial is a **natural number**. In the expansion of the power $(x+y)^n$ we let $C_k^n \in \mathbb{N}$ denote the coefficient of the monomial $x^{n-k}y^k$, $k = 0, 1, 2, \ldots, n$.

Summarizing, so far we have the following

$$(x+y)^n = \sum_{k=0}^{n} C_k^n x^{n-k} y^k = C_0^n x^n + C_1^n x^{n-1} y + \cdots + C_{n-1}^n x y^{n-1} + C_n^n y^n.$$

Thus, it remains to determine the coefficients C_k^n for all $k = 0, 1, 2, \ldots, n$. To do this we take a look at the detailed chart:



Since we are after the coefficients, we highlighted them by using boldface (even for the coefficient 1). This is called the **Pascal Triangle** after the French mathematician and philosopher Blaise Pascal (1623–1662).

History

The Binomial Formula and the Pascal Triangle were known about two millennia before Pascal first published them in the Western world. The earliest known record for the general Binomial Formula (with any power) is from the Indian mathematician Pingala (around 200 BCE) from the Vedic period. Another Indian mathematician Halayudha (around the 10-th century CE) wrote a commentary on Pingala's work which contains the description of the Pascal Triangle. The next few centuries have witnessed several independent discoveries of these in Persia (Al-Karajī (953–c. 1029) and Omar Khayyám (1048–1131)) and in China (Jia Xian (c. 1010–1070) and Yang Hui (1238–1298)).

After a quick glance we realize that along the two sides of the triangle the coefficients are always 1, that is we have $C_0^n = C_n^n = 1$. More importantly, in the interior of the triangle, at each location of a monomial, the coefficient is the sum of the coefficients of its top two neighbors in the row above. For example, the coefficient 10 of the monomial $10x^3y^2$ is the sum of the coefficients of the two neighbors above, $4x^3y$ and $6x^2y^2$. This is indicated by arrows pointing in

the southeastern and southwestern directions. The arrows actually carry another meaning; the southwestern arrow always means multiplication by x, and the southeastern arrow means multiplication by y.

It is easy to see why this is true if we take a look at the general pattern at an interior location:

$$\begin{array}{lccc} (x+y)^n = \ldots & C_{k-1}^n x^{n-(k-1)} y^{k-1} & + & C_k^n x^{n-k} y^k \\ & \searrow & & \swarrow \\ (x+y)^{n+1} = \ldots & & C_k^{n+1} x^{(n+1)-k} y^k & \end{array}$$

The binomial $(x+y)^{n+1}$ is obtained from the previous binomial $(x+y)^n$ via multiplication by $(x+y)$:

$$(x+y)^{n+1} = (x+y)^n (x+y) = (x+y)^n x + (x+y)^n y.$$

Thus, the monomials in the expansion of $(x+y)^{n+1}$ are obtained from the monomials in the expansion of $(x+y)^n$ via multiplications by x and y (and combining like terms). To obtain the monomial $C_k^{n+1} x^{(n+1)-k} y^k$, the monomial $C_k^n x^{n-k} y^k$ needs to be multiplied by x (southwestern arrow), and the monomial $C_{k-1}^n x^{n-(k-1)} y^{k-1}$ needs to be multiplied by y (southeastern arrow). There are no other sources in the top row to contribute to the monomial in the bottom.

As a byproduct, we also see the inductive relation

$$C_k^{n+1} = C_k^n + C_{k-1}^n.$$

This understanding of the coefficients of the Pascal Triangle is useful for low values of n. To obtain a better (non-inductive) formula for C_k^n, we need to go back to our original expansion

$$(x+y)^n = \overbrace{(x+y)}^{1} \overbrace{(x+y)}^{2} \overbrace{(x+y)}^{3} \cdots \overbrace{(x+y)}^{n}.$$

On the right-hand side there are n parentheses. To form a term in the expansion, within each bracket we need to choose an x or a y. The term obtained this way contributes to C_k^n if and only if we choose y exactly k times, and consequently x exactly $(n-k)$ times. Thus C_k^n is the **number of ways** k elements (the y's) can be selected out of n elements (the brackets). If we mark the brackets by the first n positive integers $1, 2, 3, \ldots, n$ as above, then C_k^n is the number of k-element subsets of the set $\{1, 2, 3, \ldots, n\}$. Because of this interpretation, the binomial coefficient C_k^n is usually spelled as "n **choose** k" and denoted by

$$C_k^n = \binom{n}{k}, \quad k = 0, 1, 2, \ldots, n.$$

Notice that, in particular, we have $\binom{n}{0} = \binom{n}{n} = 1$.

History
The notation C_k^n reflects the combinatorial meaning, "combinations" or "choices." The symbol $\binom{n}{k}$ is due to the Austrian mathematician and physicists Andreas Freiherr von Ettingshausen (1796–1878) used in his book *Die combinatorische Analysis als Vorebereitungslehre der theoretischen höhern Mathematik* published in 1826.

With this, our **Binomial Formula** takes the final form

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

Replacing the indeterminate y by its negative in the Binomial Formula above, we obtain

$$(x-y)^n = \binom{n}{0} x^n - \binom{n}{1} x^{n-1} y + \binom{n}{2} x^{n-2} y^2 - \cdots + (-1)^n \binom{n}{n} y^n.$$

Remark In the future, it will be convenient to define $\binom{n}{k} = 0$ if $k > n$ or $k < 0$. With this, $\binom{n}{k}$ is defined for all integers $k, n \in \mathbb{Z}$.

There are several immediate properties of the binomial coefficients. First of all, if we select a k-element subset from $\{1, 2, 3, \ldots, n\}$ then, automatically, the $(n-k)$-element complement, the set of elements that have not been selected, becomes well-defined. Thus, the number of k-element subsets and the number of $(n-k)$-element subsets are the same:

$$\binom{n}{k} = \binom{n}{n-k}, \quad k = 0, 1, 2, \ldots, n.$$

Looking back at the Pascal Triangle, we see that this means that it is symmetric with respect to its middle vertical axis.

With our new notation, the **inductive** relation above for the coefficients takes the form

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

Actually, this also follows easily from our new interpretation of the binomial coefficients. The binomial coefficient on the left-hand side is the number of ways a k-element subset can be selected from a set of $n+1$ elements $\{1, 2, 3, \ldots, n, n+1\}$, say. There are two kinds of k-element subsets here. First, there are those which do not contain the last element $n+1$. The number of this kind of subsets is $\binom{n}{k}$. Second, there are those which contain $n+1$. The number of this kind of subsets is $\binom{n}{k-1}$ since, once $n+1$ is selected, we need to select only $k-1$ additional elements. The inductive formula follows.

We now tackle our basic question: Is there a non-inductive formula for the binomial coefficients?

To answer this question we need to take a more careful look at the selection process. As before, we let our base set be $\{1, 2, 3, \ldots, n\}$. To obtain a k-element subset we need to select the first element. This can be done n ways. Next, we select the second element. This can be done $n - 1$ ways since the selection of the first element reduced the amount of choices by one. These selections are independent so that the number of ways to select the first element and then the second is $n(n - 1)$. We continue this way up to k elements, $k = 1, 2, \ldots, n$, and realize that the number of possible selections is

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

where we used the factorial notation (Example 0.4.2).

We now realize that this is not exactly what we want since the selection process was carried out in an **order**; that is, we know which element was the first, the second, etc. and the k-th. In other words, this product is the number of **ordered** sequences of k-elements of the set $\{1, 2, 3, \ldots, n\}$. Thus, each k-element subset (with no order) is over-counted by $k!$ times, the number of permutations of a k-element set. We obtain

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}, \quad k = 0, 1, \ldots, n,$$

where, for consistency, we must have $0! = 1$.

Remark The quartic binomial formula

$$(x - y)^4 = x^4 - 4x^3y + 6x^2y^2 - 4xy^3 + y^4, \quad x, y \in \mathbb{R},$$

(with $-y$ in place of y) gives a (somewhat lesser known) sharpening of the AM-GM inequality

$$\frac{a+b}{2} - \sqrt{ab} \geq \frac{(a-b)^2}{4(a+b)}, \quad 0 < a, b \in \mathbb{R}.$$

Indeed, eliminating the denominator and simplifying, this is equivalent to

$$a^2 + 6ab + b^2 \geq 4(a+b)\sqrt{ab}, \quad 0 < a, b \in \mathbb{R}.$$

Now, the substitution $a = x^2$ and $b = y^2$ reduces this to $(x - y)^4 \geq 0$.

Example 6.3.3 How many ways can n one dollar bills be distributed amongst k people so that each person receives at least one dollar?

We line up the n one dollar bills in a row, and partition them by placing $k - 1$ separators between them.[12] Since there are $n - 1$ gaps between the adjacent bills

[12]The graphical interpretation of this and similar combinatorial problems is usually termed as "stars and bars," as advocated by the Croatian-American mathematician Willibald Srećko Feller

that can receive separators, the number of ways to distribute the money amongst k people is $\binom{n-1}{k-1}$.

Three variations on the theme are as follows:

Example 6.3.4 Given $n \in \mathbb{N}_0$, find the number of solutions $x_1, x_2, \ldots, x_k \in \mathbb{N}_0$ to the equation $x_1 + x_2 + \cdots + x_k = n$.

We move up the values of the indeterminates by one, $x_i' = x_i + 1 \in \mathbb{N}$, $i = 1, 2, \ldots, k$, and realize that the modified equation $x_1' + x_2' + \cdots + x_k' = n + k$ patterns the previous example. The number of solutions is therefore $\binom{n+k-1}{k-1}$.

Example 6.3.5 How many distinct monomials do we get when we expand $(x_1 + x_2 + \cdots + x_k)^n$?

Every term in the expansion is of the form $x_1^{a_1} x_2^{a_2} \cdots x_k^{a_k}$, where $a_1, a_2, \ldots, a_n \in \mathbb{N}_0$ with $a_1 + a_2 + \cdots + a_n = n$. The previous example gives the answer as $\binom{n+k-1}{k-1}$.

Example 6.3.6 Let $k, n \in \mathbb{N}$. How many natural numbers $x_1, x_2, \ldots, x_k \in \mathbb{N}$ satisfy the equation $x_1 \cdot x_2 \cdots x_k = 10^n$?

The factors must have the form $x_i = 2^{a_i} \cdot 5^{b_i}$, $a_i, b_i \in \mathbb{N}_0$, $i = 1, 2, \ldots, k$. The exponents must satisfy the equations $a_1 + a_2 + \cdots + a_k = n$ and $b_1 + b_2 + \cdots + c_k = n$. The previous example gives the number of solutions as $\binom{n+k-1}{k-1}^2$.

We now briefly return to maps between finite sets. Recall that in Example 0.4.1 we determined the number of **injective** maps $X \to Y$, $|X| = m$ and $|Y| = n$, $m \leq n$, $m, n \in \mathbb{N}$, as

$$n(n-1)\cdots(n-m+1) = \frac{n!}{(n-m)!} = \binom{n}{m} m!,$$

where we used the binomial coefficient formula above.[13]

Example 6.3.7 The number of **surjective** maps $X \to Y$, $|X| = m$ and $|Y| = n$, $m \geq n$, $m, n \in \mathbb{N}$, is

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

The number of all maps $X \to Y$ is n^m (Example 0.4.1). (This corresponds to $k = 0$ in the sum above.) To derive the stated formula, we will count the number of maps $X \to Y$ that are **not** surjective.

Letting $Y = \{1, 2, \ldots, n\}$, for $i = 1, 2, \ldots, n$, we denote by A_i the set of maps $X \to Y$ that **miss** $i \in Y$ (that is, i is not in the range). The set of **all**

(1907–1970). In our case, the n one dollar bills are represented by stars, and the separators are the bars.

[13] We also reverted to m instead of k for consistency.

non-surjective maps $X \to Y$ is therefore $\bigcup_{i=1}^{n} A_i$. By the Principle of Inclusion-Exclusion (Example 0.4.4), we have

$$\left|\bigcup_{i=1}^{n} A_i\right| = \sum_{\emptyset \neq J \subset \{1,\ldots,n\}} (-1)^{|J|+1} \left|\bigcap_{j \in J} A_j\right|.$$

Now, for $\emptyset \neq J \subset \{1, \ldots, n\}$, the intersection $\bigcap_{j \in J} A_j$ is the set of all maps that miss the subset J. These maps therefore must map into the complement $Y \setminus J$. The number of maps $X \to Y \setminus J$ is $(n - |J|)^m = \left|\bigcap_{j \in J} A_j\right|$. By the discussion on the binomial coefficient above, for each $k = 1, \ldots, n$, the number of k-element subsets $J \subset Y$, $|J| = k$, is $\binom{n}{k}$. Putting everything back into the sum above, we obtain

$$\left|\bigcup_{i=1}^{n} A_i\right| = \sum_{k=1}^{n-1} (-1)^{k+1} \binom{n}{k} (n-k)^m.$$

(Note that the term corresponding to $k = n$ vanishes.) Subtracting this from n^m ($k = 0$), the stated formula follows.

Example 6.3.8 Let $n \in \mathbb{N}$. How many **distinct** monomials do we get when we expand

$$(x_1 + x_2 + x_3 + x_4 + \cdots + x_{2n-1} + x_{2n})(x_1 - x_2 + x_3 - x_4 + \cdots + x_{2n-1} - x_{2n})?$$

Using the difference of squares identity, this expression can be written as

$$\begin{aligned}&((x_1 + x_3 + \cdots + x_{2n-1}) + (x_2 + x_4 + \cdots + x_{2n}))\\ &\quad \times ((x_1 + x_3 + \cdots + x_{2n-1}) - (x_2 + x_4 + \cdots + x_{2n}))\\ &\quad = (x_1 + x_3 + \cdots + x_{2n-1})^2 - (x_2 + x_4 + \cdots + x_{2n})^2.\end{aligned}$$

Expanded, each square on the right-hand side contains n perfect squares of the respective indeterminates, and $\binom{n}{2} = n(n-1)/2$ hybrid products of two distinct indeterminates. Since the two sets of indeterminates in the two squares are disjoint, we obtain the total of $2(n + n(n-1)/2) = n(n+1)$ monomials.

Example 6.3.9 (Revisited) We return to the limit $\lim_{n\to\infty} \sqrt[n]{n} = 1$ of Example 3.2.8 and give a new proof by the Binomial Formula.

Let $a_n = \sqrt[n]{n} - 1$, $n \in \mathbb{N}$. Note that, for $n \geq 2$, a_n is positive. We need to show that $\lim_{n\to\infty} a_n = 0$.

By the Binomial Formula, we have

$$n = (1 + a_n)^n = \sum_{k=0}^{n} \binom{n}{k} a_n^k > \binom{n}{2} a_n^2 = \frac{n(n-1)}{2} a_n^2, \quad 2 \leq n \in \mathbb{N}.$$

Rewriting this, we obtain

$$0 < a_n < \sqrt{\frac{2}{n-1}}, \quad 2 \leq n \in \mathbb{N}.$$

By the monotonicity of the limit, we have

$$0 \leq \lim_{n\to\infty} a_n \leq \lim_{n\to\infty} \sqrt{\frac{2}{n-1}} = 0.$$

The example follows.

The binomial formula has some interesting special cases.

Letting $x = y = 1$, we obtain

$$2^n = \sum_{j=0}^{n} \binom{n}{j} = \binom{n}{0} + \binom{n}{1} + \binom{n}{2} + \cdots + \binom{n}{n-1} + \binom{n}{n}.$$

Actually, this can be seen directly as follows. The right-hand side is the sum of k-element subsets of the set $\{1, 2, 3, \ldots, n\}$ for all $k = 0, 1, 2, \ldots, n$. This sum is then the number of **all** subsets of $\{1, 2, 3, \ldots, n\}$ (regardless the number of elements in the subsets). On the other hand, selecting a subset from $\{1, 2, 3, \ldots, n\}$ amounts to make n decisions: Choose 1 or not, choose 2 or not, etc. choose n or not (for this subset). Each decision has two outcomes, "yes" or "no," so that the total number of decisions to select a subset is $\overbrace{2 \cdot 2 \cdot 2 \cdots 2}^{n \text{ times}} = 2^n$. This is the number on the left-hand side.

Another substitution, $x = 1$ and $y = -1$, gives the alternating sum

$$0 = \sum_{j=0}^{n} (-1)^j \binom{n}{j} = \binom{n}{0} - \binom{n}{1} + \binom{n}{2} - \cdots + (-1)^{n-1}\binom{n}{n-1} + (-1)^n \binom{n}{n}.$$

The binomial coefficients satisfy many identities; some of these we defer to the exercises at the end of this chapter.

Example 6.3.10 Let X be a set of $n \in \mathbb{N}$ elements. Recall that a relation R on X is a subset $R \subset X \times X$. How many relations are there of the form $R = A \times B$, where $A \subset B \subset X$?

We need to count the pairs (A, B) of subsets of X such that $A \subset B$. Let $|B| = k$, $k = 0, 1, \ldots, n$. The number of subsets B of X is $\binom{n}{k}$. Once B is chosen, the number of subsets A of B is 2^k. With this, we obtain that the number of pairs (A, B) with $A \subset B$ is $\sum_{k=0}^{n} \binom{n}{k} \cdot 2^k$. By the Binomial Formula ($x = 1$ and $y = 2$), this is equal to 3^n.

Example 6.3.11 (Derangements) A permutation $f : X \to X$ of a set X of n elements, $2 \leq n \in \mathbb{N}$, is called a **derangement** if no element in X stays fixed under f; that is, we have $f(x) \neq x$ for all $x \in X$. Determine the number D_n of derangements of X.

For $x \in X$, let A_x denote the set of permutations that fix the element x. The total number of derangements is then

$$D_n = n! - \left| \bigcup_{x \in X} A_x \right|,$$

since the number of all permutations of X is $n!$ (Example 0.4.2). On the other hand, by the Principle of Inclusion-Exclusion (Example 0.4.4), we have

$$\left| \bigcup_{x \in X} A_x \right| = \sum_{\emptyset \neq J \subset X} (-1)^{|J|+1} \left| \bigcap_{z \in J} A_z \right|.$$

For a given $\emptyset \neq J \subset X$, the set $\bigcap_{z \in J} A_z$ consists of all permutations that fix the elements in J (and permute the rest of the elements in $X \setminus J$). Hence

$$\left| \bigcap_{z \in J} A_z \right| = (n - |J|)!$$

Since, for a given $i = 1, \ldots, n$, the number of **subsets** $J \subset X$ having $i = |J|$ elements is $\binom{n}{i}$, we obtain

$$\left| \bigcup_{x \in X} A_x \right| = \sum_{i=1}^{n} (-1)^{i+1} \binom{n}{i} (n - i)!$$

Finally, subtracting this from $n!$ as above, we arrive at the total number of derangements of X as

$$D_n = \sum_{i=0}^{n} (-1)^i \binom{n}{i} (n - i)! = n! \sum_{i=0}^{n} \frac{(-1)^i}{i!}.$$

History
The study of derangements originated in the the work *Essay d'analyse sur les jeux de hazard* by Pierre Rémond de Montmort (1678–1719) published in 1708. He determined the number of derangements in 1713, and so did his friend Nicholas Bernoulli (1687–1759) around the same time.

Exercises

6.3.1. Derive the identities (with appropriate ranges of the indeterminates):

$$\textbf{i.}\ \binom{n-1}{k} - \binom{n-1}{k-1} = \frac{n-2k}{n}\binom{n}{k}$$

$$\textbf{ii.}\ \binom{n}{j}\binom{n-j}{k} = \binom{n}{k}\binom{n-k}{j}$$

$$\textbf{iii.}\ \binom{n}{k}\binom{k}{l} = \binom{n}{l}\binom{n-l}{k-l}$$

6.3.2. Derive the identities (with appropriate ranges of the indeterminates):[14]

$$\textbf{i.}\ \sum_{j=0}^{n} j\binom{n}{j} = n2^{n-1}$$

$$\textbf{ii.}\ \sum_{j=0}^{n} j^2\binom{n}{j} = (n+n^2)2^{n-2}$$

$$\textbf{iii.}\ \sum_{j=0}^{k} \binom{m}{j}\binom{n-m}{k-j} = \binom{n}{k}$$

$$\textbf{iv.}\ \sum_{j=0}^{n} \binom{n}{j}^2 = \binom{2n}{n}$$

$$\textbf{v.}\ \sum_{m=j}^{n} \binom{m}{j}\binom{n-m}{k-j} = \binom{n+1}{k+1}$$

$$\textbf{vi.}\ \sum_{m=j}^{n} \binom{m}{j} = \binom{n+1}{j+1}$$

$$\textbf{vii.}\ \sum_{j=0}^{n} \binom{m+j}{j} = \binom{m+n+1}{n}$$

[14] These identities are referred to by various names. Some reflect the author, some the location of the entries in the Pascal Triangle. For example, iii. is called the Vandermonde-convolution, vi. is the column-sum property, vii. is the SE-diagonal sum property, and viii. is the NW-diagonal sum property. Note, finally, that these identities are interrelated, for example, iii. implies iv., v. implies vi., etc.

viii. $\displaystyle\sum_{j=0}^{m}\binom{n-j}{m-j}=\binom{n+1}{m}$

ix. $\displaystyle\sum_{j=k}^{n}\binom{n}{j}\binom{j}{k}=2^{n-k}\binom{n}{k}$

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

6.3.3. Show that, for any polynomial $p(x)$ of degree $\leq n$, $n\in\mathbb{N}$, we have

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

6.3.4. Show that

$$\sum_{j=0}^{[n/2]}\binom{n-j}{j}=F_{n+1},$$

where F_n is the nth Fibonacci number.

6.3.5. Use the Binomial Formula to show $1<a^{1/n}<1+a/n$, $1<a\in\mathbb{R}$, $n\in\mathbb{N}$. Conclude that $\lim_{n\to\infty}\sqrt[n]{a}=1$, $0<a\in\mathbb{R}$.

6.3.6. How many arrangements can seven cards have from a deck of standard playing cards (Example 0.1.6) with strictly increasing rank such that the fourth card is a 7, and no consecutive cards have the same suite?

6.3.7. Derive the following inductive formula for the number of derangements D_n (Example 6.3.11):

$$D_n=(n-1)\left(D_{n-1}+D_{n-2}\right),\quad 2\leq n\in\mathbb{N},\quad D_0=1,\quad D_1=0.$$

6.3.8. Let X be a set of $n\in\mathbb{N}$ elements. Show that the number of ordered pairs (A,B) of subsets of X with $A\subset B$ is 3^n.

6.4 Factoring Polynomials

Factoring a polynomial is the reverse of the process of expanding polynomials; factoring a polynomial amounts to express it as a product of polynomials of **lesser degree**. The polynomials appearing in the product are called **factors**. A factor is always understood to be non-constant; a polynomial of positive degree.

We call a polynomial **reducible** if it can be factored, and **irreducible** if it cannot be factored. A simple application of Peano's Principle of Induction is that every polynomial possesses a **complete factorization**; that is, it can be written as a product of irreducible factors.

Since the complexity of polynomials increases very quickly with the degree, factorization is a very important technique in the study of polynomials. For example, if factorization is available then the problem of finding the roots of a polynomial of a single indeterminate is reduced to that of the factors.

The somewhat crude definition of factorization above is riddled with more subtle issues. For example, we have seen (Example 6.2.1) that the simple quartic binomial $x^4 + 1$ has the factorization $(x^2 - \sqrt{2}x + 1)(x^2 + \sqrt{2}x + 1)$. The original quartic polynomial has integer coefficients, but, in the factors, the irrational number $\sqrt{2}$ appeared. This means that even if we started with a polynomial with integral coefficients, or rational coefficients if we insist on a field, at the end we obtained polynomial factors whose coefficients are **not** integral, in fact, not even rational numbers. We see that if we allow only rational coefficients then the quartic polynomial above is irreducible, but if we allow real coefficients then it is reducible.

We say that our polynomial is **irreducible over** $\mathbb{Q}$ and **reducible over** $\mathbb{R}$. What we learned from this example is that whether a polynomial is reducible or irreducible depends on the field that the coefficients reside in.

Remark The quadratic polynomial $x^2 + 1$ is irreducible over $\mathbb{R}$ since if it were reducible then it would have linear factors, and any linear factor would have a real root. This root would also be a root of the original quadratic polynomial which is impossible since $x^2 + 1 \geq 1$ for all $x \in \mathbb{R}$. On the other hand, it is possible to extend $\mathbb{R}$ to a larger field, the so-called **field of complex numbers** $\mathbb{C}$, and if we allow our coefficients to venture out from $\mathbb{R}$ to $\mathbb{C}$ then we do have the (complete) factorization $x^2 + 1 = (x + \mathrm{i})(x - \mathrm{i})$, where i is the complex unit satisfying $\mathrm{i}^2 = -1$. (The factorization above actually points to the way to **define** the field $\mathbb{C}$.)

As an interesting byproduct, we see that, unlike the factorization $x^2 - y^2 = (x - y)(x + y)$, the polynomial $x^2 + y^2$ is irreducible over $\mathbb{R}$. Indeed, if $x^2 + y^2$ were reducible then, substituting $y = 1$ in the factorization, $x^2 + 1$ would also be reducible; a contradiction.

We just touched upon a fundamental question of algebra: When factoring, how much flexibility do we allow for the coefficients to change (fields)?

We agree that all factorizations will take place in the real number field $\mathbb{R}$. The study of factorizations over the complex field (notably the so-called Fundamental Theorem of Algebra), and, more generally, the study of how the fields change under factorizations belongs to Galois Theory.[15]

There are many beautiful methods and tricks in polynomial factorization. In this section we discuss some basic factoring methods.

[15] For a much more detailed account, see the author's *Glimpses of Algebra and Geometry*, 2nd ed. Springer, New York, 2002.

History
Polynomial factorization using modern symbolism (representing indeterminates and constants by symbols) could not possibly have come into existence before the 17th century. The first algorithm for factoring polynomials is due to the German mathematician Hermann Schubert (1848–1911). Kronecker not only rediscovered the original algorithm of Schubert, but also extended it to polynomials with several indeterminates. Kronecker also realized that for factorization the field for the coefficients often needs to be extended.

Example 6.4.1 Factor the polynomial $x^4 - 20x^2 + 4$ over the integers $\mathbb{Z}$.

The idea is to use the quadratic binomial formula to write this as the difference of squares:

$$x^4 - 20x^2 + 4 = x^4 - 4x^2 + 4 - 16x^2 = (x^2 - 2)^2 - (4x)^2$$
$$= (x^2 - 4x - 2)(x^2 + 4x - 2).$$

An interesting byproduct of this is the fact that, for all $n \in \mathbb{N}$, the number $n^4 - 20n^2 + 4$ is always composite.[16] Indeed, by the above, we have

$$n^4 - 20n^2 + 4 = (n^2 + 4n - 2)(n^2 - 4n - 2),$$

and neither factors are equal to ± 1. ($n^2 \pm 4n - 2 = \pm 1$ would mean $n(n \pm 4) = 2 \pm 1$, which are impossible for $n \in \mathbb{N}$.)

The simplest factoring techniques include identifying **common multiples** and **grouping** monomials within the polynomial. We begin here with a simple example as follows:

Example 6.4.2 Factor the cubic polynomial $x^3 - x^2 + x - 1$.

First Solution. We pair the first two terms and the last two terms. This gives $x^2(x - 1) + 1(x - 1)$. Hence $(x - 1)$ is a common factor, and we arrive at $x^3 - x^2 + x - 1 = (x - 1)(x^2 + 1)$.

Second Solution. We write this polynomial as $-(1 - x + x^2 - x^3)$ and recognize a finite geometric series with ratio $-x$ (in the parentheses). After simplification, the Finite Geometric Series Formula gives $x^4 - 1 = (x + 1)(x^3 - x^2 + x - 1)$. On the other hand, the polynomial on the left-hand side can be written as a difference of squares

$$x^4 - 1 = (x^2)^2 - 1 = (x^2 - 1)(x^2 + 1) = (x - 1)(x + 1)(x^2 + 1).$$

Finally, we cancel the initial factor $(x + 1)$, and arrive at the factorization $x^3 - x^2 + x - 1 = (x - 1)(x^2 + 1)$.

Example 6.4.3 Show that $4x - x^4 \leq 3$, $x \in \mathbb{R}$.

The crux here is to factor the quartic polynomial $p(x) = x^4 - 4x + 3$ by adding and subtracting suitable terms

[16] See also the Crux Mathematicorum (Canadian Mathematical Society), June/July 1978.

$$p(x) = x^4 - 4x + 3 = x^4 - 2x^2 + 1 + 2x^2 - 4x + 2 = (x^2 - 1)^2 + 2(x - 1)^2 \geq 0.$$

The example follows.

The next example is somewhat subtle and will be important later:

Example 6.4.4 Factor the quintic polynomial $x^5 + x - 1$.

This polynomial does not have monomials of degrees 2, 3, and 4. We insert them in opposite pairs

$$x^5 + x - 1 = x^5 - x^4 + x^3 + x^4 - x^3 + x^2 - x^2 + x - 1.$$

We now group and factor

$$\begin{aligned} x^5 + x - 1 &= (x^5 - x^4 + x^3) + (x^4 - x^3 + x^2) - (x^2 - x + 1) \\ &= x^3(x^2 - x + 1) + x^2(x^2 - x + 1) - (x^2 - x + 1) \\ &= (x^3 + x^2 - 1)(x^2 - x + 1). \end{aligned}$$

Remark We may wonder if the last product is the complete factorization of the quintic $x^5 + x - 1$. It is not. While the (second) quadratic factor is irreducible (over $\mathbb{R}$), the (first) cubic factor can further be split into a linear factor and another quadratic factor. We will discuss this later in more details.

For polynomials of several indeterminates, we sporadically encounter factorization problems where we can use our basic identities above. Here we assemble a few illustrative examples.

Example 6.4.5 Factor the polynomial $x^4 - y^4$.

We use the difference of squares identity as follows:

$$x^4 - y^4 = (x^2)^2 - (y^2)^2 = (x^2 - y^2)(x^2 + y^2) = (x - y)(x + y)(x^2 + y^2).$$

A more illuminating example is the following:

Example 6.4.6 Factor the quartic polynomial $x^4 + y^4$.

We may initially be discouraged by noticing that, with the substitutions $a = x^2$ and $b = y^2$, our polynomial can be written as $x^4 + y^4 = (x^2)^2 + (y^2)^2 = a^2 + b^2$, and we have seen above that $a^2 + b^2$ is irreducible.

To introduce a different idea, we add and subtract the term $2x^2y^2$, group, and use our basic identities:

$$\begin{aligned} x^4 + y^4 &= x^4 + 2x^2y^2 + y^4 - 2x^2y^2 = (x^2)^2 + 2x^2y^2 + (y^2)^2 - 2x^2y^2 \\ &= (x^2 + y^2)^2 - (\sqrt{2}xy)^2 = (x^2 + y^2 - \sqrt{2}xy)(x^2 + y^2 + \sqrt{2}xy). \end{aligned}$$

Factoring polynomials of higher degree with simple structure is based on reducing their monomials to lower degrees.

Example 6.4.7 Factor the sextic polynomial $x^6 - y^6$.

We calculate

$$x^6 - y^6 = (x^3)^2 - (y^3)^2 = (x^3 - y^3)(x^3 + y^3)$$
$$= (x - y)(x^2 + xy + y^2)(x + y)(x^2 - xy + y^2),$$

where in the last step we used our basic cubic identities.

Example 6.4.8 Factor the quartic polynomial $x^4 + x^2y^2 + y^4$.

We can use the method of Example 6.4.6 as follows:

$$x^4 + x^2y^2 + y^4 = (x^2 + y^2)^2 - x^2y^2 = (x^2 + xy + y^2)(x^2 - xy + y^2).$$

A different method is the following. Substituting $a = x^2$ and $b = y^2$, our polynomial becomes $x^4 + x^2y^2 + y^4 = a^2 + ab + b^2$. This is the quadratic factor in the identity $a^3 - b^3 = (a-b)(a^2+ab+b^2)$. Returning to our original indeterminates x and y, we thus have

$$x^6 - y^6 = (x^2 - y^2)(x^4 + x^2y^2 + y^4) = (x - y)(x + y)(x^4 + x^2y^2 + y^4).$$

On the other hand, by the previous example, we have

$$x^6 - y^6 = (x - y)(x^2 + xy + y^2)(x + y)(x^2 - xy + y^2).$$

Comparing these two results, we arrive at the factorization

$$x^4 + x^2y^2 + y^4 = (x^2 + xy + y^2)(x^2 - xy + y^2).$$

Factorization is an indispensable tool in solving equations with several indeterminates. The following example illustrates this.

Example 6.4.9 Find all integer solutions $x, y, z \in \mathbb{Z}$ of the equation[17]

$$x^3 - y^3 + z^3 = (x - y + z)^3.$$

We rewrite this as

$$x^3 - y^3 = (x - y + z)^3 - z^3$$

[17] A variant of this problem is in the Crux Mathematicorum (Canadian Mathematical Society), April 1979.

and factor

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

This gives $x = y$, and

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

Expanding and simplifying, the equation reduces to $(z-y)(z+x) = 0$. We conclude that the general solution is $x = y$, or $y = z$, or $x = -z$, and the missing variable is arbitrary.

Example 6.4.10 Given that $x^2+y^2+z^2 = 1$, $x, y, z \in \mathbb{R}$, what is the minimum value of $xy+yz+zx$?[18]

We have

$$0 \le (x+y+z)^2 = x^2+y^2+z^2+2(xy+yz+zx) = 1+2(xy+yz+zx).$$

Hence the minimum value is $-1/2$.

Exercises

6.4.1. Factor the polynomial $x^3y^3-x^3-y^3+1$.
6.4.2. For a given $a \in \mathbb{R}$, solve $(x+1)(x+a)(x+a+2)(x+2a+1) = a^2$.

6.5 The Division Algorithm for Polynomials

In Section 1.3 we introduced and studied the division algorithm for integers. There is also a division algorithm for polynomials.

Division Algorithm[19] **(Polynomials).** *For any polynomials $n(x)$ and $d(x) \neq 0$, there exist unique polynomials $q(x)$ and $r(x)$ such that*

$$n(x) = q(x) \cdot d(x) + r(x),$$

[18]This example is usually treated in multivariate calculus as a simple example of the Lagrange multipliers method. It was also posed as a problem (without the use of calculus) in the $MA\Theta$ National Convention, 1987.

[19]Sometimes called "Euclidean Division." Since the proof captures the pivotal step of the associated computational algorithm, usually termed as the "Long Division Algorithm," and also due to the close analogy with integers, we kept the term "Division Algorithm" for polynomials as well.

where

$$r(x) = 0 \quad \text{or} \quad \deg r(x) < \deg d(x).$$

Remark The polynomial $n(x)$ is the **dividend** and the non-zero polynomial $d(x)$ is the **divisor**. Upon division we obtain the **quotient** $q(x)$ and the **remainder** $r(x)$ satisfying the division algorithm formula above.

Proof We may assume that $\deg d(x) > 0$. The proof of existence is by induction with respect to the degree of the dividend $n(x)$. (If $n(x) = 0$ then $q(x) = r(x) = 0$.)

If $\deg n(x) = 0$, $n(x) \neq 0$, then $q(x) = 0$ and $r(x) = n(x)$, and the division algorithm formula follows.

For the general induction step $0, 1, 2, \ldots, n-1 \Rightarrow n$ assume that the division algorithm formula holds for all polynomials $n(x)$ with $\deg n(x) < n$, $n \in \mathbb{N}$.

Let $n(x)$ be a polynomial of degree n. We set

$$n(x) = a_n x^n + a_{n-1} x^{n-1} + \cdots + a_0, \quad a_n \neq 0,$$

and

$$d(x) = b_m x^m + b_{m-1} x^{m-1} + \cdots + b_0, \quad b_m \neq 0.$$

If $n < m$, then $q(x) = 0$ and $n(x) = r(x)$ satisfy the division algorithm formula. Thus, we may assume $n \geq m$.

We have

$$\frac{a_n}{b_m} x^{n-m} d(x) = a_n x^n + \text{lower order terms}$$

Since the leading term of this polynomial is the **same** as that of $n(x)$, the polynomial

$$n(x) - \frac{a_n}{b_m} x^{n-m} d(x)$$

has degree less than n. The induction hypothesis applies, and we have

$$n(x) - \frac{a_n}{b_m} x^{n-m} d(x) = q'(x) \cdot d(x) + r(x),$$

where either $r(x) = 0$ or $\deg r(x) < \deg d(x)$. Rearranging, we obtain

$$n(x) = \left(\frac{a_n}{b_m} x^{n-m} + q'(x) \right) d(x) + r(x).$$

Existence of the division algorithm follows with

$$q(x) = \frac{a_n}{b_m} x^{n-m} + q'(x).$$

To show unicity, assume that

$$n(x) = q(x) \cdot d(x) + r(x) = q'(x) \cdot d(x) + r'(x),$$

where $r(x)$ and $r'(x)$ are either zero or have degrees less than the degree of $d(x)$. These give

$$(q(x) - q'(x))d(x) = r'(x) - r(x).$$

The degree of the polynomial on the right-hand side is less than the degree of $d(x)$. The only way this is possible is that $q(x) = q'(x)$. This implies $r(x) = r'(x)$. Unicity of the division algorithm follows.

Starting with a dividend $n(x)$ and a divisor $d(x)$, the process that results in the quotient $q(x)$ and remainder $r(x)$ is via the well-known **Long Division Algorithm**. This algorithm is based on progressively matching the leading terms of $n(x)$ and its successors with the leading term of $d(x)$, and it is essentially contained in the main induction step of the proof above.

Example 6.5.1 What is the sum of all $n \in \mathbb{Z}$ such that $n^2 + 2n + 2$ divides $n^3 + 4n^2 + 4n - 14$?

We replace n by the real indeterminate $x \in \mathbb{R}$ to obtain polynomials. We divide the polynomial $x^3 + 4x^2 + 4x - 14$ by $x^2 + 2x + 2$ using long division:

$$\begin{array}{rr} & x \;+2 \\ \hline x^2+2x+2\big) & x^3+4x^2+4x-14 \\ & -x^3-2x^2-2x \\ \hline & 2x^2+2x-14 \\ & -2x^2-4x\;\;-4 \\ \hline & -2x-18 \end{array}$$

In terms of the original $n \in \mathbb{Z}$, this gives

$$n^3 + 4n^2 + 4n - 14 = (n^2 + 2n + 2)(n + 2) - (2n + 18), \quad n \in \mathbb{Z}.$$

The divisibility requirement implies $n^2 + 2n + 2 | 2n + 18$, and hence $|2n + 18| \geq |n^2 + 2n + 2|$ or $2n + 18 = 0$. Since $n^2 + 2n + 2 = (n + 1)^2 + 1 > 0$, the inequality reduces to $\pm(2n + 18) > n^2 + 2n + 1$. The negative sign is clearly not realized, so that (with the positive sign) we end up with $-4 \leq n \leq 4$. Of these values the divisibility condition gives $n = -4, -2, -1, 0, 1, 4$. In addition, $2n + 18 = 0$ gives $n = -9$, and this also satisfies the divisibility condition. With these the sum is -11.

Example 6.5.2 Perform the multiplication in the shortest[20] possible way in expanding the product:

$$(1+x+x^2+x^3+x^4+x^5)(1-x+x^2-x^3+x^4-x^5).$$

We use the Finite Geometric Series Formula as follows:

$$1+x+x^2+x^3+x^4+x^5=\frac{x^6-1}{x-1} \quad \text{and} \quad 1-x+x^2-x^3+x^4-x^5=-\frac{x^6-1}{x+1},$$

where the second formula is obtained from the first by replacing x with its negative $-x$. Multiplying, we obtain

$$(1+x+x^2+x^3+x^4+x^5)(1-x+x^2-x^3+x^4-x^5)=-\frac{(x^6-1)^2}{x^2-1}=-\frac{x^{12}-2x^6+1}{x^2-1}.$$

We divide $x^{12}-2x^6+1$ by x^2-1 using long division, and get

$$x^{12}-2x^6+1=(x^2-1)(x^{10}+x^8+x^6-x^4-x^2-1).$$

Since we have a zero remainder, we arrive at

$$(1+x+x^2+x^3+x^4+x^5)(1-x+x^2-x^3+x^4-x^5)=-x^{10}-x^8-x^6+x^4+x^2+1.$$

The special case of the Long Division Algorithm when the divisor is **linear** is of great interest. In this case the process can be compressed into a much shorter algorithm called **Synthetic Division**.

If $d(x)=x-c$, $c\in\mathbb{R}$, then, for a given dividend $n(x)$ of degree n, the Division Algorithm gives

$$n(x)=(x-c)q(x)+r,$$

where the remainder $r\in\mathbb{R}$ must be a **constant** (since the divisor is linear).

We now let $n(x)=a_nx^n+a_{n-1}x^{n-1}+\cdots+a_1x+a_0$ and $q(x)=b_{n-1}x^{n-1}+b_{n-2}x^{n-2}+\cdots+b_1x+b_0$ and calculate

$$\begin{aligned}
&a_nx^n+a_{n-1}x^{n-1}+\cdots+a_1x+a_0\\
&\quad=(x-c)(b_{n-1}x^{n-1}+b_{n-2}x^{n-2}+\cdots+b_1x+b_0)+r\\
&\quad\quad=b_{n-1}x^n+(b_{n-2}-cb_{n-1})x^{n-1}+\cdots+(b_0-cb_1)x+(r-cb_0).
\end{aligned}$$

[20] Expanding and using the Cauchy Product Rule would amount to work out 36 terms.

A simple comparison of coefficients gives

$$\begin{aligned} a_n &= b_{n-1} \\ a_{n-1} & \\ &\cdots \\ a_1 &= b_0 - cb_1 \\ a_0 &= r - cb_0. \end{aligned}$$

Inverting, we obtain

$$\begin{aligned} b_{n-1} &= a_n \\ b_{n-2} &= a_{n-1} + cb_{n-1} \\ &\cdots \\ b_0 &= a_1 + cb_1 \\ r &= a_0 + cb_0. \end{aligned}$$

The whole process with all these data can be conveniently tabulated as follows:

$$\begin{array}{c|ccccccc} & a_n & & a_{n-1} & & a_{n-2} \cdots & a_1 & & a_0 \\ & & & + \downarrow & & + \downarrow \cdots & + \downarrow & & + \downarrow \\ c & \downarrow & & cb_{n-1} & & cb_{n-2} \cdots & cb_1 & & cb_0 \\ \hline & b_{n-1} & \nearrow & b_{n-2} & \nearrow & b_{n-3} \cdots & b_0 & \nearrow & r \end{array}$$

The quotient can then be reconstructed from the bottom register as $q(x) = b_{n-1}x^{n-1} + b_{n-2}x^{n-2} + \cdots + b_1x + b_0$ while the remainder r appears as the last entry.

Example 6.5.3 What is the largest $n \in \mathbb{N}$ such that $n^3 - 100$ is divisible by $n - 10$?

Once again we replace $n \in \mathbb{N}$ by the real indeterminate $x \in \mathbb{R}$. We use synthetic division to divide the cubic polynomial $x^3 - 100$ by $x - 10$. We obtain[21]

$$\begin{array}{r|cccc} & 1 & 0 & 0 & -100 \\ 10 & & 10 & 100 & 1000 \\ \hline & 1 & 10 & 100 & 900 \end{array}$$

This gives

$$\frac{x^3 - 100}{x - 10} = x^2 + 10x + 100 + \frac{900}{x - 10}.$$

Going back to $x = n \in \mathbb{N}$, we see that $n = 910$.

[21] Note the somewhat different layout of the synthetic division in LaTex.

A polynomial $p(x)$ can be evaluated at a number $c \in \mathbb{R}$ by substitution to obtain $p(c)$. The polynomial $p(x)$ can also be divided by $x - c$, and the remainder r will be a constant (since the divisor is linear). By the Remainder Theorem, these two numbers are equal.

Remainder Theorem. *Let $c \in \mathbb{R}$. When a polynomial $p(x)$ is divided by the linear polynomial $x - c$, then the remainder of the division is equal to $p(c)$.*

Proof This is an immediate consequence of the Division Algorithm

$$p(x) = (x - c)q(x) + r.$$

Substituting $x = c$, we obtain $r = p(c)$.

A typical application of the Remainder Theorem is to obtain the value of a polynomial at a number by performing a usually faster synthetic division.

Example 6.5.4 (Revisited) In Example 6.1.2 we can use synthetic division to obtain the values of the polynomial $x^2 + x + 41$ at $c = 38, 39, 40$ as follows:

$$\begin{array}{r|rrr} 38 & 1 & 1 & 41 \\ & & 38 & 1482 \\ \hline & 1 & 39 & 1523 \end{array} \qquad \begin{array}{r|rrr} 39 & 1 & 1 & 41 \\ & & 39 & 1560 \\ \hline & 1 & 40 & 1601 \end{array} \qquad \begin{array}{r|rrr} 40 & 1 & 1 & 41 \\ & & 40 & 1640 \\ \hline & 1 & 41 & 1681 \end{array}$$

In the special case when the remainder is zero, $r = p(c) = 0$, then the divisor $x - c$ becomes a factor of $p(x)$. This, the so-called **Factor Theorem**, is of great importance since it provides a link between the roots of a polynomial and its linear factors.

Factor Theorem. *A number $c \in \mathbb{R}$ is a root of a polynomial $p(x)$ if and only if $x - c$ divides $p(x)$.*

The Factor Theorem along with synthetic division can be used to obtain some of our earlier identities. To illustrate this we return to Example 6.4.2:

Example 6.5.5 (Revisited) Derive the complete factorization of the cubic polynomial $x^3 - x^2 + x - 1$.

Clearly, $x = 1$ is a root since $1^3 - 1^2 + 1 - 1 = 0$. We now use synthetic division as follows

$$\begin{array}{r|rrrr} 1 & 1 & -1 & 1 & -1 \\ & & 1 & 0 & 1 \\ \hline & 1 & 0 & 1 & 0 \end{array}$$

The coefficients of the quotient are displayed in the bottom register, $q(x) = x^2 + 1$, and the remainder is zero. This gives the factorization

$$x^3 - x^2 + x - 1 = (x - 1)(x^2 + 1).$$

As for another simple example, 1 is clearly a root of the polynomial $p(x) = x^n - 1$. Performing synthetic division

$$\begin{array}{c|cccccc} & 1 & 0 & 0 & \cdots & 0 & -1 \\ 1 & & 1 & 1 & \cdots & 1 & 1 \\ \hline & 1 & 1 & 1 & \cdots & 1 & 0 \end{array}$$

we obtain the quotient $q(x) = x^{n-1} + x^{n-2} + \cdots + x + 1$. All these can be compactly expressed via the factorization

$$x^n - 1 = (x-1)(x^{n-1} + x^{n-2} + \cdots + x^2 + x + 1).$$

Dividing by $x - 1$ (assuming $x \neq 1$) and moving up the value of n by one, we rediscover the **Finite Geometric Series Formula**

$$1 + x + x^2 + \cdots + x^{n-1} + x^n = \frac{1 - x^{n+1}}{1 - x}.$$

As demonstrated previously, this formula has many beautiful applications. As another illustrative example, a quick look gives the following identity

$$(x+1)(1+x^2+x^4+\cdots+x^{2n-2}) = (x^n+1)(1+x+x^2+\cdots+x^{n-1}), \quad n \in \mathbb{N}.$$

Indeed, multiplication by x in the first factor on the left-hand side gives the odd power terms, while multiplication by x^n in the first factor on the right-hand side gives the portion of the geometric sequence from the exponents n to $2n - 1$.

This identity can also be obtained by a less ad hoc way as follows. The Finite Geometric Series Formula gives

$$x^n - 1 = (x-1)(1 + x + x^2 + \cdots + x^{n-1}).$$

Replacing x by x^2, we also have

$$\begin{aligned} x^{2n} - 1 = (x^2)^n - 1 &= (x^2 - 1)(1 + x^2 + (x^2)^2 + \cdots + (x^2)^{n-1}) \\ &= (x^2 - 1)(1 + x^2 + x^4 + \cdots + x^{2n-2}). \end{aligned}$$

Combining these with $x^{2n} - 1 = (x^n - 1)(x^n + 1)$ and $x^2 - 1 = (x - 1)(x + 1)$, the identity above follows.

The next example is a direct consequence of this:[22]

Example 6.5.6 For what $n \in \mathbb{N}$ (if any) is $1 + x + x^2 + \cdots + x^{n-1}$ a factor of $1 + x^2 + x^4 + \cdots + x^{2n-2}$?

[22] See also the Mathematical Olympiad Program, 1997.

By the identity above, $1 + x^2 + x^4 + \cdots + x^{2n-2}$ is divisible by $1 + x + x^2 + \cdots + x^{n-1}$ if and only if -1 is a root of $x^n + 1$ if and only if $n \in \mathbb{N}$ is **odd**.

An immediate and important consequence of the Factor Theorem is that a degree n polynomial $p(x)$ can have **at most** n (real) roots.

Before showing this we introduce the following definition. A root c of a polynomial $p(x)$ has **multiplicity** $m \in \mathbb{N}$ if

$$p(x) = (x - c)^m q(x) \text{ and } q(c) \neq 0.$$

By the Factor Theorem, $m \in \mathbb{N}$ is the **largest** integer such that $(x-c)^m$ divides $p(x)$ (with zero remainder). Clearly, the quotient polynomial $q(x)$ has degree $n - m$.

The process of dividing the polynomial by the root factor can be performed inductively. If c_1 is a root of $p(x)$ with multiplicity m_1, then we have

$$p(x) = (x - c_1)^{m_1} p_1(x),$$

where the quotient $p_1(x)$ (renamed) has degree $n - m_1$. Now, if c_2 is another root of $p(x)$ (different from c_1), then we have

$$p(c_2) = (c_2 - c_1)^{m_1} p_1(c_2) = 0.$$

Since $c_1 \neq c_2$ we see that c_2 is a root of $p_1(x)$. If c_2 is of multiplicity m_2 (as a root of $p_1(x)$ and hence also as a root of $p(x)$) then, dividing by the corresponding root factor, we obtain

$$p(x) = (x - c_1)^{m_1} (x - c_2)^{m_2} p_2(x),$$

where the quotient $p_2(x)$ is of degree $n - m_1 - m_2$. This process must end after finitely many steps, and we obtain

$$p(x) = (x - c_1)^{m_1} (x - c_2)^{m_2} \cdots (x - c_k)^{m_k} q(x),$$

where $q(x)$ has no real roots. Since the degree of $q(x)$ is $n - m_1 - m_2 - \cdots - m_k \geq 0$, we obtain

$$m_1 + m_2 + \cdots + m_k \leq n.$$

This is actually a stronger statement than the one we made above: A degree n polynomial has at most n roots **counted with multiplicity**.

An illustrative example for roots with multiplicity is as follows:

Example 6.5.7 For $n \in \mathbb{N}$, consider the degree $n + 1$ polynomial $p(x) = x^{n+1} - (n + 1)x + n$. Clearly, $c = 1$ is a root. We perform synthetic division of $p(x)$ by the corresponding root factor $x - 1$:

$$\begin{array}{c|ccccccc} & 1 & 0 & 0 & \cdots & 0 & -(n+1) & n \\ 1 & & 1 & 1 & \cdots & 1 & 1 & -n \\ \hline & 1 & 1 & 1 & \cdots & 1 & -n & 0 \end{array}$$

We obtain

$$p(x) = x^{n+1} - (n+1)x + n = (x-1)(x^n + x^{n-1} + \cdots + x - n).$$

We see that the quotient still has $c = 1$ as a root.

Performing yet another synthetic division, we obtain

$$\begin{array}{c|ccccccc} & 1 & 1 & 1 & \cdots & 1 & 1 & -n \\ 1 & & 1 & 2 & \cdots & n-2 & n-1 & n \\ \hline & 1 & 2 & 3 & \cdots & n-1 & n & 0 \end{array}$$

We arrive at the factorization

$$p(x) = x^{n+1} - (n+1)x + n = (x-1)^2(x^{n-1} + 2x^{n-2} + 3x^{n-3} + \cdots + (n-1)x + n).$$

Since $c = 1$ is not a root of the quotient, we conclude that it is a root of $p(x)$ with multiplicity 2.

As an interesting consequence, we obtain

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

where T_n, $n \in \mathbb{N}$, is the nth triangular number discussed in Section 0.4.

A somewhat more involved variation on the theme (of the last limit) is the following:

Example 6.5.8 Given $m < n$, $m, n \in \mathbb{N}$, calculate the limit

$$\lim_{x \to 1} \left(\frac{m}{x^m - 1} - \frac{n}{x^n - 1} \right).$$

We rewrite the expression in the limit as follows

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

The crux is that the polynomial numerator $m(x^n - 1) - n(x^m - 1)$ has $x = 1$ as a root with multiplicity 2. First, we use the Finite Geometric Series Formula to divide the numerator by $x - 1$ and obtain

$$\begin{aligned}
&\frac{m}{x^m-1}-\frac{n}{x^n-1}\\
&=\frac{m(x^{n-1}+x^{n-2}+\cdots+x+1)-n(x^{m-1}+x^{m-2}+\cdots+x+1)}{(x-1)(x^{n-1}+x^{n-2}+\cdots+x+1)(x^{m-1}+x^{m-2}+\cdots+x+1)}\\
&=\frac{m(x^{n-1}+x^{n-2}+\cdots+x^m)-(n-m)(x^{m-1}+x^{m-2}+\cdots+x+1)}{(x-1)(x^{n-1}+x^{n-2}+\cdots+x+1)(x^{m-1}+x^{m-2}+\cdots+x+1)}
\end{aligned}$$

Second, we use synthetic division to divide the numerator of the last expression by $x-1$ and obtain the quotient

$$\begin{aligned}
&mx^{n-2}+2mx^{n-3}+\cdots+(n-m)mx^{m-1}\\
&\qquad+(n-m)(m-1)x^{m-2}+(n-m)(m-2)x^{m-3}+\cdots+(n-m).
\end{aligned}$$

This, and the factor $(x^{n-1}+x^{n-2}+\cdots+x+1)(x^{m-1}+x^{m-2}+\cdots+x+1)$ in the denominator in the last expression are non-zero at $x=1$. We obtain

$$\begin{aligned}
&\lim_{x\to 1}\left(\frac{m}{x^m-1}-\frac{n}{x^n-1}\right)\\
&=\frac{m(1++2+\cdots+(n-m))+(n-m)((m-1)+(m-2)+\cdots+1)}{nm}.
\end{aligned}$$

We now use the formula $1+2+\cdots+k=k(k+1)/2$, $k\in\mathbb{N}$, for the nth triangular number T_n in two instances (Section 0.4), for $k=n-m$ and $k=m-1$, and finally obtain

$$\lim_{x\to 1}\left(\frac{m}{x^m-1}-\frac{n}{x^n-1}\right)=\frac{m\cdot\frac{(n-m)(n-m+1)}{2}+(n-m)\cdot\frac{(m-1)m}{2}}{nm}=\frac{n-m}{2}.$$

There are many problems in mathematical contests involving some given values of a polynomial and asking to find the value of the polynomial in yet another value. Although this seems to relate to the Lagrange interpolation polynomial in Example 6.1.1, the solution is often effected by constructing another polynomial. The following example illustrates this.

Example 6.5.9 Let $a,b\in\mathbb{R}$, and $p(x)$ a degree n polynomial such that $p(1)=p(2)=\cdots=p(n)=a$ and $p(n+1)=b$. Find $p(0)$.

By the first condition, $1,2,\ldots,n$ are roots of the polynomial $q(x)=p(x)-a$. Since $q(x)$ also has degree n, we have $q(x)=c(x-1)(x-2)\cdots(x-n)$, where $c\in\mathbb{R}$ is the leading coefficient of $q(x)$. Evaluating $q(x)$ at $n+1$, we obtain $q(n+1)=c\cdot n!=b$, and hence $c=b/n!$. This gives $q(x)=(b/n!)(x-1)(x-2)\cdots(x-n)$. Finally, we arrive at $p(0)=a+q(0)=a+b(-1)(-2)\cdots(-n)/n!=a+(-1)^n\cdot b$.

Example 6.5.10 Determine $a, b \in \mathbb{R}$ such that the quartic polynomial $p(x) = x^4 - 24x^3 + 54x^2 + ax + b$ has **two** double roots.[23]

Letting $r, s \in \mathbb{R}$ denote the double roots, by assumption, we have

$$p(x) = x^4-24x^3+54x^2+ax+b = (x-r)^2(x-s)^2 = (x^2-2rx+r^2)(x^2-2sx+s^2).$$

Expanding and comparing coefficients, we obtain

$$r + s = 12 \quad \text{and} \quad r^2 + 4rs + s^2 = 54.$$

Subtracting the square of the first equation from the second, we get $rs = -45$. This, along with the first equation, give $r = 15$ and $s = -3$ (or $r = -3$ and $s = 15$). Thus, our polynomial rewrites as

$$p(x)=x^4-24x^3+54x^2+ax+b = (x-15)^2(x+3)^2=(x^2-30x+225)(x^2+6x+9).$$

Once again, expanding and comparing coefficients, we obtain

$$a = -30 \cdot 9 + 225 \cdot 6 = 1080 \quad \text{and} \quad b = 225 \cdot 9 = 2025.$$

As a simple application of the Division Algorithm, we now claim that, for **any** polynomial $n_0(x)$ of degree $\leq k - 1$ and $c \in \mathbb{R}$, there exist $A_1, A_2, \ldots, A_k \in \mathbb{R}$ such that

$$\frac{n_0(x)}{(x-c)^k} = \frac{A_1}{x-c} + \frac{A_2}{(x-c)^2} + \cdots + \frac{A_k}{(x-c)^k}.$$

This is a special case of the **partial fraction decomposition** to be discussed in Section 9.2.

For the proof, we eliminate the fractions by multiplying both sides by $(x - c)^k$, and obtain the equivalent form

$$n_0(x) = A_1(x-c)^{k-1} + A_2(x-c)^{k-2} + \cdots + A_{k-1}(x-c) + A_k.$$

Now, it is clear that the coefficients $A_1, A_2, \ldots, A_k \in \mathbb{R}$ are those of the expansion of the polynomial $n_0(x + c)$ into powers of x. The claim follows.

Exercises

6.5.1. Find all real roots of the sextic polynomial

$$p(x) = x^6 - x^5 - 3x^4 + 2x^3 + 3x^2 - x - 1.$$

[23] A similar problem (to calculate only $a + b$) was in the $MA\Theta$ National Convention, 1991.

6.5.2. Find all natural numbers $n \in \mathbb{N}$ such that $n + 5$ divides $n^2 + 15$.
6.5.3. For $n \in \mathbb{N}$, let

$$p(x) = x^{n+2} + x^{n+1} - (n+1)^2 x^2 + (2n(n+1) - 1)x - n^2.$$

Show that $p(x) = (x-1)^3 q(x)$ where

$$q(x) = 1^2 \cdot x^{n-1} + 2^2 \cdot x^{n-2} + \cdots + (n-1)^2 \cdot x + n^2.$$

In particular

$$\lim_{x \to 1} \frac{x^{n+2} + x^{n+1} - (n+1)^2 x^2 + (2n(n+1) - 1)x - n^2}{(x-1)^3}$$
$$= 1^2 + 2^2 + \cdots + (n-1)^2 + n^2 = \frac{n(n+1)(2n+1)}{6}.$$

6.5.4. Factor the polynomial $x^{10} + x^5 + 1$ into a product of two factors.

6.6 Symmetric Polynomials

Polynomials can be composed to form other polynomials. In general, if

$$p_1(x_1, \ldots, x_n),\ p_2(x_1, \ldots, x_n),\ p_m(x_1, \ldots, x_n),$$

are polynomials in the indeterminates $x_1, \ldots, x_n$, and $q(u_1, \ldots, u_m)$ is a polynomial in the indeterminates $u_1, \ldots, u_m$, then we can form the composition

$$q(p_1(x_1, \ldots, x_n), \ldots, p_m(x_1, \ldots, x_n)).$$

This composition is a polynomial is the indeterminates $x_1, \ldots, x_n$.

We have already seen simple examples of this. For $n \in \mathbb{N}$, the polynomial $(x + y)^n$ (in the binomial formula) is the composition of the linear polynomial $p(x) = x + y$ and the power function $\mathfrak{p}_n(u) = u^n$. In another example, the polynomial $(1 + x/365)^{365}$ is the composition of the linear polynomial $p(x) = 1 + x/365$ and the power function $\mathfrak{p}_{365}(u) = u^{365}$. In this section we will discuss an important application for symmetric polynomials.

A polynomial $p(x_1, \ldots, x_n)$ is called **symmetric** if it remains the same under any permutation of the indeterminates; that is, if $p(x_{\pi(1)}, \ldots, x_{\pi(n)}) = p(x_1, \ldots, x_n)$, for any permutation $\pi : \{1, \ldots, n\} \to \{1, \ldots, n\}$.

Example 6.6.1 The polynomial $(x + y)^n$, $n \in \mathbb{N}$, is symmetric. The polynomial $x^2/a^2 + y^2/b^2 - 1$ is symmetric if and only if $a = b$. In three indeterminates, the polynomial $x^3 + y^3 + z^3 - 3xyz$ is symmetric while $x^n + y^n - z^n$, $n \in \mathbb{N}$, is not.

We define the **elementary symmetric polynomials** $s_k(x_1, \ldots, x_n)$, $k = 1, \ldots, n$, in the n indeterminates $x_1, \ldots, x_n$ by

$$s_k(x_1, \ldots, x_n) = \sum_{1 \le j_1 < \cdots < j_k \le n} x_{j_1} \cdots x_{j_k}.$$

(The sum is over all products of k-element subsets of the indeterminates $x_1, \ldots, x_n$.)

More explicitly, we have

$$\begin{aligned} s_1(x_1, \ldots, x_n) &= \sum_{j=1}^{n} x_j = x_1 + \cdots + x_n \\ s_2(x_1, \ldots, x_n) &= \sum_{1 \le j < k \le n} x_j x_k = x_1 x_2 + \cdots + x_{n-1} x_n \\ &\ldots \quad \ldots \\ s_n(x_1, \ldots, x_n) &= x_1 \cdots x_n, \end{aligned}$$

and we add the constant polynomial $s_0(x_1, \ldots, x_n) = 1$ for completeness.

We now introduce the concept of homogeneity for polynomials that will be useful in many instances in the future. A polynomial $p(x_1, \ldots, x_n)$ is called **homogeneous of degree** $d \in \mathbb{N}_0$ if

$$p(tx_1, \ldots, tx_n) = t^d p(x_1, \ldots, x_n), \quad t \in \mathbb{R}.$$

Clearly, for $k = 0, 1, \ldots, n$, the elementary symmetric polynomial $s_k(x_1, \ldots, x_n)$ is homogeneous of degree k.

It is also clear that any polynomial $p(x_1, \ldots, x_n)$ can be written uniquely as the sum of homogeneous polynomials (of different degrees). We call these the homogeneous components of $p(x_1, \ldots, x_n)$. The degree d homogeneous component of a polynomial $p(x_1, \ldots, x_n)$ is simply the sum of all degree d monomials in $p(x_1, \ldots, x_n)$.

Finally, since permuting the indeterminates in a monomial does not change its degree, in the decomposition of a **symmetric** polynomial $p(x_1, \ldots, x_n)$ into homogeneous components, each homogeneous component is also symmetric.

Fundamental Theorem on Symmetric Polynomials. *Let $p(x_1, \ldots, x_n)$ be a symmetric polynomial. Then there exists a unique polynomial $q(u_1, \ldots, u_n)$ such that*

$$p(x_1, \ldots, x_n) = q(s_1(x_1, \ldots, x_n), \ldots, s_n(x_1, \ldots, x_n)).$$

Proof By the observation above, without loss of generality, we may restrict ourselves to **homogeneous** symmetric polynomials.

The proof is by Peano's Principle of Induction with respect to both the number of variables $n \in \mathbb{N}$ and the degree $d \in \mathbb{N}$ (of homogeneity). The theorem clearly holds for $n = 1$ and any degree $d \in \mathbb{N}_0$, and also for any $n \in \mathbb{N}$ and degrees $d = 0, 1$.

Assume that the theorem holds for all homogeneous symmetric polynomials of degree less than d, $2 \leq d \in \mathbb{N}$, and having less than n, $2 \leq n \in \mathbb{N}$, indeterminates.

Let $p(x_1, \ldots, x_n)$ be a degree d homogeneous symmetric polynomial. We first split the polynomial as

$$p(x_1, \ldots, x_n) = p_0(x_1, \ldots, x_n) + x_1 \cdots x_n \cdot p_1(x_1, \ldots, x_n),$$

where $p_0(x_1, \ldots, x_n)$ is the sum of those monomials in $p(x_1, \ldots, x_n)$ that have at least one indeterminate from $\{x_1, \ldots, x_n\}$ **missing**. We call $p_0(x_1, \ldots, x_n)$ the **lacunary part** of $p(x_1, \ldots, x_n)$. Since the rest of the monomials have all the indeterminates $x_1, \ldots, x_n$ present, these monomials are multiples of the product $x_1 \cdots x_n$. Thus, the splitting above follows. Note that if $d < n$ then $p_1 = 0$.

Clearly, the lacunary part $p_0(x_1, \ldots, x_n)$ is itself symmetric, and hence so is $p_1(x_1, \ldots, x_n)$.

Moreover, the sum of the monomials in the lacunary part $p_0(x_1, \ldots, x_n)$ in which the indeterminate x_n is missing is the polynomial $p(x_1, \ldots, x_{n-1}, 0)$.

It is an important fact that the polynomial $p(x_1, \ldots, x_{n-1}, 0)$ **uniquely** determines the lacunary part $p_0(x_1, \ldots, x_n)$. In other words, if $p'(x_1, \ldots, x_n)$ is another symmetric homogeneous polynomial (of degree d) such that $p(x_1, \ldots, x_{n-1}, 0) = p'(x_1, \ldots, x_{n-1}, 0)$, then we have $p_0(x_1, \ldots, x_n) = p'_0(x_1, \ldots, x_n)$. This follows from symmetry. Indeed, consider any monomial in $p_0(x_1, \ldots, x_n)$. It has (at least) one of the indeterminates missing, x_i, $i = 1, \ldots, n$, say, and therefore any permutation that carries i to n, also carries the respective monomial to one of the monomials in $p(x_1, \ldots, x_{n-1}, 0)$. Now, since $p(x_1, \ldots, x_{n-1}, 0) = p'(x_1, \ldots, x_{n-1}, 0)$, this transformed monomial also appears in $p'(x_1, \ldots, x_{n-1}, 0)$. The inverse of the permutation carries this back to the original monomial, and we see that this monomial is also in $p'_0(x_1, \ldots, x_n)$. The claim follows.

Since the polynomial $p_0(x_1, \ldots, x_{n-1}, 0)$ contains only $n - 1$ indeterminates, the induction hypothesis applies. Thus, we have

$$p_0(x_1, \ldots, x_{n-1}, 0) = q_0(s_1(x_1, \ldots, x_{n-1}), \ldots, s_{n-1}(x_1, \ldots, x_{n-1}))$$

for some polynomial $q_0(v_1, \ldots, v_{n-1})$.

Consider now the polynomial

$$r(x_1, \ldots, x_n) = q_0(s_1(x_1, \ldots, x_n), \ldots, s_{n-1}(x_1, \ldots, x_n))$$

in the indeterminates $x_1, \ldots, x_n$, where we moved up the number of variables in the elementary symmetric polynomials. This polynomial is symmetric and homogeneous of degree d. Moreover, we have

$$r(x_1, \ldots, x_{n-1}, 0) = p(x_1, \ldots, x_{n-1}, 0),$$

since $s_k(x_1, \ldots, x_{n-1}) = s_k(x_1, \ldots, x_{n-1}, 0)$, $k = 0, \ldots, n-1$. As shown above, this implies that the **lacunary part** $r_0(x_1, \ldots, x_n)$ of $r(x_1, \ldots, x_n)$ is equal to $p_0(x_1, \ldots, x_n)$. Hence, the difference $p(x_1, \ldots, x_n) - r(x_1, \ldots, x_n)$ has no lacunary part, and thereby it is a multiple of $x_1 \cdots x_n$. By the induction hypothesis, this implies that $p(x_1, \ldots, x_n) - r(x_1, \ldots, x_n)$ is a polynomial of the elementary symmetric polynomials in the indeterminates $x_1, \ldots, x_n$. Since $r(x_1, \ldots, x_n)$ is a polynomial of the elementary symmetric polynomials, so is $p(x_1, \ldots, x_n)$. The general induction step is complete.

The theorem follows.

Viète Relations. *If $r_1, r_2, \ldots, r_n \in \mathbb{R}$ are the roots*[24] *(with multiplicity) of a polynomial $p(x) = a_n x^n + a_{n-1} x^{n-1} + \cdots + a_1 x + a_0$, then we have*

$$s_k(r_1, r_2, \ldots, r_n) = (-1)^k \frac{a_{n-k}}{a_n}, \quad k = 1, 2, \ldots, n.$$

Proof We use the Factor Theorem to write $p(x)/a_n$ as

$$x^n + \frac{a_{n-1}}{a_n} x^{n-1} + \cdots + \frac{a_k}{a_n} x^k + \cdots + \frac{a_1}{a_n} x + \frac{a_0}{a_n} = (x - r_1)(x - r_2) \cdots (x - r_n).$$

We now expand the right-hand side as follows. We first number each pair of parentheses

$$\overbrace{(x - r_1)}^{1} \overbrace{(x - r_2)}^{2} \cdots \overbrace{(x - r_n)}^{n},$$

thus forming n brackets. To make a term in the expansion, from each bracket, we need to choose the indeterminate x or the negative of the respective root, and then multiply these choices together. The term obtained this way is of the form

$$(-1)^{n-k} r_{j_1} \cdots r_{j_{n-k}} x^k,$$

where $1 \le j_1 < \cdots < j_{n-k} \le n$ mark those brackets from which the corresponding root is chosen (and thereby the indeterminate x is chosen from the complementary brackets). For fixed $k = 0, \ldots, n$, the sum of these coefficients is exactly $(-1)^{n-k} s_{n-k}(r_1, \ldots, r_n) = a_k/a_n$. Swapping k and $n-k$, the Viète relations follow.

There are literally hundreds of mathematical contest problems centered around the Viète relations. Some, in addition, exploit the simple fact that, if $0 \neq r \in \mathbb{R}$ is a root of a polynomial $p(x) = a_n x^n + a_{n-1} x^{n-1} + \cdots + a_1 x + a_0$ then the reciprocal $1/r$ is a root of the polynomial $x^n p(1/x) = a_n + a_{n-1} x + \cdots + a_1 x^{n-1} + a_0 x^n$. The next example illustrates this.

[24] The statement holds for complex roots as well.

Example 6.6.2 Find all real solutions $x, y, z, w \in \mathbb{R}$ of the system[25]

$$x + y = z + w \quad \text{and} \quad \frac{1}{x} + \frac{1}{y} = \frac{1}{z} + \frac{1}{w}.$$

Consider the cubic polynomial $p(t) = t^3 + at^2 + bt + c$, $a, b, c \in \mathbb{R}$, with roots $x, y, -z$. The Factor Theorem gives

$$p(t) = (t - x)(t - y)(t + z).$$

By the first Viète relation, we have

$$x + y - z = w = -a.$$

Using the remark above, the reciprocals $1/x, 1/y, -1/z$ are roots of the cubic polynomial $t^3 p(1/t) = ct^3 + bt^2 + at + 1$. The first Viète relation now gives

$$\frac{1}{x} + \frac{1}{y} - \frac{1}{z} = \frac{1}{w} = -\frac{b}{c}.$$

Using these, our original polynomial becomes

$$p(t) = t^3 - wt^2 + bt - bw = t^2(t - w) + b(t - w) = (t - w)(t^2 + b).$$

This shows that the only solution to the system is w for one of quantities $x, y, -z$ while the other two are the opposites $\pm\sqrt{-b}$. The example follows.

We now introduce the **power sums**

$$p_k(x_1, \ldots, x_n) = x_1^k + \cdots + x_n^k, \quad k \in \mathbb{N}.$$

These are homogeneous symmetric polynomials in the indeterminates $x_1, \ldots, x_n$, and, by the Fundamental Theorem on Symmetric Polynomials above, they can be expressed as polynomials in the elementary symmetric polynomials as indeterminates. The precise statement is the following:

Newton–Girard Formulas. *Let $k, n \in \mathbb{N}$. For $k \leq n$, we have*

$$p_k(x_1, \ldots, x_n) = (-1)^{k-1} k s_k(x_1, \ldots, x_n) + \sum_{i=1}^{k-1} (-1)^{k-i-1} s_{k-i}(x_1, \ldots, x_n) p_i(x_1, \ldots, x_n),$$

[25] A similar problem was in the William Lowell Putnam Mathematical Competition, May 1977. An elementary solution (simpler than the one given in the text) is to realize that $xy = zw$, and make various quadratic expressions in the use of the first equation.

and, for $k > n$, we have

$$p_k(x_1, \ldots, x_n) = \sum_{i=k-n}^{k-1} (-1)^{k-i-1} s_{k-i}(x_1, \ldots, x_n) p_i(x_1, \ldots, x_n).$$

Proof Let $k \in \mathbb{N}$. Consider the monic degree k polynomial $p(x)$ (in the single indeterminate x) with roots $r_1 = x_1, \ldots, r_k = x_k$. By the Factor Theorem and the Viète Relations, we have

$$p(x) = (x - x_1) \cdots (x - x_k) = \sum_{i=0}^{k} (-1)^{k-i} s_{k-i}(x_1, \ldots, x_k) x^i.$$

Substituting $x = x_j$, $j = 1, \ldots, k$, we obtain

$$0 = \sum_{i=0}^{k} (-1)^{k-i} s_{k-i}(x_1, \ldots, x_k) x_j^i.$$

Now, summing up with respect to $j = 1, \ldots, k$ gives

$$0 = (-1)^k k s_k(x_1, \ldots, x_k) + \sum_{i=1}^{k} (-1)^{k-i} s_{k-i}(x_1, \ldots, x_k) p_i(x_1, \ldots, x_k).$$

(Note that p_0 is not defined.)

Splitting off the kth term $p_k(x_1, \ldots, x_k)$ in the sum, and rearranging, we arrive at the Newton–Girard formula for $k = n$:

$$p_k(x_1, \ldots, x_k) = (-1)^{k-1} k s_k(x_1, \ldots, x_k) + \sum_{i=1}^{k-1} (-1)^{k-i-1} s_{k-i}(x_1, \ldots, x_k) p_i(x_1, \ldots, x_k).$$

This identity immediately gives the second Newton–Girard formula for $n < k$ (n indeterminates $x_1, \ldots, x_n$ and kth power) by simply setting $x_{n+1} = \ldots = x_k = 0$ because then $s_{k-i}(x_1, \ldots, x_n) = 0$ for $n < k - i$.

The first Newton–Girard formula also follows from this by showing that the coefficients of the respective monomials in each side of the formula match. This matching follows because, for $k \leq n$, every monomial appearing it the formula contains at most k indeterminates, and, setting $n - k$ complementary indeterminates to be zero, the respective coefficient can be extracted from the formula for the reduced number of (k) indeterminates. The theorem follows.

Using the Newton–Girard formulas recursively, the power sums $p_k(x_1, \ldots, x_n)$ can be expressed as polynomials in the elementary symmetric polynomials. Suppressing the indeterminates, the first few cases are as follows:

$$p_1 = s_1$$

$$p_2 = s_1^2 - 2s_2$$

$$p_3 = s_1^3 - 3s_2s_1 + 3s_3$$

$$p_4 = s_1^4 - 4s_2s_1^2 + 4s_3s_1 + 2s_2^2 - 4s_4$$

$$p_5 = s_1^5 - 5s_2s_1^3 + 5s_3s_1^2 + 5s_2^2s_1 - 5s_4s_1 - 5s_3s_2 + 5s_5.$$

History
As the name suggests, the Viète relations were discovered by the French mathematician François Viète (for positive coefficients) and then Albert Girard (1595–1632) in general. The Newton–Gerard identities above have been discovered by Newton around 1666. He was apparently unaware of the earlier work by Girard, who discovered in 1629 the first four formulas for p_k, $k = 1, 2, 3, 4$, as above.

Example 6.6.3 Let $x, y, z \in \mathbb{R}$ such that[26]

$$x + y + z = 1, \quad x^2 + y^2 + z^2 = 3, \quad x^3 + y^3 + z^3 = 7.$$

Find the value of the 5th power sum $x^5 + y^5 + z^5$.

The indeterminates are x, y, z so that $n = 3$. The system of equations above give $p_1 = 1$, $p_2 = 3$, $p_3 = 7$. We need to find p_5.

The first three identities above can be solved for the elementary symmetric polynomials. We obtain $s_1 = 1$, $s_2 = -1$, $s_3 = 1$. (In particular, by the Viète Relations, x, y, x are roots of the cubic polynomial $t^3 - t^2 - t - 1$, but we do not need this fact.) Now the second Newton–Girard formula can be used recursively for $k = 4, 5$ to obtain $p_4 = s_3p_1 - s_2p_2 + s_1p_3 = 1 + 3 + 7 = 11$, and $p_5 = s_3p_2 - s_2p_3 + s_1p_4 = 3 + 7 + 11 = 21$.
The example follows.

Returning to the Viète Relations, as a simple application, we now derive[27] the Quadratic Formula which gives the roots of the quadratic equation $ax^2+bx+c = 0$, $a \neq 0$, in terms of the coefficients $a, b, c \in \mathbb{R}$.

We begin by assuming that the quadratic polynomial $p(x) = ax^2 + bx + c$ has two real roots r_1 and r_2 (which may coincide). By the Viète relations, we have

$$s_1(r_1, r_2) = r_1 + r_2 = -\frac{b}{a} \quad \text{and} \quad s_2(r_1, r_2) = r_1r_2 = \frac{c}{a}.$$

By the Fundamental Theorem on Symmetric Polynomials, any symmetric polynomial can be written as a polynomial in $s_1(x_1, x_2)$ and $s_2(x_1, x_2)$. We try this for the symmetric polynomial $(x_1 - x_2)^2$. We calculate

[26] A similar problem was in the USA Mathematical Olympiad in 1973.

[27] The typical proof uses the completing the square technique.

$$(x_1 - x_2)^2 = x_1^2 - 2x_1x_2 + x_2^2 = x_1^2 + 2x_1x_2 + x_2^2 - 4x_1x_2$$
$$= (x_1 + x_2)^2 - 4x_1x_2 = s_1(x_1, x_2)^2 - 4s_2(x_1, x_2).$$

Substituting $x_1 = r_1$ and $x_2 = r_2$, the Viète relations now give

$$(r_1 - r_2)^2 = s_1(r_1, r_2)^2 - 4s_2(r_1, r_2) = \left(-\frac{b}{a}\right)^2 - 4\frac{c}{a} = \frac{b^2 - 4ac}{a^2}.$$

Taking the square root of both sides we obtain

$$r_1 - r_2 = \pm\frac{\sqrt{b^2 - 4ac}}{a}.$$

Combining this with the first Viète relation, we arrive at the Quadratic Formula

$$r_1, r_2 = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}.$$

In this formula the expression $b^2 - 4ac$ is called the **discriminant** of the quadratic equation $ax^2 + bx + c = 0$, and it is usually denoted by D. Its name comes from the fact that it determines the **number** of real solutions to the quadratic equation as

$D > 0$ if and only if there are two real solutions;

$D = 0$ if and only if there is one real solution;

$D < 0$ if and only if there are no real solutions.

Remark If $D = b^2 - 4ac < 0$ then the Quadratic Formula gives

$$r_1, r_2 = \frac{-b \pm \mathrm{i}\sqrt{4ac - b^2}}{2a}$$

as complex conjugate roots.

For the next beautiful (and somewhat striking) example we need the fact that, over the reals $\mathbb{R}$, an irreducible polynomial is either linear or quadratic. This can be shown using basic complex arithmetic. By the above, a quadratic polynomial is irreducible if and only if its discriminant is negative. Thus, by the Factor Theorem, every polynomial over the reals $\mathbb{R}$ is the product of linear and irreducible quadratic factors.

Example 6.6.4 Assume that $p(x)$ is a polynomial such that $p(x) \geq 0$ for **all** $x \in \mathbb{R}$. Then, we have $p(x) = a(x)^2 + b(x)^2$ for some polynomials $a(x)$ and $b(x)$.

We write the complete decomposition of $p(x)$ into distinct powers of irreducible factors as

$$p(x) = (x - c_1)^{m_1} \cdots (x - c_k)^{m_k} \cdot (x^2 + p_1 x + q_1)^{n_1} \cdots (x^2 + p_l x + q_l)^{n_l}.$$

As for the **linear** factors, we observe that $p(x) \geq 0$, $x \in \mathbb{R}$, implies that all the exponents $m_1, \ldots, m_k$ are even numbers. Mimicking the desired pattern $a(x)^2 + b(x)^2$, for $i = 1, \ldots, k$, we write

$$(x - c_i)^{m_i} = ((x - c_i)^2 + 0^2)^{m_i/2}.$$

As for the **quadratic** factors, since they are all irreducible, their respective discriminants are negative. For $j = 1, \ldots, l$, we have[28]

$$x^2 + p_j x + q_j = \left(x + \frac{p_j}{2}\right)^2 + \frac{4q_j - p_j^2}{4} = \left(x + \frac{p_j}{2}\right)^2 + \left(\frac{\sqrt{4q_j - p_j^2}}{2}\right)^2,$$

where the square root is defined since the discriminant $p_j^2 - 4q_j < 0$.

Finally, each pair of products of sum of squares can be written as a single sum of squares using the identity

$$(a^2 + c^2)(b^2 + d^2) = (ab + cd)^2 + (ad - bc)^2.$$

(See also Section 5.3.) Using this repeatedly, the entire factored $p(x)$ can be turned into a single sum of squares. The example follows.

In the following example we return to the Monotone Convergence Theorem. We use it for an inductively defined sequence via a quadratic polynomial.

Example 6.6.5 Let $0 < c \in \mathbb{R}$, and $(a_n)_{n \in \mathbb{N}_0}$ a real sequence defined by $a_0 = 0$, and $a_n = c + a_{n-1}^2$, $n \in \mathbb{N}$. Show that $\lim_{n \to \infty} a_n$ exists if and only if $c \leq 1/4$.

Assume first that $\lim_{n \to \infty} a_n = L$ exists. Taking the limit in the inductive definition of the sequence we obtain $L = c + L^2$. This is a quadratic equation in L, so that L exists (as a real number) if and only if the discriminant $D = 1 - 4c \geq 0$. This gives $c \leq 1/4$.

Conversely, assume that $0 < c \leq 1/4$. We show, by induction with respect to $n \in \mathbb{N}$, that the sequence $(a_n)_{n \in \mathbb{N}_0}$ is (strictly) increasing and bounded above. (Clearly, $a_n > 0$, $n \in \mathbb{N}$.) Indeed, $a_1 - a_0 = c > 0$, and, for the general induction step $n \Rightarrow n + 1$, we have $a_{n+1} - a_n = a_n^2 - a_{n-1}^2 = (a_n - a_{n-1})(a_n + a_{n-1}) > 0$, $n \in \mathbb{N}$. For boundedness, we claim $a_n \leq 1/2$, $n \in \mathbb{N}_0$. For the general induction step $n \Rightarrow n + 1$, we have $a_{n+1} = c + a_n^2 \leq 1/4 + (1/2)^2 = 1/2$, $n \in \mathbb{N}$. The claim follows.

By the Monotone Convergence Theorem, the limit $\lim_{n \to \infty} a_n$ exists.

[28] This is the so-called **completing the square** technique; equivalent to the Quadratic Formula.

Example 6.6.6 Let $m, n \in \mathbb{N}$, and assume that n is odd.[29] If both roots of the quadratic polynomial $p(x) = x^2 - nx + m$ are prime numbers, then show that $n - 2$ must be a prime and $m = 2(n - 2)$.

Denoting the roots by r, s, we have $p(x) = x^2 - nx + m = (x - r)(x - s)$. Hence $r + s = n$ and $rs = m$. The crux is that n is odd so that one of the prime roots, s, say, must be even. Thus, $s = 2$. This gives $n = r + 2$, and hence $n - 2$ must be a prime. Finally, we have $m = rs = 2(n - 2)$. The example follows.

Example 6.6.7 Show that there are no positive integer solutions $a, b \in \mathbb{N}$ for the equation

$$a^2 + 2a = b^2 + b.$$

We treat this as a quadratic equation in a. The Quadratic Formula gives

$$a = \frac{-2 \pm \sqrt{4 + 4(b^2 + b)}}{2} = -1 \pm \sqrt{1 + b + b^2}.$$

The crux is that $b < \sqrt{1 + b + b^2} < b + 1$, so that, for $b \in \mathbb{N}$, the square root $\sqrt{1 + b + b^2}$ cannot be an integer.

The next example is sometimes termed as the most challenging problem ever created for mathematical contests. It was Problem #6 in the International Mathematical Olympiad in 1988. The usual technique to solve this problem, presented below, is sometimes (and recently) called **Viète jumping** or **root flipping**. It actually belongs to the reduction theory of quadratic forms, and has been known at least since the late eighteenth century. We will give another geometric solution to this problem in Section 8.4 using hyperbolas.

Example 6.6.8 Let $0 < a, b \in \mathbb{N}$ such that $ab - 1$ divides $a^2 + b^2$. Show that

$$\frac{a^2 + b^2}{ab + 1}$$

is a perfect square. (For example, $a = 8$ and $b = 2$.)

Assume not. Then there exist $a, b \in \mathbb{N}$ such that

$$c = \frac{a^2 + b^2}{ab + 1} \in \mathbb{N}$$

is **not** a square. For this $c \in \mathbb{N}$, consider the set

$$A_c = \left\{ u + v \;\middle|\; c = \frac{u^2 + v^2}{uv + 1} \in \mathbb{N},\ u, v \in \mathbb{N} \right\}$$

By the above, $a + b \in A_c$, in particular, A_c is non-empty.

[29] A special case ($n = 63$) was a problem in the American Mathematics Competitions, 2002.

By assumption $c \neq 1$. In addition, if $c = 2$ then $a^2 + b^2 = 2ab + 2$ gives $(a - b)^2 = 2$. This cannot happen since $\sqrt{2}$ is not an integer.

Thus, from now on we may assume $3 \leq c \in \mathbb{N}$.

Let $a_0 + b_0 = \inf A_c$, $a_0, b_0 \in \mathbb{N}$. Without loss of generality, we may assume $a_0 < b_0$. (Note that $a_0 \neq b_0$ since otherwise we would have $2a_0^2 - ca_0^2 - c = 0$, and this cannot happen since $c \geq 3$.)

We now replace b_0 in the equation

$$c = \frac{a_0^2 + b_0^2}{a_0 b_0 + 1}$$

by the indeterminate y. Multiplying out by the denominator, it follows that this modified equation is equivalent to the condition that $y = b_0$ is a root of the quadratic polynomial[30]

$$p(y) = y^2 - ca_0 y + (a_0^2 - c).$$

If $b_0' \in \mathbb{R}$ is the other root, then the Viète relations give

$$b_0 + b_0' = ca_0 \quad \text{and} \quad b_0 b_0' = a_0^2 - c,$$

or equivalently

$$b_0' = ca_0 - b_0 \quad \text{and} \quad b_0' = \frac{a_0^2 - c}{b_0}.$$

By the first equation, $b_0' \in \mathbb{Z}$, and, by the second, $b_0' \neq 0$ (since c is not a perfect square). Since

$$p(b_0') = {b_0'}^2 - ca_0 b_0' + a_0^2 - c = {b_0'}^2 - c(a_0 b_0' + 1) + a_0^2 = 0,$$

it also follows that $b_0' > 0$. (If $b_0' < 0$ then $a_0 b_0' + 1 \leq 0$, and we would have $a_0 = b_0' = 0$.)

Summarizing, $b_0' \in \mathbb{N}$, and we obtain $a_0 + b_0' \in A_c$.

Finally, by $a_0 < b_0$ and the second Viète relation, we have

$$b_0' = \frac{a_0^2 - c}{b_0} < \frac{a_0^2}{b_0} < b_0.$$

This gives $a_0 + b_0' < a_0 + b_0$; a contradiction to the minimality of $a_0 + b_0$. The claim follows.

[30] Although a and b play symmetric roles, the choice of the indeterminate y (and not x) is justified by the geometric content of the problem to be discussed in Section 8.4.

Remark With somewhat more involved calculations one can derive an inductive formula for a sequence $(a_n)_{n\in\mathbb{N}_0}$ of natural numbers such that the consecutive terms satisfy the equation

$$\frac{a_n^2 + a_{n+1}^2}{a_n a_{n+1} + 1} = g^2, \quad g = \gcd(a_n, a_{n+1}),\ n \in \mathbb{N}_0.$$

This can be solved, and we obtain

$$a_n = \sum_{i=0}^{n/2} (-1)^{i+n/2} \binom{n/2+i}{n/2-i} g^{4i+1}, \text{ if } n \text{ is even}$$

$$a_n = \sum_{i=0}^{(n-1)/2} (-1)^{i+(n-1)/2} \binom{1+(n-1)/2+i}{(n-1)/2-i} g^{4i+3}, \text{ if } n \text{ is odd.}$$

The first few values are tabulated as follows:

n	a_n
0	g
1	g^3
2	$g^5 - g$
3	$g^7 - 2g^3$
4	$g^9 - 3g^5 + g$
5	$g^{11} - 4g^7 + 3g^3$
6	$g^{13} - 5g^9 + 6g^5 - g$

Example 6.6.9 Find all integers $a, b \in \mathbb{Z}$ satisfying[31]

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

First, if $a = 0$ then $b^3 = (-b)^3$, so that $b = 0$ as well. If $b = 0$, then all $a \in \mathbb{Z}$ satisfy the equation. Thus, from now on, we may assume $a \neq 0 \neq b$.

Expanding and factoring the difference of the left-hand and right-hand sides, we obtain

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

Since $b \neq 0$, our equation reduces to

$$2b^2 - a^2 b - 3ab + 3a^2 - a = 2b^2 - a(a+3)b + a(3a-1) = 0,$$

[31] A similar problem was in the USA Mathematical Olympiad, 1987.

where we rearranged the terms to obtain a quadratic polynomial in the indeterminate b (with coefficients in the indeterminate a). The Quadratic Formula gives

$$b = \frac{a(a+3) \pm \sqrt{D(a)}}{4},$$

with discriminant

$$D(a) = a^2(a+3)^2 - 8a(3a-1) = a(a^3 + 6a^2 - 15a + 8).$$

Now, the crux is that for integral solution, $D(a)$ must be a perfect square. We now observe that the sum of the coefficients of the cubic polynomial in the parentheses is zero. Therefore $a = 1$ is a root, and $a - 1$ is a factor. Preforming synthetic division, we have

$$\begin{array}{r|rrrr} & 1 & 6 & -15 & 8 \\ 1 & & 1 & 7 & -8 \\ \hline & 1 & 7 & -8 & 0 \end{array}$$

This gives

$$D(a) = a(a-1)(a^2 + 7a - 8) = a(a-1)^2(a+8),$$

where the last factoring is either by another synthetic division or by simple inspection. Discarding the perfect square $(a-1)^2$, we obtain that, for integral solutions, we must have $a(a+8) = c^2$, $c \in \mathbb{Z}$. Since $a(a+8) = a^2 + 8a = (a+4)^2 - 4^2 = c^2$, this is equivalent to $(a+c+4)(a-c+4) = 16$. In addition, for integral b, the quadratic formula above gives $4|a(a+3) \pm (a-1)c$.

By divisibility, the possible cases are easy to enumerate. The possible pairs $(a + c + 4, a - c + 4)$ are $\pm(1, 16)$, $\pm(2, 8)$, $\pm(4, 4)$, $\pm(8, 2)$, $\pm(16, 1)$. The first and last cannot happen. The remaining cases are tabulated as follows:

$(a+c+4, a-c+4)$	a	b	c
$(2, 8)$	1	2	-3
$(-2, -8))$	-9	12, 42	3
$(4, 4)$	0	0	0
$(-4, -4)$	-8	20	0
$(8, 2)$	1	2	3
$(-8, -2)$	-9	42, 12	-3

Exercises

6.6.1. Solve the system

$$x^3 + y^3 = 1 \quad \text{and} \quad x^4 + y^4 = 1.$$

6.6.2. Solve the system of equations[32]

$$x + y + z = 3, \quad x^2 + y^2 + z^2 = 3, \quad x^3 + y^3 + z^3 = 3.$$

6.6.3. A quadratic polynomial with integer coefficients has one rational root. Show that the other root is also rational. Give an example of a cubic polynomial for which this is not true.

6.6.4. Define the **discriminant** D of the reduced cubic polynomial $p(x) = x^3 + px + q$ as

$$D = (r_1 - r_2)^2(r_2 - r_3)^2(r_3 - r_1)^2.$$

where r_1, r_2, r_3 are the roots of $p(x)$.[33] Derive the formula

$$\begin{aligned} D &= -4(r_1r_2 + r_2r_3 + r_3r_1)^3 - 27(r_1r_2r_3)^2 \\ &= -4p^3 - 27q^2 = -108\left(\left(\frac{p}{3}\right)^3 + \left(\frac{q}{2}\right)^2\right). \end{aligned}$$

6.6.5. Let $p(x) = ax^2 - bx + c$ be a quadratic polynomial with $0 < a \in \mathbb{R}$, and $b, c \in \mathbb{R}$ (note the sign change), and assume that the roots are real and distinct. Then the roots are contained in the interval $(0, 1)$ if and only if $b, c > 0$, $b < 2a$, $c < a$, and $4ac < b^2 < (a + c)^2$. (Notice that the last inequality means that $b/2$ is strictly between the geometric and arithmetic means of a and c.)

6.6.6. For what $c \in \mathbb{R}$ does the cubic polynomial $p(x) = x^3 + cx^2 + 2cx + c^2 - 1$ have exactly one real root?

6.7 The Cauchy–Schwarz Inequality

As a prominent application of the Quadratic Formula, we now derive the general **Cauchy–Schwarz inequality**:

$$(a_1b_1 + a_2b_2 + \cdots + a_nb_n)^2 \le (a_1^2 + a_2^2 + \cdots + a_n^2)(b_1^2 + b_2^2 + \cdots + b_n^2)$$

valid for any $a_1, a_2, \ldots, a_n, b_1, b_2, \ldots, b_n \in \mathbb{R}$, $n \in \mathbb{N}$. (Note that the special case $n = 2$ has already been derived in Section 5.3.)

[32] This was a problem in the USA Mathematical Olympiad, 1973; a straightforward solution uses the Newton–Girard formulas.

[33] For analogy, the discriminant D of the quadratic polynomial $x^2 + px + q$ is $D = (r_1 - r_2)^2 = p^2 - 4q$, where r_1, r_2 are the roots.

For a proof, consider the quadratic polynomial

$$p(x) = (a_1 + b_1x)^2 + (a_2 + b_2x)^2 + \cdots + (a_n + b_nx)^2.$$

Note that $p(x)$ is non-negative and can have at most one root. Consequently, the discriminant D of $p(x)$ is non-positive. Expanding and grouping the like terms, we obtain

$$p(x) = (b_1^2+b_2^2+\cdots+b_n^2)x^2+2(a_1b_1+a_2b_2+\cdots+a_nb_n)x+(a_1^2+a_2^2+\cdots+a_n^2).$$

Therefore the discriminant is

$$D = 4(a_1b_1+a_2b_2+\cdots+a_nb_n)^2-4(a_1^2+a_2^2+\cdots+a_n^2)(b_1^2+b_2^2+\cdots+b_n^2) \leq 0.$$

Rearranging, the Cauchy–Schwarz inequality follows.

Remark The proof above also shows that **equality** holds in the Cauchy–Schwarz inequality if and only if there exists $x_0 \in \mathbb{R}$ such that $a_1 + b_1x_0 = a_2 + b_2x_0 = \ldots = a_n + b_nx_0 = 0$.

History
The inequality above was discovered by Cauchy in 1821. It has been generalized to an inequality for integrals by the Russian mathematician Viktor Bunyakovsky (1804–1889) in 1859, and subsequently this generalization was rediscovered by the German mathematician Hermann Amandus Schwarz (1843–1921) in 1888. Because of this, it is sometimes called the Cauchy–Bunyakovsky–Schwarz inequality.

There are literally hundreds of applications of the Cauchy–Schwarz inequality. We give a few examples.

Example 6.7.1 Let $A, B, C \in \mathbb{R}$ satisfying[34] $AC = B^2$. Assume $0 < a_1, a_2, \ldots, a_n \in \mathbb{R}$, $n \in \mathbb{N}$, such that

$$\sum_{i=1}^{n} a_i = A, \quad \sum_{i=1}^{n} a_i^2 = B, \quad \sum_{i=1}^{n} a_i^3 = C.$$

We then have $n = A^2/B$ and $a_1 = \ldots = a_n = B/A$.

Indeed, the Cauchy–Schwarz inequality gives

$$\left(\sum_{i=1}^{n} a_i\right)\left(\sum_{i=1}^{n} a_i^3\right) \geq \left(\sum_{i=1}^{n} a_i^2\right)^2$$

(since $\sqrt{a_i}\sqrt{a_i^3} = a_i^2$, $i = 1, \ldots, n$). Now, by condition, $AC = B^2$, so that equality holds. We obtain

$$\sqrt{a_1} + x_0\sqrt{a_1}^3 = \ldots = \sqrt{a_n} + x_0\sqrt{a_n}^3 = 0,$$

[34]This is a generalization of a problem in the Iranian Mathematics Competition, 1997.

for some $x_0 \in \mathbb{R}$. This gives $a_1 = \ldots = a_n = a$, where $a = -1/x_0$. With this, we have $A = na$, $B = na^2$, and $C = na^3$. Hence $a = B/A$ and $n = A^2/B$.

In the next example the Viète relations are combined with the Cauchy–Schwarz inequality:

Example 6.7.2 Show that if all the roots of the monic polynomial[35]

$$p(x) = x^n + a_{n-1}x^{n-1} + a_{n-2}x^{n-2} + \cdots + a_1 x + a_0, \quad a_0, a_1, \ldots, a_{n-1} \in \mathbb{R},$$

are **real** then

$$a_{n-2} \le \frac{n-1}{2n} a_{n-1}^2.$$

Let $r_j \in \mathbb{R}$, $1 \le j \le n$, be the roots of $p(x)$. We need the first two Viète relations

$$a_{n-1} = -\sum_{j=1}^{n} r_j \quad \text{and} \quad a_{n-2} = \sum_{1 \le j < k \le n} r_j r_k.$$

We calculate

$$\begin{aligned} 2a_{n-2} = 2 \sum_{1 \le j < k \le n} r_j r_k &= \left(\sum_{j=1}^{n} r_j \right)^2 - \left(\sum_{j=1}^{n} r_j^2 \right) \\ &= a_{n-1}^2 - \frac{1}{n} (1 + 1 + \cdots + 1) \left(r_1^2 + r_2^2 + \cdots + r_n^2 \right) \\ &\le a_{n-1}^2 - \frac{1}{n} (r_1 + r_2 + \cdots + r_n)^2 = \frac{n-1}{n} a_{n-1}^2, \end{aligned}$$

where we used the Cauchy–Schwarz inequality.[36] The claim follows.

Example 6.7.3 Let $p(x)$ be a polynomial with **positive** coefficients. Show that if

$$p\left(\frac{1}{x}\right) \ge \frac{1}{p(x)}$$

holds for $x = 1$ then it also holds for all $0 < x \in \mathbb{R}$.

Let

$$p(x) = a_n x^n + \cdots + a_1 x + a_0, \quad 0 < a_0, a_1, \ldots, a_n \in \mathbb{R}.$$

[35] The special case $n = 5$ was a problem in the USA Mathematical Olympiad, 1983.

[36] In the second equality we can also use the Newton–Girard formula $p_2(r_1, \ldots, r_n) = a_{n-1}^2 - 2a_{n-2}$ along with the Viète relations.

Substituting $x = 1$ gives $p(1) \geq 1$. Using our condition and the Cauchy–Schwarz inequality, for $0 < x \in \mathbb{R}$, we calculate

$$p(x)p\left(\frac{1}{x}\right) = (a_n x^n + \cdots + a_1 x + a_0)\left(\frac{a_n}{x^n} + \cdots + \frac{a_1}{x} + a_0\right)$$
$$\geq (a_n + \cdots + a_1 + a_0)^2 = p(1)^2 \geq 1.$$

The example follows.

Example 6.7.4 (Nesbitt Inequality) For $0 < a, b, c \in \mathbb{R}$, we have

$$\frac{a}{b+c} + \frac{b}{c+a} + \frac{c}{a+b} \geq \frac{3}{2}$$

with equality if and only if $a = b = c$.

To derive this, we first add 1 to each fraction, and obtain the equivalent form

$$\frac{a+b+c}{b+c} + \frac{a+b+c}{c+a} + \frac{a+b+c}{a+b} \geq \frac{9}{2}.$$

This can be written as

$$2(a+b+c)\left(\frac{1}{b+c} + \frac{1}{c+a} + \frac{1}{a+b}\right) \geq 9,$$

or equivalently

$$((b+c) + (c+a) + (a+b))\left(\frac{1}{b+c} + \frac{1}{c+a} + \frac{1}{a+b}\right) \geq 9.$$

Now, letting $a_1 = \sqrt{b+c}$, $a_2 = \sqrt{c+a}$, $a_3 = \sqrt{a+b}$, and $b_1 = 1/a_1 = 1/\sqrt{b+c}$, $b_2 = 1/a_2 = 1/\sqrt{c+a}$, $b_3 = 1/a_3 = 1/\sqrt{a+b}$, this last inequality turns into the Cauchy–Schwarz inequality for $n = 3$.

Example 6.7.5 For $a, b, c \in \mathbb{R}$, show that $2a^2 + 3b^2 + 6c^2 \geq (a+b+c)^2$.

Indeed, since $1/2 + 1/3 + 1/6 = 1$, we have

$$2a^2 + 3b^2 + 6c^2 = \left(\frac{1}{2} + \frac{1}{3} + \frac{1}{6}\right)(2a^2 + 3b^2 + 6c^2) \geq (a+b+c)^2;$$

yet another form of the Cauchy–Schwarz inequality (with $a_1 = 1/\sqrt{2}$, $a_2 = 1/\sqrt{3}$, $a_3 = 1/\sqrt{6}$, and $b_1 = \sqrt{2}a$, $b_2 = \sqrt{3}b$, $b_3 = \sqrt{6}c$).

Remark The following inequality is a trivial consequence of the Cauchy–Schwarz inequality:[37]

For positive real numbers $0 < x_1, x_2, \ldots, x_n, y_1, y_2, \ldots, y_n \in \mathbb{R}$, $n \in \mathbb{N}$, we have

$$\frac{x_1^2}{y_1} + \frac{x_2^2}{y_2} + \cdots + \frac{x_n^2}{y_n} \geq \frac{(x_1 + x_2 + \cdots + x_n)^2}{y_1 + y_2 + \cdots + y_n}.$$

Indeed, this follows by the substitution $a_i = x_i/\sqrt{y_i}$ and $b_i = \sqrt{y_i}$, $i = 1, 2, \ldots, n$, into the Cauchy–Schwarz inequality.

We close this section by a brief note on the **Chebyshev sum inequality** due to the Russian mathematician Pafnuty Chebyshev (1821–1894):

Given real numbers $a_1, a_2, \ldots, a_n, b_1, b_2, \ldots, b_n \in \mathbb{R}$, $n \in \mathbb{N}$, such that

$$a_1 \geq a_2 \geq \cdots \geq a_n \quad \text{and} \quad b_1 \geq b_2 \geq \cdots \geq b_n,$$

then we have

$$\frac{a_1 b_1 + a_2 b_2 + \cdots + a_n b_n}{n} \geq \frac{a_1 + a_2 + \cdots + a_n}{n} \cdot \frac{b_1 + b_2 + \cdots + b_n}{n}.$$

Remark If the inequality signs are reversed in **one** sequence of inequalities in the assumptions, then the reverse inequality sign holds in the Chebyshev sum inequality. This is clear; if, for example, $a_1 \leq a_2 \leq \cdots \leq a_n$ but $b_1 \geq b_2 \geq \cdots \geq b_n$, then we apply the Chebyshev sum inequality to $-a_1 \geq -a_2 \geq \cdots \geq -a_n$, etc.

The proof of the Chebyshev sum inequality is simple. We have

$$0 \leq \sum_{j=1}^{n} \sum_{k=1}^{n} (a_j - a_k)(b_j - b_k) = 2n \sum_{j=1}^{n} a_j b_j - 2 \sum_{j=1}^{n} a_j \sum_{k=1}^{n} b_k.$$

The initial double sum on the left-hand side is non-negative since, for each $1 \leq j, k \leq n$, the factors $a_j - a_k$ and $b_j - b_k$ (if non-zero) are simultaneously positive or negative. Expanding, we arrive at the right-hand side. Rearranging, the Chebyshev sum inequality follows.

Example 6.7.6 A trivial consequence of the Chebyshev sum inequality ($a_i = b_i$, $i = 1, 2, \ldots, n$) is the following:

$$a_1^2 + a_2^2 + \cdots + a_n^2 \geq \frac{(a_1 + a_2 + \cdots + a_n)^2}{n}, \quad 0 < a_1, a_2, \ldots, a_n \in \mathbb{R}.$$

[37] Due to its usefulness in some mathematical contest problems, this is sometimes called the Titu–Engel–Sedrakyan inequality after Titu Andreescu (1965 –), Arthur Engel (1928 –), and Nairi Sedrakyan (1961 –).

Setting $a_i = x_i^{2^k}$, $0 < x_i \in \mathbb{R}$, $i = 1, 2, \ldots, n$, in the inequality above, we obtain

$$p_{2^{k+1}}(x_1, x_2, \ldots, x_n) \geq \frac{p_{2^k}^2(x_1, x_2, \ldots, x_n)}{n}, \quad k \in \mathbb{N}_0,$$

where $p_l(x_1, x_2, \ldots, x_n) = x_1^l + x_2^l + \cdots + x_n^l$, $l \in \mathbb{N}$, is the lth power sum.

A simple induction with respect to $k \in \mathbb{N}$ gives

$$p_{2^k}(x_1, x_2, \ldots, x_n) \geq \frac{p_1^{2^k}(x_1, x_2, \ldots, x_n)}{n^{2^k-1}}, \quad k \in \mathbb{N}.$$

In terms of $0 < x_1, x_2, \ldots, x_n \in \mathbb{R}$, this rewrites as

$$x_1^{2^k} + x_2^{2^k} + \cdots + x_n^{2^k} \geq \frac{(x_1 + x_2 + \cdots + x_n)^{2^k}}{n^{2^k-1}}, \quad k \in \mathbb{N}.$$

An often quoted special case is $n = k = 2$:

$$x^4 + y^4 \geq \frac{(x+y)^4}{8}, \quad 0 < x, y \in \mathbb{R}.$$

Exercises

6.7.1. Derive the following generalization of the Nesbitt inequality (Example 6.7.4).

Let $0 < a_1, a_2, \ldots, a_n \in \mathbb{R}$, $n \in \mathbb{N}$, and $s = a_1 + a_2 + \cdots + a_n$. We have

$$\frac{a_1}{s - a_1} + \frac{a_2}{s - a_2} + \cdots + \frac{a_n}{s - a_n} \geq \frac{n}{n-1}.$$

6.7.2. For $0 < a, b, c \in \mathbb{R}$, derive the inequality

$$\left(a + \frac{1}{b}\right)\left(b + \frac{1}{c}\right)\left(c + \frac{1}{a}\right) \geq 8$$

with equality if and only if $a = b = c = 1$.

6.7.3. Show that, for $0 < a, b, c \in \mathbb{R}$, we have

$$\frac{a^2}{b^2} + \frac{b^2}{c^2} + \frac{c^2}{a^2} \geq \frac{b}{a} + \frac{c}{b} + \frac{a}{c},$$

with equality if and only if $a = b = c$.

Chapter 7
Polynomial Functions

"In our days Scipione del Ferro of Bologna has solved the case of the cube and first power equal to a constant, a very elegant and admirable accomplishment. . . . In emulation of him, my friend Niccolò Tartaglia of Brescia, wanting not to be outdone, solved the same case when he got into a contest with his [Scipione's] pupil, Antonio Maria Fior, and, moved by my many entreaties, gave it to me."
in Ars Magna by Gerolamo Cardano (1501–1576)

In this chapter we enrich our algebraic point of view of polynomials by considering them as functions. We develop first order analysis (critical points and monotonicity) for graphs of polynomial functions using synthetic division applied to difference quotients. We treat the difference quotient of a polynomial as a rational function with a removable singularity at the point where the quotient is taken. Removing the singularity then takes us directly to the concept of the derivative without taking limits. We discuss the special case of cubic polynomials in great details. In the second half of this chapter we return to algebra and study the roots of polynomials, once again with full details of the cubic case. We finish this chapter by the somewhat more advanced topic of multivariate factoring. Some of the material here is also preparatory to the general AM-GM inequality to be discussed in Section 9.5.

7.1 Polynomials as Functions

Recall that a polynomial $p(x)$ with indeterminate $x \in \mathbb{R}$ defines a polynomial function $p : \mathbb{R} \to \mathbb{R}$ with variable $x \in \mathbb{R}$. Using classical terminology, we write $y = p(x)$ with $x \in \mathbb{R}$. In this section we assemble a few important facts about polynomial functions.

First, any polynomial function is defined everywhere; that is, the domain of definition is always $\mathbb{R}$.

G. Toth, *Elements of Mathematics*, Undergraduate Texts in Mathematics,
https://doi.org/10.1007/978-3-030-75051-0_7

Let $p : \mathbb{R} \to \mathbb{R}$ be a degree n polynomial function given by

$$y = p(x) = \sum_{k=0}^{n} a_k x^k = a_n x^n + a_{n-1} x^{n-1} + \cdots + a_1 x + a_0, \quad a_0, a_1, \ldots, a_n \in \mathbb{R}, \ \ a_n \neq 0,$$

where in the expanded form we used descending powers.

We first claim

$$\lim_{x \to \infty} p(x) = \pm\infty,$$

where the choice of $\pm$ depends on the sign of the leading coefficient a_n.

Indeed, first recall

$$\lim_{x \to \infty} p(x) = \lim_{u \to 0^+} p\left(\frac{1}{u}\right).$$

Using this, we calculate

$$\begin{aligned}
\lim_{u \to 0^+} p\left(\frac{1}{u}\right) &= \lim_{u \to 0^+} \left(\frac{a_n}{u^n} + \frac{a_{n-1}}{u^{n-1}} + \cdots + \frac{a_1}{u} + a_0\right) \\
&= \lim_{u \to 0^+} \frac{1}{u^n} \lim_{u \to 0^+} (a_n + u a_{n-1} + \cdots + u^{n-1} a_1 + u^n a_0) \\
&= a_n \lim_{u \to 0^+} \frac{1}{u^n} = \pm\infty.
\end{aligned}$$

The claim follows.

The limit at negative infinity can be obtained by taking opposites:

$$\lim_{x \to -\infty} p(x) = \lim_{x \to \infty} p(-x) = \pm\infty.$$

Next, recall the **difference quotient** from Section 4.3:

$$\mathfrak{m}_p(x, c) = \frac{p(x) - p(c)}{x - c}, \quad x \neq c, \ \ x, c \in \mathbb{R}.$$

The difference quotient is a rational expression in the indeterminate x with domain of definition being all real numbers except c.

We claim that, away from c, it is actually a polynomial. Indeed, pairing up the kth monomials in $p(x)$ and in $p(c)$ with $k = 1, 2, \ldots, n$, and factoring, for $x \neq c$, we calculate

$$\mathfrak{m}_p(x, c) = \sum_{k=0}^{n} \frac{a_k x^k - a_k c^k}{x - c} = \sum_{k=0}^{n} a_k \frac{x^k - c^k}{x - c}$$

$$= \sum_{k=1}^{n} a_k \frac{(x-c)(x^{k-1} + x^{k-2}c + \cdots + xc^{k-2} + c^{k-1})}{x-c}$$

$$= \sum_{k=1}^{n} a_k (x^{k-1} + x^{k-2}c + \cdots + xc^{k-2} + c^{k-1}).$$

The claim follows.

The crux of this computation is that, although the difference quotient is undefined at $x = c$, the right-hand side, being a polynomial in x, can be evaluated at $x = c$. Evaluating the right-hand side at $x = c$ amounts to taking the limit of the difference quotient and obtain the derivative:

$$p'(c) = \lim_{x \to c} \mathfrak{m}_p(x, c) = \sum_{k=1}^{n} a_k \lim_{x \to c} (x^{k-1} + x^{k-2}c + \cdots + xc^{k-2} + c^{k-1}) = \sum_{k=1}^{n} k a_k c^{k-1}.$$

We can therefore extend the definition of the difference quotient and **define**

$$\mathfrak{m}_p(c, c) = \sum_{k=1}^{n} k a_k c^{k-1} = n a_n c^{n-1} + (n-1) a_{n-1} c^{n-2} + \cdots + 2a_2 c + a_1.$$

With this, the difference quotient $\mathfrak{m}_p$ becomes a **polynomial** in the indeterminates x and c. As a byproduct, we also see that $\mathfrak{m}_p(c, c)$ is the **derivative** $p'(c)$ of the polynomial function p at c.

Remark The tangent line to the unit parabola $y = x^2$ has the property that it is the unique non-vertical line that meets the parabola only at the point (c, c^2). This geometric condition gives the slope of the tangent line as $2c$. Indeed, combining the equation of the line $y - c^2 = m(x - c)$ through (c, c^2) with $y = x^2$, we obtain $x^2 - c^2 = (x - c)(x + c) = m(x - c)$, and it follows that the unique intersection requires $m = 2c$. Note that the tangent line meets the first axis at $c/2$, the midpoint of the first coordinate of the point of tangency (c, c^2), and the origin.

For $y = p(x) = x^2$, our formula above also gives $p'(c) = 2c$. Therefore these two concepts coincide. We can arrive at the same conclusion about tangent lines drawn to ellipses and hyperbolas which we can use to derive their reflective properties. More about this in Chapter 8.

We now return to our polynomial function p. Recall that c is a **critical point** of p if $p'(c) = 0$. Geometrically this means that the tangent line is horizontal. Since $p'(x)$ is a degree $n - 1$ polynomial in the indeterminate x, the Factor Theorem implies that p has **at most** $n - 1$ critical points.

Let c be a critical point of p. By definition, c is a root of the difference quotient $\mathfrak{m}_p(x, c)$ viewed as a polynomial in the indeterminate x. Thus, by the Factor Theorem, $(x - c)$ is a factor of $\mathfrak{m}_p(x, c)$, and we have $\mathfrak{m}_p(x, c) = (x - c)q(x)$. Here the quotient $q(x)$ is a degree $n - 2$ polynomial (with the dependence on c

suppressed). Using the definition of $\mathfrak{m}_p(x, c)$ above, this can be written as

$$\frac{p(x) - p(c)}{x - c} = (x - c)q(x).$$

Multiplying out and rearranging, we arrive at

$$p(x) = q(x)(x - c)^2 + p(c).$$

Retracing our steps, we see that the converse also holds; that is, c is a critical point of p if and only if this equality holds (for some polynomial function q).

Assuming $q(c) \neq 0$, the equation $y = q(c)(x - c)^2 + p(c)$ is the equation of a **parabola**.[1] We have

$$\lim_{x \to c} \left| \frac{p(x) - (q(c)(x - c)^2 + p(c))}{(x - c)^2} \right| = \lim_{x \to c} |q(x) - q(c)| = 0.$$

We obtain that this parabola "best approximates" the graph $G(p)$ at $(c, p(c))$.

We have now come to the fundamental problem of understanding the large-scale behavior of polynomials. The graph of a linear function (degree one polynomial) is a line. A linear function with non-zero slope is automatically one-to-one, thereby it always has an inverse. The graph of a quadratic (degree two) polynomial is a parabola which fails the horizontal intersection property, thereby a quadratic polynomial function is not one-to-one, and has no inverse. If we restrict the parabola to one of its branches, then a single branch does satisfy the horizontal intersection property, and thereby the corresponding function has an inverse.

We now ask the following general question: To what extent does the one-to-one property fail for polynomials, and how can we analyze this failure to obtain a geometric description of the graph?

To answer this question we start again with our polynomial p (of degree $n \geq 2$), and assume that p fails to be injective. This means that is there exist $x' < x''$ such that $p(x') = p(x'')$. We restrict p to the closed interval $[x', x'']$. Since p is continuous, it assumes its supremum (infimum) at a point c of the open interval (x', x''):

$$p(x) \leq p(c) \quad \text{for all} \quad x \in [x', x''].$$

(For infimum, the inequality sign is reversed.)

By the Fermat Principle, c is a critical point of p.

Remark Alternatively, we can also use polynomial division; we can divide $p(x)$ by $(x - c)^2$ and obtain $p(x) = (x - c)^2 q(x) + mx + b$, with the remainder $mx + b$ of

[1] A parabola with vertical symmetry axis is defined as the graph of the polynomial function $y = ax^2 + bx + c$, $0 \neq a, b, c \in \mathbb{R}$. See Section 8.2.

degree ≤ 1. We claim that $m = 0$ and $b = p(c)$, so that, by the above, c is a **critical point** of p. Indeed, since at c the polynomial p assumes its supremum, we have

$$p(x) = (x-c)^2 q(x) + mx + b \leq mc + b = p(c), \quad \text{for all} \quad x \in (x', x'').$$

Rearranging and simplifying, we obtain $(x - c)\,[(x - c)q(x) + m] \leq 0$. Now, assuming $m \neq 0$, since $\lim_{x\to c}(x - c)q(x) = 0$, for x close enough to c, the expression in the square brackets will have the same sign (positive or negative) as m. On the other hand, depending on which side of c is x, the difference $x - c$ is positive or negative. Thus, the left-hand side of the inequality above can be made positive or negative with x arbitrarily close to c. We see that the inequality above cannot hold for $m \neq 0$. Hence $p(x) = (x-c)^2 q(x) + b$. Finally, substituting $x = c$, we obtain $p(c) = b$. The claim follows.

In summary, we see that between $x' < x''$ with $p(x') = p(x'')$ there is a critical point c at which p assumes an extremum on the closed interval $[x', x'']$.

Now let x' and x'' be **consecutive** critical points of p. The polynomial function p restricted to the interval $[x', x'']$ must be one-to-one since otherwise, by the construction above, there would be a critical point in the open interval (x', x''), and x' and x'' would not be consecutive. Since p is one-to-one, it must be strictly monotonic. The same argument applies for the infinite closed intervals before the first and after the last critical points of p. The following transparent picture of the graph $G(p)$ of p emerges: At the critical points the graph $G(p)$ has horizontal tangents. Between consecutive critical points and before the first and after the last critical points the polynomial function p is strictly increasing or decreasing.

Example 7.1.1 An important sequence $\{e_n\}_{n\in\mathbb{N}_0}$ of polynomial functions, playing a paramount importance in Newton's treatment of the natural exponential function, is defined by

$$e_n(x) = \sum_{k=0}^{n} \frac{x^k}{k!} = 1 + \frac{x}{1!} + \frac{x^2}{2!} + \cdots + \frac{x^n}{n!}, \quad x \in \mathbb{R},\ n \in \mathbb{N}.$$

By definition, $0! = 1$, and with this we set $e_0(x) = x^0/0! = 1$, $x \in \mathbb{R}$. Clearly, $e_n(x)$ is a polynomial of degree $n \in \mathbb{N}_0$. Taking the derivative, we obtain the characteristic property of e_n:

$$e_n'(c) = e_{n-1}(c), \quad c \in \mathbb{R},\ n \in \mathbb{N}.$$

In this example we show the following: (1) For $n \in \mathbb{N}$ odd, e_n has no critical points. Moreover, e_n is strictly increasing and has a unique negative root; (2) For $n \in \mathbb{N}$ even, e_n has a unique critical point $c < 0$ at which it attains its absolute minimum. Moreover, $e_n(c) = c^n/n! > 0$, so that e_n is everywhere positive.

We will use Peano's Principle of Induction to prove (1)-(2). For $n \in \mathbb{N}$ odd, we let $n = 2k - 1$, $k \in \mathbb{N}$, and, for $n \in \mathbb{N}$ even, we let $n = 2k$, $k \in \mathbb{N}$. We will proceed with a two-step induction with respect to $k \in \mathbb{N}$ to derive (1)–(2).

Let $k = 1$. Then $e_1(x) = 1 + x$, $x \in \mathbb{R}$, is linear, and therefore has no critical points; it is strictly increasing, and has a unique negative root at -1.

Let $k = 2$. Then $e_2(x) = 1 + x + x^2/2$, $x \in \mathbb{R}$, is quadratic, and has a unique critical point at $c = -1$ at which it attains its absolute minimum. Moreover, $e_2(-1) = (-1)^2/2 = 1/2 > 0$, so that e_2 is everywhere positive. The initial step of the induction is complete.

We now turn to the general induction step $k - 1 \Rightarrow k$.

(1) First, consider e_{2k-1}. As noted above, we have $e'_{2k-1} = e_{2k-2}$. By the induction hypothesis, $e_{2k-2} = e_{2(k-1)}$ is everywhere positive, and hence so is e'_{2k-1}. In particular, e_{2k-1} cannot have critical points, and so it must be strictly monotonic. Moreover, since e_{2k-1} is an odd degree polynomial with positive leading coefficient $1/(2k-1)!$, we have $\lim_{x\to\pm\infty} e_{2k-1}(x) = \pm\infty$. We conclude that e_{2k-1} is strictly increasing. By the Intermediate Value Theorem, e_{2k-1} must have a root which, by strict monotonicity, must be unique. Finally, since $0 < e_{2k-1}(x)$ for $0 \leq x \in \mathbb{R}$, we see that this root must be negative. The induction is complete in this case.

(2) Second, consider e_{2k}. We have $e'_{2k} = e_{2k-1}$. By the previous case, e_{2k-1} has a unique root at $c < 0$, say, and it is the unique critical point of e_{2k}. Since e_{2k} is an even degree polynomial with positive leading coefficient $1/(2k)!$, we have $\lim_{x\to\pm\infty} e_{2k}(x) = \infty$. Since there are no critical points on the intervals $(-\infty, c)$ and (c, ∞), the limit relation implies that e_{2k} is strictly decreasing on $(-\infty, c)$, and strictly increasing on (c, ∞). It follows that e_{2k} assumes its absolute minimum at c. Since c is a critical point of e_{2k}, we have $e'_{2k}(c) = e_{2k-1}(c) = 0$, and hence $e_{2k}(c) = e_{2k-1}(c) + c^{2k}/(2k)! = c^{2k}/(2k)! > 0$. We obtain that e_{2k} is everywhere positive.

The general induction step is complete. The example follows.

We close this section by discussing the critical points in more detail for cubic polynomials.

Example 7.1.2 Let a cubic polynomial function be given by[2]

$$y = p(x) = x^3 + px^2 + qx + r,$$

where, for simplicity and without loss of generality, we assume that the leading coefficient is one. In degree three there are at most two critical points. As discussed above, for each critical point c, we have

$$p(x) = x^3 + px^2 + qx + r = (x - c)^2(x - s) + p(c),$$

where the linear quotient takes the form $q(x) = x - s$ for some $s \in \mathbb{R}$. There are three equations connecting the unknown s with the coefficients of $p(x)$ and c.

[2]Note the unfortunate double appearance of the symbol p. We will keep the polynomial $p(x)$ and the coefficient p separate.

We only need $p = -2c - s$, and this can be obtained by expanding the right-hand side of the equation above (and comparing the coefficients for the quadratic terms involving x^2). The critical points are solutions of the quadratic equation $p'(c) = 3c^2 + 2pc + q = 0$. The two values of c can be obtained by the Quadratic Formula

$$c = \frac{-p \pm \sqrt{p^2 - 3q}}{3}.$$

We are primarily interested in the case when there are exactly two roots; that is, when the discriminant is positive: $p^2 > 3q$. For each of the two values of c we can estimate the location of s relative to c. First, let c be the smaller root. Using the formula for s obtained above, we calculate

$$s - c = -p - 3c = -p + p + \sqrt{p^2 - 3q} = \sqrt{p^2 - 3q} > 0.$$

Thus, we have $c < s$. Looking back to the original factorization of $p(x)$, we see that, as long as $x < s$, we have $p(x) - p(c) = (x - c)^2(x - s) \leq 0$ with sharp inequality for $x \neq c$ only.

Summarizing, we see that on the interval $(-\infty, s)$ which includes c we have $p(x) \leq p(c)$, with sharp inequality for $x \neq c$ only. We conclude that $p(x)$, restricted to $(-\infty, s)$ has a (unique) **maximum** at c. The horizontal line given by $y = p(c)$ is **tangent** to the graph of p.

The case of the larger root c is similar. We obtain that our cubic polynomial, restricted to the interval (s, ∞) (with s corresponding to this larger root c) has a unique **minimum** at c. The horizontal line given by $y = p(c)$ is **tangent** to the graph of p.

The quadratic equation for c above has a unique solution if and only if the discriminant is zero: $p^2 = 3q$. In this case $c = -p/3$, so that we have $s = -p - 2c = -p + 2p/3 = -p/3 = c$. We see that in this case our cubic reduces to

$$p(x) = x^3 + px^2 + qx + r = (x - c)^3 + p(c).$$

Geometrically, this means that the graph of our cubic is obtained from the graph of the third power function $\mathfrak{p}_3$ by translation.

Finally, there are no critical points if and only if $p^2 < 3q$. By the discussion above, it follows that our cubic polynomial is **strictly monotonic** with no horizontal tangent line.

Exercises

7.1.1. Analyze the graphs of the cubic polynomial functions:

$$(a) \quad y = x^3 - 6x^2 + 4; \quad (b) \quad y = x^3 - 3x^2 + 3x + 1; \quad (c) \quad y = x^3 - x + 1.$$

7.2 Roots of Cubic Polynomials

The general cubic **polynomial** is of the form $a_3x^3 + a_2x^2 + a_1x + a_0$ with $a_0, a_1, a_2, a_3 \in \mathbb{R}$ and $a_3 \neq 0$. Setting this polynomial equal to zero and dividing by a_3, we obtain the equation for the general monic cubic:

$$p(x) = x^3 + ax^2 + bx + c = 0,$$

where we renamed the coefficients as $a, b, c \in \mathbb{R}$.

In this section we discuss real solutions of cubic equations.
By the Factor Theorem, a cubic polynomial has at most three roots. In addition, there must be at least one real root. This is a direct consequence of the Intermediate Value Theorem, since $\lim_{x\to\pm\infty} p(x) = \pm\infty$. Algebraically, as noted previously, this also follows from the fact that, over the real numbers $\mathbb{R}$, the irreducible polynomials have degree one or two, so that any cubic polynomial must have a linear factor, and thereby a real root. Once this root is obtained we can use the Factor Theorem to divide by the corresponding root factor and reduce our cubic equation to a quadratic equation whose solutions we already analyzed via the Quadratic Formula.

Returning to the general cubic above, we use the substitution $x \mapsto x - a/3$ and calculate

$$\begin{aligned}
&\left(x - \frac{a}{3}\right)^3 + a\left(x - \frac{a}{3}\right)^2 + b\left(x - \frac{a}{3}\right) + c \\
&\quad = x^3 - ax^2 + \frac{a^2}{3}x - \frac{a^3}{27} + ax^2 - \frac{2a^2}{3}x + \frac{a^3}{9} + bx - \frac{ab}{3} + c \\
&\quad = x^3 + \left(b - \frac{a^2}{3}\right)x + \frac{2a^3}{27} - \frac{ab}{3} + c.
\end{aligned}$$

Letting

$$p = b - \frac{a^2}{3} \quad \text{and} \quad q = \frac{2a^3}{27} - \frac{ab}{3} + c,$$

we obtain the so-called **reduced** cubic equation

$$x^3 + px + q = 0.$$

The trivial case $x^3 = 0$ can obviously be excluded so that we may assume that p and q do not vanish simultaneously. Moreover, if $p = 0$, then $x = \sqrt[3]{-q}$ is a root, and if $q = 0$, then $x = 0$ is a root. Thus, from now on we may assume that p and q do not vanish.

The crux to solve the reduced cubic equation is to write the sum of the first two terms $x^3 + px$ as the product $x(x^2 + p)$ and **match** it with the factors of the left-hand side of the cubic identity

$$(u+v)(u^2-uv+v^2)=u^3+v^3.$$

This gives

$$x=u+v \quad \text{and} \quad x^2+p=u^2-uv+v^2.$$

We now eliminate x by squaring both sides of the first equation and substituting the result into the second. After simplification, we arrive at $3uv+p=0$.
Based on this, we **introduce** two indeterminates u and v satisfying

$$x=u+v \quad \text{and} \quad 3uv=-p.$$

Since u and v play symmetric roles, these amount to the so-called **Viète substitution**

$$x=w-\frac{p}{3w},$$

where w is either u or v. (Note that w does not vanish since $uv=-p/3\neq 0$.
On the other hand, returning to our matching above, the reduced cubic can be written as

$$u^3+v^3+q=0.$$

In terms of the single indeterminate w, our reduced cubic takes the form

$$w^3-\left(\frac{p}{3}\right)^3\frac{1}{w^3}+q=0.$$

Multiplying by w^3 and rearranging we arrive at the sextic equation

$$w^6+qw^3-\left(\frac{p}{3}\right)^3=0.$$

This is a quadratic equation in w^3. The Quadratic Formula gives

$$w^3=\frac{-q\pm\sqrt{q^2+4(p/3)^3}}{2}=-\frac{q}{2}\pm\sqrt{\left(\frac{q}{2}\right)^2+\left(\frac{p}{3}\right)^3}.$$

At this point, in order to stay within real numbers $\mathbb{R}$, we assume

$$\left(\frac{q}{2}\right)^2+\left(\frac{p}{3}\right)^3\geq 0.$$

Taking the cube root, we have

$$w = \sqrt[3]{-\frac{q}{2} \pm \sqrt{\left(\frac{q}{2}\right)^2 + \left(\frac{p}{3}\right)^3}}.$$

Since u and v play symmetric roles, swapping them if necessary, we obtain

$$u = \sqrt[3]{-\frac{q}{2} - \sqrt{\left(\frac{q}{2}\right)^2 + \left(\frac{p}{3}\right)^3}} \quad \text{and} \quad v = \sqrt[3]{-\frac{q}{2} + \sqrt{\left(\frac{q}{2}\right)^2 + \left(\frac{p}{3}\right)^3}}.$$

Finally, using $x = u + v$, we arrive at the **Cubic Formula** giving a solution of the cubic equation:

$$x = \sqrt[3]{-\frac{q}{2} - \sqrt{\left(\frac{q}{2}\right)^2 + \left(\frac{p}{3}\right)^3}} + \sqrt[3]{-\frac{q}{2} + \sqrt{\left(\frac{q}{2}\right)^2 + \left(\frac{p}{3}\right)^3}}.$$

Equivalently, using the Viète substitution

$$x = \sqrt[3]{-\frac{q}{2} - \sqrt{\left(\frac{q}{2}\right)^2 + \left(\frac{p}{3}\right)^3}} - \frac{p}{3\sqrt[3]{-\frac{q}{2} - \sqrt{\left(\frac{q}{2}\right)^2 + \left(\frac{p}{3}\right)^3}}}.$$

Note that the equivalency of these formulas also follows from rationalizing the denominator in the last algebraic fraction.

Example 7.2.1 Is there a real number whose cube is 1 more than the number itself? (For square instead of cube, this is the golden number and its negative reciprocal; see Example 3.1.2.)

The number x must satisfy the equation $x^3 = x + 1$, and it is therefore a root of the cubic equation

$$x^3 - x - 1 = 0.$$

We have $p = -1$ and $q = -1$ so that

$$\left(\frac{q}{2}\right)^2 + \left(\frac{p}{3}\right)^3 = \frac{1}{2^2} - \frac{1}{3^3} = \frac{23}{108} = \frac{69}{18^2}.$$

Using this in the Cubic Formula above, we obtain

$$x = \sqrt[3]{\frac{1}{2} - \frac{\sqrt{69}}{18}} + \sqrt[3]{\frac{1}{2} + \frac{\sqrt{69}}{18}} = \frac{1}{6}\sqrt[3]{108 - 12\sqrt{69}} + \frac{1}{6}\sqrt[3]{108 + 12\sqrt{69}}.$$

Returning to the main line, the computations above were performed with the understanding that the critical expression

$$\left(\frac{q}{2}\right)^2 + \left(\frac{p}{3}\right)^3$$

is **non-negative**.
If this expression is **positive**, then the Cubic Formula gives one real root as above. As noted previously, once a real root is found, the remaining two roots are given by the Quadratic Formula. It can be shown that, in this case, the other two roots are complex.
If this expression is **zero**, then, once again, the Cubic Formula gives $w = \sqrt[3]{-q/2} = -\sqrt[3]{q/2}$, and hence the real root $-2\sqrt[3]{q/2}$. Moreover, in this case, $\sqrt[3]{q/2}$ is another root of **multiplicity two**. (Indeed, for $w = -\sqrt[3]{q/2}$, we have $(x - 2w)(x + w)^2 = x^3 - 3w^2x - 2w^3 = x^3 + px + q$.)

The following example shows that the expression provided by the Cubic Formula may not be of the simplest form.

Example 7.2.2 [3] Show that

$$\sqrt[3]{5\sqrt{2}+7} - \sqrt[3]{5\sqrt{2}-7} = 2.$$

A simple matching shows that the left-hand side is the Cubic Formula for $p = 3$ and $q = -14$. Thus, it is a (real) root of the cubic equation $x^3 + 3x - 14$. A simple check shows that $x = 2$ is a root of this polynomial. Since $(q/2)^2 + (p/3)^3 = 7^2 + 1 = 50 > 0$, this is the only real root. The equality follows.

Alternatively, synthetic division with $x - 2$ gives

$$\begin{array}{r|rrr} & 1 & 0 & 3 & -14 \\ 2 & & 2 & 4 & 14 \\ \hline & 1 & 2 & 7 & 0 \end{array}$$

Hence, we have the factorization $x^3 + 3x - 14 = (x - 2)(x^2 + 2x + 7)$. The discriminant of the quadratic factor is $4 - 28 = -24 < 0$. This means that $x = 2$ is the only real root.

Example 7.2.3 Solve the cubic equations:

$$(a)\ x^3 + 3x - 1 = 0; \quad (b)\ x^3 - 27x + 54 = 0; \quad (c)\ x^3 - x^2 + 1 = 0.$$

In (a), $p = 3$ and $q = -1$, so that we have

$$\sqrt{\left(\frac{q}{2}\right)^2 + \left(\frac{p}{3}\right)^3} = \sqrt{\frac{1}{4} + 1} = \sqrt{\frac{5}{4}} = \frac{\sqrt{5}}{2}.$$

[3] Inspired by a problem in the Kettering University Mathematics Olympiad, 2007. Similar problems abound in mathematical contests.

Using the Cubic Formula we obtain the real root

$$\sqrt[3]{\frac{1-\sqrt{5}}{2}}+\sqrt[3]{\frac{1+\sqrt{5}}{2}}.$$

The other two roots are complex.

In (b), we have $p=-27$ and $q=54$ and hence

$$\left(\frac{q}{2}\right)^2+\left(\frac{p}{3}\right)^3=27^2-9^3=(3^3)^2-(3^2)^3=0.$$

The Cubic Formula gives the real root

$$2\sqrt[3]{-\frac{54}{2}}=-2\sqrt[3]{27}=-6.$$

By the above, 3 is another real root of multiplicity two.

In (c), we first realize that the cubic polynomial is not in a reduced form. The substitution $x\mapsto x+1/3$ reduces our equation to the form

$$x^3-\frac{1}{3}x+\frac{25}{27}=0.$$

(The original coefficients $a=-1$, $b=0$, $c=1$ transform into $p=b-a^2/3=-1/3$ and $q=2a^3/27-ab/3+c=-2/27+1=25/27$.) We have

$$\sqrt{\left(\frac{q}{2}\right)^2+\left(\frac{p}{3}\right)^3}=\frac{\sqrt{5^4-2^2}}{2\cdot 3^3}=\frac{\sqrt{69}}{18}.$$

Continuing our computations, we have

$$\sqrt[3]{-\frac{q}{2}\pm\sqrt{\left(\frac{q}{2}\right)^2+\left(\frac{p}{3}\right)^3}}=\sqrt[3]{-\frac{5^2}{2\cdot 3^3}\pm\frac{\sqrt{69}}{2\cdot 3^2}}=\sqrt[3]{\frac{-5^2\pm 3\sqrt{69}}{2\cdot 3^3}}=\frac{\sqrt[3]{-100\pm 12\sqrt{69}}}{6}.$$

Finally, substituting this into the Cubic Formula, we obtain that the real root of our original cubic polynomial is

$$\frac{1}{3}-\frac{1}{6}\sqrt[3]{100+12\sqrt{69}}-\frac{1}{6}\sqrt[3]{100-12\sqrt{69}}.$$

The other two roots are complex.

History

The history of solving cubic equations is very complex and can be traced back to ancient times in Babylonia, Egypt, Greece, India, and China. In addition, the Persian mathematician and poet Omar Khayyàm found geometric solutions by intersecting hyperbolas and parabolas with a circle.

The Italian mathematician Scipione del Ferro (1465–1526) discovered an algebraic method to solve cubic equations (for $p > 0$ and $q < 0$) but nurtured it as a secret right before his death when he revealed it to his student, Antonio Fior. Shortly afterwards in about 1530, upon learning that another Italian mathematician, Niccolò Tartaglia (1500–1557) claimed to have solved the problem, Fiore challenged him to a contest. When Fiore was defeated, Tartaglia became well-known in mathematical circles in Italy. This drew the attention of yet another Italian mathematician, Gerolamo Cardano, who eventually persuaded Tartaglia to reveal the solution to him provided that he would not publish it. About six years later, upon having seen del Ferro's solution predating Tartaglia's, in 1545 Cardano did publish it in his *Ars Magna* giving credit to both del Ferro and Tartaglia. (See the epitaph of this chapter above.) The solution above, using the auxiliary indeterminates u and v, is the one in his book. The single substitution with the indeterminate w above is due to François Viète.

Remark Finally, we briefly discuss the case

$$\left(\frac{q}{2}\right)^2 + \left(\frac{p}{3}\right)^3 < 0.$$

Taking the square root, we obtain the purely imaginary complex number as

$$\sqrt{\left(\frac{q}{2}\right)^2 + \left(\frac{p}{3}\right)^3} = \mathrm{i}\sqrt{-\left(\frac{q}{2}\right)^2 - \left(\frac{p}{3}\right)^3},$$

where i is the complex unit, and the radical expression is real since the radicand is now positive. With this, so far we have

$$-\frac{q}{2} \pm \mathrm{i}\sqrt{-\left(\frac{q}{2}\right)^2 - \left(\frac{p}{3}\right)^3}.$$

Now, one needs to take the cubic root of these as complex numbers. Complex arithmetic shows that there are actually three distinct cubic roots of a single non-zero complex number. Corresponding to the two signs $\pm$, we thus have the total of six cubic roots. Finally, it turns out that these six complex numbers are paired up to obtain **three distinct real roots** in this case.

History
The apparent subtlety in the Cubic Formula is that in the three distinct real root case we have to recourse to complex arithmetic to recover the roots. In the 16th century complex numbers were unknown. Although the *Ars Magna* implicitly contains an example of the use of square roots of negative numbers, namely $(5+\sqrt{-15})(5-\sqrt{-15}) = 40$, Cardano himself never applied the Cubic Formula in this case.

The Cubic Formula obtained in this section provides a real root of a reduced cubic in an explicit algebraic expression (involving square and cubic roots) in the coefficients. As it has been recognized by Viète, another cubic formula can be obtained using the sine and the inverse sine functions. Although this is a transcendental (non-algebraic) method, the advantage of this formula is that in the case of three real roots it gives all of them in a single formula without having to recourse to complex arithmetic. We will discuss this in Section 11.3.

Exercises

7.2.1. Solve $x^3 - 2x^2 + 3x - 1$.
7.2.2. Simplify

$$\sqrt[3]{\frac{5\sqrt{33}-27}{18}} - \sqrt[3]{\frac{5\sqrt{33}+27}{18}}.$$

7.3 Roots of Quartic and Quintic Polynomials

In Sections 6.6 and 7.2 we demonstrated that quadratic and cubic equations can be solved by **root formulas**, algebraic expressions with the coefficients of the polynomials as indeterminates.

The question naturally arises: Is there a root formula for polynomials of higher degree?

History
Even in the early 16th century, contemporaneously with del Ferro, Tartaglia, and Cardano, there have been attempts to solve quartic (degree four) polynomial equations. In fact, working as a servant in Cardano's household and soon recognized for his brilliance in mathematics, Lodovico Ferrari (1522–1565) found a general solution for quartic equations. In yet another public contest he defeated Tartaglia, and his solution of the quartics found its way to Cardano's *Ars Magna* along with the del Ferro-Tartaglia solution of cubics. Ferrari's solution relies heavily on the Cubic Formula. In fact, to any quartic polynomial one can associate a cubic polynomial, the so-called **cubic resolvent**, and, using the roots of this resolvent, a simple algebraic trick gives the roots of the original quartic equation.
Although there is a closed formula for the roots of quartic equations, it is long and complex. Since the algebraic tools to discuss this are best done over the complex number field (not known in Ferrari's lifetime), we will not pursue this path any further.[4]
During the next two and a half centuries finding the root formula for quintic (degree five) polynomials eluded the mathematicians. Finally, in 1823 Niels Henrik Abel (1802–1829) gave a proof that no such formula exists. This result is usually called the **Ruffini-Abel Theorem** in recognition of an earlier, but incomplete, attempt by Paolo Ruffini (1765–1822). The key understanding of the break from degree four to five was provided by Évariste Galois (1811–1832), and the corresponding theory (solving many other classical problems) is known as **Galois Theory**.

As noted above, there is a root formula for quartic equations. In special cases, however, it is often easier to look for a splitting the quartic polynomial into two quadratic factors. The following example illustrates this.

Example 7.3.1 Show that the quartic polynomial $p(x) = x(x+1)(x+2)(x+3) + 1$ is the square of a quadratic polynomial and has two real roots each with multiplicity 2.

[4]See the author's *Glimpses of Algebra and Geometry*, 2nd ed. Springer, New York, 2002.

We calculate

$$p(x)=x(x+1)(x+2)(x+3)+1=x(x+3)\cdot(x+1)(x+2)+1=(x^2+3x)\cdot(x^2+3x+2)+1.$$

Letting $u = x^2 + 3x + 1$, we have

$$p(x) = (u-1)(u+1)+1 = u^2 - 1 + 1 = u^2 = \left(x^2+3x+1\right)^2.$$

The Quadratic Formula gives the two real roots $(-3\pm\sqrt{5})/2$, each of multiplicity 2.

Remark A simple consequence of the example above is that the numbers

$$n(n+1)(n+2)(n+3)+1, \quad n \in \mathbb{N},$$

are **perfect squares**.
One may ask[5] whether this holds if the number of consecutive factors is other than four; that is, for what $2 \le m \in \mathbb{N}$ is $n(n+1)(n+2)\cdots(n+m)+1$ is a perfect square for all (or some) $n \in \mathbb{N}$.
For $m = 2$ the answer is "no;" that is, no number of the form $n(n+1)+1$, $n \in \mathbb{N}$, is a perfect square. Indeed, letting $n(n+1)+1 = a^2$, $a \in \mathbb{N}$, we have $n(n+1) = a^2 - 1 = (a-1)(a+1)$. This gives $(a-1)a < n(n+1)$, and hence $a - 1 < n$. Moreover, $n(n+1) < a(a+1)$, and hence $n < a$. Combining these, we obtain $a - 1 < n < a$, which is impossible.
For $m = 3$, we have $2\cdot3\cdot4+1 = 5^2$, $4\cdot5\cdot6+1 = 11^2$, and $55\cdot56\cdot57+1 = 419^2$. It turns out that these are the only cases with perfect squares, but the proof of this is beyond the scope of this book.[6]
For $m = 4$, the answer is "no" up to $n \le 10^4$.
A related problem is to ask for what $n \in \mathbb{N}$ is the number $n!+1$ a perfect square. This is called the Brocard problem dating from 1876–1885. The pairs (n, m), $n, m \in \mathbb{N}$, satisfying $n! + 1 = m^2$ are called Brown pairs. Up until 2019, there were only three Brown pairs known: $(4, 5)$, $(5, 11)$ and $(7, 71)$. Paul Erdös and others conjectured that these are the only Brown pairs. At present this problem is unsolved. Up to $n \le 10^{15}$ the conjecture is true.

Returning to the main line, given a (monic) degree n polynomial equation

$$x^n + a_{n-1}x^{n-1} + a_{n-2}x^{n-2} + \cdots + a_1x + a_0 = 0,$$

the initial substitution

$$x \mapsto x - \frac{a_{n-1}}{n}$$

[5]The author is indebted to one of the reviewers for raising this question.

[6]This problem can be reformulated to finding the integer points on the elliptic curve $y^2 = x^3 - x + 1$ (with $x = n - 1$).

has the effect of eliminating the second highest degree monomial. We used this for $n = 3$ to obtain the reduced cubic, and, for $n = 2$, this is also the content of the completing the square technique for quadratic equations.

This substitution is the first of the so-called **Tschirnhaus transformations** which meant to reduce the original polynomial into a simpler form in which one or several coefficients vanish. The construction of Tschirnhaus transformations that eliminate lower degree monomials requires advanced tools of algebra, and they are of "exponentially" increasing complexity.

History
The original intention of Ehrenfried Walther von Tschirnhaus (1651–1708) in 1683 about the transformations that were named after him was to obtain solutions of polynomial equations by reducing them to simple ones in which all but a few coefficients vanish. Tschirnhaus himself believed (erroneously) that with these transformations any degree polynomial equations can be solved.

Although there is no root formula for quintic polynomials, using Tschirnhaus transformations, one can reduce a general quintic polynomial equation to the form

$$x^5 + px + q = 0.$$

This is the so-called **Bring-Jerrard** form. It is named after Erland Bring (1736–1798), and George Jerrard (1804–1863) (who was reluctant to accept Abel's negative resolution of the problem of quintic equations). They showed independently that this reduction is possible.

Employing yet another scaling (a suitable constant multiple of the indeterminate x), the Bring-Jerrard form can further be reduced to the form

$$x^5 + x - c = 0.$$

A root of this polynomial is called an **ultraradical**, denoted by $\sqrt[*]{c}$. Thus, the result of Bring and Jerrard can be concisely stated that the general quintic equation can be solved by root formulas that **include** ultraradicals.

Note that some specific ultraradicals can be expressed by root formulas. The following example illustrates this.

Example 7.3.2 We have

$$\sqrt[*]{1} = -\frac{1}{3} + \frac{1}{6}\sqrt[3]{100 + 12\sqrt{69}} + \frac{1}{6}\sqrt[3]{100 - 12\sqrt{69}}.$$

Recall from Example 6.4.4 the factorization

$$x^5 + x - 1 = (x^3 + x^2 - 1)(x^2 - x + 1).$$

According to Example 7.2.3 (c) (with $-x$ in place of x) the first cubic factor has the given root. The claim follows.

Exercise

7.3.1. Under what condition on $a, b, c \in \mathbb{R}$ can the quartic polynomial

$$p(x) = ax^4 + bx^3 + cx^2 + bx + a$$

be factored? Note the symmetry in the sequence of the coefficients a, b, c, b, a.

7.4 Polynomials with Rational Coefficients

In most of the previous examples the polynomial $p(x)$ in question had a rational (or even integral) root $c \in \mathbb{Q}$, and, using synthetic division, we found a factorization $p(x) = (x - c)q(x)$. The question arises whether there is a simple (arithmetical) test to find rational roots of a polynomial with rational (or even integer) coefficients.

A solution to this problem was provided by Gauss:

Rational Root Theorem *If $c = a/b \in \mathbb{Q}$ with $a, b \in \mathbb{Z}$, $b \neq 0$, is a rational root of a degree n polynomial*

$$p(x) = a_n x^n + a_{n-1} x^{n-1} + \cdots + a_1 x + a_0$$

with integer coefficients $a_0, a_1, \ldots, a_n \in \mathbb{Z}$, $a_n \neq 0$, then a divides a_0 and b divides a_n.

Proof We may assume that the fraction a/b is irreducible; that is, $\gcd(a, b) = 1$. We substitute a/b into the equation $p(x) = 0$ and obtain

$$a_n \left(\frac{a}{b}\right)^n + a_{n-1} \left(\frac{a}{b}\right)^{n-1} + \cdots + a_1 \frac{a}{b} + a_0 = 0.$$

Multiplying through by b^{n-1} we have

$$a_n \frac{a^n}{b} + a_{n-1} a^{n-1} + \cdots + a_1 a b^{n-2} + a_0 b^{n-1} = 0.$$

This shows that $a_n a^n / b$ must be an integer. Therefore b divides $a_n a^n$. Since a and b are relatively prime, we obtain that b divides a_n. (See Corollary to Proposition 1.3.1.)

Returning to our numerical equation above, we multiply through b^n/a and obtain

$$a_n a^{n-1} + a_{n-1} a^{n-2} b + \cdots + a_1 b^{n-1} + a_0 \frac{b^n}{a} = 0.$$

This shows that $a_0 b^n/a$ is an integer so that a divides $a_0 b^n$. Since $\gcd(a, b) = 1$, we obtain that a divides a_0. The Rational Root Theorem follows.

Example 7.4.1 Factor the quintic polynomial completely:

$$6x^5 - 17x^4 - x^3 + 26x^2 - 17x + 3.$$

By the Rational Root Theorem, if $r = a/b \in \mathbb{Q}$, $a, b \in \mathbb{Z}$, $(a, b) = 1$, is a rational root, then $a|3$ and $b|6$. These give the following possibilities:

$$r = \frac{a}{b} = \pm 1, \pm 3, \pm\frac{1}{2}, \pm\frac{2}{3}, \pm\frac{1}{3}, \pm\frac{1}{6}.$$

First, we immediately see that $r = 1$ is a root (since the sum of the coefficients is zero). Performing synthetic division, we obtain

$$p(x) = (x - 1)(6x^4 - 11x^3 - 12x^2 + 14x - 3).$$

Second, synthetic divisions reveal that $r = 1/2$ and $r = 1/3$ are roots. Performing them consecutively, we obtain

$$\begin{array}{c|ccccc} \frac{1}{2} & 6 & -11 & -12 & 14 & -3 \\ & & 3 & -4 & -8 & 3 \\ \hline & 6 & -8 & -16 & 6 & 0 \end{array} \quad \text{and then} \quad \begin{array}{c|cccc} \frac{1}{3} & 3 & -4 & -8 & 3 \\ & & 1 & -1 & -3 \\ \hline & 3 & -3 & -9 & 0 \end{array}.$$

Summarizing, we have so far the following:

$$p(x) = (x-1)(2x-1)(x-1/3)(3x^2-3x-9) = (x-1)(2x-1)(3x-1)(x^2-x-3).$$

For the last quadratic quotient, the Quadratic Formula gives the two roots as $r = (1 \pm \sqrt{13})/2$, both irrational numbers. With these the complete factorization is as follows:

$$p(x) = (x - 1)(2x - 1)(3x - 1)\left(x - \frac{1 + \sqrt{13}}{2}\right)\left(x - \frac{1 - \sqrt{13}}{2}\right).$$

Example 7.4.2 [7] Let $p(x)$ be a polynomial with integer coefficients. Show that if $p(0)$ and $p(1)$ are odd numbers, then $p(x)$ has no integral root.

Let

$$p(x) = a_n x^n + a_{n-1}x^{n-1} + \cdots + a_1 x + a_0, \quad a_0, a_1, \ldots, a_{n-1}, a_n \in \mathbb{Z}.$$

[7] This was a problem in the Canadian Mathematical Olympiad, 1971.

By assumption, $p(0) = a_0$ and $p(1) = a_0 + a_1 + \cdots + a_n$ are odd numbers. This implies that $a_1 + \cdots + a_n$ is an **even** number.

Assume, on the contrary, that m is an integral root of $p(x)$; that is, we have $p(m) = a_n m^n + a_{n-1} m^{n-1} + \cdots + a_1 m + a_0 = 0$. By the Rational Root Theorem, $m | a_0$ so that m must be an odd number (since a_0 is odd). Since $a_1 + \cdots + a_n$ is even, there must be an even number of coefficients a_k, $k = 1, \ldots, n$, that are odd. Since m is odd, this implies that there must also be an even number of odd terms in the sum $a_n m^n + a_{n-1} m^{n-1} + \cdots + a_1 m$. Therefore this sum is an even number. This sum, however, is equal to $-a_0$, an odd number. This is a contradiction.

It is important to emphasize that, in solving (reduced) cubics with integer coefficients and with three real roots (discussed at the end of Section 7.2), we first should look for rational roots, and apply the Rational Root Theorem. If there is a rational root, then, by synthetic division, we can bypass the often tedious arithmetic of the Cubic Formula. The following example illustrates this point.

Example 7.4.3 Solve the cubic equation $x^3 - 2x - 1 = 0$.

We have $p = -2$ and $q = -1$ so that

$$\left(\frac{q}{2}\right)^2 + \left(\frac{p}{3}\right)^3 = \frac{1}{4} - \frac{8}{27} = -\frac{5}{108} < 0.$$

Instead of getting into complex arithmetic, the Rational Root Theorem gives ± 1 as the only candidates for rational roots. Substituting, we see that -1 is indeed a root. Performing synthetic division, we obtain the factorization

$$x^3 - 2x - 1 = (x + 1)(x^2 - x - 1).$$

The roots of the quadratic factor are the golden number τ and $-1/\tau$. (See Example 3.1.2.) With these, we have the complete factorization

$$x^3 - 2x - 1 = (x + 1)\,(x - \tau)\,(x + 1/\tau)\,.$$

We now return to the original setting and discuss a special case; factorization of a quadratic polynomial with integer coefficients:

$$ax^2 + bx + c, \quad a \neq 0,\ a, b, c \in \mathbb{Z}.$$

We **assume** that our trinomial can be factored as

$$ax^2 + bx + c = \frac{(ax)^2 + abx + ac}{a} = \frac{(ax + s)(ax + t)}{a},$$

where s, t are **integers**. Expanding the last numerator, we see that this factorization is possible if and only if s and t satisfy the equations

$$st = ac \quad \text{and} \quad s + t = b.$$

Since all ingredients here are integers, the first equation says that s and t **divide** ac. Since any integer has only finitely many divisors, we can compile the list of admissible pairs (s, t). As s and t play a symmetric role, we may assume $|s| \leq |t|$. (Note that s and t may well be negative integers.) Once this list is compiled, it is a simple matter to check which one satisfies the second linear constraint.

This technique, relying on the divisors of ac, is also called the **AC method**.

A natural question is the following: Under what condition (on the trinomial) does the AC method work?

First, if it works, then, by the factorization above, $-s/a$ and $-t/a$ are roots of our trinomial. Since they are rational numbers, we see that a necessary condition for the AC method to work is that the trinomial has **rational roots**.

We now claim that the converse is also true: If the original quadratic equation has rational roots, then $s, t \in \mathbb{Z}$ exist and the AC method works.

This follows from the Quadratic Formula. Indeed, if the roots are rational, then the square root of the discriminant, $\sqrt{D}$, must also be rational. But the discriminant $D = b^2 - 4ac$ is a non-negative integer, and we showed in Section 2.1 that the square root of a non-negative integer is rational if and only if the integer itself is a perfect square. Thus, we have $D = b^2 - 4ac = d^2$ for some $d \in \mathbb{N}_0$. Rearranging, we have $4ac = b^2 - d^2 = (b - d)(b + d)$. The crux is that this equality implies that 4 divides $(b - d)(b + d)$, so that one of the factors and hence **both** $b - d$ and $b + d$ have to be even numbers. ($b - d$ is even if and only if $b + d = (b - d) + 2d$ is even.) Now the Quadratic Formula gives the roots as

$$r = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a} = \frac{-b \pm \sqrt{d^2}}{2a} = \frac{-b \pm d}{2a}.$$

By what we concluded above, the numerators $-(b - d)$ and $-(b + d)$ are even integers. Dividing by 2, we conclude that the roots are $-s/a$ and $-t/a$, where $s = (b - d)/2$ and $t = (b + d)/2$ are integers. Therefore the AC method works. (Note that, in terms of s, t, the discriminant is $D = b^2 - 4ac = (s + t)^2 - 4st = (s - t)^2$ so that $d = |s - t|$.)[8]

Example 7.4.4 Factor $12x^2 + 7x - 10$ using the AC method.

Since $ac = -120$ has many divisors, we first compile the list of all positive divisors of 120:

$$1, 2, 3, 4, 5, 6, 8, 10, 12, 15, 20, 24, 30, 40, 60, 120.$$

Thus we have the following table for the pairs (s, t) with $st = -120$ and $|s| \leq |t|$:

[8]The AC method is tedious and has very limited applicability. (It is unclear why this method plays such a paramount role in teaching basic algebra in schools.) Not only do the coefficients a, b, c have to be integers (or rational numbers at worst), but the AC method works if and only if the roots are rational numbers.

$$(\pm 1, \mp 120) \quad (\pm 2, \mp 60) \quad (\pm 3, \mp 40) \quad (\pm 4, \mp 30)$$
$$(\pm 5, \mp 24) \quad (\pm 6, \mp 20) \quad (\pm 8, \mp 15) \quad (\pm 10, \mp 12).$$

The only pair which satisfies $s + t = 7$ is $(-8, 15)$. We split the linear term $7x$ and group

$$\begin{aligned} 12x^2 + 7x - 10 = 12x^2 - 8x + 15x - 10 &= (12x^2 - 8x) + (15x - 10) \\ &= 4x(3x - 2) + 5(3x - 2) = (4x + 5)(3x - 2). \end{aligned}$$

The factorization is complete.

A famous negative case is the content of the following:

Example 7.4.5 Show that the cubic polynomial $p(x) = 8x^3 - 6x - 1$ has no rational roots.

As before, the possible rational roots $r = a/b$ are

$$\pm 1, \pm\frac{1}{2}, \pm\frac{1}{4}, \pm\frac{1}{8}.$$

Now synthetic division shows that none of these are roots of $p(x)$.

Remark The significance of this example lies in the fact that $\cos(\pi/9)$ is a root of this polynomial. (See Section 11.3.) This will imply that $\pi/3$ cannot be trisected by straightedge and compass.

Exercises

7.4.1. The equation $x^3 = 15x + 4$ appears in the 1570 edition of the Ars Magna. Find all three roots.

7.4.2. Find a real root of the cubic equation $x^3 + 6x - 20 = 0$.

7.5 Factoring Multivariate Polynomials

Factoring multivariate polynomials is often more difficult then factoring polynomials of a single indeterminate. In this section we assemble a sequence of examples starting with simple and ending with complex factoring. Whenever instructive, we will determine the zero-set of the respective polynomial.

Example 7.5.1 Factor the quartic polynomial

$$p(x, y) = x^3 y - x y^3.$$

Clearly xy is a common factor of the two monomials. We thus have

$$p(x, y) = xy(x^2 - y^2) = xy(x - y)(x + y).$$

Since the factors are linear they are irreducible, so that this is the complete decomposition of $p(x, y)$ into irreducible components.

The corresponding polynomial equation $p(x, y) = 0$ has a simple geometric interpretation on $\mathbb{R}^2$. The vanishing of each factor is represented by a line in $\mathbb{R}^2$. The equations $y = 0$ and $x = 0$ describe the first and second coordinate axes. The equations $x \pm y = 0$ correspond to the two perpendicular lines that meet at the origin and have slopes ± 1. Altogether, the entire zero-set is the union of four lines arranged in a symmetric pattern.

Example 7.5.2 Factor the quartic polynomial

$$p(x, y) = x^2y^2 - x^2 - y^2 + 1.$$

This polynomial is **biquadratic**, a quadratic polynomial in the indeterminates x^2 and y^2. This motivates us to set $a = x^2$ and $b = y^2$ and factor the quadratic polynomial in a and b as $ab - a - b + 1 = (a - 1)(b - 1)$. Returning to our original indeterminates, we obtain the complete factorization

$$p(x, y) = (x^2 - 1)(y^2 - 1) = (x - 1)(x + 1)(y - 1)(y + 1).$$

The equation $p(x, y) = 0$ on $\mathbb{R}^2$ is represented by four lines, the extensions of the four sides of the square with vertices $(\pm 1, \pm 1)$.

Example 7.5.3 Factor the biquadratic polynomial

$$p(x, y) = x^4 + y^4 - 2x^2y^2 - 2x^2 - 2y^2 + 1.$$

First Solution. As in the previous example, setting $a = x^2$ and $b = y^2$, we need to factor the expression

$$a^2 + b^2 - 2ab - 2a - 2b + 1.$$

One is tempted to use the binomial identity $(a - b)^2 = a^2 - 2ab + b^2$ but this does not match the linear terms $-2a - 2b = -2(a + b)$. Insisting on the presence of the expression $a + b$, we split the term $-2ab$ into $2ab - 4ab$, rearrange, and rewrite this as

$$a^2+2ab+b^2-2(a+b)+1-4ab = (a+b)^2-2(a+b)+1-4ab = (a+b-1)^2-4ab.$$

We can factor this at the expense of introducing the square roots of a and b, and appealing to the difference of squares identity. Instead, we now go back to our original indeterminates x and y and calculate

$$\begin{aligned} p(x, y) &= (x^2 + y^2 - 1)^2 - 4x^2y^2 = (x^2 + y^2 - 1)^2 - (2xy)^2 \\ &= (x^2 + y^2 - 2xy - 1)(x^2 + y^2 + 2xy - 1) = ((x - y)^2 - 1)((x + y)^2 - 1) \\ &= (x - y - 1)(x - y + 1)(x + y - 1)(x + y + 1). \end{aligned}$$

Second Solution. This time we let $a = x + y$ and $b = x - y$. Squaring, we have

$$a^2 = x^2 + y^2 + 2xy \quad \text{and} \quad b^2 = x^2 + y^2 - 2xy.$$

We calculate

$$a^2b^2 = (x^2+y^2+2xy)(x^2+y^2-2xy) = (x^2+y^2)^2-4x^2y^2 = x^4+y^4-2x^2y^2,$$

and

$$a^2 + b^2 = 2x^2 + 2y^2.$$

Using these the original polynomial rewrites as

$$p(x, y) = a^2b^2 - a^2 - b^2 + 1.$$

We now notice that this is precisely the polynomial of the previous example (in the indeterminates a and b). We thus have

$$p(x, y) = (a-1)(a+1)(b-1)(b+1) = (x+y-1)(x+y+1)(x-y-1)(x-y+1).$$

In both cases the equation $p(x, y) = 0$ is geometrically represented in $\mathbb{R}^2$ as the line extensions of the four sides of the square with vertices $(\pm 1, 0)$ and $(0, \pm 1)$.

Example 7.5.4 Factor the biquadratic polynomial

$$p(x, y, z) = x^4 + y^4 + z^4 - 2x^2y^2 - 2y^2z^2 - 2z^2x^2.$$

First Solution. We first notice that $p(x, y, z)$ is homogeneous since all monomials have degree 4. A simple method to reduce the number of indeterminates is to "dehomogenize" $p(x, y, z)$ by dividing by z^4, say, and changing to the new indeterminates $u = x/z$ and $v = y/z$. We obtain

$$\frac{p(x, y, z)}{z^4} = p(u, v, 1) = u^4 + v^4 - 2u^2v^2 - 2u^2 - 2v^2 + 1.$$

By the previous example, this factors as

$$u^4 + v^4 - 2u^2v^2 - 2u^2 - 2v^2 + 1 = (u - v - 1)(u - v + 1)(u + v - 1)(u + v + 1).$$

Reverting to the original indeterminates and multiplying through by z^4, we obtain

$$p(x, y, z) = (x - y - z)(x - y + z)(x + y - z)(x + y + z).$$

Second Solution. A direct solution is based on breaking the cyclic symmetry $x \mapsto y \mapsto z \mapsto x$ as follows:

$$\begin{aligned} p(x, y, z) &= x^4 + y^4 + z^4 - 2x^2y^2 + 2y^2z^2 - 2z^2x^2 - 4y^2z^2 \\ &= (x^2 - y^2 - z^2)^2 - 4y^2z^2 \\ &= (x^2 - y^2 - z^2 - 2yz)(x^2 - y^2 - z^2 + 2yz) \\ &= (x^2 - (y + z)^2)(x^2 - (y - z)^2) \\ &= (x - y - z)(x + y + z)(x - y + z)(x + y - z). \end{aligned}$$

Factorization may be a critical tool in deriving inequalities. The following simple example illustrates this:

Example 7.5.5 Show that, for $0 < a, b \in \mathbb{R}$, we have

$$a^3 + b^3 \geq ab(a + b).$$

Indeed, this holds because of the factorization

$$a^3 + b^3 - ab(a + b) = (a - b)^2(a + b) \geq 0.$$

Example 7.5.6 Factor the cubic polynomial

$$p(x, y) = 2x^3 - 6xy^2 - 3x^2 - 3y^2 + 1.$$

We first isolate the terms that contain the indeterminate y:

$$p(x, y) = 2x^3 - 3x^2 + 1 - 3(2x + 1)y^2.$$

Next, we notice that the cubic polynomial formed by the first three monomials has $-1/2$ as a root. Hence $(2x + 1)$ is a common factor:

$$p(x, y) = (2x + 1)(x^2 - 2x + 1 - 3y^2).$$

We now have

$$p(x, y) = (2x + 1)((x - 1)^2 - 3y^2) = (2x + 1)(x - \sqrt{3}y - 1)(x + \sqrt{3}y - 1).$$

Although the presence of the "irrationality"[9] $\sqrt{3}$ may indicate some complexity, the geometric characterization of the zero-set $p(x, y) = 0$ is simple and elegant. The

[9] As a nineteenth century mathematician would call it.

zero-set is represented by three lines; the vertical line given by $x = -1/2$, and two additional lines given by $y = \pm 1/\sqrt{3}(x-1)$ and meeting at the point $(1, 0)$. The vertical line cuts out two additional intersection points $(-1/2, \pm\sqrt{3}/2)$. Calculating distances, we realize that these three points are the vertices of an equilateral triangle inscribed into the unit circle $\mathbb{S}$. We conclude that $p(x, y) = 0$ represents the union of three lines which are the extensions of the sides of this triangle.

Example 7.5.7 Factor the quartic polynomial

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

First Solution. We group various terms and calculate

$$p(x, y) = x^2y^2 + 2x^2y + 2xy^2 + x^2 + 2xy + y^2 = x^2(y^2 + 2y + 1) + 2xy^2 + 2xy + y^2$$
$$= x^2(y+1)^2 + 2xy(y+1) + y^2 = (x(y+1) + y)^2 = (xy + x + y)^2.$$

Second Solution. For a less "ad hoc" approach, we notice that $p(x, y)$ is symmetric, so that the Fundamental Theorem on Symmetric Polynomials applies. Using the elementary symmetric polynomials $s_1(x, y) = x + y$ and $s_2(x, y) = xy$, we obtain

$$p(x, y) = x^2y^2 + 2xy(x + y) + (x + y)^2 = s_2^2 + 2s_2s_1 + s_1^2 = (s_2 + s_1)^2.$$

Returning to our original indeterminates x, y, we arrive at $p(x, y) = (xy + x + y)^2$.

We now turn to more complex cubic polynomials:

Example 7.5.8 Factor the cubic polynomial

$$p(x, y, z) = x^3 + y^3 + z^3 - 3xyz.$$

First Solution. This polynomial is **homogeneous** of degree three; that is, for $t \in \mathbb{R}$, we have $p(tx, ty, tz) = t^3p(x, y, z)$, and **symmetric**. This indicates that the factors, if any, should have similar properties. The simplest symmetric homogeneous expressions (up to degree two) in the indeterminates x, y, z are

$$x + y + z, \quad x^2 + y^2 + z^2, \quad xy + yz + zx.$$

To form cubic expressions, we calculate

$$(x + y + z)(x^2 + y^2 + z^2) = x^3 + y^3 + z^3 + xy^2 + yx^2 + yz^2 + zy^2 + zx^2 + xz^2,$$

and

$$(x + y + z)(xy + yz + zx) = 3xyz + xy^2 + yx^2 + yz^2 + zy^2 + zx^2 + xz^2.$$

Subtracting, we obtain

$$(x+y+z)(x^2+y^2+z^2)-(x+y+z)(xy+yz+zx)=x^3+y^3+z^3-3xyz.$$

Since $x+y+z$ is a common factor, we arrive at the factorization

$$p(x,y,z)=x^3+y^3+z^3-3xyz=(x+y+z)(x^2+y^2+z^2-xy-yz-zx).$$

Second Solution. Once again a less "ad hoc" method is to observe that our polynomial $p(x,y,z)$ is **symmetric**, and apply the Fundamental Theorem of Symmetric Polynomials. In three indeterminates x, y, z it says that any symmetric polynomial $p(x,y,z)$ can be uniquely written as a polynomial of the **elementary symmetric polynomials** s_1, s_2, s_3 as indeterminates, where

$$s_1(x,y,z)=x+y+z;\ \ s_2(x,y,z)=xy+yz+zx;\ \ s_3(x,y,z)=xyz.$$

For our cubic, by homogeneity, the only possibility is

$$p(x,y,z)=As_1^3(x,y,z)+Bs_1(x,y,z)s_2(x,y,z)+Cs_3(x,y,z)$$

with appropriate constants $A, B, C \in \mathbb{R}$. Comparing coefficients, we find that $A=1$, $B=-3$, and $C=0$. With these, we have

$$\begin{aligned}p(x,y,z)&=s_1^3(x,y,z)-3s_1(x,y,z)s_2(x,y,z)\\&=s_1(x,y,z)(s_1(x,y,z)^2-3s_2(x,y,z))\\&=(x+y+z)((x+y+z)^2-3(xy+yz+zx))\\&=(x+y+z)(x^2+y^2+z^2-xy-yz-zx).\end{aligned}$$

The (double of the) last quadratic factor in the previous example can be written in another symmetric form:

$$\begin{aligned}&2(x^2+y^2+z^2-xy-yz-zx)\\&\quad=(x^2-2xy+y^2)+(y^2-2yz+z^2)+(z^2-2zx+x^2)\\&\quad=(x-y)^2+(y-z)^2+(z-x)^2.\end{aligned}$$

This is the sum of three squares so that it is non-negative. In particular, it is zero if and only if $x=y=z$.

Using this in the example above, for $x, y, z \geq 0$, we obtain

$$x^3+y^3+z^3-3xyz\geq 0,$$

or equivalently

$$xyz \leq \frac{x^3 + y^3 + z^3}{3},$$

with equality if and only if $x = y = z$. Finally, letting $x = \sqrt[3]{a}$, $y = \sqrt[3]{b}$, $z = \sqrt[3]{c}$, we arrive at the following:

$$\sqrt[3]{abc} \leq \frac{a + b + c}{3}, \quad a, b, c \geq 0.$$

Equality holds if and only if $a = b = c$. This is the **AM-GM inequality in three indeterminates**.

A variation on the theme is the following:

Example 7.5.9 Factor the cubic polynomial

$$p(x, y, z) = (x - y)^3 + (y - z)^3 + (z - x)^3.$$

The form of $p(x, y, z)$ suggests to introduce the new indeterminates $a = x - y$, $b = y - z$ and $c = z - x$. Their sum automatically vanishes:

$$a + b + c = (x - y) + (y - z) + (z - x) = 0.$$

According to the factorization in Example 7.5.8 above, we have

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

In terms of our original indeterminates x, y, z, this then gives

$$(x - y)^3 + (y - z)^3 + (z - x)^3 - 3(x - y)(y - z)(z - x) = 0.$$

With this we arrive at the factorization

$$p(x, y, z) = (x - y)^3 + (y - z)^3 + (z - x)^3 = 3(x - y)(y - z)(z - x).$$

There are several beautiful applications of the factorization of the cubic polynomial $x^3 + y^3 + z^3 - 3xyz$ in Example 7.5.8. We give here two:

Example 7.5.10 Let $0 \neq r \in \mathbb{R}$ such that $\sqrt[3]{r} + 1/\sqrt[3]{r} = a \in \mathbb{R}$. Calculate $r^3 + 1/r^3$ in terms of a.

Letting $x = \sqrt[3]{r}$, $y = 1/\sqrt[3]{r}$, $z = -a$, we have $x+y+z = 0$ so that Example 7.5.8 gives

$$(\sqrt[3]{r})^3 + 1/(\sqrt[3]{r})^3 + (-a)^3 = 3 \cdot \sqrt[3]{r} \cdot 1/\sqrt[3]{r} \cdot (-a) = -3a.$$

We obtain $r + 1/r = a^3 - 3a$. Applying the same identity again, we have

$$r^3 + \frac{1}{r^3} + (-a^3 + 3a)^3 = 3 \cdot r \cdot \frac{1}{r} \cdot (-a^3 + 3a) = -3a^3 + 9a.$$

This gives

$$r^3 + 1/r^3 = (a^3 - 3a)^3 - 3a(a^2 - 3) = a(a^2 - 3)(a^2(a^2 - 3)^2 - 3).$$

Example 7.5.11 Let $a, b, c \in \mathbb{R}$ be such that $a + b + c = 0$. Solve the following equation for $t \in \mathbb{R}$:

$$\sqrt[3]{t - a} + \sqrt[3]{t - b} + \sqrt[3]{t - c} = 0.$$

Using the identity above for $x = \sqrt[3]{t - a}$, $y = \sqrt[3]{t - b}$, $z = \sqrt[3]{t - c}$, the equation to be solved is equivalent to

$$(t - a) + (t - b) + (t - c) - 3\sqrt[3]{(t - a)(t - b)(t - c)} = 0.$$

Since $a + b + c = 0$, this gives $t = \sqrt[3]{(t - a)(t - b)(t - c)}$. Taking the cube of both sides, we obtain $t^3 = (t - a)(t - b)(t - c)$. Expanding, the cubic and quadratic terms cancel, and we arrive at $t(ab + bc + ca) = abc$. If $ab + bc + ca \neq 0$, then the unique solution is $t = abc/(ab + bc + ca)$. If $ab + bc + ca = 0$, then there is no solution if $abc \neq 0$, and all real numbers are solutions if $abc = 0$. The example follows.

Example 7.5.12 Factor the cubic polynomial

$$p(x, y, z) = x^3 + y^3 + z^3 - (x + y + z)^3.$$

Since $p(x, y, z)$ is symmetric, as a first step, using the result of the second solution of Example 7.5.8, we can write it in terms of the elementary symmetric polynomials s_1, s_2, s_3 as

$$p(x, y, z) = -3(s_1(x, y, z)s_2(x, y, z) - s_3(x, y, z)).$$

We now calculate (by breaking the symmetry):

$$\begin{aligned} s_1(x, y, z)s_2(x, y, z) - s_3(x, y, z) &= (x + y + z)(xy + yz + zx) - xyz \\ &= (x + y)(xy + yz + zx) + z(xy + yz + zx) - xyz \\ &= (x + y)(xy + yz + zx) + (x + y)z^2 \\ &= (x + y)(xy + yz + zx + z^2) = (x + y)(y + z)(z + x). \end{aligned}$$

With this we finally arrive at the factorization

$$p(x, y, z) = x^3 + y^3 + z^3 - (x + y + z)^3 = -3(x + y)(y + z)(z + x).$$

To finish this section we return to the AM-GM inequality in three indeterminates, and show two simple applications:

Example 7.5.13 Show that, for $0 < a, b, c \in \mathbb{R}$, we have

$$\frac{a}{b} + \frac{b}{c} + \frac{c}{a} \geq 3.$$

Indeed, we have

$$\frac{a}{b} + \frac{b}{c} + \frac{c}{a} \geq 3\sqrt[3]{\frac{a}{b}\frac{b}{c}\frac{c}{a}} = 3.$$

Example 7.5.14 Show that, for $0 < a, b, c \in \mathbb{R}$, we have

$$(a^2b + b^2c + c^2a)(ab^2 + bc^2 + ca^2) \geq 9a^2b^2c^2.$$

We use the AM-GM inequality for each factor on the left-hand side as follows

$$(a^2b + b^2c + c^2a)(ab^2 + bc^2 + ca^2) \geq (3\sqrt[3]{a^3b^3c^3})(3\sqrt[3]{a^3b^3c^3}) = 9a^2b^2c^2.$$

The example follows.

Exercises

7.5.1. Factor the binomial $x^{10} - y^{10}$.

7.5.2. Find the **number** of solutions of integral quadruples (a, b, c, d), $a, b, c, d \in \mathbb{Z}$, satisfying $ab + cd = ac + bd = ad + bc = -2$.

7.6 The Greatest Common Factor

The **greatest common factor (gcf)** of two polynomials $a(x)$ and $b(x)$, at least one of which is non-zero, is the polynomial of largest degree that divides both $a(x)$ and $b(x)$. It is usually denoted[10] by $\gcf(a(x), b(x))$. It is uniquely determined by $a(x)$

[10] When discussing $\gcf(a(x), b(x))$, we always tacitly assume that at least one of the polynomials $a(x)$ or $b(x)$ is non-zero.

and $b(x)$ **up to a non-zero constant multiple**. Alternatively, adding the requirement that the gcf must be monic (with leading coefficient 1) it becomes unique.

Remark The greatest common divisor of two integers is sometimes called the greatest common factor. For clarity, we keep the two concepts separate, so that the greatest common factor applies only for polynomials.

The Euclidean Algorithm for integers (and its proof) (Section 1.3) can be transplanted almost verbatim to obtain the Euclidean Algorithm to find the gcf of two polynomials. The only change is that, instead of keeping track of the numerical values of the remainders, we need to keep track of their degrees. With this, the Euclidean algorithm to find $\gcf(a(x), b(x))$ for two polynomials $a(x)$ and $b(x)$ with $\deg a(x) \geq \deg b(x)$ is as follows:

$$\begin{aligned}
a(x) &= b(x)q_1(x) + r_1(x), \quad && \deg r_1(x) < \deg b(x)\\
b(x) &= r_1(x)q_2(x) + r_2(x), \quad && \deg r_2(x) < \deg r_1(x)\\
r_1(x) &= r_2(x)q_3(x) + r_3(x), \quad && \deg r_3(x) < \deg r_2(x)\\
r_2(x) &= r_3(x)q_4(x) + r_4(x), \quad && \deg r_4(x) < \deg r_3(x)\\
&\cdots \quad \cdots \quad \cdots\\
r_{n-3}(x) &= r_{n-2}(x)q_{n-1}(x) + r_{n-1}(x), \quad && \deg r_{n-1}(x) < \deg r_{n-2}(x)\\
r_{n-2}(x) &= r_{n-1}(x)q_n(x).
\end{aligned}$$

As before, we set the indices such that $r_n(x) = 0$. Thus, we have

$$\gcf(a(x), b(x)) = r_{n-1}(x).$$

Example 7.6.1 Find $\gcf(x^3 - 3x^2 + 3x - 2, x^2 - 5x + 6)$.

Using long divisions, a straightforward computation gives

$$\begin{aligned}
x^3 - 3x^2 + 3x - 2 &= (x^2 - 5x + 6)(x + 2) + 7x - 14\\
x^2 - 5x + 6 &= (7x - 14)\left(\frac{1}{7}x - \frac{3}{4}\right).
\end{aligned}$$

Hence $\gcf(x^3 - 3x^2 + 3x - 2, x^2 - 5x + 6) = 7x - 14$.

A final general remark. As before, systematic elimination of the intermediate remainders $r_1(x), r_2(x), \ldots, r_{n-2}(x)$ gives the following: There exist polynomials $k(x)$ and $l(x)$ such that we have

$$\gcf(a(x), b(x)) = k(x) \cdot a(x) + l(x) \cdot b(x).$$

Example 7.6.2 If an **irreducible** polynomial $p(x)$ divides a product $a(x) \cdot b(x)$ of two polynomials, then $p(x)$ divides $a(x)$ or $b(x)$.

Assume that $p(x)$ does not divide $a(x)$. Since $p(x)$ is irreducible, $\gcf(p(x), a(x)) = 1$. By the remark above, there exist polynomials $k(x)$ and $l(x)$ such that $1 = \gcf(p(x), a(x)) = k(x)p(x) + l(x)a(x)$. Multiplying through by $b(x)$, we get $b(x) = k(x)p(x)b(x) + l(x)a(x)b(x)$. Now, since $p(x)$ divides $a(x)b(x)$, it must divide the sum on the right-hand side, and hence $b(x)$. The example follows.

Exercise

7.6.1. Calculate

(*a*) $\gcf(x^4 - x^3 + 7x^2 - 6x + 6, x^5 - x^4 + x^3 + 3x^2 - 3x + 3)$;

(*b*) $\gcf(x^5 - 5x^4 + x^3 + 6x^2 - 30x + 6, x^6 - 5x^5 + x^4 + 3x^2 - 15x + 3)$.

Chapter 8
Conics

"Eratosthenes, in his work entitled Platonicus relates that, when the god proclaimed to the Delians through the oracle that, in order to get rid of a plague, they should construct an altar double that of the existing one, their craftsmen fell into great perplexity in their efforts to discover how a solid could be made the double of the similar solid..."
Theon of Smyrna (c. 70–c. 135) quoting Eratosthenes

In this short chapter, we give a complete and elementary classification of conics without using linear algebraic tools. We derive many classical properties of them with applications and full historical details. We show how parabolas can be used to give a geometric interpretation of the Babylonian method of extracting square roots. Finally, we use symmetry properties of hyperbolas to present a geometric proof of the famous 1988 International Mathematical Olympiad problem discussed in Chapter 6 (Example 6.6.8).

8.1 The General Conic

Conics, or **quadratic curves**, are important examples of plane curves possessing many elegant geometric properties. The classical (geometric) term "conic (section)" is because these curves are intersections of the surface of a right circular double cone with a plane. The (algebraic) term "quadratic curve" is due to the fact that they can be represented as the zero-set $\{(x, y) \in \mathbb{R}^2 \mid p(x, y) = 0\}$ of a **quadratic polynomial** $p(x, y)$ in two indeterminates x and y. Although we pursue here an algebraic approach we retain the geometric term "conic."

A conic is **non-degenerate** if the representing quadratic polynomial $p(x, y)$ is **irreducible**; that is, if it does not factor into a product of two linear factors. A degenerate conic is a pair of intersecting or parallel lines (including the case when the two lines coincide). In addition, the **single point**, and the **empty set** are also considered degenerate conics.

G. Toth, *Elements of Mathematics*, Undergraduate Texts in Mathematics,
https://doi.org/10.1007/978-3-030-75051-0_8

Example 8.1.1 The zero-set of the polynomial $(ax - by - c)(a'x - b'y - c')$, $a^2 + b^2 > 0$, $a'^2 + b'^2 > 0$ is a pair of lines (with various incidence properties discussed in Section 5.2). The zero-set of the polynomial $x^2 + y^2$ consists of the origin 0 only. Finally, the zero-sets of the polynomials $x^2 + 1$ and $y^2 + 1$ are the empty set. These are all degenerate conics.

Remark Over the complex number field $\mathbb{C}$, a quadratic polynomial $p(x, y)$ (with real coefficients) is called **absolutely irreducible** if it does not factor into complex linear factors. A conic is called non-degenerate if the associated polynomial is absolutely irreducible. In the examples above the conics are all reducible over $\mathbb{C}$: $x^2 + y^2 = (x + \mathrm{i}y)(x - \mathrm{i}y)$, $x^2 + 1 = (x + \mathrm{i})(x - \mathrm{i})$ and $y^2 + 1 = (y + \mathrm{i})(y - \mathrm{i})$. Staying within the real number system, we will not use this terminology.

We now begin the study of non-degenerate conics. We split a general quadratic polynomial $p(x, y)$ in two indeterminates x, y into **homogeneous components** as

$$p(x, y) = p_2(x, y) + p_1(x, y) + p_0, \quad p_0 \in \mathbb{R},$$

where the subscripts stand for the degree. Expanding, we have

$$p_2(x, y) = Ax^2 + By^2 + Cxy, \quad p_1(x, y) = Ux + Vy, \quad p_0 = K.$$

where $A, B, C, U, V, K \in \mathbb{R}$ and A, B, C do not vanish simultaneously.

To reduce the complexity of the polynomial $p(x, y)$ we will perform several substitutions.

First, we let $(a_0, b_0) \in \mathbb{R}^2$ such that $a_0^2 + b_0^2 = 1$, and introduce the change of variables

$$x \mapsto a_0x - b_0y \quad \text{and} \quad y \mapsto b_0x + a_0y.$$

We pause here to discuss the geometric meaning of this. By assumption, the point $Q = (a_0, b_0)$ is on the unit radius circle $\mathbb{S}$ (with center at the origin 0). The point Q is uniquely determined by the angle measure $\theta = \alpha_0(\ell)$, where ℓ is the half-line with end-point at 0 and containing Q. (Recall that $\alpha_0(\ell) = \mu(\angle \ell_+ 0 \ell)$, where ℓ_+ is the positive first axis.) We view the substitution above as a transformation $R_\theta : \mathbb{R}^2 \to \mathbb{R}^2$ given by

$$R_\theta(P) = (a_0x - b_0y, b_0x + a_0y), \quad P = (x, y) \in \mathbb{R}^2.$$

We claim that R_θ preserves the Cartesian distance d; that is, we have

$$d(R_\theta(P_0), R_\theta(P_1)) = d(P_0, P_1), \quad P_0, P_1 \in \mathbb{R}^2.$$

Indeed, for $P_0 = (x_0, y_0)$ and $P_1 = (x_1, y_1)$, we calculate

$$\begin{aligned} &d(R_\theta(P_0), R_\theta(P_1))^2 \\ &\quad = ((a_0x_0 - b_0y_0) - (a_0x_1 - b_0y_1))^2 + ((b_0x_0 + a_0y_0) - (b_0x_1 + a_0y_1))^2 \\ &\quad\quad = (a_0(x_0 - x_1) - b_0(y_0 - y_1))^2 + (b_0(x_0 - x_1) + a_0(y_0 - y_1))^2 \\ &\quad\quad\quad = a_0^2(x_0 - x_1)^2 + b_0^2(y_0 - y_1)^2 + b_0^2(x_0 - x_1)^2 + a_0^2(y_0 - y_1)^2 \\ &\quad\quad\quad\quad = (x_0 - x_1)^2 + (y_0 - y_1)^2 = d(P_0, P_1)^2. \end{aligned}$$

The claim follows.

Clearly, R_θ fixes the origin: $R_\theta(0) = 0$. Next, we claim that R_θ preserves the orientation; in fact, we have

$$\omega(0, R_\theta(P_0), R_\theta(P_1)) = \omega(0, P_0, P_1), \quad P_0, P_1 \in \mathbb{R}^2.$$

Using the previous notations, we calculate

$$\begin{aligned} \omega(0, R_\theta(P_0), R_\theta(P_1)) &= (a_0x_0 - b_0y_0)(b_0x_1 + a_0y_1) - (a_0x_1 - b_0y_1)(b_0x_0 + a_0y_0) \\ &= a_0^2x_0y_1 - b_0^2x_1y_0 - (a_0^2x_1y_0 - b_0^2x_0y_1) \\ &= x_0y_1 - x_1y_0 = \omega(0, P_0, P_1). \end{aligned}$$

The positive first axis ℓ_+ is sent by R_θ to the half-line ℓ with end-point 0 and containing Q. Since R_θ preserves distances and orientation, it follows easily from the Birkhoff Postulate of Similarity that R_θ is the (positive) **rotation** with angle θ about the origin 0. We also see that $R_{\pi/2} = S_0$ is the (positive) quarter-turn about the origin 0.

Remark Rotation about any point O with angle θ can be obtained as the composition[1] $R_{\theta,O} = T_O \circ R_\theta \circ T_{-O}$.

We now return to our conics, and apply the substitution above (algebraically), or perform the rotation R_θ (geometrically). Since the components of the substitution are homogeneous of degree 1, it follows that the homogeneous components of $p(x, y)$ are transformed independently.

More specifically, for the degree 2 component, we have

$$\begin{aligned} &p_2(a_0x - b_0y, b_0x + a_0y) = \\ &\quad = A(a_0x - b_0y)^2 + B(b_0x + a_0y)^2 + C(a_0x - b_0y)(b_0x + a_0y) \\ &\quad\quad = (Aa_0^2 + Bb_0^2 + Ca_0b_0)x^2 + (Ba_0^2 + Ab_0^2 - Ca_0b_0)y^2 \\ &\quad\quad\quad + \Big(C(a_0^2 - b_0^2) - 2(A - B)a_0b_0\Big)\,xy. \end{aligned}$$

[1] It is a simple fact that any transformation in the plane that preserves distances and the orientation is either a rotation or a translation. We will not need this.

We claim that the quantity $C^2 - 4AB$ is unchanged by this substitution.[2] This can be shown by direct computation:

$$\begin{aligned}\Big(C(a_0^2-b_0^2)-2(A-B)a_0b_0\Big)^2 &-4(Aa_0^2+Bb_0^2+Ca_0b_0)(Ba_0^2+Ab_0^2-Ca_0b_0)\\ &= C^2(a_0^2-b_0^2)^2 - 8ABa_0^2b_0^2 - 4AB(a_0^4+b_0^4) + 4C^2a_0^2b_0^2\\ &= C^2(a_0^2+b_0^2)^2 - 4AB(a_0^2+b_0^2)^2 = C^2 - 4AB.\end{aligned}$$

For the degree 1 component, we have

$$\begin{aligned}p_1(a_0x - b_0y, b_0x + a_0y) &= U(a_0x - b_0y) + V(b_0x + a_0y)\\ &= (Ua_0 + Vb_0)x + (Va_0 - Ub_0)y.\end{aligned}$$

We claim that the quantity $U^2 + V^2$ is unchanged by this substitution. Indeed, we have

$$(Ua_0 + Vb_0)^2 + (Va_0 - Ub_0)^2 = U^2(a_0^2 + b_0^2) + V^2(a_0^2 + b_0^2) = U^2 + V^2.$$

Finally, the degree 0 component $p_0 = K$ clearly stays the same.

We use this substitution to eliminate the hybrid term Cxy in $p_2(x, y)$. By the computation above, this term vanishes if and only if

$$2(A - B)a_0b_0 = C(a_0^2 - b_0^2).$$

Squaring both sides and adding

$$4(A - B)^2a_0^2b_0^2 = C^2(a_0^2 - b_0^2)^2 = C^2(a_0^2 + b_0^2)^2 - 4C^2a_0^2b_0^2 = C^2 - 4C^2a_0^2b_0^2.$$

This gives

$$C^2 = 4\Big((A - B)^2 + C^2\Big)a_0^2b_0^2.$$

We may assume $C \neq 0$ since otherwise there is no hybrid term in the original polynomial. (If $C = 0$ and $A = B(\neq 0)$, then, as we will see later, the original equation $p(x, y) = A(x^2 + y^2) + Ux + Vy + K = 0$ gives either a circle, a point, or the empty set. If $C = 0$ and $A \neq B$, then $a_0b_0 = 0$, and therefore θ is an integer multiple of $\pi/2$.)

[2] $AB - (C/2)^2$ is the determinant of the quadratic form $p_2(x, y)$. With somewhat more linear algebra, its invariance under linear isometries follows from general facts about quadratic forms. As always, we prefer to give an elementary proof.

We obtain

$$a_0 b_0 = \pm \frac{|C|}{2\sqrt{(A-B)^2 + C^2}}.$$

This gives

$$2|a_0 b_0| \leq 1$$

with equality if and only if $A = B$.

Since $a_0^2 + b_0^2 = 1$, the **individual** values of a_0 and b_0 can be recovered from the value of $a_0 b_0$ as above via the equations

$$(a_0 + b_0)^2 = 1 + 2a_0 b_0, \quad (a_0 - b_0)^2 = 1 - 2a_0 b_0.$$

We have

$$a_0 + b_0 = \pm\sqrt{1 + 2a_0 b_0}, \quad a_0 - b_0 = \pm\sqrt{1 - 2a_0 b_0},$$

and hence

$$a_0 = \frac{\pm\sqrt{1 + 2a_0 b_0} \pm \sqrt{1 - 2a_0 b_0}}{2}, \quad b_0 = \frac{\pm\sqrt{1 + 2a_0 b_0} \mp \sqrt{1 - 2a_0 b_0}}{2}.$$

From now on, we assume that this substitution has been performed and the hybrid term has been eliminated. We now rename the new coefficients by reverting to the original notation and restart our study with the (transformed) conic given by the zero-set of the polynomial

$$p(x, y) = Ax^2 + By^2 + Ux + Vy + K = 0, \quad A^2 + B^2 > 0, \ A, B, U, V, K \in \mathbb{R},$$

where $Ax^2 + By^2$ and $Ux + Vy$ stand for the transformed (and renamed) degree 2 and degree 1 components.

We now split our treatment into three cases according to whether AB is zero, positive, or negative.

Case I $AB = 0$. We may assume $B = 0$, since otherwise we swap the indeterminates $x \leftrightarrow y$ (corresponding geometrically to reflection in the line given by the equation $x - y = 0$).

Since $A \neq 0$, we can write the polynomial as

$$p(x, y) = Ax^2 + Ux + Vy + K = A\left(x + \frac{U}{2A}\right)^2 + Vy + \left(K - \frac{U^2}{4A}\right).$$

We now perform another substitution, $x \mapsto x + U/(2A)$, corresponding to the translation T_Z, $Z = (U/(2A), 0) \in \mathbb{R}^2$, and rename the constant $K \mapsto$

$K - U^2/(4A)$. With these, we arrive at the following transformed equation

$$Ax^2 + Vy + K = 0.$$

We claim that $V = 0$ leads to degeneracy. Indeed, if $V = 0$, then $x^2 = -K/A$ and we have three cases: (a) $K/A = 0$, the single line given by $x = 0$ (the second coordinate axis); (b) $K/A < 0$, parallel lines given by $x = \pm\sqrt{-K/A}$; (c) $K/A > 0$, the empty set.

Thus, we have $V \neq 0$. Our equation now takes the form $x^2 + (V/A)(y + K/V) = 0$. We perform now the final substitution $y \mapsto y + K/V$ corresponding to the translation T_W, $W = (0, K/V) \in \mathbb{R}^2$, and introduce the new constant $d = -V/(4A) \neq 0$.

With these we arrive at the normal form of the **parabola**

$$y = \frac{1}{4d}x^2.$$

Finally, note that $d > 0$ can be assumed since otherwise we perform the substitution $y \mapsto -y$ corresponding to reflection in the first coordinate axis.

Summarizing, and going back to the beginning of our study, we obtain that, up to rotations, translations, and reflections, a **non-degenerate** conic given by $p(x, y)$ above with $p_2(x, y) = Ax^2 + By^2 + Cxy$ and satisfying $C^2 - 4AB = 0$ has the normal form of a parabola. Since all these transformations preserve the Cartesian distance, the metric properties, that is, properties that can be expressed in terms of the distance, will remain unchanged.

Cases II–III $AB \neq 0$. We can write the polynomial as

$$\begin{aligned} p(x, y) &= Ax^2 + By^2 + Ux + Vy + K \\ &= A\left(x + \frac{U}{2A}\right)^2 + B\left(x + \frac{V}{2B}\right)^2 + \left(K - \frac{U^2}{4A} - \frac{V^2}{4B}\right). \end{aligned}$$

We now perform the substitution $x \mapsto x + U/(2A)$ and $y \mapsto y + V/(2B)$ corresponding to the translation T_W, $W = (U/(2A), V/(2B)) \in \mathbb{R}^2$, and rename the constant $K \mapsto K - U^2/(4A) - V^2/(4B)$.

With these, we arrive at the following transformed equation:

$$Ax^2 + By^2 + K = 0, \quad A, B \neq 0.$$

Case II $AB > 0$. We may assume $A, B > 0$ since otherwise we change all coefficients to their negatives.

We claim that $K \geq 0$ leads to degeneracy. Indeed, if $K = 0$, then the conic reduces to the origin, and if $K > 0$, then it is the empty set. Both are degenerate cases.

Thus, we have $K < 0$. We now introduce the new constants $a = \sqrt{-K/A}$ and $b = \sqrt{-K/B}$. With these, the equation becomes

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} = 1.$$

This is the normal form of the **ellipse**. For $a = b = r$, it is the equation of the circle with radius r (and center at the origin 0). Otherwise, it is customary to assume $a > b$ since in the opposite case we can perform the simultaneous swapping $x \leftrightarrow y$ and $a \leftrightarrow b$.

Summarizing, and going back to the beginning of our study, we obtain that, up to rotations, translations, and reflections, a **non-degenerate** conic given by $p(x, y)$ above with $p_2(x, y) = Ax^2 + By^2 + Cxy$ and satisfying $C^2 - 4AB < 0$ has the normal form of an ellipse. Once again, since all these transformations preserve the Cartesian distance, the metric properties will remain unchanged.

Case III $AB < 0$. We may assume $A > 0 > B$ since otherwise we change all coefficients to their negatives.

We now introduce the new constants $a = \sqrt{|K|/A}$ and $b = \sqrt{-|K|/B}$. As before, $K = 0$ leads to degeneracy, so that we may assume $a \neq 0 \neq b$. With these, the equation becomes

$$\frac{x^2}{a^2} - \frac{y^2}{b^2} = \pm 1.$$

This is the normal form of the **hyperbola**. It is customary to eliminate the ambiguity of ± 1 on the right-hand side (due to the sign of K) and assume that it is 1, since otherwise we perform the simultaneous swapping $x \leftrightarrow y$ and $a \leftrightarrow b$.

Summarizing, and going back to the beginning of our study, we obtain that, up to rotations, translations, and reflections, a **non-degenerate** conic given by $p(x, y)$ above with $p_2(x, y) = Ax^2 + By^2 + Cxy$ and satisfying $C^2 - 4AB > 0$ has the normal form of a hyperbola. Once again, since all these transformations preserve the Cartesian distance, the metric properties will remain unchanged.

This finishes our classification of (non-degenerate) conics.

History

Hippocrates of Chios (c. 470–410 BCE) was the first to discover that the **Delian problem** of doubling the cube (of the altar of Apollo) noted in Section 3.2 (and also in the epitaph of this chapter) can be reformulated to solving two mean proportions between the original, a, and the doubled, $2a$, volumes (of the altars). In other words, one needs to solve simultaneously any two of the equations $a/x = x/y = y/(2a)$.

According to ancient sources, Menaechmus (380–320 BCE) of Thracian Chersonese, a Greek mathematician and geometer, a friend of Plato and student of Eudoxus, was the discoverer of the conic sections and the use of the parabola and the hyperbola to solve the Delian problem of doubling the cube. More specifically, Hippocrates' mean proportions give rise to the system $x^2 = ay$, $y^2 = 2ax$, $xy = 2a^2$. Geometrically, this amounts to intersect any two of the parabolas or the hyperbola given by these equations. Solving these, we obtain $x = \sqrt[3]{2} \cdot a$ and $y = \sqrt[3]{4} \cdot a$. This is equivalent to construct geometrically a line segment of length $\sqrt[3]{2}$.

Archimedes in his *Quadrature of the Parabola* determined the area of a parabolic sector made by a chord of a parabola. In another work he demonstrated how to use conic sections to split the sphere into two spherical sections with given volume ratio.
Apollonius in his only surviving work of the *Conics* (in eight volumes) made an extensive study of conic sections. Most of Apollonius' work survived in Arabic translations. As noted previously (Section 7.2), the Persian mathematician and poet Omar Khayyàm studied the intersections of hyperbolas and parabolas with a circle.

Exercise

8.1.1. Assume that $p > 0$ and $q \neq 0$ in the cubic equation $x^3 + px + q = 0$. Consider the parabola and the circle given by $y = x^2/\sqrt{p}$ and $y^2 + x(x + q/p) = 0$. The parabola and the circle clearly intersect at the origin $(0, 0)$ and at another point. Show that the first coordinate of the second intersection is a root of the cubic equation.

8.2 Parabolas

A characteristic geometric property of the parabola is that it is the set of points **equidistant** to a line Δ and a point $F \notin \Delta$. We call Δ the **directrix** and F the **focus** of the parabola. The line through F and perpendicular to Δ is the **axis** of the parabola.

Reflection to the axis fixes F and carries Δ to itself. Points equidistant to F and Δ are carried to equidistant points. It follows that this reflection carries the parabola to itself, therefore it is the **symmetry axis** of the parabola. Finally, the midpoint between F and the intersection point of Δ and the axis is called the **vertex** of the parabola.

Letting $d(F, \Delta) = 2d$, $0 < d \in \mathbb{R}$, up to a rotation and a translation, we may arrange that the directrix Δ is given by the equation $y = -d$, and the focus F is given by $F = (0, d)$. With this, the symmetry axis is the second axis, and the vertex is at the origin. (See Figure 8.1.)
Now, using the distance formula between a point and a line, a point $P = (x, y)$ is equidistant to Δ and F if and only if we have

$$\sqrt{x^2 + (y - d)^2} = y + d,$$

or equivalently, $x^2 + (y - d)^2 = (y + d)^2$. Expanding and simplifying, we obtain the normal equation of the parabola

$$y = \frac{1}{4d}x^2, \quad d > 0.$$

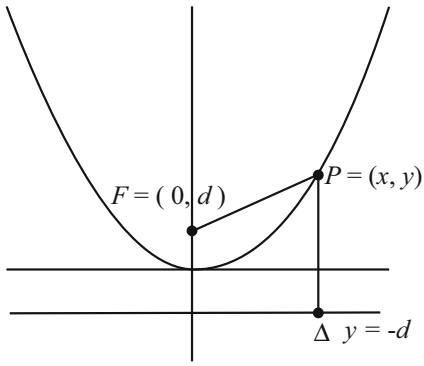


Fig. 8.1 The focus-directrix property of the parabola.

This matches with the equation of the parabola obtained in the previous section algebraically.

History
The term "parabola" (meaning "application (to areas)") is due to Apollonius.

Example 8.2.1 Let $\triangle[A, B, C]$ be a (non-degenerate) triangle with vertices A, B, C, and ℓ a line not parallel to any of the sides of the triangle. Show that there is a unique parabola passing through the vertices A, B, C whose symmetry axis is parallel to ℓ.

We can perform a rotation on the entire configuration such that the rotated ℓ will become vertical. Let the vertices of the rotated triangle be $(x_1, y_1), (x_2, y_2), (x_3, y_3) \in \mathbb{R}^2$. Clearly, x_1, x_2, x_3 are all distinct. Therefore, the Lagrange interpolation polynomial in Example 6.1.1 for these (non-collinear) points is a quadratic polynomial. The graph of the corresponding polynomial function is a parabola which solves the problem.

Parabolas appear in a myriad of applications, and it is convenient to introduce yet two additional equations for them.

First, translating the normal parabola by a translation T_W, $W = (u, v)$, the normal equation transforms into the following

$$y - v = \frac{1}{4d}(x - u)^2, \quad d > 0.$$

This is the equation of the parabola with vertical symmetry axis (given by $x = u$) and vertex at $W = (u, v)$. Here we also allow d to be negative (by reflecting first the normal parabola to the first axis).

Second, expanding and simplifying, we obtain the equation

$$y = ax^2 + bx + c, \quad a \neq 0, \ a, b, c \in \mathbb{R}.$$

This equation of the parabola (with vertical axis) connects the parabola as a geometric object with polynomial algebra as the right-hand side is the general form

of a quadratic polynomial. Comparing coefficients, it follows that

$$d = \frac{1}{4a} \quad u = -\frac{b}{2a} \quad v = \frac{4ac - b^2}{4a}.$$

This gives the vertex $W = (u, v)$ in terms of the coefficients a, b, c as

$$W = \left(-\frac{b}{2a}, \frac{4ac - b^2}{4a} \right).$$

This shows that, for $a > 0$, resp. $a < 0$, the minimum, resp. maximum, occurs at $x = -b/(2a)$ and the minimum, resp. maximum, value is $y = (4ac - b^2)/(4a)$.

The Quadratic Formula

$$r = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}$$

gives the first coordinates of the (possible) intersection of the parabola with the first axis.

Example 8.2.2 [3] Suppose that a parabola has vertex $(u, v) \in \mathbb{Q} \times \mathbb{Q}$ (with u, v rational), $u \neq 1$, $v < 0$, and equation $y = ax^2 + bx + c$, $a, b, c \in \mathbb{R}$, where $a > 0$ and $a + b + c \in \mathbb{Z}$. Show that for the minimum possible value of $a + b + c$ the number a must be rational. Find a in terms of u and v.

By the formulas above, we have $u = -b/2a$ and $v = (4ac - b^2)/(4a)$. These can be solved easily for b and c, and we obtain

$$b = -2au \quad \text{and} \quad c = \frac{b^2}{4a} + v = au^2 + v.$$

Hence

$$a + b + c = a - 2au + au^2 + v = a(1 - 2u + u^2) + v = a(1 - u)^2 + v \in \mathbb{Z}.$$

By assumption, this is an integer, so that $a \in \mathbb{Q}$ (since $u, v \in \mathbb{Q}$ and $u \neq 1$). In addition, since $a > 0$, the minimum value of the fraction on the right-hand side occurs when it is equal to $[v]$, the greatest integer of v. This gives $a = ([v] - v)/(1 - u^2)$.

Returning to the main line, performing the linear change of indeterminates $x \mapsto 4dx$, $y \mapsto 4dy$, the standard equation of the parabola is transformed into the equation of the **unit parabola** $y = x^2$.

[3] A special case of this was a problem in the American Invitational Mathematics Examination, 2011.

This transformation preserves the distance only **up to scaling** with scaling factor $4d$, but it preserves all geometric quantities that we are going to discuss here such as angles, midpoints, tangents, etc. Therefore the proof of any statement about metric properties of the parabola can immediately be reduced to the special case of the unit parabola.

As a first application, note that given a point P_0 on a parabola, there is a unique line ℓ_0 not parallel to the axis that meets the parabola only at P_0. This line ℓ_0 is called the **tangent line** to the parabola at P_0. Moreover, ℓ_0 is uniquely determined by the fact that it intersects the tangent line ℓ to the vertex at the **midpoint** of the vertex itself and the projection of P_0 to ℓ along the axis. Finally, a simple byproduct is that any line not parallel to the axis and not tangent to the parabola must be a **secant** (unless it avoids the parabola altogether); that is, it intersects the parabola at exactly two points.

Recall now that this has been remarked in Section 7.1 for the unit parabola given by $y = x^2$. Since the transformations above preserve all metric properties, including tangency, it follows that the same statements hold in the case of an arbitrary parabola.

Next, we derive the **reflective property of the parabola**: If a light ray parallel to the axis hits the parabola, then it is reflected to the focus.

To make this statement more precise we need to define how a parabola reflects light. By definition, if a light ray hits the parabola at a point P, then it reflects the ray according to the Principle of Shortest Distance **with respect to the tangent line to the parabola** at P.

For a change, we give a geometric proof of the reflective property for the unit parabola given by $y = x^2$. Let the vertical light ray hit the parabola at the point P. We let F be the focus of the parabola, and D, resp. Q, the (vertical) projections of P to the first axis, resp. to the directrix. As we have seen above, the tangent line to the parabola at P meets the first axis at the midpoint M of the origin 0 and the projected point D. (See Figure 8.2.)

Consider now the triangle $\triangle[P, F, Q]$. By the characteristic property of the parabola above, this triangle is **isosceles** since $d(P, F) = d(P, Q)$. Moreover, by

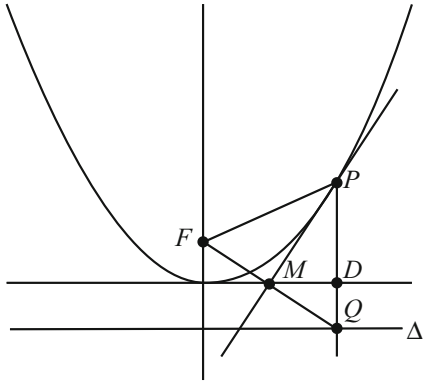


Fig. 8.2 Reflective property of the parabola.

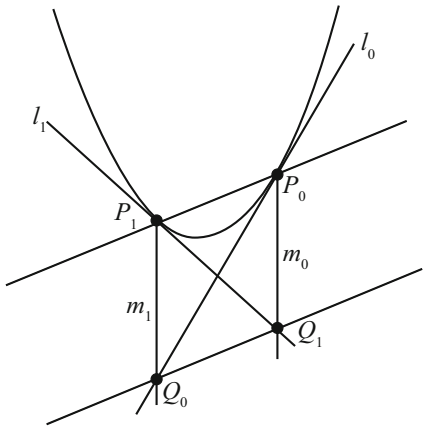


Fig. 8.3 The two-points-two-tangents property.

the position of the parabola, the distance of the points F and Q from the first axis is the **same**. Thus, M is also the midpoint of the side $[F, Q]$ of the triangle opposite to P. By the **pons asinorum** (Section 5.5), the tangent line is the **altitude** line of this triangle since it meets this side at the midpoint M, and this altitude line also bisects the angle at P. But the congruent two halves of these angles (and their opposites) are the angles at which the tangent line meets the (incident) vertical line, and the line extension of the line segment $[P, F]$. Thus, by the Principle of Shortest Distance, the reflected ray passes along this line extension, and thereby it must pass through the focus F. The reflective property of the parabola follows.

As noted above, the parabola has many interesting metric properties. We only discuss here the so-called **two-points-two-tangents property**: Given two points P_0 and P_1 on a parabola, let ℓ_0, resp. ℓ_1 be the tangent lines containing P_0, resp. P_1. (See Figure 8.3.) Moreover, let m_0, resp. m_1, be the lines parallel to the axis of the parabola through P_0, resp. P_1. Finally, let $Q_0 = \ell_0 \cap m_1$ and $Q_1 = \ell_1 \cap m_0$. Then the secant line through P_0 and P_1 is **parallel** to the line through Q_0 and Q_1.

As before, it is enough to show this for the unit parabola given by $y = x^2$. We let $P_0 = (x_0, x_0^2)$ and $P_1 = (x_1, x_1^2)$. The equation of the line m_0 is $x = x_0$, and that of m_1 is $x = x_1$. As shown above, the equation of the tangent line at P_0 is $y = x_0^2 + 2x_0(x - x_0)$, and the equation of the tangent line at P_1 is $y = x_1^2 + 2x_1(x - x_1)$. Intersecting, we obtain $Q_0 = (x_1, x_0^2 + 2x_0(x_1 - x_0))$ and $Q_1 = (x_0, x_1^2 + 2x_1(x_0 - x_1))$.

With these, we calculate the slope through Q_0 and Q_1 as

$$\frac{x_1^2 + 2x_1(x_0 - x_1) - x_0^2 - 2x_0(x_1 - x_0)}{x_0 - x_1} = \frac{x_0^2 - x_1^2}{x_0 - x_1} = x_0 + x_1.$$

This is the slope through P_0 and P_1. The claim follows.

We now return to an old problem of extracting square roots from positive integers.

Remark (Geometric Interpretation of the Babylonian Method) Recall that in Section 5.4 we described the Babylonian Method on how to approximate the square root of a natural number $a \in \mathbb{N}$ by rational numbers. Here we give a geometric interpretation of this.

Our starting point is that $\sqrt{a}$ is the positive solution of the polynomial equation $x^2 = a$. Geometrically, this solution is obtained by considering the graph of the unit parabola $y = x^2$, intersecting it with the horizontal line $y = a$, and taking the first coordinate of the intersection point in the first quadrant. To avoid the trivial case, from now on we assume that $a \in \mathbb{N}$ is not a square, so that $\sqrt{a}$ is an irrational number.

We construct an infinite sequence of positive rational numbers $(q_n)_{n\in\mathbb{N}_0}$ such that the points (q_n, a), $n \in \mathbb{N}_0$, approach the intersection point $(\sqrt{a}, a)$ as follows. We first choose $q_0 > 0$ arbitrarily. Then $q_1, q_2, \ldots, q_n, \ldots$ will be given inductively in the sense that given the nth term q_n with $n \geq 0$, we will derive a formula for the next member q_{n+1} in terms of q_n.

Thus, assume that the positive rational number q_n is given. We draw a tangent line to the unit parabola $y = x^2$ at (q_n, q_n^2) and intersect it with the horizontal line $y = a$. By definition, the first coordinate of the intersection is q_{n+1}.

The equation of the tangent line to the unit parabola through (q_n, q_n^2) is $y = q_n^2 + 2q_n(x - q_n)$. Intersecting this tangent line with the horizontal line given by $y = a$ amounts to substitute $y = a$ to this equation and solve for x to obtain q_{n+1}. An easy computation gives $q_{n+1} = (1/2)(q_n + a/q_n)$, $n \in \mathbb{N}_0$. This is the Babylonian recurrence formula postulated and studied in Section 5.4.

Returning to the main line, sometimes the solution of a geometric problem relies on simple factoring as in the following:

Example 8.2.3 What is the radius of the largest disk that can be dropped inside (the graph of) the unit parabola such that the disk touches the vertex?

The unit parabola is given by the graph of the equation $y = x^2$ in the Cartesian plane $\mathbb{R}^2$. By symmetry, we may assume that the center of the disk is on the positive second axis. Letting $r > 0$ to be its radius, the boundary circle of the disk contains the origin (the vertex of the parabola), and hence its center must be at $(0, r)$. Thus this circle is given by the equation $x^2 + (y - r)^2 = r^2$. Substituting $y = x^2$, and expanding and simplifying, we obtain $y^2 + (1 - 2r)y = 0$. Factoring, we arrive at $y(y - (2r - 1)) = 0$. This shows that $y = 0$ is a solution. This we already know since the circle contains the origin. Now the crux is that this is the only solution of this equation since the circle touches the parabola only at the origin. Thus, we have $2r - 1 = 0$, and the radius is $r = 1/2$.

Exercises

8.2.1. Let ℓ be the set of intersection points of any tangent line to a parabola and the perpendicular line through the focus to the tangent line. Use the reflective

property of the parabola to show that ℓ is the tangent line through the vertex of the parabola.

8.2.2. Derive the **four-points property** of the parabola: Let $P_1 = (x_1, y_1)$, $P_2 = (x_2, y_2)$, $P_3 = (x_3, y_3)$, $P_4 = (x_4, y_4)$ be four points on a parabola given by $y = ax^2 + bx + c$, $a \neq 0$, $a, b, c \in \mathbb{R}$. Let Q_1 be the intersection of the secant line through P_2 and P_3 with the line given by $x = x_1$, and similarly, let Q_2 be the intersection of the secant line through P_1 and P_4 with the line given by $x = x_2$. Prove that the line through Q_1 and Q_2 is parallel to the secant line through P_3 and P_4.

8.2.3. Show that if two tangent lines to a parabola are perpendicular, then their intersection point lies on the directrix.

8.2.4. Prove that the midpoints of parallel chords of a parabola fill a half-line parallel to the axis of the parabola.

8.2.5. Let $P_1 = (x_1, y_1)$ and $P_2 = (x_2, y_2)$ be two points on a parabola given by $y = ax^2+bx+c$ such that the midpoint of P_1 and P_2 is the origin. Determine the coordinates x_1, x_2, y_1, y_2 of P_1 and P_2 in terms of a, b, c.

8.3 Ellipses

A characteristic metric property of the ellipse is that it is the set of all points the sum of whose distances from two fixed points, called the foci, is constant.[4] More precisely, let the (not necessarily distinct) foci be $F_\pm \in \mathbb{R}^2$. Given a positive real number greater than $d(F_+, F_-)$, the distance between the foci, consider the set of points P on the plane whose **sum of distances** $d(P, F_+) + d(P, F_-)$ is equal to this number. This set is called the **ellipse** with foci $F_\pm$. (When the foci coincide the ellipse becomes a circle.)

The line containing the two foci is called the **focal axis**. The midpoint of the foci is the **center** of the ellipse. The line perpendicular to the focal axis and passing through the center is called the **conjugate axis**. The focal and conjugate axes are symmetry axes of the ellipse. These axes intersect the ellipse in two antipodal pairs of **vertices**. As a slight misnomer, the distance of a vertex on the focal axis from the center is called the **semimajor axis**, and the distance of a vertex on the conjugate axis is the **semiminor axis**. (See Figure 8.4.)

Letting the distance between the foci as $2c$ with $0 < c \in \mathbb{R}$, we derive a the normal equation of the ellipse when the foci are symmetrically placed on the first axis, $F_\pm = (\pm c, 0)$, and the sum of the distances of the variable point $P = (x, y)$ from the foci is $2a$ with $c < a \in \mathbb{R}$. Using the Cartesian distance formula, the

[4] This so-called pins-and-string method is simple and instructive. Take a wooden board with two pins, and attach a string to the pins hanging loosely between them. Tighten the string with a marker to form a wedge, slide it along (keeping the string tight) tracing and marking a curve on the board. This way we obtain an ellipse with foci at the two pins.

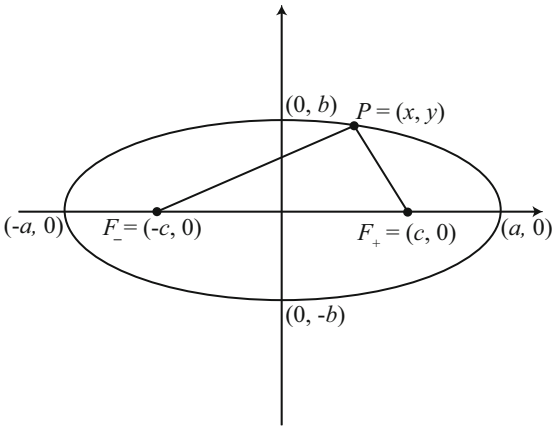


Fig. 8.4 The pin-and-string property of the ellipse.

defining equality $d(P, F_+) + d(P, F_-) = 2a$ is

$$\sqrt{(x-c)^2 + y^2} + \sqrt{(x+c)^2 + y^2} = 2a.$$

We now calculate as follows:

$$\begin{aligned}
\sqrt{(x+c)^2 + y^2} &= 2a - \sqrt{(x-c)^2 + y^2} \\
(x+c)^2 + y^2 &= 4a^2 - 4a\sqrt{(x-c)^2 + y^2} + (x-c)^2 + y^2 \\
a\sqrt{(x-c)^2 + y^2} &= a^2 - cx \\
a^2((x-c)^2 + y^2) &= (a^2 - cx)^2 \\
a^2x^2 + a^2c^2 + a^2y^2 &= a^4 + c^2x^2 \\
(a^2 - c^2)x^2 + a^2y^2 &= a^2(a^2 - c^2) \\
\frac{x^2}{a^2} + \frac{y^2}{a^2 - c^2} &= 1.
\end{aligned}$$

Since $a > c$, we can let $b = \sqrt{a^2 - c^2} > 0$. With this we arrive at the **normal equation of the ellipse**

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} = 1, \quad F_\pm = (\pm c, 0), \quad a^2 = b^2 + c^2.$$

This matches with the normal form of the ellipse obtained in the previous section algebraically. For an ellipse in normal position as above, the focal axis is the first axis (and the conjugate axis is the second), a is the semimajor axis, b is the semiminor axis, and we have $a > b$. The center of the ellipse is at the origin.

If in the normal equation above we have $a < b$, then the focal axis is the second axis and all the roles are reversed. For $a = b = r$, the ellipse reduces to the circle with radius r and center at the origin.

Note that, using a translation, the equation of the ellipse with center at $O = (u, v)$ and symmetry axes parallel to the coordinate axes is given by

$$\frac{(x-u)^2}{a^2} + \frac{(y-v)^2}{b^2} = 1.$$

History
The term "ellipse" (meaning "omission" in applications of areas as noted above) is due to Apollonius. In the post ancient Greek era ellipses came into primary focus in 1609 when Johannes Kepler derived his first law of planetary motion: A planet orbits around the Sun in an elliptical orbit with the Sun in one of the foci.

Example 8.3.1 [5] Let S_1 and S_2 be the circles such that S_2 is contained in the interior of S_1. Show that the set E of the centers of the circles internally tangent to S_1 and externally tangent to S_2 is an ellipse.

Let S_1, resp. S_2, have centers and radii (u_1, v_1) and r_1, resp. (u_2, v_2) and r_2, so that they are given by the equations

$$(x-u_1)^2 + (y-v_1)^2 = r_1^2, \quad \text{resp.} \quad (x-u_2)^2 + (y-v_2)^2 = r_2^2.$$

Let S be a circle internally tangent to S_1 and externally tangent to S_2. If $(u, v) \in E$ is the center and r is the radius of S, then the tangency conditions give[6]

$$\sqrt{(u-u_1)^2 + (v-v_1)^2} = r_1 - r \quad \text{and} \quad \sqrt{(u-u_2)^2 + (v-v_2)^2} = r - r_2.$$

Squaring and subtracting, we obtain

$$2u(u_2-u_1) + 2v(v_2-v_1) + u_1^2 - u_2^2 + v_1^2 - v_2^2 = 2r(r_2-r_1) + r_1^2 - r_2^2.$$

The crux is that this is a **linear** equation in the indeterminates r and u, v (with $u_1, u_2, v_1, v_2, r_1, r_2$ being constants). Hence, expressing r in terms of u, v and substituting into the square of the first tangency condition

$$(u-u_1)^2 + (v-v_1)^2 = (r_1-r)^2,$$

we obtain a quadratic equation in u and v. Thus, E is a conic. (It is possible to write down this explicit equation for E but we will not need this.) This conic is non-

[5] This example generalizes the first part of a numerical problem in the American Invitational Mathematics Examination, 2005.

[6] Clearly, a common tangent line of two circles is perpendicular to the line passing through the centers of the circles, and hence the point of tangency and the two centers are collinear; see Section 5.5.

degenerate (as it clearly has at least two points), and it is bounded (as it is contained in the interior of S_1). In view of the classification of the conics in Section 8.1, we see that E must be an ellipse.

We now fix a point P_0 on the ellipse in normal position and claim that there is a unique line that meets the ellipse only at this point. This line is called the **tangent line** to the ellipse at P_0. More precisely, for $P_0 = (x_0, y_0)$, we claim that the equation of this tangent line is

$$\frac{x_0}{a^2}\, x + \frac{y_0}{b^2}\, y = 1.$$

All these statements follow from the corresponding statements for the special case of the unit circle ($a = b = 1$). Indeed, the linear change of the indeterminates $x \mapsto ax$, $y \mapsto by$ transforms the normal equation of the ellipse to the equation of the unit circle. It preserves lines and the tangency condition, and, along with $x_0 \mapsto ax_0$, $y_0 \mapsto by_0$, it transforms the equation of the tangent line above to the equation $x_0 x + y_0 y = 1$. This, however, is the equation of the tangent line to the unit circle at the point (x_0, y_0) as was derived in Section 5.5. The claim follows.

The **reflective property** of the ellipse states that a light ray emitted at one of the foci is reflected to the other. Unlike the case of the parabola, this is much simpler to show, and follows from the Principle of Shortest Distance. One only needs to observe that for an ellipse in normal position as above, the interior of the ellipse consists of those points P for which the sum of distances $d(P, F_+) + d(P, F_-) < 2a$, and the exterior of the ellipse consists of those points P for which the sum of distances $d(P, F_+) + d(P, F_-) > 2a$. Now, if the light ray emitted at F_+ hits the ellipse at a point P, then $d(P, F_+) + d(P, F_-) = 2a$, and for any other point Q on the tangent line, being in the exterior of the ellipse, we have $d(Q, F_+)+d(Q, F_-) > 2a$. Thus, by the Principle of Shortest Distance, the angle of incidence and the angle of reflection at P with respect to the tangent line are equal. The reflective property of the ellipse follows.

History

The so-called "whisper galleries" are large elliptical rooms in which a person, standing at one of the foci, can hear the conversation of other people near the other focus. The most prominent example is the National Statuary Hall in the US Capitol, where Quincy Adams allegedly used this to eavesdrop political conversations.

Example 8.3.2 Given an ellipse, show that the product of the distances of the two foci from any tangent line to the ellipse is equal to the square of the semiminor axis; in particular, this product is a constant (that is, it does not depend on the choice of the tangent line).

We may assume that the ellipse is given by the normal equation $x^2/a^2 + y^2/b^2 = 1$ with foci $F_\pm = (\pm c, 0)$, $a^2 = b^2 + c^2$. Let ℓ be a tangent line to the ellipse at a point $P_0 = (x_0, y_0)$ given by the equation $(x_0/a^2)x + (y_0/b^2)y = 1$ as above. We now use the formula of the distance of a point from a line (Section 5.5) as

$$d(F_\pm, \ell) = \frac{\left|\pm(x_0/a^2)c - 1\right|}{\sqrt{x_0^2/a^4 + y_0^2/b^4}}.$$

Using this, and $x_0^2/a^2 + y_0^2/b^2 = 1$, we now calculate

$$d(F_+, \ell)d(F_-, \ell) = \frac{1 - (x_0^2/a^4)c^2}{x_0^2/a^4 + y_0^2/b^4} = \frac{1 - (x_0^2/a^4)(a^2 - b^2)}{x_0^2/a^4 + (1 - x_0^2/a^2)/b^2} = b^2.$$

The example follows. (A more illuminating solution will be given in Section 11.3.)

Example 8.3.3 [7] An ellipse is tangent to the first axis. Express the length of the semimajor axis in terms of the foci $F_\pm$.

Let P_0 be the point of tangency on the first axis. By the Principle of Shortest Distance, for P any point in the first axis, the sum of distances $d(P, F_-)+d(P, F_+)$ is minimal for $P = P_0$. This means that, reflecting F_+ to the first axis to obtain F'_+, we have $P_0 \in [F_-, F'_+]$. Hence, for the length of the semimajor axis, we have

$$2a = d(F_-, P_0) + d(P_0, F_+) = d(F_-, P_0) + d(P_0, F'_+) = d(F_-, F'_+).$$

Returning to the main line, the ellipse possesses two **directrices** $\Delta_\pm$; they form a pair of parallel lines perpendicular to the focal axis and having distance a^2/c from the center, where $2c$ is the distance between the foci, and $2a$ is the sum of the distances of a generic point on the ellipse to the foci. Each directrix has the property that the ellipse is the set of points P such that

$$\frac{d(P, F_\pm)}{d(P, \Delta_\pm)} = e,$$

where $e = c/a < 1$ is the **eccentricity** of the ellipse, the position of the focus as a fraction to the semimajor axis. Thus, the parabola can be viewed as a conic with eccentricity $e = 1$ while the ellipse has eccentricity $e < 1$.

As usual, it is enough to show this for the ellipse in normal position as above. The equation of the directrices is $x = \pm a^2/c$. By symmetry, we can restrict ourselves to the directrix Δ_+. For $P = (x, y) \in \mathbb{R}^2$, we have

$$d(P, F_+)^2 = (x - c)^2 + y^2 \quad \text{and} \quad d(P, \Delta_+)^2 = \left(x - \frac{a^2}{c}\right)^2.$$

We write the square of the eccentricity ratio above as

$$d(P, F_+)^2 - \frac{c^2}{a^2} d(P, \Delta_+)^2 = 0.$$

[7] This example is inspired by a numerical special case of the American Invitational Mathematics Examination, 1985.

Substituting, we obtain

$$(x-c)^2 + y^2 - \frac{c^2}{a^2}\left(x - \frac{a^2}{c}\right)^2 = 0.$$

Expanding and simplifying, we arrive at

$$\left(1 - \frac{c^2}{a^2}\right)x^2 + y^2 + c^2 - a^2 = 0.$$

Finally, using $c^2 = a^2 - b^2$, the equation of the normal ellipse follows.

Example 8.3.4 In this example we derive the **trammel construction** for the ellipse: If the end-points of a segment are moved along two intersecting lines, a fixed point on the segment traces an arc of an ellipse.

We first make a reduction step. Recall that an ellipse in general position is given by the zero-set of a quadratic polynomial $p(x, y)$ such that the coefficients of the degree two homogeneous component $Ax^2 + By^2 + Cxy$ satisfies $C^2 - AB < 0$.

For given $r, s \in \mathbb{R}$, $r \neq s$, we perform the substitution $x \mapsto rx$ and $y \mapsto sy$. The degree two homogeneous component of the transformed polynomial $p(rx, sy)$ has the form $A(rx)^2 + B(sy)^2 + C(rx)(sy) = r^2Ax^2 + s^2By^2 + rsCxy$. The condition for the ellipse becomes $(rsC)^2 - (r^2A)(s^2B) = r^2s^2(C^2 - AB) < 0$. It follows that this change of variables transforms an ellipse into another ellipse. Being linear, this transformation[8] sends lines to lines. Moreover, as simple computation shows, it preserves the affine parametrization; in particular, it preserves the **ratio** of distances along a line.

We now turn to the trammel construction. By performing a translation and a rotation, we may assume that the intersecting lines are given by $y = \pm mx$ with slope $0 < m \in \mathbb{R}$. Performing the substitution above, we obtain the lines $y = \pm(r/s)mx$. Now, we choose $s/r = m$ so that the transformed intersecting lines become perpendicular. Since ellipses transform to ellipses, it follows that we may assume that the intersecting lines are perpendicular. Finally, performing yet another rotation, we may assume that these lines are the first and second axes, and movement of the line segment takes place in the first quadrant.

In an intermediate position the line segment is the hypotenuse of a right triangle with horizontal and vertical sides. The point $P = (x, y)$ on the line segment in an intermediate position splits the hypotenuse into two line segments that are the hypotenuses of two similar right triangles. Assuming that P splits the line segment of length $a + b$, $0 < a, b \in \mathbb{R}$, in the ratio $a \div b$, the Pythagorean Theorem along with the Birkhoff Postulate on Similarity gives $\sqrt{a^2 - x^2}/a = y/b$. Squaring and simplifying, the normal equation of the ellipse follows.

[8] These are called affine transformations, and the geometry based on these is called affine geometry.

Exercises

8.3.1. Show that, for the normal ellipse given by $x^2/a^2 + y^2/b^2 = 1$, $a > b$, the intersection points of perpendicular pairs of tangent lines lie on the circle $x^2 + y^2 = a^2 + b^2$.

8.3.2. Show that the midpoints of parallel chords of an ellipse lie on a diameter, a chord through the center.

8.4 Hyperbolas

As in the case of ellipses, we start with two distinct points $F_\pm$ which we call foci. Given a positive real number less than $d(F_+, F_-)$, consider the set of points P on the plane whose absolute value of the **difference of distances**, $|d(P, F_+) - d(P, F_-)|$, is equal to this number. This set is called the **hyperbola** with foci $F_\pm$. The line containing the two foci is called the **focal axis**. The midpoint of the foci is the **center** of the hyperbola. The line perpendicular to the focal axis through the center is the **conjugate axis**. The focal and conjugate axes are symmetry axes of the hyperbola. The hyperbola meets the focal axis in two points. The conjugate axis is disjoint from the hyperbola and it separates the hyperbola into two **branches**. It is possible to obtain a more precise description of the metric properties of the hyperbola in this general setting, but it will be much simpler to work these out for the hyperbola in a specific position.

We set the distance between the foci to be $2c$, $c > 0$. Applying a translation and a rotation, we set the foci on the first axis in a symmetric position: $F_\pm = (\pm c, 0)$. We derive an equation of the hyperbola in this **normal position**. For a hyperbola in normal position, the focal axis is the first axis and the conjugate axis is the second. The center of the hyperbola is at the origin.

We let $a < c$ such that $|d(P, F_+) - d(P, F_-)| = 2a$. Using the Cartesian distance formula with $P = (x, y)$, this condition gives

$$\sqrt{(x-c)^2 + y^2} - \sqrt{(x+c)^2 + y^2} = \pm 2a.$$

By a minor modification in the computation for the ellipse, we obtain

$$\frac{x^2}{a^2} - \frac{y^2}{c^2 - a^2} = 1.$$

Since $a < c$, we can let $0 < b = \sqrt{c^2 - a^2}$. With this, we arrive at the normal equation of the hyperbola

$$\frac{x^2}{a^2} - \frac{y^2}{b^2} = 1, \quad a^2 + b^2 = c^2.$$

The left-hand side of the equation of the hyperbola can be factored

$$\left(\frac{x}{a} - \frac{y}{b}\right)\left(\frac{x}{a} + \frac{y}{b}\right) = 1.$$

The factors **vanish** on a pair of lines $\ell_\pm$ intersecting at the origin. They are given by the equations

$$\frac{x}{a} \pm \frac{y}{b} = 0.$$

We arrange the signs so that $\ell_\pm$ has slope $\pm b/a$.

These two lines split the plane $\mathbb{R}^2$ into four angular regions. These regions are given by **any two** of the four inequalities $x/a \pm y/b \gtreqless 0$. Each angular region contains exactly one of the positive or negative coordinate axes. For example, the region that contains the positive first axis is given by $x/a \pm y/b \geq 0$, and the angular region that contains the negative first axis is given by $x/a \pm y/b \leq 0$. Since the equation of the hyperbola implies $(x/a + y/b)(x/a - y/b) > 0$, it follows that the hyperbola is in the interior of the opposite pair of angular regions that contain the positive and negative first axes.

Consider the rectangle R with vertices $(\pm a, \pm b)$. The two lines $\ell_\pm$ pass through the two antipodal pairs of vertices of R. By the equation of the hyperbola, we have $x^2/a^2 - 1 = y^2/b^2 \geq 0$. This gives $x^2 \geq a^2$, or equivalently, $x \geq a$ or $x \leq -a$. It follows that the hyperbola meets R only at the two boundary points $(\pm a, 0)$. (See Figure 8.5.)

Our present goal is to describe the relationship between the hyperbola and the two lines $\ell_\pm$. To do this, we first construct a "parametrization" of the hyperbola via the affine parametrization on ℓ_+ given by the points $P_0 = (0, 0)$, the origin, and $P_1 = (a, b)$, the northeast corner of the rectangle R. Recall that this affine parametrization is defined by $P_t = (1 - t)P_0 + tP_1 \leftrightarrow t$, $t \in \mathbb{R}$. In our case, we have $P_t = (at, bt)$, $t \in \mathbb{R}$.

We consider the pencil of parallel lines containing ℓ_-. Since each member of this pencil meets ℓ_+ at a unique point, the pencil **itself** can be parametrized by the affine

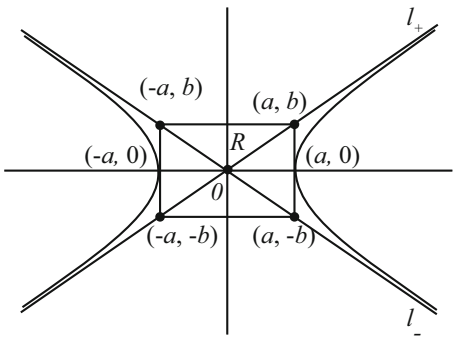


Fig. 8.5 The hyperbola in normal position.

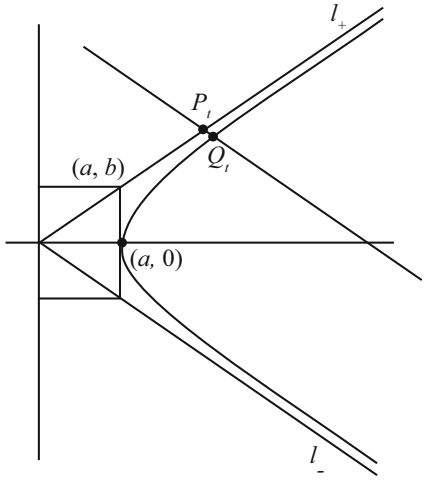


Fig. 8.6 Parametrization of the hyperbola.

coordinate on ℓ_+. More specifically, the line in this pencil with parameter $t \in \mathbb{R}$ meets ℓ_+ in $P_t = (at, bt)$, and it is given by the equation

$$\frac{x}{a} + \frac{y}{b} = 2t.$$

Now, the crux is that, for $t \neq 0$, this line meets the hyperbola at exactly one point Q_t, say. (See Figure 8.6.) By the factored form of the hyperbola above, this point can be obtained by solving the system

$$\frac{x}{a} + \frac{y}{b} = 2t \quad \text{and} \quad \frac{x}{a} - \frac{y}{b} = \frac{1}{2t}.$$

Solving for x and y, we obtain the coordinates of Q_t as follows[9]

$$x = \frac{a}{2}\left(2t + \frac{1}{2t}\right) \quad \text{and} \quad y = \frac{b}{2}\left(2t - \frac{1}{2t}\right).$$

Clearly, the converse also holds: Any point on the hyperbola is the intersection of a member of the pencil with a **non-zero** parameter. Hence this is a parametrization of the hyperbola. The positive parameters describe the branch of the hyperbola contained in the quadrants $I \cup IV$, while the negative parameters describe the branch in $II \cup III$.

Finally, we calculate the distance between points of the same parameter on the line ℓ_+ and on the hyperbola. For $0 \neq t \in \mathbb{R}$, we have

[9]These formulas show that the hyperbola can be conveniently parametrized by the hyperbolic cosine and sine functions. See the set of exercises after Exercise 10.3.3 at the end of Section 10.3.

$$d(P_t, Q_t)^2 = \left(at - \frac{a}{2}\left(2t + \frac{1}{2t}\right)\right)^2 + \left(bt - \frac{b}{2}\left(2t - \frac{1}{2t}\right)\right)^2 = \frac{a^2+b^2}{16t^2} = \frac{c^2}{16t^2}.$$

Hence, we obtain $d(P_t, Q_t) = c/(4|t|)$, $0 \neq t \in \mathbb{R}$.

At this point we introduce the concept of **asymptote**. We need some preparations.

First, given a set $E \in \mathbb{R}^2$ and a point $P \in \mathbb{R}^2$, we define the distance between P and E as the infimum $d(P, E) = \inf\{d(P, Q) \mid Q \in E\}$. Second, let ℓ be a half-line with end-point P_0. Choose another point $P_1 \in \ell$, and let $P_t = (1-t)P_0 + tP_1$, $0 \leq t \in \mathbb{R}$, be the corresponding affine parametrization of ℓ.

With these, we say that the half-line ℓ is an asymptote of E if $\lim_{t\to\infty} d(P_t, E) = 0$. Clearly, this concept does not depend on the choice of $P_1 \in \ell$.

History

The term "asymptote" was introduced by Apollonius, and its literal meaning is a derivative of the negative infinitive "not falling together." Note that in our definition the asymptote can intersect the curve itself (as is often the case for horizontal and oblique asymptotes of functions).

Returning to our hyperbola in normal position, we denote by H the hyperbola as a subset of the plane $\mathbb{R}^2$.

We claim that the half-line $\ell'_+ \subset I$ of ℓ_+ with end-point at the origin is an asymptote of the hyperbola. Indeed, by our computation above, we have

$$0 \leq \lim_{t\to\infty} d(P_t, H) \leq \lim_{t\to\infty} \frac{c}{4t} = 0.$$

The claim follows.

By the fourfold symmetry of the hyperbola, we obtain that all four half-lines of $\ell_\pm$ (with common end-point at the origin) are asymptotes of the hyperbola.

Since all properties above are metric, our entire description can be transplanted to a hyperbola in general position.

We now fix a point $P_0 = (x_0, y_0)$ on the hyperbola in normal position, and, in analogy with the equation of the tangent line to the ellipse, we consider the line given by the equation

$$\frac{x_0}{a^2}\, x - \frac{y_0}{b^2}\, y = 1.$$

Clearly, P_0 is on this line. We claim that P_0 is the only intersection point of this line and the hyperbola, and that the branch of the hyperbola through P_0 is on **one side** of this line. We call this the **tangent line** to the hyperbola at the point P_0.

To prove the claim, assume that $P = (x, y)$ is an arbitrary point on the hyperbola on the same branch as P_0. We let $p_0^\pm = x_0/a \pm y_0/b$ and $p^\pm = x/a \pm y/b$. Since P_0 and P are on the hyperbola, we have $p_0^- p_0^+ = 1$ and $p^- p^+ = 1$. In addition, we have

$$p_0^- p^+ + p_0^+ p^- = \left(\frac{x_0}{a} - \frac{y_0}{b}\right)\left(\frac{x}{a} + \frac{y}{b}\right) + \left(\frac{x_0}{a} + \frac{y_0}{b}\right)\left(\frac{x}{a} - \frac{y}{b}\right) = 2\left(\frac{x_0}{a^2}\, x - \frac{y_0}{b^2}\, y\right).$$

By symmetry, we can now restrict ourselves to the "right" branch of the hyperbola (contained in the angular sector given by $x/a \pm y/b > 0$). This gives $p_0^{\pm} > 0$ and $p^{\pm} > 0$. We can now use the AM-GM-inequality:

$$\frac{x_0}{a^2}\, x - \frac{y_0}{b^2}\, y = \frac{p_0^- p^+ + p_0^+ p^-}{2} \geq \sqrt{p_0^- p^+ p_0^+ p^-} = 1$$

with equality if and only if $p_0^- p^+ = p_0^+ p^-$, that is, if and only if $p_0^+ = p^+$ and $p_0^- = p^-$, and finally, if and only if $P_0 = P$.

This gives us two conclusions. First, the tangent line meets the hyperbola at one point (P_0) only. Second, the right branch of the hyperbola is on one side of the tangent line. The claim follows.

Continuing the analogy with parabolas and ellipses, we now note the **reflective property** of the hyperbola. It states that a light ray toward a focus reflects off toward the other focus. This is clearly equivalent to saying that, at an arbitrary point P_0 of the hyperbola, the tangent line through P_0 bisects the angle formed by the half-lines through the foci $F_{\pm}$ with common end-point P_0.

As usual, we may restrict ourselves to the hyperbola in normal position with $2c$ being the distance between the foci, and $2a$ the difference of the distances of a generic point on the hyperbola to the foci.

By symmetry, we may assume that $P_0 \in I$. We let ℓ_0 denote the **angular bisector** of the angle $\angle F_+ P_0 F_-$ and show that ℓ_0 is tangent to the hyperbola; that is, if $P_0' \in \ell_0$, $P_0' \neq P_0$, then P_0' cannot be on the hyperbola.

Let $Q \in [F_-, P_0]$ such that $d(F_-, Q) = 2a$. Since $d(P_0, F_-) - d(P_0, F_+) = 2a$, the point Q exists, and we also have $d(P_0, Q) = d(P_0, F_+)$. By the triangle inequality, we have

$$d(P_0', F_-) < d(P_0', Q) + d(Q, F_-) = d(P_0', Q) + 2a = d(P_0', F_+) + 2a,$$

where the sharp inequality holds because the points F_-, Q, P_0' are not collinear, and, in the last equality, we used the Birkhoff Postulate on Similarity applied to the triangles $\triangle P_0 Q P_0'$ and $\triangle P_0 F_+ P_0'$. This gives $d(P_0', F_-) - d(P_0', F_+) < 2a$. Hence the point P_0' cannot be on the hyperbola. The reflective property of the hyperbola follows.

Just like the ellipse, the hyperbola also possesses two **directrices** $\Delta_{\pm}$; they form a pair of parallel lines perpendicular to the focal axis and having distance a^2/c from the center, where $2c$ is the distance between the foci, and $2a$ is the difference of the distances of a generic point on the hyperbola to the foci. Each directrix has the property that the hyperbola is the set of points P such that

$$\frac{d(P, F_{\pm})}{d(P, \Delta_{\pm})} = e,$$

where $e = c/a > 1$ is the **eccentricity** of the hyperbola.

As usual, it is enough to show this for the hyperbola in normal position as above. The proof is almost verbatim to the case of the ellipse. The equation of the directrices is $x = \pm a^2/c$. By symmetry, we can restrict ourselves to the directrix Δ_+. For $P = (x, y) \in \mathbb{R}^2$, we have

$$d(P, F_+)^2 = (x-c)^2 + y^2 \quad \text{and} \quad d(P, \Delta_+)^2 = \left(x - \frac{a^2}{c}\right)^2.$$

We write the square of the eccentricity ratio above as

$$d(P, F_+)^2 - \frac{c^2}{a^2} d(P, \Delta_+)^2 = 0.$$

Substituting, we obtain

$$(x-c)^2 + y^2 - \frac{c^2}{a^2}\left(x - \frac{a^2}{c}\right)^2 = 0.$$

Expanding and simplifying, we arrive at

$$\left(1 - \frac{c^2}{a^2}\right) x^2 + y^2 + c^2 - a^2 = 0.$$

Finally, using $c^2 = a^2 + b^2$, the equation of the normal hyperbola follows.

Remark The three non-degenerate conics, the parabola, ellipse, and hyperbola, can be united by the eccentricity as follows. We let $F = (f, 0)$ be a focal point, and assume that the conic with eccentricity $e > 0$ contains the origin. We let a directrix Δ be given by the equation $x = -f/e$. Then, the set of points $P = (x, y) \in \mathbb{R}^2$ satisfying $d(P, F) = e \cdot d(P, \Delta)$ is the following

$$(x-f)^2 + y^2 = e^2\left(x + \frac{f}{e}\right)^2 = (ex + f)^2.$$

Simplifying, we obtain $x^2(e^2-1) + 2f(e+1)x - y^2 = 0$. For $e = 1$ this gives the parabola; for $e < 1$, the ellipse; and for $e > 1$, the hyperbola. In the last two cases the center is at $(f/(1-e), 0)$.

History
The focus-directrix property of the parabola, ellipse, and hyperbola is due to Pappus of Alexandria (c. 290–c. 350 BCE).

Example 8.4.1 (Revisited) Recall Example 6.6.8: Let $0 < a, b \in \mathbb{N}$ such that $ab-1$ divides $a^2 + b^2$. Show that $(a^2 + b^2)/(ab+1)$ is a perfect square.

We now give a geometric solution to this problem.

Let

$$\frac{a^2+b^2}{ab+1} = c \in \mathbb{N}.$$

Our method is to find $a, b \in \mathbb{N}$ in terms of the (fixed) natural number c. Multiplying out, we obtain $a^2 + b^2 - cab - c = 0$. As before, we may assume that $c \geq 3$. We introduce two indeterminates, x and y, and consider the **conic**

$$x^2 + y^2 - cxy - c = 0, \quad 3 \leq c \in \mathbb{N}.$$

The problem is to study the positive **integral points** $(a, b) \in \mathbb{N} \times \mathbb{N}$ on this conic in the first quadrant I.

Since $C^2 - AB = c^2 - 1 > 0$ and $c \geq 3$, the conic is a (non-degenerate) hyperbola H. (Notice that, for $c = 2$, the conic reduces to a pair of parallel lines, a degenerate conic.) Note that, due to symmetry with respect to the interchange $x \leftrightarrow y$, the equation $x - y = 0$ gives the **conjugate** axis, and therefore $x + y = 0$ gives the **focal** axis. Therefore, the "upper branch" H_+ of the hyperbola is contained in the half-plane given by $y > x$, whereas the "lower branch" H_- is contained in the half-plane given by $y < x$.

The change of variables

$$x \mapsto \frac{y+x}{\sqrt{2}} \quad \text{and} \quad y \mapsto \frac{y-x}{\sqrt{2}},$$

(corresponding to rotation by angle $\pi/4$) transforms this conic to

$$\left(\frac{1}{2} + \frac{1}{c}\right) x^2 - \left(\frac{1}{2} - \frac{1}{c}\right) y^2 = 1.$$

This is a hyperbola in normal form. The asymptotes of this hyperbola are given by the equations

$$\sqrt{\frac{1}{2} + \frac{1}{c}} \cdot x \pm \sqrt{\frac{1}{2} - \frac{1}{c}} \cdot y = 0.$$

Hence, inverting the change of variables above, the asymptotes of our original quartic are given by

$$\left(\sqrt{\frac{1}{2} + \frac{1}{c}} \pm \sqrt{\frac{1}{2} - \frac{1}{c}}\right) \cdot x - \left(\sqrt{\frac{1}{2} + \frac{1}{c}} \mp \sqrt{\frac{1}{2} - \frac{1}{c}}\right) \cdot y = 0.$$

This shows an important feature that the asymptotes are contained in the union of the first and the third quadrants.

As before, we start with a positive integral point $(a_0, b_0) \in \mathbb{N} \times \mathbb{N}$ in the first quadrant I. By symmetry, we may assume $a_0 < b_0$, that is, we have $(a_0, b_0) \in H_+$, the upper branch of the hyperbola H. Replacing b_0 by the indeterminate y, we consider the quadratic equation

$$y^2 - ca_0 y + a_0^2 - c = 0.$$

By construction, b_0 is a solution. The first Viète relation gives another solution $b_0' = ca_0 - b_0 \in \mathbb{Z}$. Since $(a_0, b_0) \in H_+$ and $(a_0, b_0') \in H$, we must have $(a_0, b_0') \in H_-$, the lower branch of H.

We claim that either the integral point (a_0, b_0') is in the interior of the first quadrant ($b_0' > 0$), or c is a perfect square.

Indeed, assume $b_0' \leq 0$. Then we have

$$b_0'^2 - c(a_0 b_0' + 1) + a_0^2 = 0.$$

Clearly, $b_0' < 0$ cannot happen. Thus, $b_0' = 0$, and this gives $c = a_0^2$, a perfect square. The claim follows.

If c is a perfect square, then we are done. Otherwise, by the above, (a_0, b_0') is in the interior of the first quadrant I. We can now perform reflection to the line given by $y = x$. This reflection swaps the two branches of the hyperbola H (and maps the interior of the first quadrant to itself). Since $(a_0, b_0') \in H_-$ we obtain $(b_0', a_0) \in H_+$, still in the interior of the first quadrant I. Since $a_0 < b_0$ this point (b_0', a_0) has a smaller second coordinate than (a_0, b_0).

Since these points have positive integral coordinates, repeating this, the process must end in finitely many steps, and we obtain that c is a perfect square. The example follows.

Remark In Section 2.1 we discussed Pell's equation $x^2 - d \cdot y^2 = 1$, where $d \in \mathbb{N}$ is a non-square integer. We showed that Brahmagupta's identity provides an inductive method to obtain all positive integral solutions in the form of an infinite sequence of pairs $(x_k, y_k) \in \mathbb{N} \times \mathbb{N}$, $k \in \mathbb{N}_0$, starting from the fundamental solution (x_0, y_0). In our present geometric terms, Pell's equation defines a hyperbola, and the solutions in the first quadrant are integral points on this hyperbola.

Exercises

8.4.1. Show that, for the normal hyperbola given by $x^2/a^2 - y^2/b^2 = 1$, $a > b$, the intersection points of perpendicular pairs of tangent lines lie on the circle $x^2 + y^2 = a^2 - b^2$.

8.4.2. Show that the midpoints of parallel chords of a hyperbola lie on a line through the center.

8.4.3. A hyperbola given by the equation $y = a/(x - b) + c$, $a \neq 0$, is uniquely determined by three points $P_i = (x_i, y_i)$, $i = 1, 2, 3$ with different first and second coordinates: $x_i \neq x_j$, $y_i \neq y_j$, $i \neq j$, $i, j = 1, 2, 3$. Show that an equation of the hyperbola through these three points is

$$\frac{y - y_1}{x - x_1} \cdot \frac{x - x_2}{y - y_2} = \frac{y_3 - y_1}{x_3 - x_1} \cdot \frac{x_3 - x_2}{y_3 - y_2}.$$

8.4.4. Let P_0 be a point on the normal hyperbola given by $x^2/a^2 - y^2/b^2 = 1$. Assume that the line segment $[O, P_0]$ with O, the center of the hyperbola, is a diagonal of a parallelogram whose other two vertices P_0' and P_0'' lie on the asymptotes. Show that the tangent line through P_0 is **parallel** to the other diagonal $[P_0', P_0'']$.

8.4.5. Derive the following analogue of Example 8.3.1 for hyperbolas: Let C_1 and C_2 be two disjoint circles on the plane $\mathbb{R}^2$. Then the set of centers of circles that are externally tangent to C_1 and C_2 comprise a branch of a hyperbola.

8.4.6. Find an equilateral triangle whose vertices are on the graph of the hyperbola $y = 1/x$.

Chapter 9
Rational and Algebraic Expressions and Functions

"'Every minute dies a man, Every minute one is born;' I need hardly point out to you that this calculation would tend to keep the sum total of the world's population in a state of perpetual equipoise, whereas it is a well-known fact that the said sum total is constantly on the increase. I would therefore take the liberty of suggesting that in the next edition of your excellent poem the erroneous calculation to which I refer should be corrected as follows: 'Every moment dies a man, And one and a sixteenth is born.' I may add that the exact figures are 1.067, but something must, of course, be conceded to the laws of metre."
Charles Babbage, from a letter to Alfred, Lord Tennyson.

As a natural continuation of the study of polynomials, in this chapter we introduce and discuss rational and algebraic expressions in a wide variety of settings. One of the main objectives of this chapter is to present the partial fraction decomposition in complete details; this is accompanied by a few Olympiad level problems. Asymptotes, briefly alluded to in treating hyperbolas in Section 8.4, are fully and rigorously developed here. Another main objective of this chapter is to extend the AM-GM inequality (Sections 5.4, 7.5) to the multivariate harmonic-geometric-arithmetic-quadratic mean inequalities.

As pointed out by Gelfand, the AM-GM inequality along with its extensions is a cornerstone of analysis. It has a beautiful geometry which was known to the ancient Greeks, and it appears in a myriad problem such as multivariate extremal problems, factorization problems, etc. Amongst the literally hundreds of mathematical contest problems involving these means we chose a representative sample to demonstrate the principal methods. The lesser known permutation (arrangement) inequality is also introduced here pointing out that it implies all the other classical inequalities such as the AM-GM, Cauchy–Schwarz (Sections 5.3, 6.7), and Chebyshev (Section 6.7) inequalities. Finally, we give a detailed (and somewhat more advanced) account on the greatest integer function along with some of Ramanujan's formulas, and the Hermite identity.

G. Toth, *Elements of Mathematics*, Undergraduate Texts in Mathematics,
https://doi.org/10.1007/978-3-030-75051-0_9

9.1 Rational Expressions and Rational Functions

A mathematical expression constructed from real numbers and an indeterminate x using the operations of **addition**, **multiplication**, and **division** is called a **rational expression**. Any rational expression can be transformed into a **rational fraction**, a fraction with polynomial numerator and denominator. The typical notation for a rational expression in the indeterminate x is $q(x)$. With this, a rational fraction is of the form $q(x) = n(x)/d(x)$, where $n(x)$ is the polynomial **numerator**, and $d(x)$ is the polynomial **denominator**.

The definition of rational expression can be naturally extended to expressions in several indeterminates $x, y, z \ldots$, and $x_1, x_2, x_3, \ldots, x_n$ with $n \in \mathbb{N}$, etc., and we obtain rational expressions $q(x, y)$, $q(x, y, z)$, $q(x_1, x_2, x_3, \ldots, x_n)$, etc.

Remark The terminology for rational expressions is somewhat different from that of polynomial expressions. This is because in rational expressions the replacement of the indeterminate by an entity is rare, and, if needed, it can be specified at its occurrence.

Rational expressions can be **evaluated** on (real) numbers by substitution; that is, by performing the operations that the rational expression is made up on numbers instead of indeterminates. Rational expressions $q(x)$, $q(x, y)$, etc. evaluated on specific numbers $a, b \in \mathbb{R}$ are denoted by $q(a)$, $q(a, b)$, etc.

Since division by zero is undefined, unlike the case of polynomial expressions, rational expressions may not be defined for all (real) values of the indeterminates. The **domain of definition** of a rational expression is the (maximal) set of values of the indeterminates for which the rational expression is defined. In particular, the domain of definition of a rational fraction is the set of values of the indeterminates for which the denominator does not vanish. The domain of definition of a rational expression $q(x)$, $q(x, y)$, etc. is denoted by $D(q(x))$, $D(q(x, y))$, etc.

A **rational function** is a function of the form $y = q(x)$, $z = q(x, y)$, etc., where $q(x)$, $q(x, y)$, etc. are rational expressions. The **domain** of a rational function is the domain of definition of the corresponding rational expression. Functionally, we denote a rational function by $q : \mathbb{R} \to \mathbb{R}$, $q : \mathbb{R}^2 \to \mathbb{R}$, etc., even though the domain of q may not be the whole $\mathbb{R}$, $\mathbb{R}^2$, etc.

The domain of definition applies only to the **specific form** of the rational expression. It may change when the rational expression undergoes algebraic manipulations.

Example 9.1.1 Consider the rational expression

$$q(x) = \frac{x^5 + x^4 + x^3 + x^2 + x + 1}{x + 1}$$

with domain of definition $D(q(x)) = \{x \in \mathbb{R} \mid x \neq -1\}$.

Using the Finite Geometric Series Formula, we may be tempted to reduce the complexity of $q(x)$ as

$$\frac{x^5+x^4+x^3+x^2+x+1}{x+1}=\frac{(x-1)(x^5+x^4+x^3+x^2+x+1)}{(x-1)(x+1)}=\frac{x^6-1}{x^2-1}.$$

However, simple, the final rational expression is **not** equal to $q(x)$ since its domain of definition is $\{x \in \mathbb{R} \mid x \neq \pm 1\}$. Restricting to this latter domain, however, the two rational expressions become equal.

Example 9.1.2 (The Fibonacci Sequence via Continued Fractions) Consider the sequence of (finite) continued fractions

$$q_1(x) = 1 + \frac{1}{x},\ q_2(x) = 1 + \frac{1}{1+\frac{1}{x}},\ q_3(x) = 1 + \frac{1}{1+\frac{1}{1+\frac{1}{x}}},\ q_4(x) = 1 + \frac{1}{1+\frac{1}{1+\frac{1}{1+\frac{1}{x}}}}, \ldots$$

The pattern of this sequence is that any term can be obtained from the previous by the **inductive formula**

$$q_{n+1}(x) = q_n\left(1 + \frac{1}{x}\right), \quad n \in \mathbb{N}.$$

Writing the members of this sequence as **rational fractions**, we have

$$q_1(x) = \frac{x+1}{x},\ q_2(x) = \frac{2x+1}{x+1},\ q_3(x) = \frac{3x+2}{2x+1},\ q_4(x) = \frac{5x+3}{3x+2},\ q_5(x) = \frac{8x+5}{5x+3}, \ldots$$

The general pattern of the coefficients of these rational fractions is easy to recognize. The coefficients are members of the sequence $0, 1, 1, 2, 3, 5, 8, \ldots$, and every member of this sequence is obtained as the sum of the previous two. This is the Fibonacci sequence discussed previously in Example 3.1.2. Our observation on the coefficients of the rational fractions can be written as

$$q_n(x) = \frac{F_{n+1}x + F_n}{F_n x + F_{n-1}}, \quad n \in \mathbb{N}.$$

Indeed, we can verify that this is correct using Peano's Principle of Induction. For the initial step $n = 1$, we have

$$q_1(x) = \frac{F_2 x + F_1}{F_1 x + F_0} = \frac{x+1}{x},$$

and the formula holds. For the general induction step $n \Rightarrow n+1$, we assume that the formula is valid for n, $n \in \mathbb{N}$. We calculate

$$q_{n+1}(x) = q_n\left(1 + \frac{1}{x}\right) = \frac{F_{n+1}\left(1+\frac{1}{x}\right) + F_n}{F_n\left(1+\frac{1}{x}\right) + F_{n-1}} = \frac{F_{n+1}(x+1) + F_n x}{F_n(x+1) + F_{n-1}x}$$

$$= \frac{(F_{n+1} + F_n)x + F_{n+1}}{(F_n + F_{n-1})x + F_n} = \frac{F_{n+2}x + F_{n+1}}{F_{n+1}x + F_n}.$$

The general induction step is completed, and the formula follows.

Dividing the numerator and denominator by F_n, and using the ratios $r_n = F_{n+1}/F_n$, we obtain

$$q_n(x) = \frac{r_n x + 1}{x + r_n - 1}.$$

Since $\lim_{n\to\infty} r_n = \tau$, the golden number, we anticipate that

$$\lim_{n\to\infty} q_n(x) = \frac{\tau x + 1}{x + \tau - 1} = \tau,$$

where we used $\tau - 1 = 1/\tau$. We claim that this holds for $x \neq 1 - \tau$.

Remark Note that, for $x = 1 - \tau$, we have

$$q_n(1-\tau) = q_n(-1/\tau) = \frac{F_{n+1}(-1/\tau) + F_n}{F_n(1-\tau) + F_{n-1}} = \frac{F_{n+1}(-1/\tau) + F_n}{F_{n+1} - \tau F_n} = -\frac{1}{\tau} = 1-\tau.$$

To show the claim, let $x \neq 1 - \tau$. Setting $\delta = |x + \tau - 1| > 0$, we choose $N \in \mathbb{N}$ such that, for $n \geq N$, we have $|x + r_n - 1| > \delta/2$. (This is possible since $\lim_{n\to\infty} r_n = \tau$.)

For $n \geq N$, we now calculate

$$\begin{aligned} |q_n(x) - \tau| &= \left| \frac{r_n x + 1}{x + r_n - 1} - \frac{\tau x + 1}{x + \tau - 1} \right| \\ &= \frac{|(r_n x + 1)(x + \tau - 1) - (\tau x + 1)(x + r_n - 1)|}{|x + r_n - 1|\,|x + \tau - 1|} \\ &\leq \frac{2}{\delta^2 \tau} |\tau x + 1||x - \tau||r_n - \tau|. \end{aligned}$$

This gives

$$\lim_{n\to\infty} |q_n(x) - \tau| \leq \frac{2}{\delta^2 \tau} |\tau x + 1||x - \tau| \lim_{n\to\infty} |r_n - \tau| = 0.$$

The claim follows.

Recalling now the original definition of $q_n(x)$, we arrive at the so-called (infinite) **continued fraction**

$$\tau = 1 + \cfrac{1}{1 + \cfrac{1}{1 + \ddots}}.$$

We now turn to examples of rational expressions with several indeterminates:

Example 9.1.3 Show that, for $0 < x, y \in \mathbb{R}$, we have

$$(x + y)\left(\frac{1}{x} + \frac{1}{y}\right) \geq 4.$$

Indeed, multiplying both sides by xy, after simplification, we obtain $(x + y)^2 \geq 4xy$, or equivalently $(x - y)^2 \geq 0$. The inequality follows.

We define the **harmonic mean** (HM) of two positive real numbers x and y by

$$\frac{2}{\frac{1}{x} + \frac{1}{y}}.$$

The example above can be paraphrased by saying that the harmonic mean is always less than or equal to the arithmetic mean:

$$\frac{2}{\frac{1}{x} + \frac{1}{y}} \leq \frac{x + y}{2}.$$

Using the AM-GM inequality (Section 5.4), we can actually derive a stronger statement. For $x, y > 0$, we have the **GM-HM Inequality**

$$\frac{2}{\frac{1}{x} + \frac{1}{y}} \leq \sqrt{xy}.$$

Indeed, reducing the complex fraction and taking the reciprocal of both sides, this becomes the AM-GM inequality.

We conclude this section by two somewhat more involved examples of rational fractions in three indeterminates:

Example 9.1.4 Simplify the following rational expression

$$\frac{(x-a)(x-b)}{(c-a)(c-b)} + \frac{(x-b)(x-c)}{(a-b)(a-c)} + \frac{(x-c)(x-a)}{(b-c)(b-a)},$$

where $a, b, c \in \mathbb{R}$ are distinct.

Notice that, under the cyclic permutation $a \mapsto b \mapsto c \mapsto a$, the three terms transform into each other cyclically, and the sum remains the same. Keeping a, b, c fixed, this is a quadratic polynomial. The leading coefficient is

$$\frac{1}{(c-a)(c-b)} + \frac{1}{(a-b)(a-c)} + \frac{1}{(b-c)(b-a)}$$

$$= -\frac{a-b}{(a-b)(b-c)(c-a)} - \frac{b-c}{(a-b)(b-c)(c-a)} - \frac{c-a}{(a-b)(b-c)(c-a)} = 0.$$

Thus, this expression is either linear or constant. On the other hand, substituting $x = a$, $x = b$, and $x = c$, we invariably get 1. We conclude that the original rational expression is identically 1.

Example 9.1.5[1] If $1 \neq x, y, z \in \mathbb{R}$ such that $xyz = 1$, then show that

$$\left(1 + \frac{1}{x-1}\right)^2 + \left(1 + \frac{1}{y-1}\right)^2 + \left(1 + \frac{1}{z-1}\right)^2 \geq 1.$$

As in previous examples, it is convenient to homogenize the rational fractions on the left-hand side by the substitutions

$$x = \frac{a^2}{bc}, \quad y = \frac{b^2}{ca}, \quad z = \frac{c^2}{ab}, \quad abc \neq 0,$$

such that $a^2 \neq bc$, $b^2 \neq ca$, $c^2 \neq ab$. (Note that with this substitution $xyz = 1$ is automatically satisfied.) After simplification, we obtain

$$\frac{a^4}{(a^2-bc)^2} + \frac{b^4}{(b^2-ca)^2} + \frac{c^4}{(c^2-ab)^2} \geq 1.$$

The Cauchy–Schwarz inequality gives

$$(a^2+b^2+c^2)^2 \leq \left((a^2-bc)^2 + (b^2-ca)^2 + (c^2-ab)^2\right) \times \left(\frac{a^4}{(a^2-bc)^2} + \frac{b^4}{(b^2-ca)^2} + \frac{c^4}{(c^2-ab)^2}\right).$$

With this, it remains to show that

$$(a^2+b^2+c^2)^2 \geq (a^2-bc)^2 + (b^2-ca)^2 + (c^2-ab)^2.$$

Simplifying and rearranging, we obtain

$$a^2(b+c)^2 + 2abc(b+c) + b^2c^2 \geq 0.$$

This holds, however, since the left-hand side is a monic quadratic polynomial in the expression $a(b+c)$ with discriminant $(2bc)^2 - 4b^2c^2 = 0$.

[1] An equivalent problem was in the International Mathematical Olympiad, 2008. There are many solutions to this problem.

Exercises

9.1.1. Simplify the fraction[2]

$$\frac{(x+1)^4+4x^4}{x^2+(2x+1)^2}.$$

9.1.2. Let $f(x) = 1/(1-x)$, $x \neq 1$. Show that $f(f(f(x))) = x$, $x \neq 0, 1$.

9.1.3. Recall from Example 4.3.1 that a function $f : X \to \mathbb{R}$ is **even** if whenever $x \in X$ then we also have $-x \in X$ and $f(-x) = f(x)$. The function f is **odd** if whenever $x \in X$ then we also have $-x \in X$ and $f(-x) = -f(x)$. (a) Show that a polynomial $p : \mathbb{R} \to \mathbb{R}$ is even if and only if $p(x)$ consists of even degree monomials only, and it is odd if and only $p(x)$ consists of odd degree monomials only. (b) Show that any real function $f : X \to \mathbb{R}$ with $X \subset \mathbb{R}$ can be written as the sum of even and odd functions $f = f_0 + f_1$, where

$$f_0(x) = \frac{f(x)+f(-x)}{2} \quad \text{and} \quad f_1(x) = \frac{f(x)-f(-x)}{2}.$$

Here the common domain of definition of f_0 and f_1 is the set $X \cap (-X)$, where $-X = \{-x \,|\, x \in X\}$. (c) Write the rational functions $1/(1+x)$ and $1/(x^4+x)$ as sums of even and odd functions.

9.2 The Partial Fraction Decomposition

We start with rational fractions, $n_1(x)/d_1(x)$ and $n_2(x)/d_2(x)$, whose denominators $d_1(x)$ and $d_2(x)$ are **relatively prime**, that is, they have **no common factors**. (As before, a factor is understood to be a non-constant polynomial.) For simplicity, we will assume that the necessary polynomial divisions have been performed, and the quotients have been discarded, so that $n_1(x)/d_1(x)$ and $n_2(x)/d_2(x)$ are **proper**:[3] $\deg n_1(x) < \deg d_1(x)$ and $\deg n_2(x) < \deg d_2(x)$. We can write

$$\frac{n_1(x)}{d_1(x)} + \frac{n_2(x)}{d_2(x)} = \frac{n_1(x)d_2(x) + n_2(x)d_1(x)}{d_1(x)d_2(x)}.$$

After adding, the rational fraction on the right-hand side is also proper.

The **partial fraction decomposition** is the exact opposite of this. We start with a **proper** rational fraction $n(x)/d(x)$, $\deg n(x) < \deg d(x)$, and **assume** that the

[2] A special numerical case $x = 2013$ was part of a problem in the 2013 British Math Olympiad.

[3] As usual, deg denotes the degree of the respective polynomial.

denominator $d(x)$ splits into a product of two relatively prime polynomials: $d(x) = d_1(x) \cdot d_2(x)$, $\gcf(d_1(x), d_2(x)) = 1$, $\deg d_1(x), \deg d_2(x) < \deg d(x)$. We then **seek** polynomials $n_1(x)$ and $n_2(x)$ such that

$$\frac{n(x)}{d(x)} = \frac{n_1(x)}{d_1(x)} + \frac{n_2(x)}{d_2(x)}, \quad \deg n_1(x) < \deg d_1(x), \ \deg n_2(x) < \deg d_2(x).$$

We claim that $n_1(x)$ and $n_2(x)$ exist.

As we noted at the end of Section 7.6, as a consequence of the Euclidean Algorithm for polynomials, there exist polynomials $m_1(x)$ and $m_2(x)$ such that

$$m_1(x)d_1(x) + m_2(x)d_2(x) = \gcf(d_1(x), d_2(x)) = 1,$$

where we used that $d_1(x)$ and $d_2(x)$ are relatively prime. (Note that the gcf is determined up a non-zero constant multiple.)

Multiplying through by $n(x)/d(x)$, we obtain

$$\frac{n(x)m_1(x)d_1(x)}{d(x)} + \frac{n(x)m_2(x)d_2(x)}{d(x)} = \frac{n(x)}{d(x)}.$$

Using $d(x) = d_1(x)d_2(x)$, and canceling the common factors, we arrive at

$$\frac{n(x)m_2(x)}{d_1(x)} + \frac{n(x)m_1(x)}{d_2(x)} = \frac{n(x)}{d(x)}.$$

We now perform polynomial divisions. We divide $n(x)m_2(x)$ by $d_1(x)$ to obtain a quotient $q_1(x)$ and a remainder $n_1(x)$. Similarly, we divide $n(x)m_1(x)$ by $d_2(x)$ to obtain a quotient $q_2(x)$ and a remainder $n_2(x)$. By the Division Algorithm, we have

$$q_1(x) + \frac{n_1(x)}{d_1(x)} + q_2(x) + \frac{n_2(x)}{d_2(x)} = \frac{n(x)}{d(x)}$$

with

$$\deg n_1(x) < \deg d_1(x) \quad \text{and} \quad \deg n_2(x) < \deg d_2(x).$$

Now the crux is that all rational fractions are proper so that the polynomial sum $q_1(x) + q_2(x)$ must be zero. We obtain

$$\frac{n_1(x)}{d_1(x)} + \frac{n_2(x)}{d_2(x)} = \frac{n(x)}{d(x)}.$$

This concludes the proof that the partial fraction decomposition holds.

Example 9.2.1 Let $a, b, m \in \mathbb{N}$. Reduce the infinite sum

$$\sum_{n=1}^{\infty} \frac{n(b^m - a^m) + mb^m}{n(n+m)} \frac{a^n}{b^n}$$

to a finite sum.

The crux is to decompose the rational fraction

$$\frac{x(b^m - a^m) + mb^m}{x(x+m)}$$

into partial fractions. Since $\gcf(x, x+m) = 1$, $m \in \mathbb{N}$, the only possible partial fractions are of the form A/x and $B/(x+m)$, where $A, B \in \mathbb{R}$. We therefore write

$$\frac{x(b^m - a^m) + mb^m}{x(x+m)} = \frac{A}{x} + \frac{B}{x+m}.$$

Eliminating the denominators, we obtain

$$x(b^m - a^m) + mb^m = A(x+m) + Bx.$$

Since this holds for any value of the indeterminate x, we have

$$A + B = b^m - a^m \quad \text{and} \quad mA = mb^m.$$

This is easily resolved yielding $A = b^m$ and $B = -a^m$. Returning to our rational fraction, we thus have

$$\frac{x(b^m - a^m) + mb^m}{x(x+m)} = \frac{b^m}{x} - \frac{a^m}{x+m}.$$

For $x = n$, $n \in \mathbb{N}$, we substitute this into the infinite sum and calculate

$$\begin{aligned}
\sum_{n=1}^{\infty} \frac{n(b^m - a^m) + mb^m}{n(n+m)} \frac{a^n}{b^n} &= \sum_{n=1}^{\infty} \left(\frac{b^m}{n} - \frac{a^m}{n+m} \right) \frac{a^n}{b^n} \\
&= \sum_{n=1}^{\infty} \frac{1}{n} \frac{a^n}{b^{n-m}} - \sum_{n=1}^{\infty} \frac{1}{n+m} \frac{a^{n+m}}{b^n} = \sum_{n=1}^{\infty} \frac{1}{n} \frac{a^n}{b^{n-m}} - \sum_{n=m+1}^{\infty} \frac{1}{n} \frac{a^n}{b^{n-m}} \\
&= \sum_{n=1}^{m} \frac{1}{n} \frac{a^n}{b^{n-m}} = b^m \sum_{n=1}^{m} \frac{1}{n} \left(\frac{a}{b} \right)^n = b^m \left(\frac{a}{b} + \frac{1}{2} \left(\frac{a}{b} \right)^2 + \cdots + \frac{1}{m} \left(\frac{a}{b} \right)^m \right),
\end{aligned}$$

a finite sum.

Remark A reader well-versed in calculus will no doubt realize that the last sum in the parentheses is a partial sum of the power series expansion (at zero) of the natural logarithm $-\ln(1-x)$ for $x = a/b$.

Returning to the main line, the partial fraction decomposition above generalizes to decompositions with finitely many partial fractions in a straightforward manner using Peano's Principle of Induction.

More specifically, given a proper rational fraction $n(x)/d(x)$, we can split the denominator $d(x)$ into a product of **maximum number** of mutually relatively prime factors

$$d(x) = d_1(x)d_2(x)\cdots d_k(x), \quad \deg d_1(x), \deg d_2(x), \ldots, \deg d_k(x) < \deg n(x),$$

(assuming that there are at least two), and obtain the partial fraction decomposition

$$\frac{n(x)}{d(x)} = \frac{n_1(x)}{d_1(x)} + \frac{n_2(x)}{d_2(x)} + \cdots + \frac{n_k(x)}{d_k(x)},$$

where the partial fractions on the right-hand side are all proper.

The next question that we need to answer is the following: What are the possible (general) forms of the mutually relatively prime denominators $d_1(x), d_2(x), \ldots, d_k(x)$?

The answer depends on the field that the coefficients of the polynomials reside in. In our case, this is the field of real numbers $\mathbb{R}$. To answer this question we first consider the finer splitting of $d(x)$ into irreducible factors (as opposed to splitting $d(x)$ into relatively prime factors). For simplicity, from now on we assume that $d(x)$ is monic (has leading coefficient equal to 1) as are the factors in any decompositions of $d(x)$ into products. (A non-unit leading coefficient can always be absorbed into the numerator $n(x)$ of the fraction $n(x)/d(x)$.)

We now recall that the irreducible polynomials with real coefficients are either linear or quadratic. By our assumption, they are also monic so that they must be of the form $x - c$ with $c \in \mathbb{R}$, or $x^2 + px + q$ with $p, q \in \mathbb{R}$ such that the discriminant $p^2 - 4q < 0$.

Returning to our denominator $d(x)$ and its decomposition into relatively prime factors, we see that, corresponding to these two cases, the **relatively prime factors** must be powers of the irreducible factors:

$$(x-c)^m \qquad \text{and} \qquad (x^2+px+q)^n, \quad p^2-4q<0, \quad m,n \in \mathbb{N}.$$

We call m and n the **multiplicity** of the respective irreducible factor.

We first discuss the **multiplicity one** cases. Since partial fractions must be proper, we obtain that, corresponding to these two cases, in multiplicity 1 they must be of the form

$$\frac{A}{x-c} \qquad \text{and} \qquad \frac{Ax+B}{x^2+px+q}, \quad A, B \in \mathbb{R}.$$

Example 9.2.2 Find the infinite sum

$$\sum_{n=2}^{\infty} \frac{1}{n(n^2-1)}.$$

We view the general term of the series as a rational function in the indeterminate x (instead of n), and decompose it into partial fractions:

$$\frac{1}{x(x^2-1)} = \frac{1}{(x-1)x(x+1)} = \frac{A}{x-1} + \frac{B}{x} + \frac{C}{x+1}.$$

Eliminating the denominators, we obtain

$$1 = Ax(x+1) + B(x^2-1) + Cx(x-1).$$

This gives

$$A + B + C = 0, \quad A - C = 0, \quad B = -1.$$

This can be easily resolved to obtain $A = 1/2$, $B = -1$, $C = 1/2$. We thus have

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

We substitute this into the series and expand[4]

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

The crux is that the middle term $-2/n$ in the nth parentheses cancels the third term $1/((n-1)+1)$ in the previous parentheses, and the first term $1/((n+1)-1)$ in the next parentheses. Hence everything cancels in this sum[5] except three surviving terms $1 - 1 + 1/2 = 1/2$. Thus, we obtain

$$\sum_{n=2}^{\infty} \frac{1}{n(n^2-1)} = \frac{1}{4}.$$

Remark When the denominator splits into a product of mutually relatively prime (irreducible) linear factors there is a simpler method to find the coefficients. After we arrive at the equation $1 = Ax(x+1) + B(x^2-1) + Cx(x-1)$, letting $x = 1$ we

[4] For technical convenience, we doubled the sum.

[5] Sums with this property are called **telescopic**.

find $A = 1/2$, letting $x = 0$ we find $B = -1$, and letting $x = -1$ we find $C = 1/2$. We see that, by letting x equal to the roots of the irreducible factors, we can solve for the remaining coefficients. This process, however, starts losing its effectiveness when the linear factors of the denominator have multiplicity greater than 1.

For the case of linear (irreducible) factors, what we have done so far can be summarized in the following general setting. Assume that the denominator of a proper rational fraction $n(x)/d(x)$ splits into a product of **distinct** linear factors

$$d(x) = (x - c_1)(x - c_2) \cdots (x - c_k)$$

with all roots $c_1, c_2, \ldots, c_k \in \mathbb{R}$ distinct. Then we have the partial fraction decomposition

$$\frac{n(x)}{d(x)} = \frac{n(x)}{(x - c_1)(x - c_2) \cdots (x - c_k)} = \frac{A_1}{x - c_1} + \frac{A_2}{x - c_2} + \cdots + \frac{A_k}{x - c_k},$$

where $A_1, A_2, \ldots, A_k \in \mathbb{R}$.

Remark [6] Recall the Lagrange interpolation polynomial $\ell(x)$ introduced in Example 6.1.1. It is a polynomial of degree $< n$ uniquely defined by n distinct numbers $x_1, x_2, \ldots, x_n \in \mathbb{R}$, $2 \leq n \in \mathbb{N}$, and $y_i \in \mathbb{R}$, $i = 1, 2, \ldots, n$, such that $\ell(x_i) = y_i$, $i = 1, 2, \ldots, n$. The definition of $\ell(x)$ can be paraphrased in terms of the partial fraction decomposition as

$$\frac{\ell(x)}{(x - x_1)(x - x_2) \cdots (x - x_n)} = \frac{y_1/z_1}{x - x_1} + \frac{y_2/z_2}{x - x_2} + \cdots + \frac{y_n/z_n}{x - x_n},$$

where

$$z_i = \prod_{\substack{j=1 \\ j \neq i}}^{n} (x_i - x_j).$$

Returning to the main line, we now discuss the case of quadratic irreducible factors.

Example 9.2.3 Determine the partial fraction decomposition of the rational fraction

$$\frac{2x^3}{x^4 + x^2 + 1}.$$

According to Example 6.4.8 (for $y = 1$) the denominator decomposes as

$$x^4 + x^2 + 1 = (x^2 + x + 1)(x^2 - x + 1).$$

[6] The reader is indebted to one of the reviewers for having this pointed out.

Note that both quadratic factors are irreducible since their (common) discriminant is $1 - 4 = -3 < 0$. The partial fraction decomposition is

$$\frac{2x^3}{x^4 + x^2 + 1} = \frac{2x^3}{(x^2 + x + 1)(x^2 - x + 1)} = \frac{Ax + B}{x^2 + x + 1} + \frac{Cx + D}{x^2 - x + 1}.$$

As usual, we eliminate all denominators and obtain

$$\begin{aligned} 2x^3 &= (Ax + B)(x^2 - x + 1) + (Cx + D)(x^2 + x + 1) \\ &= (A + C)x^3 + (-A + B + C + D)x^2 + (A - B + C + D)x + (B + D). \end{aligned}$$

This gives

$$A + C = 2, \quad -A + B + C + D = 0, \quad A - B + C + D = 0, \quad B + D = 0.$$

This system of linear equations is easily solved, and we obtain $A = B = C = 1$ and $D = -1$. Substituting these back to the original decomposition, we finally arrive at the following:

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

In retrospect, this partial fraction decomposition is also clear form the identity $x^3 \pm 1 = (x \pm 1)(x^2 \mp x + 1)$.

Finally, an illustrative example for the "hybrid case" is as follows:

Example 9.2.4 Determine the partial fraction decomposition of the rational fraction

$$\frac{x^2 - 2}{x^3 + 2x^2 + 2x + 1}.$$

We factor the denominator by grouping

$$x^3+2x^2+2x+1 = (x^3+1)+(2x^2+2x) = (x+1)(x^2-x+1)+2x(x+1) = (x+1)(x^2+x+1).$$

The quadratic factor is irreducible (over $\mathbb{R}$) since its discriminant is $1-4 = -3 < 0$. The partial fraction decomposition is

$$\frac{x^2 - 2}{x^3 + 2x^2 + 2x + 1} = \frac{A}{x + 1} + \frac{Bx + C}{x^2 + x + 1}, \quad A, B, C \in \mathbb{R}.$$

The usual computations give

$$A + B = 1, \quad A + B + C = 0, \quad A + C = -2,$$

and finally we obtain $A = -1$, $B = 2$, $C = -1$. Hence, we have

$$\frac{x^2 - 2}{x^3 + 2x^2 + 2x + 1} = -\frac{1}{x + 1} + \frac{2x - 1}{x^2 + x + 1}.$$

We summarize the multiplicity one case as follows. Assume that the denominator of a proper rational fraction $n(x)/d(x)$ splits into a product of **distinct** linear and **distinct** quadratic factors

$$\begin{aligned} d(x) = &(x - c_1)(x - c_2) \cdots (x - c_k) \\ &\times (x^2 + p_1 x + q_1)(x^2 + p_2 x + q_2) \cdots (x^2 + p_l x + q_l), \end{aligned}$$

where the roots $c_1, c_2, \ldots, c_k \in \mathbb{R}$ and the pairs $(p_1, q_1), (p_2, q_2), \ldots, (p_l, q_l) \in \mathbb{R}^2$ are distinct, and the discriminants of the quadratic factors are all negative:

$$p_1^2 - 4q_1 < 0, \quad p_2^2 - 4q_2 < 0, \quad \ldots \quad p_l^2 - 4q_l < 0.$$

Then we have the partial fraction decomposition

$$\begin{aligned} \frac{n(x)}{d(x)} &= \frac{n(x)}{(x - c_1)(x - c_2) \cdots (x - c_k)(x^2 + p_1 x + q_1)(x^2 + p_2 x + q_2) \cdots (x^2 + p_l x + q_l)} \\ &= \frac{A_1}{x - c_1} + \frac{A_2}{x - c_2} + \cdots + \frac{A_k}{x - c_k} \\ &\quad + \frac{B_1 x + C_1}{x^2 + p_1 x + q_1} + \frac{B_2 x + C_2}{x^2 + p_2 x + q_2} + \cdots + \frac{B_l x + C_l}{x^2 + p_l x + q_l}, \end{aligned}$$

where $A_1, A_2, \ldots, A_k, B_1, B_2, \ldots, B_l, C_1, C_2, \ldots, C_l \in \mathbb{R}$.

It remains to discuss the **higher multiplicity** cases of repeated linear and quadratic factors. If $(x - c)^k$ with $2 \leq k \in \mathbb{N}$ is a relatively prime factor in the factorization of $d(x)$, then, in the partial fraction decomposition of the proper rational fraction $n(x)/d(x)$, the corresponding partial fraction should be $n_0(x)/(x - c)^k$, where $n_0(x)$ is a polynomial of degree $\leq k - 1$. Independent of the partial fraction decomposition, we write this as another sum of "partial fractions" as

$$\frac{n_0(x)}{(x - c)^k} = \frac{A_1}{x - c} + \frac{A_2}{(x - c)^2} + \cdots + \frac{A_k}{(x - c)^k},$$

where $A_1, A_2, \ldots, A_k \in \mathbb{R}$.

We gave a proof of this decomposition at the end of Section 6.5 as an application of the Division Algorithm for Polynomials.

Example 9.2.5 We have the partial fraction decomposition of the rational fraction

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

where the last equality is by simple inspection.

If $(x^2 + px + q)^k$ with $p^2 - 4q < 0$, $2 \leq k \in \mathbb{N}$, is a relatively prime factor in the factorization of $d(x)$, then, in the partial fraction decomposition of the proper rational fraction $n(x)/d(x)$, the corresponding partial fraction should be $n_0(x)/(x^2 + px + q)^k$, where $n_0(x)$ is a polynomial of degree $\leq 2k - 1$.

Independent of the partial fraction decomposition, we write this as another sum of "partial fractions" as

$$\frac{n_0(x)}{(x^2 + px + q)^k} = \frac{A_1x + B_1}{x^2 + px + q} + \frac{A_2x + B_2}{(x^2 + px + q)^2} + \cdots + \frac{A_kx + B_k}{(x^2 + px + q)^k},$$

where $A_1, \ldots, A_k, B_1, \ldots, B_k \in \mathbb{R}$.

Multiplying through $(x^2 + px + q)^k$, this is equivalent to

$$n_0(x) = (A_1x+B_1)(x^2+px+q)^{k-1}+(A_2x+B_2)(x^2+px+q)^{k-2}+\cdots+(A_kx+B_k).$$

To show the validity of the partial fraction decomposition above, we claim that, for **any** polynomial $n_0(x)$ of degree $\leq 2k-1$, there exist $A_1, \ldots, A_k, B_1, \ldots, B_k \in \mathbb{R}$ such that this equality holds.

Once again, this is an application of the Division Algorithm. We use Peano's Principle of Induction with respect to $k \in \mathbb{N}$.

For $k = 1$, the polynomial is linear or a constant, and the claim clearly holds.

For the general induction step $1, 2, \ldots, k - 1 \Rightarrow k$, we assume that the claim holds for any polynomial of degree $\leq 2k - 3$.

Let $n_0(x)$ be a polynomial of degree $\leq 2k - 1$. Dividing $n_0(x)$ by the degree $2k - 2$ polynomial $(x^2 + px + q)^{k-1}$, we obtain a linear quotient $A_1x + B_1$ and a remainder $n_1(x)$ which is of degree $\leq 2k - 3$ or zero:

$$n_0(x) = (A_1x + B_1)(x^2 + px + q)^{k-1} + n_1(x).$$

The induction hypothesis applies to $n_1(x)$, and we have

$$n_1(x) = (A_2x + B_2)(x^2 + px + q)^{k-2} + \cdots + (A_kx + B_k).$$

The induction is complete and the claim follows.

Example 9.2.6 Determine the partial fraction decomposition of the rational fraction

$$\frac{x^3 + x}{x^4 + 2x^3 + 3x^2 + 2x + 1}.$$

The symmetric sequence of coefficients in the denominator is suggestive for the grouping

$$\begin{aligned} x^4+2x^3+3x^2+2x+1 &= (x^4+x^3+x^2)+(x^3+x^2+x)+(x^2+x+1) \\ &= x^2(x^2+x+1)+x(x^2+x+1)+(x^2+x+1) \\ &= (x^2+x+1)^2. \end{aligned}$$

Thus, we have the partial fraction decomposition

$$\frac{x^3+x}{x^4+2x^3+3x^2+2x+1} = \frac{A_1x+B_1}{x^2+x+1} + \frac{A_2x+B_2}{(x^2+x+1)^2}.$$

Eliminating the denominators, we have

$$\begin{aligned} x^3+x &= (A_1x+B_1)(x^2+x+1)+A_2x+B_2 \\ &= A_1x^3+(A_1+B_1)x^2+(A_1+B_1+A_2)x+(B_1+B_2). \end{aligned}$$

Comparing coefficients, we have

$$A_1=1, \quad A_1+B_1=0, \quad A_1+B_1+A_2=1, \quad B_1+B_2=0.$$

This can be easily solved giving $A_1 = A_2 = B_2 = 1$ and $B_1 = -1$. Finally, we arrive at the following partial fraction decomposition

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

Remark To obtain the coefficients in the partial fraction decomposition we used the brute force "method of undetermined coefficients." Other approaches, notably the so-called **Heaviside Cover-Up Method**, and, using differential calculus, yet another method, reminiscent to the **Lagrange Interpolation**, are also available.

Exercises

9.2.1. Perform the partial fraction decomposition for the following:

$$(a)\ \frac{6x^2-7x-25}{x^3+2x^2-5x-6}; \quad (b)\ \frac{x^2-x+1}{x^3-3x^2+3x-1}; \quad (c)\ \frac{4x^3+3x^2+6x}{(x^2+x+1)(x^2+1)}.$$

9.2.2. Use the method of Example 9.2.3 to show that

$$\sum_{k=1}^{n} \frac{k}{k^4+k^2+1} = \frac{1}{2}\left(1-\frac{1}{n^2+n+1}\right).$$

Conclude that we have

$$\sum_{n=1}^{\infty} \frac{n}{n^4+n^2+1} = \frac{1}{2}.$$

9.3 Asymptotes of Rational Functions

Recall that a rational function $q : \mathbb{R} \to \mathbb{R}$ is defined by a rational expression $q(x)$ in the indeterminate x via $y = q(x)$. The rational expression $q(x)$ can be brought into a rational fraction $q(x) = n(x)/d(x)$, where $n(x)$ (the numerator) and $d(x)$ (the denominator) are polynomials. The domain of definition of $q(x)$ is the set of real numbers for which the denominator $d(x)$ does not vanish. Since $d(x)$ is a polynomial, it vanishes only at its roots. By the Factor Theorem, the number of roots of $d(x)$ cannot exceed the degree, $\deg d(x)$. We conclude that a rational function $q : \mathbb{R} \to \mathbb{R}$, $y = q(x) = n(x)/d(x)$, is defined for all real numbers except at the finitely many roots of the denominator $d(x)$. We call these points the **singular points** of the rational function.

Recall that a rational function is continuous everywhere in its domain; that is, it is continuous at every non-singular point.

In this section we discuss possible **asymptotes** of graphs of rational functions.

Recall from Section 8.4 that an asymptote to a set $E \in \mathbb{R}^2$ (in our case the graph of a rational function) is a half-line ℓ with end-point P_0 which satisfies the following: Given another point $P_1 \in \ell$ with associated affine parametrization $P_t = (1-t)P_0 + tP_1$, $0 \leq t \in \mathbb{R}$, we have $\lim_{t\to\infty} d(P_t, E) = 0$.

First, we discuss **vertical asymptotes**, that is, half-line asymptotes given by $x = c$, $c \in \mathbb{R}$, $y \gtreqless 0$.

For our rational function $q : \mathbb{R} \to \mathbb{R}$ given by the fractional representation $q(x) = n(x)/d(x)$, a vertical asymptote cannot happen at a non-singular point $c \in \mathbb{R}$ since at a non-singular point we have $d(c) \neq 0$, so that $\lim_{x\to c} n(x)/d(x) = n(c)/d(c)$ exists.

It remains to consider the case when $c \in \mathbb{R}$ is a singular point of $q(x) = n(x)/d(x)$, that is, we have $d(c) = 0$.

Assume that $c \in \mathbb{R}$ is a root of the denominator $d(x)$ of **multiplicity** $k_0 \in \mathbb{N}$, so that we have

$$d(x) = (x-c)^{k_0} d_0(x), \quad d_0(c) \neq 0.$$

Let $l_0 \in \mathbb{N}_0$ be the highest power of the root factor $(x-c)$ that divides the numerator $n(x)$, that is

$$n(x) = (x-c)^{l_0} n_0(x), \quad n_0(c) \neq 0.$$

(The case $n = 0$ corresponds to $n(c) \neq 0$.) We then have

$$q(x) = \frac{n(x)}{d(x)} = \frac{(x-c)^{l_0} n_0(x)}{(x-c)^{k_0} d_0(x)}, \quad n_0(c) \neq 0 \neq d_0(c).$$

Case I $k_0 \leq l_0$. In this case, we have

$$q(x) = \frac{n(x)}{d(x)} = (x-c)^{l_0-k_0} \frac{n_0(x)}{d_0(x)}.$$

Since $n_0(c) \neq 0 \neq d_0(c)$, we have

$$\lim_{x \to c} \frac{n(x)}{d(x)} = \begin{cases} 0, & \text{if} \quad k_0 < l_0 \\ n_0(c)/d_0(c), & \text{if} \quad k_0 = l_0. \end{cases}$$

We can **define** q at c by setting $q(c) = \lim_{x \to c} q(x)$. With this the extended q becomes continuous at c. We call c a **removable** singular point for q.

Clearly, in this case, neither of the two vertical half-lines at c can be a vertical asymptote for q.

Remark Let $p : \mathbb{R} \to \mathbb{R}$ be a polynomial function and $c \in \mathbb{R}$. Recall the **difference quotient** (Section 4.3)

$$\mathfrak{m}_p(x, c) = \frac{p(x) - p(c)}{x - c}, \quad x \neq c.$$

It is a rational function in the variable x, and its only singular point is c. Since the numerator $p(x) - p(c)$ vanishes at c, this singular point is removable. The construction of the derivative $p'(c)$ amounts to "remove" this singularity, and define $\mathfrak{m}_p(x, c)$ across c. We thus see that taking the derivative of a polynomial is the same as removing the singularity of the corresponding difference quotient.

Example 9.3.1 (Revisited) The rational function $q : \mathbb{R} \to \mathbb{R}$ given by

$$q(x) = \frac{x^m - 1}{x^n - 1}, \quad m, n \in \mathbb{N},$$

has a removable singular point at $c = 1$, since

$$\lim_{x \to 1} \frac{x^m - 1}{x^n - 1} = \frac{m}{n}.$$

(See Section 4.3.)

Case II $k_0 > l_0$. In this case, we have

$$q(x) = \frac{n(x)}{d(x)} = \frac{1}{(x-c)^{k_0-l_0}} \frac{n_0(x)}{d_0(x)}, \quad n_0(c) \neq 0 \neq d_0(c).$$

If $k_0 - l_0 \in \mathbb{N}$ is **even**, then we have

$$\lim_{x \to c} q(x) = \lim_{x \to c} \frac{n(x)}{d(x)} = \pm\infty,$$

where the sign $\pm$ is according to whether $n_0(c)/d_0(c) \gtrless 0$.

If $k_0 - l_0 \in \mathbb{N}$ is **odd**, then we have the one-sided limits

$$\lim_{x \to c^{\pm}} q(x) = \lim_{x \to c^{\pm}} \frac{n(x)}{d(x)} = \pm\infty,$$

where, for the right-limit, the sign $\pm$ is according to whether $n_0(c)/d_0(c) \gtrless 0$; and, for the left-limit, the sign $\mp$ is according to whether $n_0(c)/d_0(c) \gtrless 0$.

We now claim that the vertical half-line given by $x = c$, $y \gtreqless 0$, is a vertical asymptote for q if and only if, for any of the one-sided limits, we have

$$\lim_{x \to c^{\pm}} q(x) = \pm\infty.$$

By the above, it is enough to show that $\lim_{x \to c^{\pm}} q(x) = \infty$ (with either choice of the sign) implies that the half-line given by $x = c$, $y \geq 0$, is an asymptote for q.

Letting $P_0 = (c, 0)$ and $P_1 = (c, 1)$, we parametrize the vertical half-line by $y \mapsto P_y = (1-y)P_0 + yP_1 = (c, y)$, $y \geq 0$. We now estimate the distance of the graph G_q of q from this vertical half-line as follows:

$$\begin{aligned} 0 &\leq \lim_{y \to \infty} d(G_q, P_y) = \lim_{x \to c^{\pm}} d(G_q, P_{q(x)}) = \lim_{x \to c^{\pm}} d(G_q, (c, q(x))) \\ &\leq \lim_{x \to c^{\pm}} d((x, q(x)), (c, q(x))) = \lim_{x \to c^{\pm}} |x - c| = 0. \end{aligned}$$

The claim follows.

Second, we discuss the existence of **horizontal** and **oblique asymptotes**. A horizontal asymptote is a half-line given by $y = b$, $b \in \mathbb{R}$, and $x \gtreqless 0$. An oblique asymptote is a half-line given by $y = mx + b$, $m \neq 0$, $m, b \in \mathbb{R}$, and $x \gtreqless 0$.

In both cases (allowing $m = 0$) we let $P_0 = (0, b)$ and $P_1 = (\pm 1, \pm m + b)$. With this, we have the parametrization $P_x = (1-x)P_0 + xP_1 = (\pm x, \pm mx + b)$, $x \geq 0$.

The existence of these asymptotes depends on the degree of the numerator $n(x)$ and the degree of the denominator $d(x)$ in the fractional representation $q(x) = n(x)/d(x)$.

We write the numerator and denominator in descending order

$$n(x) = a_l x^l + a_{l-1}x^{l-1} + \cdots a_1 x + a_0, \quad a_l \neq 0,\ a_0, a_1, \ldots, a_l \in \mathbb{R},$$

and

$$d(x) = b_k x^k + b_{k-1}x^{k-1} + \cdots b_1 x + b_0, \quad b_k \neq 0,\ b_0, b_1, \ldots, b_k \in \mathbb{R},$$

where $\deg n(x) = l$ and $\deg d(x) = k$.

Case I Assume $\deg n(x) = l < k = \deg d(x)$.

Dividing both the numerator and the denominator by x^l, we obtain

$$q(x) = \frac{a_l x^l + a_{l-1}x^{l-1} + \cdots a_1 x + a_0}{b_k x^k + b_{k-1}x^{k-1} + \cdots b_1 x + b_0} = \frac{a_l + a_{l-1}\frac{1}{x} + \cdots a_1 \frac{1}{x^{l-1}} + a_0 \frac{1}{x^l}}{x^{k-l}\left(b_k + b_{k-1}\frac{1}{x} + \cdots b_1 \frac{1}{x^{k-1}} + b_0 \frac{1}{x^k}\right)}.$$

Hence, we have

$$\lim_{x\to\pm\infty} q(x) = \lim_{x\to\pm\infty} \frac{a_l + a_{l-1}\frac{1}{x} + \cdots a_1 \frac{1}{x^{l-1}} + a_0 \frac{1}{x^l}}{x^{k-l}\left(b_k + b_{k-1}\frac{1}{x} + \cdots b_1 \frac{1}{x^{k-1}} + b_0 \frac{1}{x^k}\right)} = \lim_{x\to\pm\infty} \frac{a_l}{x^{k-l} b_k} = 0.$$

This is because $k - l > 0$, and $\lim_{x\to\pm\infty} 1/x^m = 0$, for $m \in \mathbb{N}$.

In this case the positive and negative first axes given by $y = 0$, $x \gtreqless 0$, are **horizontal asymptotes**. Indeed, we have

$$0 \leq \lim_{x\to\infty} d(G_q, P_x) \leq \lim_{x\to\infty} d((\pm x, q(\pm x)), (\pm x, 0)) = \lim_{x\to\pm\infty} |q(x)| = 0.$$

Case II Assume $\deg n(x) = l \geq k = \deg d(x)$.

Performing polynomial division we obtain

$$q(x) = \frac{n(x)}{d(x)} = q_0(x) + \frac{r(x)}{d(x)},$$

where $\deg q_0 = l - k \geq 0$ and $\deg r(x) < \deg d(x)$ or $r(x)$ is zero.

By Case I, we have

$$\lim_{x\to\pm\infty} \frac{r(x)}{d(x)} = 0.$$

If $l = k$, then, by the division algorithm, $q_0(x) = a_l/b_k$, constant. In this case we have the **horizontal asymptotes** given by $y = b = a_l/b_k$, $x \gtreqless 0$. Indeed, as before, we have

$$\begin{aligned} 0 \leq \lim_{x\to\infty} d(G_q, P_x) &\leq \lim_{x\to\infty} d((\pm x, q(\pm x)), (\pm x, b)) \\ &= \lim_{x\to\pm\infty} |q(x) - b| = \lim_{x\to\pm\infty} |r(x)/d(x)| = 0. \end{aligned}$$

If $l = k + 1$, then, again by the division algorithm, $q_0(x) = mx + b$ is linear with slope $m = a_l/b_k \neq 0$. In this case we have **oblique asymptotes** given by $y = mx + b$, $x \gtreqless 0$. Indeed, as before, we have

$$\begin{aligned} 0 \leq \lim_{x\to\infty} d(G_q, P_x) &\leq \lim_{x\to\infty} d((\pm x, q(\pm x)), (\pm x, \pm mx + b))) \\ &= \lim_{x\to\pm\infty} |q(x) - (mx + b)| = \lim_{x\to\pm\infty} |r(x)/d(x)| = 0. \end{aligned}$$

Finally, if $l \geq k + 2$, then $q_0(x)$ is a polynomial of degree $l - k \geq 2$. In this case, we claim that there is no asymptote.

Clearly, there cannot be any horizontal asymptote since

$$\lim_{x\to\pm\infty} q(x) = \lim_{x\to\pm\infty} q_0(x) = \pm\infty.$$

Assume now that a half-line ℓ is an oblique asymptote. We may assume that the leading coefficient of q_0 is positive (that is $\lim_{x\to\infty} q(x) = \lim_{x\to\infty} q_0(x) = \infty$), and that ℓ is given by $y = mx + b$, $x \geq 0$, where $m > 0$ (since the other cases can be treated analogously). Let ℓ' be a half-line given by $y = m'x + b$, $x \geq 0$, where $m' > m$.

Since $\deg q_0(x) \geq 2$, by the previous case, we have

$$\lim_{x\to\infty} \frac{m'x + b}{q(x)} = \lim_{x\to\infty} \frac{m'x + b}{q_0(x)} = 0.$$

Let $0 < R \in \mathbb{R}$ be such that, for $x \geq R$, we have $mx + b > 0$, $q(x) > 0$ and

$$\frac{m'x + b}{q(x)} < 1.$$

We write this last inequality as

$$(mx + b <) m'x + b < q(x), \quad x \geq R.$$

The parametrization of the half-line ℓ is given by $x \mapsto P_x = (1 - x)P_0 + xP_1 = (x, mx + b)$, $x \geq 0$. Using the inequality above and the formula for the distance of a point to a line (Section 5.5), we estimate

$$\lim_{x\to\infty} d(G_q, P_x) \geq \lim_{x\to\infty} d(\ell', P_x) = \lim_{x\to\infty} \frac{|m'x - (mx + b) + b|}{\sqrt{m'^2+1}} = \lim_{x\to\infty} \frac{(m' - m)|x|}{\sqrt{m'^2+1}} = \infty.$$

Thus, ℓ cannot be an asymptote. The claim follows.

Example 9.3.2 Determine the asymptotes of the rational function

$$q(x) = \frac{x^7 + 3x^4 - x^3 - 1}{x^6 - x^2}.$$

We first perform polynomial division and obtain

$$q(x) = x + \frac{3x^4 - 1}{x^6 - x^2}.$$

We factor the denominator as

$$x^6 - x^2 = x^2(x^4 - 1) = x^2(x^2 - 1)(x^2 + 1) = x^2(x + 1)(x - 1)(x^2 + 1).$$

Substituting this into the expression above, we obtain

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

Partial fraction decomposition gives

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

Clearly, q has vertical asymptotes at $c = 0, \pm 1$, and an oblique asymptote given by $y = x$. (The last fraction does not contribute to the asymptotic behavior.) At the vertical asymptotes, we have

$$\lim_{x \to 0} q(x) = \infty \qquad \lim_{x \to 1^{\pm}} q(x) = \pm\infty \qquad \lim_{x \to -1^{\pm}} q(x) = \mp\infty.$$

Exercises

9.3.1. Find the asymptotes of the following rational function $y = (1+2x-x^2)/(1-x^2)$.

9.3.2. Construct the graphs of the rational functions

$$(a)\ y = \frac{x + 1}{x^2 - 2x + 1}; \quad (b)\ y = \frac{1 - x}{x^3 - 4x}; \quad (c)\ y = \frac{1}{x^3} + \frac{1}{x^2}.$$

9.4 Algebraic Expressions and Functions, Rationalization

An **algebraic expression** is a mathematical expression $f(x)$ constructed from numbers and an indeterminate x under the operations of **addition**, **multiplication**, **division**, and **exponentiation by rational exponents**.

A **complex algebraic fraction** is a fraction whose numerator and denominator are both algebraic expressions. A complex algebraic fraction can be brought to

a **simple algebraic fraction** whose numerator and denominators do not contain division involving the indeterminate.

The definition of algebraic expression can be naturally extended to expressions $f(x, y)$, $f(x, y, z)$, $f(x_1, x_2, \ldots, x_n)$, etc., in several indeterminates $x, y, z \ldots$ and $x_1, x_2, \ldots, x_n, n \in \mathbb{N}$.

A single-variable **algebraic function** is defined by an algebraic expression $f(x)$ in the indeterminate x via $y = f(x)$. A multivariate algebraic function is given by $z = f(x, y)$, $w = f(x, y, z)$, etc., where $f(x, y)$, $f(x, y, z)$ are algebraic expressions.

Example 9.4.1 Derive the following algebraic limit

$$\lim_{x\to\infty}\left(\sqrt{x-\sqrt{x}}-\sqrt{x}\right)=-\frac{1}{2}.$$

We calculate

$$\begin{aligned}\lim_{x\to\infty}\left(\sqrt{x-\sqrt{x}}-\sqrt{x}\right)&=\lim_{x\to\infty}\sqrt{x}\left(\sqrt{1-1/\sqrt{x}}-1\right)\\&=\lim_{x\to\infty}\sqrt{x}\frac{(1-1/\sqrt{x})-1}{\sqrt{1-1/\sqrt{x}}+1}=-\lim_{x\to\infty}\frac{1}{\sqrt{1-1/\sqrt{x}}+1}=-\frac{1}{2}.\end{aligned}$$

Example 9.4.2 Determine the value of the algebraic expression

$$\sqrt{\left(\frac{q}{2}\right)^2+\left(\frac{p}{3}\right)^3}$$

when $p = -1/3$ and $q = 25/27$.

Whenever possible we write all natural numbers as products of primes. We substitute $p = -1/3$ and $q = 5^2/3^3$, and calculate

$$\begin{aligned}\left(\frac{q}{2}\right)^2+\left(\frac{p}{3}\right)^3&=\left(\frac{5^2}{2\cdot 3^3}\right)^2+\left(-\frac{1}{3^2}\right)^3=\frac{5^4}{2^2\cdot 3^6}-\frac{1}{3^6}\\&=\frac{5^4-2^2}{2^2\cdot 3^6}=\frac{(5^2-2)(5^2+2)}{2^2\cdot 3^6}=\frac{23}{2^2\cdot 3^3}.\end{aligned}$$

Taking the square root, we obtain

$$\sqrt{\left(\frac{q}{2}\right)^2+\left(\frac{p}{3}\right)^3}=\frac{\sqrt{23}}{6\sqrt{3}}.$$

The final answer in the last example is not the simplified (simplest) form of a radical expression. When simplifying a radical expression it is common to abide by the following rules: (1) The nth root of an expression is considered to be in

simplified form if no factors of the radicand are perfect nth powers; (2) The radicand is not a fraction; and (3) The denominator of a fraction has no radicals.

Root rationalization is a process by which one or several radicals in the denominator of a simple algebraic fraction are eliminated. Although there is a wide range of situations, the majority fall under a few cases.

The simplest case is when the numerator $r(x)$ is an algebraic expression, and the denominator is the radical expression $\sqrt[n]{p(x)}$ with $p(x)$ a polynomial.

In this case the rationalization is achieved thorough multiplying the numerator and the denominator by $\sqrt[n]{p(x)^{n-1}}$ as follows

$$\frac{r(x)}{\sqrt[n]{p(x)}} = \frac{r(x)}{\sqrt[n]{p(x)}} \cdot \frac{\sqrt[n]{p(x)^{n-1}}}{\sqrt[n]{p(x)^{n-1}}} = \frac{r(x)}{p(x)} \cdot \sqrt[n]{p(x)^{n-1}}.$$

Example 9.4.3 Rationalize the algebraic expression $1/\sqrt[3]{x^2+x+1}$.

We calculate

$$\frac{1}{\sqrt[3]{x^2+x+1}} = \frac{1}{\sqrt[3]{x^2+x+1}} \frac{\sqrt[3]{x^2+x+1}^{\,2}}{\sqrt[3]{x^2+x+1}^{\,2}} = \frac{\sqrt[3]{x^2+x+1}^{\,2}}{x^2+x+1}.$$

Another case is when the denominator is the binomial of the form $\sqrt[n]{p(x)} - \sqrt[n]{q(x)}$, where $p(x)$ and $q(x)$ are polynomials. In this case, the polynomial identity

$$u^n - v^n = (u-v)(u^{n-1} + u^{n-2}v + \cdots + uv^{n-2} + v^{n-1})$$

is employed with $u = \sqrt[n]{p(x)}$ and $v = \sqrt[n]{q(x)}$. (Note that this also covers the case $\sqrt[n]{p(x)} + \sqrt[n]{q(x)}$ with n **odd** since $\sqrt[n]{p(x)} + \sqrt[n]{q(x)} = \sqrt[n]{p(x)} - \sqrt[n]{-q(x)}$.)

The rationalization follows the pattern:

$$\frac{r(x)}{\sqrt[n]{p(x)} - \sqrt[n]{q(x)}} = \frac{r(x)}{p(x)-q(x)} \cdot \left(\sqrt[n]{p(x)}^{\,n-1} + \sqrt[n]{p(x)}^{\,n-2} \sqrt[n]{q(x)} + \cdots \right.$$
$$\left. + \sqrt[n]{p(x)} \sqrt[n]{q(x)}^{\,n-2} + \sqrt[n]{q(x)}^{\,n-1} \right).$$

Example 9.4.4 Rationalize the algebraic expression $1/\left((1+x)\left(\sqrt{1+x^2} - \sqrt{x}\right)\right)$.

We have

$$\frac{1}{(1+x)\left(\sqrt{1+x^2} - \sqrt{x}\right)} = \frac{\sqrt{1+x^2} + \sqrt{x}}{(x+1)(x^2-x+1)} = \frac{\sqrt{1+x^2} + \sqrt{x}}{x^3-1}.$$

At times we may encounter a trinomial or a more complex expression to rationalize:

Example 9.4.5 Rationalize the simple algebraic fraction $1/(1-\sqrt{x}+\sqrt{x+1})$.

The trick is to use the difference of squares formula in the following setting:

$$\begin{aligned}(1-\sqrt{x}+\sqrt{x+1})(1-\sqrt{x}-\sqrt{x+1}) &= (1-\sqrt{x})^2-\sqrt{x+1}^2\\ &= 1-2\sqrt{x}+x-(x+1)=-2\sqrt{x}.\end{aligned}$$

Using this we now calculate

$$\begin{aligned}\frac{1}{1-\sqrt{x}+\sqrt{x+1}} &= \frac{1-\sqrt{x}-\sqrt{x+1}}{(1-\sqrt{x}+\sqrt{x+1})(1-\sqrt{x}-\sqrt{x+1})}\\ &= -\frac{1-\sqrt{x}-\sqrt{x+1}}{2\sqrt{x}} = -\sqrt{x}\frac{1-\sqrt{x}-\sqrt{x+1}}{2x}.\end{aligned}$$

The **domain of definition** of an algebraic expression is the (maximal) set of values of the indeterminates for which the algebraic expression is defined. Thus, the domain of definition of a simple algebraic fraction is the set of values of the indeterminates for which the denominator does not vanish, and all the radicands under even radical signs are non-negative. As in the case of rational expressions, the domain of definition may change during simplification processes.

Example 9.4.6 Determine the domain of definition of the following algebraic expression and simplify:

$$\sqrt{\frac{\sqrt{x}+1}{\sqrt{x}-1}}-\sqrt{\frac{\sqrt{x}-1}{\sqrt{x}+1}}.$$

First, due to the presence of the radical $\sqrt{x}$, we must have $x \geq 0$. In addition, $\sqrt{x} \neq 1$ so that $x \neq 1$. Finally, $\sqrt{x}-1 > 0$, or equivalently, $x > 1$. Taking the intersection of these intervals, we see that the domain of definition is the infinite interval $(1, \infty)$. We now calculate

$$\begin{aligned}\sqrt{\frac{\sqrt{x}+1}{\sqrt{x}-1}}-\sqrt{\frac{\sqrt{x}-1}{\sqrt{x}+1}} &= \sqrt{\frac{(\sqrt{x}+1)^2}{(\sqrt{x}-1)(\sqrt{x}+1)}}-\sqrt{\frac{(\sqrt{x}-1)^2}{(\sqrt{x}+1)(\sqrt{x}-1)}}\\ &= \frac{\sqrt{x}+1}{\sqrt{x-1}}-\frac{\sqrt{x}-1}{\sqrt{x-1}} = \frac{2}{\sqrt{x-1}}.\end{aligned}$$

Exercises

9.4.1. Simplify

$$\frac{x\sqrt{y\sqrt{z}}\ y\sqrt{z\sqrt{x}}}{z\sqrt{yz\sqrt{xz}}}.$$

9.4.2. Factor $x^{3/2} - y^{3/2}$.
9.4.3. Rationalize the algebraic fraction $1/(1 - \sqrt{2} + \sqrt{3})$.

9.5 Harmonic, Geometric, Arithmetic, Quadratic Means

Just as in the case of rational expressions, algebraic expressions naturally appear in various inequalities.

Example 9.5.1 For $x, y > 0$, we have

$$\frac{\sqrt{x} + \sqrt{y}}{\sqrt{2}} \leq \sqrt{x + y} < \sqrt{x} + \sqrt{y}.$$

Indeed, squaring, and using the binomial formula, we obtain

$$\frac{x + 2\sqrt{xy} + y}{2} \leq x + y < x + 2\sqrt{xy} + y.$$

Canceling the common terms, the first inequality reduces to the AM-GM inequality. The second inequality is obvious.

Example 9.5.2 For $x, y \in \mathbb{R}$, we have

$$\frac{x + y}{2} \leq \sqrt{\frac{x^2 + y^2}{2}}.$$

We may assume $x, y > 0$. Then the inequality follows from the previous example by a simple substitution. For a change, we also derive this inequality using geometry.

First, notice that all the expressions are **positively homogeneous** (that is, replacing the indeterminates x and y by tx and ty with $t > 0$, both sides of the inequality get multiplied by t).

Therefore, we may assume that $x^2 + y^2 = 2$. This is the equation of the circle on the plane $\mathbb{R}^2$ with center at the origin and radius $\sqrt{2}$. The tangent line to this circle at the point $(1, 1)$ is given by the linear equation $x + y = 2$. (See Section 5.5.) Since the circle is on one side of its tangent line, we obtain that any point $P = (x, y)$ on this circle, satisfying $x^2 + y^2 = 2$, also satisfies $x + y \leq 2$. Equivalently, we have

$$\sqrt{\frac{x^2 + y^2}{2}} = 1 \quad \Rightarrow \quad \frac{x + y}{2} \leq 1.$$

The inequality follows.

For $x, y \in \mathbb{R}$, the quantity $\sqrt{(x^2+y^2)/2}$ is called the **Quadratic Mean** (or **Root Mean Square** or **RMS**). The inequality just derived nicely fits into the chain of inequalities that we obtained previously for the various other means (Sections 5.4 and 9.1) as follows:

$$\frac{2}{\frac{1}{x}+\frac{1}{y}} \leq \sqrt{xy} \leq \frac{x+y}{2} \leq \sqrt{\frac{x^2+y^2}{2}}, \quad x, y > 0.$$

In words

Harmonic **M**ean $\leq$ **G**eometric **M**ean $\leq$ **A**rithmetic **M**ean $\leq$ **Q**uadratic **M**ean.

The chain of inequalities above has another beautiful geometric interpretation. (See Figure 9.1.) As before, notice that every mean of two numbers x and y is positively homogeneous. To derive the chain of inequalities above, we can therefore consider inclusion relations amongst the regions $X = \{(x, y) \in I \mid XM(x, y) \leq 1\}$ on the plane, where I is the (open) first quadrant, and $XM(x, y)$ stands for the harmonic, geometric, arithmetic, and quadratic means of x and y. More specifically, we see that the inequalities above are equivalent to $Q \subset A \subset G \subset H$.

Now, the defining inequality of Q is $\sqrt{(x^2+y^2)/2} \leq 1$, or equivalently, $x^2 + y^2 \leq 2$. Restricted to the first quadrant I, Q is a quarter disk with center at the origin and radius $\sqrt{2}$. In particular, the point $(1, 1)$ is on its boundary.

Next, A is a right triangle (with right angle at the origin) since its boundary line segment is given by the equation $x + y = 2$. As noted in the previous example, this line segment is tangent to the boundary circle of Q at $(1, 1)$ so that $Q \subset A$ follows.

G is a "hyperbolic region" in the first quadrant I bounded by the branch of the hyperbola in I given by the equation $xy = 1$. Our discussion of this hyperbola in

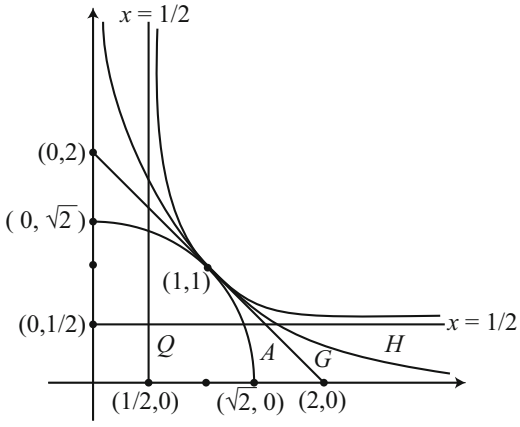


Fig. 9.1 Comparison of Means.

Section 8.4 implies that the line given by the equation $x + y = 2$ is tangent to the hyperbola at $(1, 1)$, so that $A \subset G$ follows.

The region H in the first quadrant I is bounded by a curve given by the equation $1/x + 1/y \geq 2$. We rewrite this as $(2x - 1)(2y - 1) \leq 1$. With respect to the new variables $u = 2x - 1$ and $v = 2y - 1$, we have $uv \leq 1$. The boundary curve $uv = 1$ is a hyperbola with center at $(0, 0)$ in the (u, v) variables, and therefore with center at $(1/2, 1/2)$ in the (x, y) variables. The asymptotes are $x = 1/2$ and $y = 1/2$. The line $x + y = 2$ is a common tangent to this hyperbola and $xy = 1$. Clearly $G \subset H$.

The chain of inequalities for the means follows.

Returning to the main line, recall that, as a byproduct of a cubic factoring problem, in Section 7.5 we obtained the AM-GM inequality in three indeterminates

$$\sqrt[3]{x \cdot y \cdot z} \leq \frac{x + y + z}{3}, \quad x, y, z \geq 0,$$

with equality if and only if $x = y = z$.

This indicates that the AM-GM inequality should hold for any number of indeterminates.

The precise statement is as follows. We have

$$\sqrt[n]{x_1 \cdot x_2 \cdots x_n} \leq \frac{x_1 + x_2 + \cdots + x_n}{n}, \quad x_1, \ldots, x_n \geq 0,$$

and equality holds if and only if $x_1 = x_2 = \ldots = x_n$.

We prove the general AM-GM inequality using Peano's Principle of Induction.

For $n = 1$ the AM-GM inequality is trivial. (Actually, even for $n = 2, 3$, we proved the AM-GM inequality previously.)

It remains to perform the general induction step $n \Rightarrow n+1$. To do this, we assume that the AM-GM inequality holds for n as stated above (for any $x_1, x_2, \ldots, x_n \geq 0$). We need to show that, for any $x_1, x_2, \ldots, x_n, x_{n+1} \geq 0$, we have

$$x_1 \cdot x_2 \cdots x_n \cdot x_{n+1} \leq \left(\frac{x_1 + x_2 + \cdots + x_n + x_{n+1}}{n + 1} \right)^{n+1},$$

with equality if and only if $x_1 = x_2 = \ldots = x_n = x_{n+1}$.

Let A denote the arithmetic mean in the parentheses on the right-hand side, that is

$$(n + 1)A = x_1 + x_2 + \cdots + x_n + x_{n+1}.$$

Without loss of generality we may assume that not all the numbers $x_1, x_2, \ldots, x_{n+1}$ are equal since otherwise the AM-GM inequality obviously holds. (In particular, we have $A > 0$.) Then one of these numbers is larger than A and one is smaller than A. Changing the indices, we may assume $x_n > A$ and $x_{n+1} < A$. Rearranging the defining formula for A above, we have

$$nA = x_1 + x_2 + \cdots + x_{n-1} + (x_n + x_{n+1} - A) = x_1 + x_2 + \cdots + x_{n-1} + x_n^*,$$

where

$$x_n^* = x_n + x_{n+1} - A \geq x_n - A > 0.$$

Notice the key fact that A is also the arithmetic mean of the n numbers $x_1, x_2, \ldots, x_{n-1}, x_n^*$.

We now apply the induction hypothesis, the AM-GM inequality, for these numbers as

$$A^{n+1} = A^n \cdot A \geq x_1 \cdot x_2 \cdots x_{n-1} \cdot x_n^* \cdot A,$$

where we multiplied through by A. We estimate the product of the last two factors as

$$x_n^* \cdot A - x_n \cdot x_{n+1} = (x_n + x_{n+1} - A)A - x_n \cdot x_{n+1} = (x_n - A)(A - x_{n+1}) > 0,$$

where the positivity of the factors in the last product is due to our choices of x_n and x_{n+1} above. Replacing $x_n^* \cdot A$ by the smaller product $x_n \cdot x_{n+1}$, we obtain

$$A^{n+1} = A^n \cdot A > x_1 \cdot x_2 \cdots x_{n-1} \cdot x_n \cdot x_{n+1}.$$

This is the AM-GM inequality for $n + 1$.

Finally, recall that we assumed that $x_1, x_2, \ldots, x_n, x_{n+1}$ are not all equal and we obtained here sharp inequality. This means that the equality case is also covered.

The proof of the general AM-GM inequality is complete.

Remark The general AM-GM inequality has another elementary proof. For completeness, we briefly outline this here as follows.

In the following $x_1, x_2, \ldots$ are non-negative indeterminates. For $0 \leq x_1, x_2 \in \mathbb{R}$, we have

$$x_1 x_2 = \left(\frac{x_1 + x_2}{2}\right)^2 - \left(\frac{x_1 - x_2}{2}\right)^2 < \left(\frac{x_1 + x_2}{2}\right)^2$$

unless $x_1 = x_2$. Using this and adding $0 \leq x_3, x_4 \in \mathbb{R}$, we have

$$x_1 x_2 x_3 x_4 < \left(\frac{x_1 + x_2}{2}\right)^2 \left(\frac{x_3 + x_4}{2}\right)^2 < \left(\frac{x_1 + x_2 + x_3 + x_4}{4}\right)^4,$$

unless $x_1 = x_2 = x_3 = x_4$.

Now, for $0 \leq x_1, x_2, \ldots, x_{2^m} \in \mathbb{R}$, $m \in \mathbb{N}$, by Peano's Principle of Induction

$$x_1 x_2 \cdots x_{2^m} < \left(\frac{x_1 + x_2 + \cdots + x_{2^m}}{2^m}\right)^{2^m},$$

unless $x_1 = x_2 = \ldots = x_{2^m}$.

Finally, given $0 \le x_1, \ldots, x_n \in \mathbb{R}$, $n \in \mathbb{N}$, choose $m \in \mathbb{N}$ such that $n < 2^m$. Let $0 \le x'_k = x_k$, $k = 1, \ldots, n$, and $0 \le x'_l = A = (x_1 + \cdots + x_n)/n$, $l = n+1, \ldots, 2^m$. With these, we have

$$x_1 \cdots x_n \cdot A^{2^m - n} = x'_1 \cdots x'_{2^m} < \left(\frac{x'_1 + \cdots + x'_{2^m}}{2^m} \right)^{2^m} = \left(\frac{nA + (2^m - n)A}{2^m} \right)^{2^m} = A^{2^m},$$

unless $x_1 = x_2 = \ldots = x_n$. Simplifying, the general AM-GM inequality, $\sqrt[n]{x_1 \cdots x_n} \le A = (x_1 + \cdots + x_n)/n$, follows.

We now briefly return to the Bernoulli inequality for rational exponents discussed in Section 3.2. Recall that we showed there that the Bernoulli inequality for **rational exponents** is equivalent to the monotonicity property of the sequence

$$e_n^*(s) = \left(1 + \frac{s}{n}\right)^n, \quad n \in \mathbb{N},$$

given by

$$e_n^*(s) < e_{n+1}^*(s), \quad 0 \ne s > -n, \ n \in \mathbb{N}.$$

We now show that the AM-GM-inequality actually implies **both** the Bernoulli inequality for rational exponents **and** the monotonicity property above.

First, note that the AM-GM inequality can be interpreted as a maximum principle for products: The product of n non-negative numbers $x_1, x_2, \ldots, x_n$ with a **given sum** is the largest if and only if $x_1 = x_2 = \cdots = x_n$.

With this, we show the monotonicity property above:

$$\left(1 + \frac{s}{n}\right)^n < \left(1 + \frac{s}{n+1}\right)^{n+1}, \quad 0 \ne s > -n, \ n \in \mathbb{N}.$$

Indeed, consider n copies of the non-negative number $1 + s/n$, $0 \ne s > -n$, $n \in \mathbb{N}$, and one copy of the number $1 (\ne 1 + s/n)$. These are $n + 1$ numbers. Their **product** is the left-hand side of the inequality above. Their **sum** is equal to $n+s+1$. Now consider $n+1$ copies of the non-negative number $1+s/(n+1)$. Their **product** is the right-hand side of the inequality above. Their **sum** is equal to $n + 1 + s$. By the maximum principle for products above, the monotonicity property follows.

Second, we derive the Bernoulli inequality with rational exponent $q \in \mathbb{Q}$, $0 < q < 1$, from the AM-GM inequality. We let $q = m/n$ with $0 < m < n$. In the AM-GM inequality we set $x_1 = \ldots = x_m = 1 + r$, $-1 < r \ne 0$, $r \in \mathbb{R}$, and $x_{m+1} = \ldots = x_n = 1$, and calculate

$$\begin{aligned}(1+r)^q = (1+r)^{\frac{m}{n}} = \sqrt[n]{(1+r)^m} &= \sqrt[n]{\overbrace{(1+r)\cdots(1+r)}^{m} \cdot \overbrace{1 \cdots 1}^{n-m}} \\ &< \frac{m(1+r) + (n-m)}{n} = 1 + \frac{m}{n} r = 1 + qr.\end{aligned}$$

The Bernoulli inequality follows.

In what follows we assemble a few examples of applications of the AM-GM inequality in several indeterminates.

Example 9.5.3 [7] Show that

$$n\left(\sqrt[n]{n}-1\right) < H_n = 1+\frac{1}{2}+\cdots+\frac{1}{n}, \quad 2 \le n \in \mathbb{N}.$$

This inequality is a simple application of the AM-GM inequality. We calculate

$$\begin{aligned} 1+\frac{H_n}{n} = \frac{n+H_n}{n} &= \frac{(1+1)+(1+1/2)+(1+1/3)+\cdots+(1+1/n)}{n} \\ &> \sqrt[n]{(1+1)(1+1/2)(1+1/3)\cdots(1+1/n)} \\ &= \sqrt[n]{2\cdot(3/2)\cdot(4/3)\cdots(n+1)/n} = \sqrt[n]{n+1}, \end{aligned}$$

where we used the AM-GM inequality, and noticed that the last product is telescopic. Moving the value of the radicand $n+1$ down by 1, the example follows.

Example 9.5.4 Find all monic polynomials $p(x)$ all of whose coefficients are ± 1 and all of whose roots are real.

We let

$$p(x) = x^n + a_{n-1}x^{n-1} + a_{n-2}x^{n-2} + \cdots + a_1 x + a_0, \quad a_i = \pm 1, \ i = 0\ldots, n-1.$$

Recall the estimate in Example 6.7.2

$$a_{n-2} \le \frac{n-1}{2n} a_{n-1}^2$$

which holds for any monic polynomial of degree n with real roots.

On the other hand, denoting the roots by $r_1, r_2, \ldots, r_n$, the Viète relations, the AM-GM inequality, and the Newton-Girard formula $p_2 = s_1^2 - 2s_2$ (Section 6.6), imply

$$\sqrt[n]{a_0^2} = \sqrt[n]{r_1^2 \ldots r_n^2} \le \frac{r_1^2 + \cdots r_n^2}{n} = \frac{a_{n-1}^2 - 2a_{n-2}}{n}.$$

Our conditions on the coefficients now give $a_{n-1}^2 = 1$, so that, by the first inequality, we have $a_{n-2} = -1$. On the other hand, $a_0^2 = 1$ so that the second inequality gives $n \le 3$.

For $n = 3$, equality holds in the AM-GM inequality above. Thus, $p(x)$ is a cubic polynomial with $r_1^2 = r_2^2 = r_3^2 = 1$; that is, the roots are ± 1. A simple enumeration

[7] In Section 10.3 we will derive much more precise estimates for these expressions, including the fact that both sides of this inequality grow logarithmically.

gives two possibilities $p(x) = x^3 \pm x^2 - x \mp 1$. For $n = 2$ and $n = 1$, we obtain $p(x) = x^2 \pm x - 1$ and $p(x) = x \pm 1$. The example follows.

Returning to the main line, we previously augmented the AM-GM inequality (in two indeterminates) by the harmonic and quadratic means into a chain of inequalities. The generalization to several indeterminates $x_1, x_2, \ldots, x_n > 0$ is the following:

$$\frac{n}{\frac{1}{x_1}+\frac{1}{x_2}+\cdots+\frac{1}{x_n}} \le \sqrt[n]{x_1 \cdot x_2 \cdots x_n} \le \frac{x_1+x_2+\cdots+x_n}{n} \le \sqrt{\frac{x_1^2+x_2^2+\cdots x_n^2}{n}}$$

with equalities throughout if and only if $x_1 = x_2 = \ldots = x_n$. We call this the general QM-AM-GM-HM inequality.

The first (new) inequality is an easy application of the general (middle) AM-GM inequality applied to the n indeterminates

$$\frac{x_1 \cdot x_2 \cdots x_n}{x_1}, \frac{x_1 \cdot x_2 \cdots x_n}{x_2}, \ldots, \frac{x_1 \cdot x_2 \cdots x_n}{x_n}.$$

The last inequality is an easy application of the Cauchy–Schwarz inequality of Section 6.7 (applied to $a_1{=}x_1, a_2{=}x_2, \ldots, a_n{=}x_n$ and $b_1{=}b_2 = \ldots = b_n{=}1$).

Example 9.5.5 [8] When is the quadratic mean Q_n, $n \in \mathbb{N}$, of the first n natural numbers an integer?

We have $Q_n = \sqrt{(1^2 + 2^2 + \cdots + n^2)/n} = \sqrt{(n+1)(2n+1)/6}$, where we used the formula for the sum of squares of the first n integers (before Example 3.2.12). We write this as

$$6Q_n^2 = (n+1)(2n+1)$$

and assume, from now on, that $Q_n \in \mathbb{N}$ is an integer. We first observe that $n+1$ and $2n+1$ are relatively prime. Indeed, we have $\gcd(n+1, 2n+1) = \gcd(n+1, n) = \gcd(1, n) = 1$.

Since $n + 1$ and $2n + 1$ have no common prime divisors, in view of the equation above, and apart from 2 or 3, for any prime divisor of either number, the square of this prime also divides the number. Finally, multiplying all the prime divisors of the respective numbers to form squares, since $2n + 1$ is always odd, we are left with only two cases to consider: (I) $n + 1 = 2a^2$, $2n + 1 = 3b^2$; (II) $n + 1 = 6a^2$, $2n + 1 = b^2$, for some $a, b \in \mathbb{N}$.

We can quickly rule out Case II as follows. In this case b is odd, $b = 2c + 1$, $c \in \mathbb{N}$, say. Substituting, this gives $2n + 1 = (2c + 1)^2 = 4c^2 + 4c + 1$, and hence $n = 2c^2 + 2c = 2(c^2 + c)$. In particular, n is even, and $n + 1$ is odd. This contradicts to $n + 1 = 6a^2$. Case II is not realized.

[8] Inspired by a problem in the USA Mathematical Olympiad, 1986.

Eliminating n in Case I, we obtain $(2a)^2 - 3 \cdot b^2 = 1$. This shows that the pair $(x, y) = (2a, b) \in \mathbb{N} \times \mathbb{N}$ satisfies Pell's equation

$$x^2 - 3 \cdot y^2 = 1,$$

with $d = 3$ (see Section 2.1). Since $(2, 1)$ is obviously the fundamental solution, the discussion in Section 2.1 gives all solutions in the form of the infinite sequence of pairs $(x_k, y_k) \in \mathbb{N} \times \mathbb{N}$, $k \in \mathbb{N}_0$, $(x_0, y_0) = (2, 1)$, defined inductively by

$$(x_{k+1}, y_{k+1}) = (2x_k + 3y_k, x_k + 2y_k), \quad k \in \mathbb{N}_0.$$

The first few terms of this sequence are[9]

$$(2, 1), (7, 4), (26, 15), (97, 56), (362, 209), (1351, 780), (5042, 2911).$$

Working backward to our original problem of integrality of Q_n, $n \in \mathbb{N}$, we need to extract from this sequence the terms with even first coordinate ($x = 2a$). A simple induction shows that, passing from one solution to the next, the coordinates switch parity (even to odd, and odd to even). This shows that every even term has even first coordinate.
Summarizing, we obtain that Q_n, $n \in \mathbb{N}$, is integral for the infinite sequence $\{n_k\}_{k \in \mathbb{N}_0}$, given by $n_k = x_{2k}^2/2 - 1$, $k \in \mathbb{N}_0$ (since $n+1 = 2a^2 = x^2/2$). The first few integral quadratic means are $Q_1 = 1$, $Q_{337} = 195$, $Q_{65521} = 37829$, $Q_{12710881} = 7338631$.

We now turn to a lesser known nonetheless important **Permutation Inequality**.[10] Recall from Example 0.4.2 that a permutation on the set $\{1, 2, \ldots, n\}$, $n \in \mathbb{N}$, of the first n natural numbers is a bijection $\sigma : \{1, 2, \ldots, n\} \to \{1, 2, \ldots, n\}$.

The permutation inequality states that for any two sets of n real numbers

$$x_1 \le x_2 \le \cdots \le x_n \quad \text{and} \quad y_1 \le y_2 \le \cdots \le y_n,$$

and for any permutation σ on $\{1, 2, \ldots, n\}$, we have

$$x_n y_1 + x_{n-1} y_2 + \cdots + x_1 y_n \le x_{\sigma(1)} y_1 + x_{\sigma(2)} y_2 + \cdots + x_{\sigma(n)} y_n \le x_1 y_1 + x_2 y_2 + \cdots + x_n y_n.$$

This chain of inequalities can best be interpreted in terms of permutations on $x_1, x_2, \ldots, x_n$, as follows. The permutation on the sum on the left-hand side that

[9] Note the continued fraction expansion $\sqrt{3} = 1 + \cfrac{1}{1+\cfrac{1}{2+\cfrac{1}{1+\cfrac{1}{2+\cdots}}}}$ and its convergents $1, 2, 5/3, 7/4, 19/11, 26/15, 71/41, 97/56, 265/153, 362/209, 989/571, \ldots$.

[10] Also called Rearrangement Inequality.

minimizes the permuted sums in the middle **reverses the order**: $i \mapsto n - i + 1, i = 1, 2, \ldots, n$; and the permutation on the sum on the right-hand side that maximizes the permuted sums in the middle is the **identity**: $i \mapsto i, i = 1, 2, \ldots, n$.

Finally, if strict inequalities hold

$$x_1 < x_2 < \cdots < x_n \quad \text{and} \quad y_1 < y_2 < \cdots < y_n,$$

then the order reversing permutation that minimizes all the permuted sums, and the identity permutation that maximizes all the permuted sums are both **unique**.

Remark As in the case of the Chebyshev sum inequality (Section 6.7), if the inequality signs are reversed in **one** sequence of inequalities ($x_1 \leq x_2 \leq \cdots \leq x_n$ or $y_1 \leq y_2 \leq \cdots \leq y_n$), then the reverse inequality signs hold in the permutation inequality.

Turning to the proof, once the upper bound is proved, the lower bound follows by applying the upper bound $x_1 \leq x_2 \leq \cdots \leq x_n$ replaced by $-x_n \leq -x_{n-1} \leq \cdots \leq -x_1$. Thus, it is enough to derive the upper bound. The simplest proof is by contradiction.

Let σ be a permutation on $\{1, 2, \ldots, n\}$ such that $x_{\sigma(1)}y_1 + x_{\sigma(2)}y_2 + \cdots + x_{\sigma(n)}y_n$ is maximal; and also assume that σ has the largest number of fixed points amongst all maximal sums. Assume that σ is **not the identity** permutation.

Let $1 \leq j < n$ be the first index for which $\sigma(j) \neq j$. Hence, σ is the identity permutation on $\{1, 2, \ldots, j-1\}$. (In particular, $j = n$ cannot hold since then σ would be the identity permutation on $\{1, 2, \ldots, n-1\}$, and therefore it would also fix n.)

Clearly, we have $j < \sigma(j)$, and there exists $j < k \leq n$ such that $j = \sigma(k)$. With these, we have the implications

$$j < \sigma(j) \quad \Rightarrow \quad x_j \leq x_{\sigma(j)} \quad \text{and} \quad j < k \quad \Rightarrow \quad y_j \leq y_k.$$

Expanding the product

$$0 \leq (x_{\sigma(j)} - x_j)(y_k - y_j),$$

we obtain

$$x_{\sigma(j)}y_j + x_j y_k \leq x_j y_j + x_{\sigma(j)}y_k.$$

We now define the permutation τ on $\{1, 2, \ldots, \}$ as follows.

$\tau = \sigma$ on $\{1, 2, \ldots, n\} \setminus \{j, k\}$; and $\tau(j) = \sigma(k) = j$ and $\tau(k) = \sigma(j)$.

Clearly, τ has one more fixed point, j, than σ, and, by the inequality above, the permuted sum corresponding to τ is at least as large as that of σ. This is a contradiction. The permutation inequality follows.

Finally, note that the last statement on sharp inequalities follows along the same lines replacing the inequalities by sharp ones.

Remark 1 The AM-GM inequality is a consequence of the permutation inequality.

Indeed, for $0 < x_1, x_2, \ldots, x_n \in \mathbb{R}$, let $c = \sqrt[n]{x_1 x_2 \cdots x_n}$, and define

$$a_1 = \frac{x_1}{c},\ a_2 = \frac{x_1 x_2}{c^2},\ \cdots,\ a_n = \frac{x_1 x_2 \cdots x_n}{c^n} = 1.$$

Finally, we let $b_i = 1/a_i$, $i = 1, 2, \ldots, n$. We apply the permutation inequality to the sequences $a_1, a_2, \ldots, a_n$ and $b_1, b_2, \ldots, b_n$. If we arrange the first sequence in increasing order (by some permutation), then the second sequence (similarly rearranged) will be reversely oriented. Thus, in the permutation inequality, the opposite inequality signs hold. We obtain

$$n = a_1 b_1 + a_2 b_2 + \cdots + a_n b_n \le a_1 b_n + a_2 b_1 + \cdots + a_n b_{n-1},$$

where we used the permutation that maps $1, 2, \ldots, n$ to $n, 1, 2, \ldots, n-1$. For the terms on the right-hand side, we have

$$a_1 b_n = \frac{a_1}{a_n} = \frac{x_1}{c},\ a_2 b_1 = \frac{a_2}{a_1} = \frac{x_2}{c}\ \cdots\ a_n b_{n-1} = \frac{a_n}{a_{n-1}} = \frac{x_n}{c}.$$

We obtain

$$n \le \frac{x_1 + x_2 + \cdots + x_n}{c}.$$

The AM-GM inequality follows.

Remark 2 The Chebyshev sum inequality (Section 6.7) is a direct consequence of the permutation inequality.

Indeed, let

$$a_1 \le a_2 \le \cdots \le a_n \quad \text{and} \quad b_1 \le b_2 \le \cdots \le b_n,$$

and use the permutation inequalities (for cyclic permutations on $\{1, 2, \ldots, n\}$) as follows:

$$\begin{aligned}
a_1 b_1 + a_2 b_2 + \cdots + a_n b_n &\le a_1 b_1 + a_2 b_2 + \cdots + a_n b_n \\
a_2 b_1 + a_3 b_2 + \cdots + a_1 b_n &\le a_1 b_1 + a_2 b_2 + \cdots + a_n b_n \\
&\cdots \quad \cdots \\
a_n b_1 + a_1 b_2 + \cdots + a_{n-1} b_n &\le a_1 b_1 + a_2 b_2 + \cdots + a_n b_n.
\end{aligned}$$

Adding, and factoring, we obtain

$$(a_1 + a_2 + \cdots + a_n)(b_1 + b_2 + \cdots + b_n) \le n(a_1 b_1 + a_2 b_2 + \cdots + a_n b_n).$$

The Chebyshev sum inequality follows.

Example 9.5.6 [11] Let $(a_n)_{n\in\mathbb{N}}$ be real sequence of positive numbers such that

$$\sum_{j=1}^{n} a_j = a_1 + \cdots + a_n \le C \cdot n^2, \quad n \in \mathbb{N},$$

for some constant $0 < C \in \mathbb{R}$. Show that

$$\sum_{n=1}^{\infty} \frac{1}{a_n} = \lim_{n\to\infty} \left(\frac{1}{a_1} + \cdots + \frac{1}{a_n} \right) = \infty.$$

First notice that the sequence of partial sums in the limit is strictly increasing. Therefore, it is enough to show that it is unbounded.

For $k \in \mathbb{N}$, we have

$$a_{k+1} + \cdots + a_{2k} < a_1 + \cdots + a_{2k} \le 4Ck^2.$$

Moreover, the AM-(GM)-HM inequality gives

$$\frac{k}{\frac{1}{a_{k+1}} + \cdots + \frac{1}{a_{2k}}} \le \frac{a_{k+1} + \cdots + a_{2k}}{k} < 4Ck.$$

This gives

$$\frac{1}{4C} < \frac{1}{a_{k+1}} + \cdots + \frac{1}{a_{2k}}$$

for **all** $k \in \mathbb{N}$. Applying this for $k = 2^n$, $n \in \mathbb{N}_0$, and summing up, the example follows.

As a simple application, letting $a_n = n$, $n \in \mathbb{N}$, we have

$$\sum_{j=1}^{n} a_j = a_1 + \cdots + a_n = 1 + \cdots + n = \frac{n(n+1)}{2} \le n^2.$$

The divergence of the **harmonic series**, $\sum_{n=1}^{\infty} 1/n = \infty$, follows again.

Example 9.5.7 [12] Let $0 < a, b, c \in \mathbb{R}$ such that $abc = 1$. Show that

$$a + b + c \le a^2 + b^2 + c^2.$$

[11] Inspired by a problem in the Balkan Mathematical Olympiad, 2008.

[12] This and several other examples can be treated in multivariate calculus as simple examples of the Lagrange Multipliers Method.

We make the left-hand side of the inequality homogeneous of degree 2 by multiplying by $\sqrt[3]{abc} = 1$. Using fractional exponents, we obtain

$$a^{4/3}b^{1/3}c^{1/3} + a^{1/3}b^{4/3}c^{1/3} + a^{1/3}b^{1/3}c^{4/3} \le a^2 + b^2 + c^2.$$

Now both sides of the inequality are homogeneous of degree 2, so that it should be valid for all $a, b, c \in \mathbb{R}$.

We now write the right-hand side as

$$a^2 + b^2 + c^2 = \left(\frac{2a^2}{3} + \frac{b^2}{6} + \frac{c^2}{6}\right) + \left(\frac{a^2}{6} + \frac{2b^2}{3} + \frac{c^2}{6}\right) + \left(\frac{a^2}{6} + \frac{b^2}{6} + \frac{2c^2}{3}\right),$$

and use the AM-GM inequality for each term. We have

$$\frac{2a^2}{3} + \frac{b^2}{6} + \frac{c^2}{6} = \frac{1}{6}(a^2 + a^2 + a^2 + a^2 + b^2 + c^2) \ge \sqrt[6]{a^8b^2c^2} = a^{4/3}b^{1/3}c^{1/3},$$

and analogously with the other two terms. The inequality follows.

Example 9.5.8 [13] Given $0 < a, b, c \in \mathbb{R}$, show that

$$\frac{a}{\sqrt{a^2 + 8bc}} + \frac{b}{\sqrt{b^2 + 8ca}} + \frac{c}{\sqrt{c^2 + 8ab}} \ge 1.$$

We first notice that the fractions are homogeneous in (a, b, c) (of degree 0); that is, they remain unchanged if (a, b, c) is replaced by (ka, kb, kc), $k > 0$.
This means that we can assume $abc = 1/8$, so that the inequality above reduces to

$$\frac{a}{\sqrt{a^2 + \frac{1}{a}}} + \frac{b}{\sqrt{b^2 + \frac{1}{b}}} + \frac{c}{\sqrt{c^2 + \frac{1}{c}}} = \frac{1}{\sqrt{1 + \frac{1}{a^3}}} + \frac{1}{\sqrt{1 + \frac{1}{b^3}}} + \frac{1}{\sqrt{1 + \frac{1}{c^3}}} \ge 1.$$

By monotonicity (of the three fractions on the left-hand side), we need to show that this holds if $abc \ge 1/8$.

We now change the variables as

$$x = \frac{a}{\sqrt{a^2 + \frac{1}{a}}}, \quad y = \frac{b}{\sqrt{b^2 + \frac{1}{b}}}, \quad z = \frac{c}{\sqrt{c^2 + \frac{1}{c}}}.$$

With these, we need to show

$$x + y + z < 1 \quad \Rightarrow \quad \frac{x^2y^2z^2}{(1 - x^2)(1 - y^2)(1 - z^2)} < \frac{1}{8^3},$$

[13] This is a problem by Hojoo Lee; see also the International Mathematical Olympiad, 2001.

where we changed into the contrapositive statement. We now use the AM-GM inequality (in eight indeterminates) as

$$1-x^2 > (x+y+z)^2-x^2 = y^2+z^2+xy+xy+yz+yz+zx+zx \geq 8\sqrt[8]{y^2z^2xyxyyzyzzxzx}.$$

Simplifying, we obtain

$$1 - x^2 > 8x^{1/2}y^{3/4}z^{3/4}.$$

Applying this to the other two variables, the inequality follows.

Exercises

9.5.1. In this exercise we give a geometric interpretation of the QM-AM-GM-HM inequality. (See Figure 9.2.) Let $0 < x, y \in \mathbb{R}$, and consider a line segment $[A, B]$ with $d(A, B) = x + y$ and division point $D \in [A, B]$ such that $d(A, D) = x$ and $d(B, D) = y$. Construct a semi-circle with diameter $[A, B]$ and center $O = (A + B)/2$, and let C be the intersection of this semi-circle with the line through D and perpendicular to the line extension of $[A, B]$. (a) Show that $d(C, D)$ is the geometric mean of $x = d(A, D)$ and $y = d(B, D)$, and explain why this gives the AM-GM inequality. (b) Let $[D, E]$, $E \in [O, C]$, be the altitude line of the triangle $\triangle[O, C, D]$ from vertex D. Show that $d(C, E)$ is the harmonic mean of $x = d(A, D)$ and $y = d(B, D)$, and explain why this gives the HM-GM inequality. (c) Let F be the midpoint of the semi-circle (with endpoints A and B) cut out by the radial segment perpendicular to the diameter $[A, B]$ at the midpoint O. Show that $d(D, F)$ is the quadratic mean of $x = d(A, D)$ and $y = d(B, D)$, and explain why this gives the QM-AM inequality.

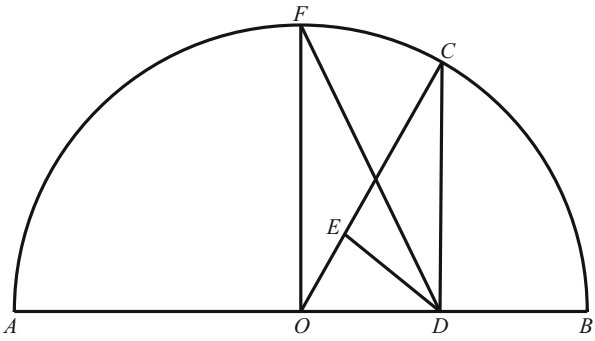


Fig. 9.2 Geometric Interpretation of the Means.

9.5.2. Let $m, n \in \mathbb{N}$. Use the general AM-GM inequality to show that the minimum of the function $f(x) = x^m + 1/x^n$, $0 < x \in \mathbb{R}$, is attained at $x = \sqrt[m+n]{n/m}$.

9.5.3. Derive the following relations amongst the means $XM(x, y)$, $0 < x, y \in \mathbb{R}$, where $X = H, G, A, Q$:

(1) $AM(x, y) = QM(\sqrt{x}, \sqrt{y})^2$;
(2) $GM(x, y)^2 = 2AM(x, y) \cdot HM(x, y)$;
(3) $AM(AM(x, y), GM(x, y)) = AM(\sqrt{x}, \sqrt{y})^2$;
(4) $GM(AM(x, y), HM(x, y)) = GM(x, y)/\sqrt{2}$;
(5) $QM(QM(x, y), GM(x, y)) = AM(x, y)$.

9.6 The Greatest Integer Function

In a few instances we previously encountered the notation $[x]$ for the greatest integer less than or equal to $x \in \mathbb{R}$. The **greatest integer** $[x]$ is actually an expression depending on the indeterminate $x \in \mathbb{R}$.

History
In his celebrated Quadratic Reciprocity Theorem Gauss introduced the square bracket notation above for the greatest integer. For any real x one can also define the smallest integer not less than x. This is usually called the **ceiling** of x denoted by $\lceil x \rceil$. Because of this duality, the Canadian computer scientist Kenneth Iverson (1820–2004) renamed the greatest integer $[x]$ of x as the **floor** with new notation $\lfloor x \rfloor$. In European textbooks one also finds the name **entier** which is "integer" in French, in honor of the French mathematician Adrien-Marie Legendre (1752–1833) who used this concept first in 1798. Finally, note that our ordinary **rounding** of a positive number x in everyday life can be expressed as $[x + 0.5]$.

We now proceed to show that $[x]$ is **not** an algebraic expression.

The usual definition of a real algebraic expression is actually wider than the one we adopted previously: An expression $f(x_1, \ldots, x_n)$ in n indeterminates $x_1, \ldots, x_n$ is called **algebraic** if it satisfies an equation $F(f(x_1, \ldots, x_n), x_1, \ldots, x_n) = 0$, where $F(x_0, x_1, \ldots, x_n)$ is an irreducible **polynomial** in the $n + 1$ indeterminates $x_0, x_1, \ldots, x_n$. This definition includes **polynomials** ($F(x_0, x_1, \ldots, x_n) = x_0 - p(x_1, \ldots, x_n)$ with $p(x_1, \ldots, x_n)$ a polynomial), **rational expressions** ($F(x_0, x_1, \ldots, x_n) = d(x_1, \ldots, x_n) \cdot x_0 - n(x_1, \ldots, x_n)$ with $n(x_1, \ldots, x_n)/d(x_1, \ldots, x_n)$ a rational expression), **root expressions** ($F(x_0, x_1, \ldots, x_n) = x_0^n - g(x_1, \ldots, x_n)$), etc., and, in general, any algebraic expression (constructed from indeterminates $x_1, \ldots, x_n$, and numbers under the operations of addition, multiplication, division, and exponentiation by rational exponents).

The main difference between this and our more restrictive definition is that the former includes roots of polynomials of degree ≥ 5 for which, according to Galois theory, there is no general root formula.

Assume now that $[x]$ is algebraic. According to this more general definition, this means that there exists a non-zero polynomial $F(x, y)$ such that $F(x, [x]) = 0$.

Expanding F, we obtain

$$p_n(x)[x]^n + p_{n-1}[x]^{n-1} + \cdots + p_1(x)[x] + p_0(x) = 0,$$

where $p_0(x), p_1(x), \ldots, p_n(x)$ are polynomials.

The principal property of the greatest integer we use here is that, for any integer $a \in \mathbb{Z}$, we have $[x] = a$ if and only if $a \leq x < a + 1, x \in \mathbb{R}$.

Thus, for a **given** $a \in \mathbb{Z}$, we have

$$p_n(x)a^n + p_{n-1}(x)a^{n-1} + \cdots + p_1(x)a + p_0(x) = 0,$$

for **any** $x \in [a, a + 1)$. Since the left-hand side is a polynomial in the indeterminate x (and thereby has only finitely many roots unless identically zero), it follows that the equation above holds for **all** $x \in \mathbb{R}$ (and thereby, for all $a \in \mathbb{Z}$).

We now fix $x \in \mathbb{R}$ and consider this equation for all $a \in \mathbb{Z}$. Since it is a polynomial of degree $\leq n$ in the indeterminate a, it has finitely many roots, so once again, this is possible only if $p_0(x) = p_1(x) = \ldots = p_n(x) = 0$. This is a contradiction, and the claim follows.

We now proceed to explore the properties of the greatest integer.

Clearly, we have $[[x]] = [x]$ and $[n + x] = [x] + n$ for all $n \in \mathbb{Z}$ and $x \in \mathbb{R}$.

In general, for addition, we have

$$[x] + [y] \leq [x + y] \leq [x] + [y] + 1, \quad x, y \in \mathbb{R}.$$

For multiplication and division, we have

$$[x] \cdot [y] \leq [x \cdot y], \quad 0 \leq x, y \in \mathbb{R},$$

and

$$\left[\frac{x}{n}\right] = \left[\frac{[x]}{n}\right], \quad n \in \mathbb{N}, \ x \in \mathbb{R}.$$

Example 9.6.1 For what $n \in \mathbb{N}$ is $[n^2/3]$ a prime?

By the Division Algorithm, we have $n = 3q + r$, $r = 0, 1, 2$, $q, r \in \mathbb{N}$. For $r = 0$, we have $[n^2/3] = [9q^2/3] = 3q^2$. This is a prime only if $q = 1$, and so $n = 3$. For $r = 1$, we have $[n^2/3] = [(3q + 1)^2/3] = [(9q^2 + 6q + 1)/3] = 3q^2 + 2q = q(3q + 2)$. This is a prime if $q = 1$, and so $n = 4$. For $r = 2$, we have $[n^2/3] = [(3q+2)^2/3] = [(9q^2+12q+4)/3] = 3q^2+4q+1 = (q+1)(3q+1)$. This is never a prime. Summarizing, we obtain $n = 3, 4$.

Example 9.6.2 [14] Solve the system of equations

$$[x] + [y] = 1 \quad \text{and} \quad x \cdot |x| + y \cdot |y| = 1.$$

[14] This and many variants are standard problems for the greatest integer; see also The Olympiad Corner, April, 1999.

We proceed to find the sets of points in the plane $\mathbb{R}^2$ defined by each of the equations.

It is clear that the first equation gives a doubly infinite sequence of squares

$$[a, a+1) \times [1-a, 2-a), \quad a \in \mathbb{Z}.$$

The second equation gives the quarter unit circle in the (closed) first quadrant **I** given by $x^2 + y^2 = 1$, $x, y \geq 0$; a half-branch of the hyperbola in the second quadrant **II** given by $-x^2 + y^2 = 1$, $x \leq 0 \leq y$; the empty set in the third quadrant **III**; and a half-branch of the hyperbola given by $x^2 - y^2 = 1$, $y \leq 0 \leq x$ in the fourth quadrant **IV**.

Clearly, the only intersection of these two sets is $(1, 0)$ and $(0, 1)$.

Example 9.6.3 [15] Show that, for $n \in \mathbb{N}$, we have

$$\left[\sqrt{n} + \sqrt{n+1}\right] = \left[\sqrt{4n+2}\right].$$

For $n \in \mathbb{N}$, squaring, we obtain

$$\left(\sqrt{n} + \sqrt{n+1}\right)^2 = 2n + 1 + 2\sqrt{n^2+n}.$$

Since

$$n^2 < n^2 + n < n^2 + n + \frac{1}{4} = \left(n + \frac{1}{2}\right)^2,$$

we get

$$4n + 1 < \left(\sqrt{n} + \sqrt{n+1}\right)^2 < 4n + 2.$$

Taking square roots, we arrive at the following

$$\sqrt{4n+1} < \sqrt{n} + \sqrt{n+1} < \sqrt{4n+2}.$$

This gives

$$\left[\sqrt{4n+1}\right] \leq \left[\sqrt{n} + \sqrt{n+1}\right] \leq \left[\sqrt{4n+2}\right].$$

[15] This is the third (and last) in the list of Ramanujan's Question 723, Papers 332, submitted to the Journal of the Indian Mathematical Society 7, p. 240; 10 pp. 357–358. It was also a problem in the William Lowell Putnam Exam, 1948. Note that Ramanujan (1887–1920) also proved that, for all $n \in \mathbb{N}$, we have $\left[\frac{n}{3}\right] + \left[\frac{n+2}{6}\right] + \left[\frac{n+4}{6}\right] = \left[\frac{n}{2}\right] + \left[\frac{n+3}{6}\right]$ and $\left[\frac{1}{2} + \sqrt{n + \frac{1}{2}}\right] = \left[\frac{1}{2} + \sqrt{n + \frac{1}{4}}\right]$.

We finally claim that equalities hold here. If not, then there would exist $m \in \mathbb{N}$ such that

$$\sqrt{4n+1} < m \leq \sqrt{4n+2},$$

or equivalently

$$4n + 1 < m^2 \leq 4n + 2.$$

This is impossible: If $m = 2k$, $k \in \mathbb{N}$, is even, then $m^2 = 4k^2$; and if $m = 2k + 1$, $k \in \mathbb{N}$, is odd, then $m^2 = (2k + 1)^2 = 4(k^2 + k) + 1$.

Example 9.6.4 [16] Show that, for $0 < x \in \mathbb{R}$ and $n \in \mathbb{N}$, we have

$$\sum_{k=1}^{n} \frac{[kx]}{k} \leq [nx].$$

We use induction with respect to $n \in \mathbb{N}$ (and fixed $0 < x \in \mathbb{R}$). For $n = 1$, the inequality is a tautology. For $n = 2$, the stated inequality is equivalent to $2[x] \leq [2x]$, and this holds by the general estimate on the greatest integer above.

The general induction step $1, 2, \ldots, n \Rightarrow n + 1$ is an elaborate rearrangement of the left-hand side of the inequality as follows.

By the induction hypothesis, we have

$$\sum_{k=1}^{n} \left(\sum_{l=1}^{k} \frac{[lx]}{l} \right) \leq \sum_{k=1}^{n} [kx], \quad k = 1, 2, \ldots, n.$$

The double sum can be rearranged as

$$\begin{aligned} \sum_{k=1}^{n} \left(\sum_{l=1}^{k} \frac{[lx]}{l} \right) &= \sum_{k=1}^{n} (n - k + 1) \frac{[kx]}{k} = (n + 1) \sum_{k=1}^{n} \frac{[kx]}{k} - \sum_{k=1}^{n} [kx] \\ &= (n + 1) \sum_{k=1}^{n} \frac{[kx]}{k} - \sum_{k=1}^{n} [(n - k + 1)x]. \end{aligned}$$

Returning to the induction hypothesis, we obtain

$$(n + 1) \sum_{k=1}^{n} \frac{[kx]}{k} \leq \sum_{k=1}^{n} ([kx] + [(n - k + 1)x]) \leq \sum_{k=1}^{n} [(n + 1)x] = n[(n + 1)x].$$

Dividing and rearranging again, the inequality follows for $n + 1$. The induction is complete, and the inequality follows.

[16] This was a problem in the USA Mathematical Olympiad, 1981.

Example 9.6.5 Derive the **Hermite identity**:[17]

$$[nx] = \sum_{k=0}^{n-1} \left[x + \frac{k}{n} \right], \quad n \in \mathbb{N}, \ x \in \mathbb{R}.$$

Let $[x] = m \in \mathbb{Z}$. By definition, we have $m \leq x < m + 1$, so that

$$nm \leq nx < nm + n.$$

Hence, there exists a unique integer $0 \leq j < n$ such that $[nx] = nm + j$. Equivalently

$$m + \frac{j}{n} \leq x < m + \frac{j+1}{n}.$$

We now introduce the integer variable $0 \leq k \leq n - 1, k \in \mathbb{N}_0$.

First, for $0 \leq k < n - j$, we have

$$m + \frac{j+k}{n} \leq x + \frac{k}{n} < m + \frac{j+k+1}{n}.$$

Since $(j + k)/n < 1$, this gives

$$m \leq x + \frac{k}{n} < m + 1,$$

or equivalently,

$$\left[x + \frac{k}{n} \right] = m = [x].$$

Second, for $n - j \leq k < n$, we have

$$m + \frac{j+k}{n} \leq x + \frac{k}{n} < m + \frac{j+k+1}{n}.$$

This gives

$$m + 1 \leq x + \frac{k}{n} < m + 2,$$

or equivalently,

$$\left[x + \frac{k}{n} \right] = m + 1 = [x] + 1.$$

[17] Due to the French mathematician Charles Hermite (1822–1911).

We now calculate

$$\sum_{k=0}^{n-1}\left[x+\frac{k}{n}\right]=\sum_{k=0}^{n-j-1}\left[x+\frac{k}{n}\right]+\sum_{k=n-j}^{n-1}\left[x+\frac{k}{n}\right]$$
$$=(n-j)[x]+j([x]+1)=n[x]+j=nm+j=[nx].$$

The Hermite identity follows.

Exercises

9.6.1. Find all natural numbers $n \in \mathbb{N}$ such that $[n/2]+[n/3]+[n/6]=n$.
9.6.2. Solve for $x \in \mathbb{R}$:

$$\left[\sqrt{\left[\sqrt{[x]}\right]}\right]=1.$$

Chapter 10
Exponential and Logarithmic Functions

"A Scottish baron has started up, his name
I cannot remember,[1] *but he has put forth*
some wonderful mode by which all necessity
of multiplications and divisions are commuted
to mere additions and subtractions."
Johannes Kepler, from a letter to Wilhelm Schickard,[2]
upon having seen a copy of Napier's
Mirifici Logarithmorum Canonis Descriptio
(Description of the Admirable Cannon of Logarithms).

Exponential and logarithmic functions (and in general all transcendental functions) can be analyzed by developing inequalities that compare them with polynomial and rational functions. This method lies in the heart of calculus as advocated by Euler, Newton, Leibniz, the Bernoulli brothers, Taylor, and others.

The most prominent applications of these inequalities are the existence and convexity properties of the exponential and logarithmic functions. We present here the two principal approaches, Newton's and Euler's, with full details. We use the method of means (Section 3.2) to derive the power series expansion of the natural exponential function without calculus. An optional section derives explicit formulas for all power sums (introduced in Section 3.2) in terms of the Bernoulli numbers. This chapter is concluded by presenting sharp estimates on the sum of reciprocals of the first n natural numbers, and a large variety of sophisticated but lesser-known limits involving natural exponents and logarithms.

10.1 The Natural Exponential Function According to Newton

In Section 3.2 we defined the power a^r for a real base $0 < a \in \mathbb{R}$ and real exponent $r \in \mathbb{R}$. We now study the resulting exponential function $y = a^x$ with domain variable $x \in \mathbb{R}$.

[1] John Napier of Merchiston (1550–1617).

[2] In 1617, the year of Napier's death.

G. Toth, *Elements of Mathematics*, Undergraduate Texts in Mathematics,
https://doi.org/10.1007/978-3-030-75051-0_10

In this chapter we begin to pursue Newton's circuitous path to real exponentiation by introducing first the natural exponential function $y = e^x$. We will follow this in later sections by taking the inverse, the natural logarithm $y = \ln(x)$, and, finally, the general exponential function $y = a^x$ with an arbitrary (positive) base a along with its inverse, $y = \log_a(x)$.

Recall from Example 7.1.1 the polynomials

$$e_n(x) = 1 + \frac{x}{1!} + \frac{x^2}{2!} + \cdots + \frac{x^n}{n!}, \quad x \in \mathbb{R},\ n \in \mathbb{N},$$

with $e_0(x) = 1$. Clearly, $e_n(x)$ has degree n (in the indeterminate $x \in \mathbb{R}$), and (rapidly decreasing) positive leading coefficient $1/n!$.

We first **assume** $x > 0$. Since

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

we have

$$e_n(x) - e_{n-1}(x) = \frac{x^n}{n!} > 0, \quad n \in \mathbb{N}.$$

Thus, the sequence $(e_1(x), e_2(x), \ldots, e_n(x), \ldots)$ is **strictly increasing**.

Keeping $x > 0$ fixed, we are interested in the growth rate of the leading term $x^n/n!$ of $e_n(x)$ as $n \to \infty$.

Let $m \in \mathbb{N}$ be a natural number such that $x < m$. For $n \geq m$, we have

$$n! = (m-1)! \cdot \overbrace{m(m+1)(m+2)\cdots n}^{n-m+1 \text{ factors}} \geq (m-1)! \cdot m^{n-m+1},$$

where in the middle product we replaced each factor $m+1, m+2, \ldots, n$ by m.

Using this, for $n \geq m$, we estimate

$$\frac{x^n}{n!} \leq \frac{x^n}{(m-1)! \cdot m^{n-m+1}} = \frac{m^{m-1}}{(m-1)!}\frac{x^n}{m^n} = \frac{m^{m-1}}{(m-1)!}\left(\frac{x}{m}\right)^n, \quad 0 < x < m.$$

We see that, for $n \geq m$, up to the constant multiple $m^{m-1}/(m-1)!$, the final upper estimate is the general member of the geometric sequence with quotient $0 < x/m < 1$. Adding up, for $n \geq m$, we arrive at the estimate

$$\begin{aligned} e_n(x) &= e_{m-1}(x) + \frac{x^m}{m!} + \frac{x^{m+1}}{(m+1)!} + \cdots + \frac{x^n}{n!} \\ &\leq e_{m-1}(x) + \frac{m^{m-1}}{(m-1)!}\left(\left(\frac{x}{m}\right)^m + \left(\frac{x}{m}\right)^{m+1} + \cdots + \left(\frac{x}{m}\right)^n\right) \\ &= e_{m-1}(x) + \frac{x^m}{m!}\left(1 + \frac{x}{m} + \left(\frac{x}{m}\right)^2 + \cdots + \left(\frac{x}{m}\right)^{n-m}\right), \quad 0 < x < m. \end{aligned}$$

In the last parentheses we have a finite geometric series with quotient $0 < x/m < 1$. Replacing it with the infinite geometric series, and applying the Infinite Geometric Series Formula, we obtain

$$\begin{aligned} e_n(x) &\le e_{m-1}(x) + \frac{x^m}{m!}\left(1 + \frac{x}{m} + \left(\frac{x}{m}\right)^2 + \cdots\right) \\ &= e_{m-1}(x) + \frac{x^m}{m!}\frac{1}{1 - x/m} = e_{m-1}(x) + \frac{x^m}{(m-1)!}\frac{1}{m-x}, \quad 0 < x < m. \end{aligned}$$

Since the upper bound is independent from $n \ge m$, we conclude that the sequence $(e_1(x), e_2(x), \ldots, e_n(x), \ldots)$ is **bounded above**. Since this sequence is strictly increasing, by the Monotone Convergence Theorem, the limit $\lim_{n\to\infty} e_n(x)$ exists. We denote this limit by

$$\exp(x) = \sum_{n=0}^{\infty} \frac{x^n}{n!} = 1 + \frac{x}{1!} + \frac{x^2}{2!} + \cdots + \frac{x^n}{n!} + \cdots, \quad x > 0.$$

Note that $\exp(x) > 1$, $x > 0$.

History
We will see below that this is the expansion of the natural exponential function $y = e^x$ into an infinite series. This approach is due to Newton in his *De analysi per aequationes numero terminorum infinitas* written in 1665. The notation $\exp(x)$ for e^x is widespread especially for in-line formulas with complex arguments x, and in generalizations of the exponential function in more general settings.

For future applications, we record here that, as a byproduct of our previous computations, we have the following lower and upper estimates

$$e_m(x) < \exp(x) \le e_{m-1}(x) + \frac{x^m}{(m-1)!}\frac{1}{m-x}, \quad 0 < x < m, \ m \in \mathbb{N}.$$

Now that $\exp(x)$ is defined for all $x > 0$ we claim that the following fundamental property holds:

$$\exp(x + y) = \exp(x) \cdot \exp(y), \quad x, y > 0.$$

To show this, we consider the general term of the series $\exp(x + y)$ (on the left-hand side):

$$\frac{(x+y)^n}{n!}.$$

We expand this using the general Binomial Formula (as in Section 6.3). We obtain

$$\frac{(x+y)^n}{n!} = \frac{\binom{n}{0}x^n + \binom{n}{1}x^{n-1}y + \cdots + \binom{n}{k}x^{n-k}y^k + \cdots + \binom{n}{n-1}xy^{n-1} + \binom{n}{n}y^n}{n!},$$

with the binomial coefficients

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}, \quad k = 0, 1, 2, \ldots, n.$$

Note that we have $\binom{n}{0} = \binom{n}{n} = 1$, $\binom{n}{1} = \binom{n}{n-1} = n$, etc., but we kept the binomial coefficients for uniformity. The general term in this binomial expansion is

$$\frac{1}{n!}\binom{n}{k}x^{n-k}y^k = \frac{1}{n!}\frac{n!}{k!(n-k)!}x^{n-k}y^k = \frac{x^{n-k}}{(n-k)!}\frac{y^k}{k!}, \quad k = 0, 1, 2, \ldots, n.$$

Substituting this into our binomial expansion, we obtain

$$\frac{(x+y)^n}{n!} = \frac{x^n}{n!} + \frac{x^{n-1}}{(n-1)!}\frac{y}{1!} + \cdots + \frac{x^{n-k}}{(n-k)!}\frac{y^k}{k!} + \cdots + \frac{x}{1!}\frac{y^{n-1}}{(n-1)!} + \frac{y^n}{n!}.$$

The right-hand side here patterns precisely the **Cauchy Product Rule** for the degree n term in the polynomial product $e_n(x) \cdot e_n(y)$. (See Section 6.2.) Since in our original infinite series n is unbounded, the fundamental property follows.

We now relax the condition on positivity of the indeterminate x. In fact, our definition of $\exp(x)$ immediately implies that $\exp(0) = 1$, and, for consistency of the fundamental property we just derived, for $x < 0$, we must define

$$\exp(x) = \frac{1}{\exp(-x)}.$$

Note that this implies that $0 < \exp(x) < 1$ for $x < 0$.

A quick check of the previous computation leading to the fundamental property shows that we have not used any sign restrictions on the indeterminates. Therefore, in general, we have

$$\exp(x + y) = \exp(x) \cdot \exp(y), \quad x, y \in \mathbb{R}.$$

(In particular, we may also keep the original definition of $e_n(x)$ as a degree n polynomial for all negative values of x.)
For $m = 1$ ($e_1(x) = x + 1$), our upper and lower estimates above give

$$(0 \leq)x \leq \exp(x) - 1 \leq \frac{x}{1-x}, \quad 0 \leq x < 1.$$

In particular, for any real **null-sequence** $(r_n)_{n\in\mathbb{N}}$ with $0 \leq r_n \in \mathbb{R}$, $n \in \mathbb{N}$, we have

$$0 = \lim_{n\to\infty} r_n \leq \lim_{n\to\infty} (\exp(r_r) - 1) \leq \lim_{n\to\infty} \frac{r_n}{1-r_n} = 0.$$

Thus, we obtain

$$\lim_{n\to\infty} \exp(r_n) = 1.$$

Since $\exp(-x) = 1/\exp(x)$, this holds for **any** real null-sequence $(r_n)_{n\in\mathbb{N}}$.

Finally, if $(r_n)_{n\in\mathbb{N}}$ is any convergent real sequence with $\lim_{n\to\infty} r_n = r \in \mathbb{R}$, then we have

$$\lim_{n\to\infty} \exp(r_n) = \lim_{n\to\infty} \exp((r_n - r) + r) = \exp(r) \lim_{n\to\infty} \exp(r_n - r) = \exp(r).$$

According to the corollary to Proposition 4.1.1, this proves **continuity** of the function $\exp : \mathbb{R} \to \mathbb{R}$.

We define the **natural exponential base** as $e = \exp(1)$; that is, we set

$$e = 1 + \frac{1}{1!} + \frac{1}{2!} + \cdots + \frac{1}{n!} + \cdots$$

Using our estimates for $\exp(x)$, for $m = 2$, we have

$$e_2(x) = 1 + x + \frac{x^2}{2} < \exp(x) \le e_1(x) + \frac{x^2}{1!}\frac{1}{2-x} = 1 + x + \frac{x^2}{2-x}$$
$$= \frac{2+x}{2-x}, \quad 0 < x < 2.$$

Substituting $x = 1$, we obtain $5/2 < e < 3$. Refining our estimates, in the next step, for $m = 3$, we have

$$1 + \frac{1}{1!} + \frac{1}{2!} + \frac{1}{3!} < e < 1 + \frac{1}{1!} + \frac{1}{2!} + \frac{1^3}{2!}\frac{1}{3-1}.$$

This gives $8/3 < e < 11/4$. Continuing this way, approximations of e up to any number of digits can be obtained; here are the first fifty:

$$2.71828182845904523536028747135266249775724709369 99\ldots$$

History
In 1873 Hermite proved that e is a **transcendental number**; that is, e is not a root of any polynomial with **rational** coefficients. The weaker statement of irrationality of e is much simpler and can be proved using basic calculus.[3] Hermite's proof was considerably simplified by Hilbert in 1902.

Using the fundamental property of $\exp(x)$ repeatedly, for $n \in \mathbb{N}$, we obtain

$$\exp(n) = \exp(\overbrace{1 + 1 + \cdots + 1}^{n}) = \exp(1) \cdot \exp(1) \cdots \exp(1) = e \cdot e \cdots e = e^n,$$

where each factor is repeated n times. Moreover, we have $\exp(-n) = 1/\exp(n) = 1/e^n = e^{-n}$, $n \in \mathbb{N}$. Thus, for all **integer** values, we have $e^n = \exp(n)$, $n \in \mathbb{Z}$.

[3]For two different proofs, see the author's *Glimpses of Algebra and Geometry*, 2nd ed. Springer, New York, 2002.

We claim that this formula extends to all rational numbers. We first calculate

$$e = \exp(1) = \exp\left(n \cdot \frac{1}{n}\right) = \exp\left(\overbrace{\frac{1}{n} + \frac{1}{n} + \cdots + \frac{1}{n}}^{n}\right) = \exp\left(\frac{1}{n}\right)^n.$$

Since $\exp(1/n) > 0$, this means that $\exp(1/n) = e^{1/n} = \sqrt[n]{e}$. Finally, for $m \in \mathbb{N}$, we have

$$\exp\left(\frac{m}{n}\right) = \exp\left(\overbrace{\frac{1}{n} + \cdots + \frac{1}{n}}^{m}\right) = \exp\left(\frac{1}{n}\right)^m = e^{\frac{m}{n}} = (\sqrt[n]{e})^m.$$

Extending this to negative fractions m/n in a straightforward way, we obtain

$$e^q = \exp(q), \quad q \in \mathbb{Q}.$$

Recall now from Section 3.2 sequential continuity of the exponentiation; that is, for any convergent rational sequence $(q_n)_{n\in\mathbb{N}}$ with $\lim_{n\to\infty} q_n = r$, we have

$$\lim_{n\to\infty} e^{q_n} = e^r.$$

Since exp is also sequentially continuous, and e^x and $\exp(x)$ are equal for $x \in \mathbb{Q}$, we obtain that, for **any real** number x, we have

$$e^x = \exp(x), \quad x \in \mathbb{R}.$$

With this, the fundamental relation takes the familiar form

$$e^{x+y} = e^x \cdot e^y, \quad x, y \in \mathbb{R}.$$

We will use the notation $\exp : \mathbb{R} \to \mathbb{R}$ for the function $y = e^x$, $x \in \mathbb{R}$, and call it the **natural exponential function**.

A few properties of the natural exponential function exp are obvious. Its domain is the set of all real numbers $\mathbb{R}$, it is strictly increasing, and its range is $(0, \infty)$, the set of all positive real numbers. Since $\lim_{x\to\infty} e^x = \infty$, we have $\lim_{x\to-\infty} e^x = \lim_{x\to\infty} e^{-x} = 0$, so that the negative first axis is a horizontal asymptote.

Some of the analytical properties of the natural exponential function follow directly from the definition. For $x \geq 0$, we automatically have

$$e_n(x) = 1 + \frac{x}{1!} + \frac{x^2}{2!} + \cdots + \frac{x^n}{n!} \leq e^x, \quad n \in \mathbb{N}.$$

More explicitly, we have the **lower estimates**

$$1 + x \leq e^x, \quad 1 + x + \frac{x^2}{2} \leq e^x, \quad 1 + x + \frac{x^2}{2} + \frac{x^3}{6} \leq e^x, \quad x \geq 0, \quad \text{etc.}$$

The lower estimates above show, in particular, that there is no polynomial upper estimate for e^x valid for all $x \geq 0$. Indeed, by the above, we have

$$\frac{x^{n+1}}{(n+1)!} \leq e^x, \quad x \geq 0,$$

so that

$$\lim_{x\to\infty} \frac{e^x}{x^n} = \infty.$$

On the other hand, for a **bounded** range of the variable, we derived the **rational upper estimates**

$$e^x \leq 1 + \frac{x}{1!} + \frac{x^2}{2!} + \cdots + \frac{x^{n-1}}{(n-1)!} + \frac{x^n}{(n-1)!}\frac{1}{n-x}, \quad 0 < x < n, \ \ n \in \mathbb{N}.$$

More explicitly, we have the **upper estimates**

$$e^x \leq 1 + \frac{x}{1-x} = \frac{1}{1-x}, \ \ 0 < x < 1,$$
$$e^x \leq 1 + x + \frac{x^2}{2-x} = \frac{2+x}{2-x}, \ \ 0 < x < 2,$$
$$e^x \leq 1 + x + \frac{x^2}{2} + \frac{x^3}{2(3-x)} = \frac{6+4x+x^2}{2(3-x)}, \ \ 0 < x < 3, \ \text{etc.}$$

Replacing x by $-x$ in the **lower estimates**, and taking reciprocals, we obtain

$$e^x \leq \frac{1}{1-x}, \ \ e^x \leq \frac{1}{1-x+x^2/2}, \ \ e^x \leq \frac{1}{1-x+x^2/2-x^3/6}, \ \ x \leq 0, \quad \text{etc.}$$

With the first upper estimate above, we arrive at

$$e^x \leq \frac{1}{1-x}, \ \ x < 1.$$

For the corresponding lower estimate, we have

$$1 + x \leq e^x, \ \ x \in \mathbb{R}.$$

Indeed, for $x \geq 0$, this is the first lower estimate; for $-1 < x < 0$, this is the consequence of the first upper estimate (with x replaced by $-x$); and, for $x \leq -1$, this is automatic since $1 + x \leq 0$.

We combine these two estimates to arrive at the **fundamental estimate** of the natural exponential function

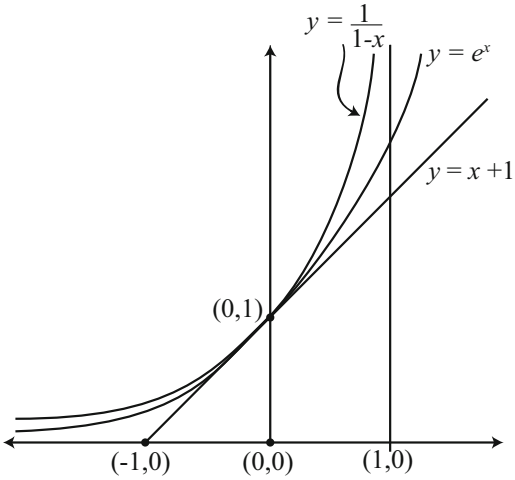


Fig. 10.1 Fundamental estimate of the natural exponential function.

$$x \le e^x - 1 \le \frac{x}{1-x}, \quad x < 1.$$

(See Figure 10.1.)

As an illustration, we now discuss the following example due to Jacob Steiner (1796–1863).

Example 10.1.1 For $x > 0$, the expression $\sqrt[x]{x}$ takes its maximum at $x = e$.

To show this we apply the previous lower bound for the natural exponential function for the number $(x - e)/e$. We have

$$\frac{x}{e} = 1 + \frac{x-e}{e} \le e^{(x-e)/e} = \frac{e^{x/e}}{e}$$

with equality if and only if $x = e$. After canceling e, we obtain $x \le e^{x/e}$. Raising both sides to the $1/x$ power, we have

$$\sqrt[x]{x} = x^{1/x} \le \left(e^{x/e}\right)^{1/x} = e^{1/e} = \sqrt[e]{e}.$$

The example follows.

Returning to the main line, we now claim that the derivative of exp at 0 is

$$\exp'(0) = \lim_{x \to 0} \frac{e^x - 1}{x} = 1.$$

Indeed, for $0 < x < 1$, the fundamental estimate above gives

$$1 \le \frac{e^x - 1}{x} \le \frac{1}{1-x}, \quad 0 < x < 1.$$

This gives the estimate for the right-limit:

$$1 \le \lim_{x\to 0^+} \frac{e^x - 1}{x} \le \lim_{x\to 0^+} \frac{1}{1-x} = 1,$$

and we arrive at the right-derivative

$$\exp'_+(0) = \lim_{x\to 0^+} \frac{e^x - 1}{x} = 1.$$

Similarly, for $x < 0$ we get

$$1 \ge \frac{e^x - 1}{x} \ge \frac{1}{1-x}, \quad x < 0,$$

giving the left-derivative

$$\exp'_-(0) = \lim_{x\to 0^-} \frac{e^x - 1}{x} = 1,$$

the claim follows.

Finally, for any $c \in \mathbb{R}$, we have

$$\exp'(c) = \lim_{x\to c} \frac{e^x - e^c}{x - c} = \lim_{x\to c} \frac{e^{x-c}e^c - e^c}{x - c} = e^c \lim_{x\to c} \frac{e^{x-c} - 1}{x - c} = e^c,$$

where, in the last equality, we used the previous limit.

We obtain that the natural exponential function is **differentiable** (at any point), and we have

$$\exp'(c) = \exp(c), \quad c \in \mathbb{R}.$$

Let $0 < x \in \mathbb{R}$ and $n \in \mathbb{N}$. We wish to calculate the **mean** (see Section 3.2) of the exponential function exp corresponding to the (equidistant) subdivision $0 = x_0 < x_1 < \cdots < x_{n-1} < x_n = x$ of the interval $[0, x]$, $x_k = kx/n$, $k = 0, 1, \ldots, n$.

We have

$$\begin{aligned}
\mathcal{A}_{\exp}(n, x) &= \frac{1}{n}\sum_{k=1}^{n} e^{k\cdot x/n} = \frac{1}{n}\sum_{k=1}^{n} \left(e^{x/n}\right)^k \\
&= \frac{e^{x/n}}{n}\left(1 + e^{x/n} + \left(e^{x/n}\right)^2 + \cdots + \left(e^{x/n}\right)^{n-1}\right) \\
&= \frac{e^{x/n}}{n} \cdot \frac{\left(e^{x/n}\right)^n - 1}{e^{x/n} - 1} = (e^x - 1) \cdot \frac{e^{x/n}/n}{e^{x/n} - 1},
\end{aligned}$$

where we used the Finite Geometric Series formula. Taking the limit, we calculate

$$\mathcal{A}_{\exp}(x) = (e^x - 1) \lim_{n\to\infty} \frac{e^{x/n}/n}{e^{x/n} - 1} = \frac{e^x - 1}{x} \lim_{n\to\infty} e^{x/n} \cdot \frac{x/n}{e^{x/n} - 1}$$
$$= \frac{e^x - 1}{x} \lim_{h\to 0} \frac{h}{e^h - 1} = \frac{e^x - 1}{x} \frac{1}{\exp'(0)} = \frac{e^x - 1}{x}.$$

Remark 1 The reader versed in calculus will no doubt recognize the Riemann sum and Riemann integral above

$$\int_0^x e^t dt = \lim_{n\to\infty} \sum_{k=1}^{n} e^{k\cdot x/n} \cdot \frac{x}{n} = x \cdot \mathcal{A}_{\exp}(x) = e^x - 1.$$

Remark 2 To complete the circle, for $n \in \mathbb{N}_0$, the inequalities

$$e_n(x) = 1 + \frac{x}{1!} + \frac{x^2}{2!} + \cdots + \frac{x^n}{n!} \le e^x, \quad 0 \le x \in \mathbb{R}$$

can be **derived** by induction from the obvious ($n = 0$) inequality $1 \le e^x$, $0 \le x \in \mathbb{R}$, by repeated application of the **mean** above. For the general induction step $n \Rightarrow n + 1$, we assume $e_n(x) \le e^x$, $0 \le x \in \mathbb{R}$, use linearity and monotonicity of the mean, and calculate

$$\mathcal{A}_{e_n}(x) = \sum_{k=0}^{n} \frac{\mathcal{A}_{\mathfrak{p}_k}(x)}{k!} = \sum_{k=0}^{n} \frac{x^k}{k! \cdot (k+1)} = \sum_{k=0}^{n} \frac{x^k}{(k+1)!} \le \mathcal{A}_{\exp}(x) = \frac{e^x - 1}{x}.$$

We used here $\mathcal{A}_{\mathfrak{p}_k}(x) = x^k/(k+1)$, $k \in \mathbb{N}_0$, as was shown in Section 3.2. Rearranging, we obtain $e_{n+1}(x) \le e^x$, $0 \le x \in \mathbb{R}$. The induction is complete, and the claim follows.

Returning to the main line, the calculation above for the mean of exp can be repeated almost verbatim for the reciprocal $1/\exp$ by replacing $0 < x \in \mathbb{R}$ with the opposite $-x < 0$ as follows

$$\mathcal{A}_{1/\exp}(n, x) = \frac{1 - e^{-x}}{x} \cdot \frac{x/n}{e^{x/n} - 1}, \quad 0 < x \in \mathbb{R},$$

and the limit

$$\mathcal{A}_{1/\exp}(x) = \frac{1 - e^{-x}}{x}, \quad 0 < x \in \mathbb{R}.$$

We now claim that the following estimate holds

$$1 - \frac{x}{1!} + \frac{x^2}{2!} - \cdots - \frac{x^{2n-1}}{(2n-1)!} < e^{-x} < 1 - \frac{x}{1!} + \frac{x^2}{2!} - \cdots + \frac{x^{2n}}{(2n)!}, \quad 0 < x \in \mathbb{R}.$$

The proof is by induction carried out by repeated application of the mean (compare with Remark 2 above), using its linearity and monotonicity for the respective functions over the interval $[0, x]$, $0 < x \in \mathbb{R}$, and finally, using the formula $\mathcal{A}_{\mathfrak{p}_k}(x) = x^k/(k+1)$, $0 < x \in \mathbb{R}$, for the mean of the power function $\mathfrak{p}_k(x) = x^k$, $x \in \mathbb{R}$, derived at the end of Section 3.2.

We begin with the obvious inequality

$$e^{-x} < 1, \quad 0 < x \in \mathbb{R}.$$

Applying the mean of both sides, we obtain

$$\frac{1 - e^{-x}}{x} < 1,$$

or equivalently $1 - x < e^{-x}$, $0 < x \in \mathbb{R}$. Applying the mean of both sides again, we obtain

$$1 - \frac{x}{2} < \frac{1 - e^{-x}}{x},$$

or equivalently

$$e^{-x} < 1 - x + \frac{x^2}{2}, \quad 0 < x \in \mathbb{R}.$$

These complete the initial step in the induction.

To perform the general induction step $n \Rightarrow n + 1$, we assume that the chain of inequalities as above hold. We take the mean of all functions as follows

$$1 - \frac{x}{2!} + \frac{x^2}{3!} - \cdots - \frac{x^{2n-1}}{(2n)!} < \frac{1 - e^{-x}}{x} < 1 - \frac{x}{2!} + \frac{x^2}{3!} - \cdots + \frac{x^{2n}}{(2n+1)!}.$$

Rearranging, we obtain the first of the chain of inequalities for $n + 1$. Repeating this, the second inequality also follows. The induction is complete and the formula follows.

The chain of inequalities just derived gives

$$\left| e^{-x} - \left(1 - \frac{x}{1!} + \frac{x^2}{2!} - \cdots - \frac{x^{2n-1}}{(2n-1)!}\right) \right| < \frac{x^{2n}}{(2n)!}, \quad 0 < x \in \mathbb{R}.$$

Since

$$\lim_{n \to \infty} \frac{x^n}{n!} = 0, \quad x \in \mathbb{R},$$

this shows that, for all $0 < x \in \mathbb{R}$, we have

$$e^{-x} = \sum_{n=0}^{\infty} (-1)^n \frac{x^n}{n!}.$$

Combining this with our previous expansion for $0 < x \in \mathbb{R}$, we obtain

$$e^x = \sum_{n=0}^{\infty} \frac{x^n}{n!}, \quad x \in \mathbb{R}.$$

Example 10.1.2 [4] Find a rational number that approximates $1/\sqrt{e}$ up to 10 decimal precision.

Since $1/\sqrt{e} = e^{-1/2}$, by the estimate above, we need to find $n \in \mathbb{N}$ such that

$$\frac{(1/2)^{2n}}{(2n)!} \leq 10^{-10},$$

or equivalently $10^{10} \leq 2^{2n} \cdot (2n)!$. Simple computation shows that $n = 6$ is the minimal value:

$$10,000,000,000 < 2^{12} \cdot (12)! = 1,961,990,553,600.$$

The approximating rational numbers is

$$\sum_{n=0}^{11} (-1)^n \frac{(1/2)^n}{n!} = \frac{49583642701}{81749606400}$$
$$= 0.606530659712142662989017155683822350489\overline{0}.$$

Exercises

10.1.1. Derive the estimate

$$e_n(1) < e < e_n(1) + \frac{1}{n \cdot n!}, \quad n \in \mathbb{N}.$$

Use this to obtain approximations of e for $n = 1, 2, 3, 4$.

10.1.2. Derive the inequality

$$e^{nx} + n(1 - e^x) \geq 1, \quad x \in \mathbb{R}, \ \ n \in \mathbb{Z}.$$

10.1.3. Prove the following:

$$\sum_{k=1}^{n} \frac{k}{(k+1)!} = \frac{1}{2!} + \frac{2}{3!} + \cdots + \frac{n}{(n+1)!} = 1 - \frac{1}{(n+1)!}, \quad n \in \mathbb{N}.$$

[4] This example needs a computer algebra system.

10.1.4. Let P_n, $n \in \mathbb{N}$, be the probability that a permutation on an n element set is a derangement. Show that $\lim_{n\to\infty} P_n = 1/e$.

10.2 The Bernoulli Numbers*

In this section we return to the problem of the pth power sum

$$s_p(n) = \sum_{k=1}^{n-1} k^p = 1^p + 2^p + \cdots + (n-1)^p, \quad p \in \mathbb{N}_0,\ 2 \le n \in \mathbb{N},$$

and show that it is a polynomial of degree $n+1$ (Section 3.2). The so-called **Bernoulli numbers** B_k, $k \in \mathbb{N}_0$, will appear naturally in the coefficients of this polynomial.

The main idea is to expand the exponential function into power series, and use the exponential identities along with the Finite Geometric Series Formula to obtain an expression for $s_p(n)$, $p \in \mathbb{N}_0$. This will then lead to a natural introduction to the Bernoulli numbers through a **generating function**.

We start with the power series expansions

$$e^{kx} = \sum_{p=0}^{\infty} k^p \frac{x^p}{p!} = 1 + k\frac{x}{1!} + k^2\frac{x^2}{2!} + \cdots + k^p\frac{x^p}{p!} + \cdots,\ k = 0, 1, \ldots, n-1,\ 2 \le n \in \mathbb{N}.$$

We sum up these with respect to $k = 0, 1, \ldots, n-1$ and obtain

$$\sum_{k=0}^{n-1} e^{kx} = 1 + \sum_{k=1}^{n-1}\sum_{p=0}^{\infty} k^p \frac{x^p}{p!} = 1 + \sum_{p=0}^{\infty}\left(\sum_{k=1}^{n-1} k^p\right)\frac{x^p}{p!} = 1 + \sum_{p=0}^{\infty} s_p(n)\frac{x^p}{p!}.$$

On the other hand, the exponential sum on the left-hand side can be evaluated by the Finite Geometric Series Formula as follows

$$\sum_{k=0}^{n-1} e^{kx} = \sum_{k=0}^{n-1} \left(e^x\right)^k = \frac{e^{nx}-1}{e^x-1}.$$

We write the last fraction as

$$\frac{e^{nx}-1}{e^x-1} = \frac{e^{nx}-1}{x}\frac{x}{e^x-1}.$$

The first factor on the right-hand side has the power series expansion

$$\frac{e^{nx}-1}{x} = \sum_{l=1}^{\infty} n^l \frac{x^{l-1}}{l!}.$$

Now, the crux is to expand the second fraction $x/(e^x - 1)$ on the right-hand side into a power series

$$\frac{x}{e^x - 1} = \sum_{k=0}^{\infty} B_k \frac{x^k}{k!},$$

with coefficients B_k, $k \in \mathbb{N}_0$, the Bernoulli numbers, **to be determined**.

Remark Note that in all our manipulations we use the power series formally, that is, disregarding convergence. In the previous section we concluded, however, that the singularity of the fraction $(e^x - 1)/x$ at $x = 0$ is removable with $\exp'(0) = 1$, and its power series expansion is convergent for all $x \in \mathbb{R}$. Therefore, the power series expansion of the reciprocal function $x/(e^x - 1)$ is also convergent for all $x \in \mathbb{R}$.

Putting everything together, we obtain

$$1 + \sum_{p=0}^{\infty} s_p(n) \frac{x^p}{p!} = \sum_{l=1}^{\infty} n^l \frac{x^{l-1}}{l!} \cdot \sum_{k=0}^{\infty} B_k \frac{x^k}{k!}.$$

We now compare coefficients. The constant terms ($p = 0, l = 1, k = 0$) give $1 + s_0(n) = 1 + (n - 1) = n = nB_0$, that is, we have $B_0 = 1$.

For $p \in \mathbb{N}$, the coefficients of the pth power ($p = l + k - 1$) give

$$\frac{s_p(n)}{p!} = \sum_{k=0}^{p} \frac{B_k}{(p - k + 1)!k!} n^{p-k+1}.$$

Multiplying through by $p!$ and converting the factorials to binomial coefficients, we obtain

$$s_p(n) = \frac{1}{p+1} \sum_{k=0}^{p} \binom{p+1}{k} B_k n^{p-k+1}, \quad p \in \mathbb{N}.$$

This proves that the power sum $s_p(n) = 1^p + 2^2 + \cdots + (n-1)^p$ is a **polynomial of degree** $p + 1$.

To obtain an inductive formula for the Bernoulli numbers we return to their definition as the coefficients in the power series expansion of the fraction $x/(e^x - 1)$. Multiplying out by the denominator, we have

$$x = \sum_{l=1}^{\infty} \frac{x^l}{l!} \sum_{k=0}^{\infty} B_k \frac{x^k}{k!}.$$

The coefficients of the linear term once again give $B_0 = 1$. For $m \in \mathbb{N}$, the coefficients of the x^{m+1} term on the right-hand side are obtained by setting $l + k = m + 1$, $k = 0, 1, \ldots, m$, and multiplying the respective terms of the two sums. We obtain

$$\sum_{k=0}^{m} \frac{B_k}{k!(m-k+1)!} = 0.$$

Multiplying through by $(m+1)!$ allows to convert the factorials into binomials. We write the resulting equality as

$$B_m = -\frac{1}{m+1} \sum_{k=0}^{m-1} \binom{m+1}{k} B_k, \quad m \in \mathbb{N}.$$

Starting with $B_0 = 1$, this equation determines the entire sequence $(B_k)_{k\in\mathbb{N}_0}$ inductively.

Note that, as a byproduct, it follows that all Bernoulli numbers are **rational**.

Another simple fact is that, with the exception of $B_1 = -1/2$, the odd Bernoulli numbers B_{2k+1} are zero for $k \in \mathbb{N}$. Indeed, this follows from the fact that the function $x/(e^x - 1) + x/2$ is even:

$$\frac{-x}{e^{-x}-1} - \frac{x}{2} = \frac{xe^x}{e^x-1} - \frac{x}{2} = \frac{x}{e^x-1} + \frac{x}{2}.$$

The first few Bernoulli numbers are tabulated as follows:

k	B_k	k	B_k
0	1	12	$-691/2730$
1	$-1/2$	14	$7/6$
2	$1/6$	16	$-3617/510$
4	$-1/30$	18	$43867/798$
6	$1/42$	20	$-174611/330$
8	$-1/30$	22	$854513/138$
10	$5/66$	24	$-236364091/2730$

Calculating the respective binomial coefficients, these give

$$\begin{aligned}
s_1(n) &= \frac{1}{2}n^2 - \frac{1}{2}n \\
s_2(n) &= \frac{1}{3}n^3 - \frac{1}{2}n^2 + \frac{1}{6}n \\
s_3(n) &= \frac{1}{4}n^4 - \frac{1}{2}n^3 + \frac{1}{4}n^2 \\
s_4(n) &= \frac{1}{5}n^5 - \frac{1}{2}n^4 + \frac{1}{3}n^3 - \frac{1}{30}n \\
s_5(n) &= \frac{1}{6}n^6 - \frac{1}{2}n^5 + \frac{5}{12}n^4 + -\frac{1}{12}n^2
\end{aligned}$$

$$s_6(n) = \frac{1}{7}n^7 - \frac{1}{2}n^6 + \frac{1}{2}n^5 - \frac{1}{6}n^3 + \frac{1}{42}n$$
$$s_7(n) = \frac{1}{8}n^8 - \frac{1}{2}n^7 + \frac{7}{12}n^6 - \frac{7}{24}n^4 + \frac{1}{12}n^2$$
$$s_8(n) = \frac{1}{9}n^9 - \frac{1}{2}n^8 + \frac{2}{3}n^7 - \frac{7}{15}n^5 + \frac{2}{9}n^3 - \frac{1}{30}n$$
$$s_9(n) = \frac{1}{10}n^{10} - \frac{1}{2}n^9 + \frac{3}{4}n^8 - \frac{7}{10}n^6 + \frac{1}{2}n^4 - \frac{3}{20}n^2$$
$$s_{10}(n) = \frac{1}{11}n^{11} - \frac{1}{2}n^{10} + \frac{5}{6}n^9 - n^7 + n^5 - \frac{1}{2}n^3 + \frac{5}{66}n$$
$$s_{11}(n) = \frac{1}{12}n^{12} - \frac{1}{2}n^{11} + \frac{11}{12}n^{10} - \frac{11}{8}n^8 + \frac{11}{6}n^6 - \frac{11}{8}n^4 + \frac{5}{12}n^2$$
$$s_{12}(n) = \frac{1}{13}n^{13} - \frac{1}{2}n^{12} + n^{11} - \frac{11}{6}n^9 + \frac{22}{7}n^7 - \frac{33}{10}n^5 + \frac{5}{3}n^3 - \frac{691}{2730}n.$$

History
Most likely it was the English mathematician and astronomer Thomas Harriot (1560–1621) who first developed symbolic formulas for sums of powers, but he did so only up to the fourth powers. In his *Academia Algebrae* published in 1631, the German mathematician Johann Faulhaber (1580–1635) derived these formulas up to the seventeenth power but he did not obtain a general pattern. Finally, Jakob Bernoulli realized that a uniform formula can be obtained by introducing a single sequence of numbers $(B_k)_{k\in\mathbb{N}_0}$, and the latter therefore was named after him. We quote here his well-known comment upon the moment of discovery as follows: "With the help of this table, it took me less than half of a quarter of an hour to find that the tenth powers of the first 1000 numbers being added together[5] will yield the sum 91, 409, 924, 241, 424, 243, 424, 241, 924, 242, 500."

Exercise

10.2.1. Define the **Bernoulli polynomials** $B_n(y)$, $n \in \mathbb{N}$, by

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

Use the Cauchy product rule to derive the formula

$$\frac{xe^{xy}}{e^x - 1} = \sum_{k=0}^{\infty} \frac{B_k(y)x^k}{k!}.$$

Show the following: (a) $B_0(y) = 1$; for $n \in \mathbb{N}$ (b) $B_n(0) = B_n$; (c) $B_n'(y) = nB_{n-1}'(y)$.

[5]This is our $s_{10}(1001)$.

10.3 The Natural Logarithm

By the results of Section 10.1, the natural exponential function $\exp : \mathbb{R} \to \mathbb{R}$ is strictly increasing onto its range $(0, \infty)$. Therefore its inverse, the **natural logarithm function** $\ln : (0, \infty) \to \mathbb{R}$ is well-defined, strictly increasing, and has range $\mathbb{R}$. In addition, the negative second axis is the vertical asymptote of the graph. Clearly, we have $\lim_{x \to 0^+} \ln(x) = -\infty$ and $\lim_{x \to \infty} \ln(x) = \infty$.

By the definition of the inverse, we have

$$e^{\ln(x)} = x, \quad x > 0,$$

and

$$\ln(e^x) = x, \quad x \in \mathbb{R}.$$

In particular, we have

$$\ln(1) = \ln(e^0) = 0 \quad \text{and} \quad \ln(e) = \ln(e^1) = 1.$$

By definition, both the natural exponential and the natural logarithm functions are one-to-one; that is, they satisfy the property: $e^x = e^y$ if and only if $x = y$, and $\ln(x) = \ln(y)$ if and only if $x = y$.

History

In 1899 the British physicist Ernest Rutherford (1871–1937) discovered that thorium, a naturally occurring radioactive chemical element, while spontaneously emanating a radioactive gas, decays into **half** of its size in the same fixed time, the so-called **half-life** $\tau (\approx 11.5$ minutes), **regardless the original amount**.

If $Q(t)$ is the amount of thorium at time $t \geq 0$, with $Q_0 = Q(0)$, the original amount, then this observation gives $Q(\tau) = Q_0/2$, $Q(2\tau) = Q_0/4$, $Q(3\tau) = Q_0/8$, etc. By a simple induction, we thus have $Q(n\tau) = Q_0/2^n = Q_0 \cdot 2^{-n}$, $n \in \mathbb{N}$. Changing to a real variable $t \geq 0$ (with the discrete values corresponding to $t = n\tau$), we obtain $Q(t) = Q_0 \cdot 2^{-t/\tau} = Q_0 \cdot e^{-t \cdot \ln(2)/\tau}$, $t \geq 0$. We write this as $Q(t) = Q_0 \cdot e^{-\lambda \cdot t}$, $t \geq 0$, where the half-life τ and the **exponential decay constant** λ are related by $\tau \cdot \lambda = \ln(2)$.

All living organisms, through consumption, contain non-radioactive carbon C^{12} and a tiny amount of the radioactive isotope C^{14}. The ratio of the amounts of C^{14} and C^{12} is approximately 10^{-12}. When the organism dies, C^{14} is no longer replenished and follows exponential decay while C^{12}, being non-radioactive, stays constant. The half-life of C^{14} is approximately 5, 730 years.

Measuring the ratio of the amounts of C^{14} and C^{12} in an organism dead for a long time, one can calculate the approximate time when the organism lived. This is **carbon dating**, invented by the American chemist and Nobel laureate Willard Libby (1908–1980).

As a famous example, the Tollund man, the naturally mummified corpse of an executed man buried in a Danish bog, had 75.7% of the atmospheric ratio of C^{12} and C^{14}. Carbon dating tells the approximate age of the Tollund man as follows. Let λ be the exponential decay constant of C^{14}. We have $\lambda = \ln(2)/\tau = \ln(2)/5730 = 0.00012096\ldots$. Hence, we have $0.757 = e^{-0.00012096 \cdot t}$. This finally gives $t = -\ln(0.757)/0.00012096 \approx 2300$ years. Note that, due to errors in measuring the amount of C^{14}, this calculation has an error of about ± 40 years.

The natural logarithm satisfies the following identities:

$$\ln(x \cdot y) = \ln(x) + \ln(y) \quad \text{and} \quad \ln\left(\frac{x}{y}\right) = \ln(x) - \ln(y), \quad x, y > 0.$$

Indeed, we have

$$e^{\ln(x \cdot y)} = x \cdot y = e^{\ln(x)} \cdot e^{\ln(y)} = e^{\ln(x)+\ln(y)}, \quad x, y > 0.$$

Taking the natural logarithm of both sides, the first identity follows. The proof of the second identity is similar.

Remark A simple induction gives the following extension of the first of the two identities above:

$$\ln(x^n) = \ln(\overbrace{x \cdot x \cdots x}^{n}) = \overbrace{\ln(x) + \ln(x) + \cdots + \ln(x)}^{n} = n \ln(x), \quad n \in \mathbb{N}.$$

History
Hailed by Pierre Simon Laplace (1749–1827) as an "admirable artifice which, by reducing to a few days the labour of many months, doubles the life of the astronomer," the logarithm was invented in 1614 by John Napier. (See also the epitaph of this chapter.) His "method of logarithms," and the logarithmic tables, the first of which was published three years later by Henry Briggs (1561–1630), was designed to reduce massive computations, especially in astronomy.

We now return to the main line and derive another characterization of the natural logarithm, due to Euler[6] as follows:

Example 10.3.1 Show that $\lim_{n\to\infty} n \cdot (\sqrt[n]{x} - 1) = \ln x, 0 < x \in \mathbb{R}$.

We may assume $x \neq 1$, since otherwise both sides of the equality are zero. The crux is to rewrite the limit in terms of the new variable $h = \ln(x)/n$ as follows:

$$\frac{1}{\ln x} \lim_{n\to\infty} n \cdot (\sqrt[n]{x} - 1) = \lim_{n\to\infty} \frac{e^{\ln x/n} - 1}{\ln x/n} = \lim_{h\to 0} \frac{e^h - 1}{h} = \exp'(0) = e^0 = 1.$$

The example follows.

In Section 10.1 we showed that the derivative of the natural exponential function exp at c is equal to $\exp'(c) = \exp(c) = e^c$. The derivative is the slope of the tangent line to the graph $G(\exp)$ at (c, e^c). Now, the graph of the inverse, the natural logarithm function ln, is obtained by reflecting the graph $G(\exp)$ to the line given by $y = x$. Upon reflection, the first and second coordinates interchange, and tangent lines of one graph map to tangent lines of the other. In particular, the slope of the reflected tangent line is the **reciprocal** of the slope of the original tangent line. We see that the slope of the (reflected) tangent line to the graph $G(\ln)$ at (e^c, c) is

[6] See also History in Section 10.5.

$1/e^c$. Reverting to the first coordinates, we obtain that the derivative of the natural logarithm function ln at c is $1/c$:

$$\ln'(c) = \frac{1}{c}, \quad 0 < c \in \mathbb{R}.$$

The next example is a generalization of Example 2.1.4.

Example 10.3.2 Let $e \leq b < c$. Show that $b^c > c^b$.

To compare $b^c \gtrless c^b$ is the same as to compare their natural logarithms $c \ln b \gtrless b \ln c$. This, in turn, amounts to compare $\ln b/b \gtrless \ln c/c$.
The crux in this example is to show that the function $f(x) = \ln x/x$, $0 < x \in \mathbb{R}$, is **strictly decreasing** for $e \leq x \in \mathbb{R}$. This will give $\ln b/b > \ln c/c$, resulting in $b^c > c^b$.
For the claimed monotonicity, we first show that f has no critical points on (e, ∞). Clearly, f is differentiable on its domain $(0, \infty)$. The derivative can be obtained by the differentiation formula for the quotient (Section 4.3) as follows

$$f'(c) = \frac{(1/c) \cdot c - \ln c}{c^2} = \frac{1 - \ln c}{c^2}, \quad 0 < c \in \mathbb{R},$$

where we used our result $\ln'(c) = 1/c$, $0 < c \in \mathbb{R}$, above. This shows that f has only one critical point at $c = e$.

As a consequence of the Fermat Principle in Section 4.3, f must be injective on (e, ∞), and, being continuous, it must be strictly monotonic. On the other hand, we have

$$\lim_{x\to\infty} f(x) = \lim_{x\to\infty} \frac{\ln x}{x} = \lim_{u\to\infty} \frac{u}{e^u} = 0.$$

It follows that f must be strictly decreasing on $[e, \infty)$. The example follows.

Remark 1 As a particular case of the example above, we have $m^n > n^m$, $3 \leq m < n$, $m, n \in \mathbb{N}$. (This is clearly equivalent to the fact that the sequence $(\sqrt[n]{n})_{n\in\mathbb{N}}$ is strictly decreasing for $3 \leq n \in \mathbb{N}$, already shown in Example 3.2.8.)
For what distinct natural numbers $m, n \in \mathbb{N}$ do we have equality[7] $m^n = n^m$? Assuming $1 \leq m < n$, by the above, this can (possibly) happen only for $m = 2$. (Clearly, $m = 1$ does not compete.) But, by Example 2.1.3, we have $2^n > n^2$, $5 \leq n \in \mathbb{N}$. This leaves us $n = 3, 4$. Since $8 = 2^3 < 3^2 = 9$, we finally end up with $n = 4$, where $2^4 = 4^2$. Summarizing, the only pair $(m, n) \in \mathbb{N} \times \mathbb{N}$, $m < n$, for which $m^n = n^m$ is $(2, 4)$.

Remark 2 As an application, and as a glimpse to integral calculus, we now calculate a (left-)Riemann sum of the function $f(x) = 1/x$, $0 \neq x \in \mathbb{R}$ over the interval $[1, a]$, $1 < a \in \mathbb{R}$. We let the subdivision $1 = x_0 < x_1 < \ldots < x_{n-1} < x_n = a$ given by $x_k = e^{k \cdot \ln a/n}$, $k = 0, \ldots, n$. We have

[7] This was also a problem (including negative integers) in the William Lowell Putnam Exam, 1960.

$$\sum_{k=1}^{n} \frac{1}{e^{(k-1)\cdot \ln a/n}} \left(e^{k\cdot \ln a/n} - e^{(k-1)\cdot \ln a/n} \right) = \sum_{k=1}^{n} \left(e^{k\cdot \ln a/n - (k-1)\cdot \ln a/n} - 1 \right)$$

$$= \sum_{k=1}^{n} \left(e^{\ln a/n} - 1 \right) = n \cdot (\sqrt[n]{a} - 1).$$

The reader versed in calculus will here recognize the limit

$$\int_1^a \frac{dx}{x} = \lim_{n\to\infty} n \cdot (\sqrt[n]{a} - 1) = \ln a, \quad 1 < a \in \mathbb{R},$$

where we used the limit in Example 10.3.1.

We now return to our estimates. Substituting $\ln(x)$ for x in our earlier lower estimate $1 + x \le e^x$ with $x \in \mathbb{R}$, and rearranging, we obtain

$$\ln(x) \le x - 1, \quad x > 0.$$

This shows that the graphs $G(\exp)$ and $G(\ln)$ are separated by a strip whose boundary consists of the tangent lines at $(1, 0)$ and $(0, 1)$ with slope 1.

The **fundamental estimate** for the natural logarithm is the following

$$\frac{x}{1+x} \le \ln(1 + x) \le x, \quad -1 < x \in \mathbb{R}.$$

The upper estimate here is just a reformulation of the upper estimate above (replacing x by $1+x$). The lower estimate follows by inverting simultaneously both sides of the previous estimate $e^x \le 1/(1-x)$, $x < 1$. This gives $\ln(x) \ge (x-1)/x$, $x > 0$. Replacing x by $x + 1$ as before, the lower estimate follows.

Remark An immediate byproduct is the limit $\lim_{x\to 1} \ln(x) = \lim_{x\to 0} \ln(1+x) = 0$. This, in turn, gives another proof of **continuity** of the natural logarithm. Indeed, let $(r_n)_{n\in\mathbb{N}}$ be a convergent positive real sequence, $0 < r_n \in \mathbb{R}$, $n \in \mathbb{N}$, with positive limit $\lim_{n\to\infty} r_n = r$, $0 < r \in \mathbb{R}$. Then we have

$$\lim_{n\to\infty} \ln(r_n) = \lim_{n\to\infty} \ln\left(\frac{r_n}{r} \cdot r\right) = \lim_{n\to\infty} \ln\left(\frac{r_n}{r}\right) + \ln(r) = \ln(r).$$

Continuity of ln follows.

For the positive range of the natural logarithm, a sharper upper bound can be obtained using the quadratic lower estimate $1 + x + x^2/2 \le e^x$ with $x \ge 0$. We substitute $\ln(x)$ for x, rearrange and obtain

$$\ln(x) + \frac{\ln(x)^2}{2} \le x - 1, \quad x \ge 1.$$

Completing the square, and rearranging we have

$$(\ln(x)+1)^2 \le 2x-1, \quad x \ge 1.$$

This gives the **sharper** upper bound $\ln(x) \le \sqrt{2x-1}-1$, $x \ge 1$, or equivalently

$$\ln(1+x) \le \sqrt{2x+1}-1, \quad x \ge 0.$$

For the next example, we let

$$e_n = e_n(1) = 1 + \frac{1}{1!} + \frac{1}{2!} + \cdots + \frac{1}{n!}, \quad e_0 = 1,$$

and recall from Section 10.1 that this is the nth partial sum of the infinite sum that defines e:

$$\lim_{n\to\infty} e_n = \lim_{n\to\infty}\left(1 + \frac{1}{1!} + \frac{1}{2!} + \cdots + \frac{1}{n!}\right) = \sum_{n=0}^{\infty} \frac{1}{n!} = e.$$

Clearly, $e_n < e$, $n \in \mathbb{N}_0$, and $\lim_{n\to\infty}(e-e_n) = 0$.

Example 10.3.3 [8] Show that

$$\lim_{n\to\infty} (1+e-e_n)^{n!} = 1 \quad \text{and} \quad \lim_{n\to\infty} (1+e-e_n)^{(n+1)!} = e.$$

We derive the first limit relation only, the second is entirely analogous. We use the fundamental estimate of the natural logarithm for $0 < e - e_n$ as

$$\frac{e-e_n}{1+e-e_n} < \ln(1+e-e_n) < e-e_n, \quad n \in \mathbb{N}_0.$$

We now use continuity of the natural logarithm function, and calculate

$$\begin{aligned} 0 \le \ln\left(\lim_{n\to\infty} (1+e-e_n)^{n!}\right) &= \lim_{n\to\infty} n! \cdot \ln(1+e-e_n) \le \lim_{n\to\infty} n! \cdot (e-e_n) \\ &= \lim_{n\to\infty} \frac{e-e_n}{1/n!}, \end{aligned}$$

where we used the fundamental estimate above. We now employ the additive Stolz–Cesàro Theorem (Section 3.4), and continue

$$\lim_{n\to\infty} \frac{e-e_n}{1/n!} = \lim_{n\to\infty} \frac{e_n - e_{n+1}}{1/(n+1)! - 1/n!} = \lim_{n\to\infty} \frac{-1/(n+1)!}{-n/(n+1)!} = \lim_{n\to\infty} \frac{1}{n} = 0.$$

[8] This is due to Virgil Nicula.

Putting everything together, we obtain

$$\ln\left(\lim_{n\to\infty}(1+e-e_n)^{n!}\right)=0.$$

The first limit follows.

Returning to the main line, we rewrite the fundamental estimate as

$$\frac{1}{1+x}\leq\frac{\ln(1+x)}{x}\leq 1,\quad -1<x\neq 0.$$

We now take the limit

$$\lim_{x\to 0}\frac{\ln(1+x)}{x}=\ln'(1)=1.$$

We write this as

$$\lim_{x\to 0}\ln\left((1+x)^{1/x}\right)=1.$$

By continuity of the natural logarithm established above, we have

$$\lim_{x\to 0}\ln\left((1+x)^{1/x}\right)=\ln\left(\lim_{x\to 0}(1+x)^{1/x}\right)=1.$$

Taking exponents, we arrive at Euler's famous limit[9]

$$\lim_{x\to 0}(1+x)^{1/x}=e.$$

Replacing x by x/n, $n\in\mathbb{N}$, with **fixed** $0\neq x\in\mathbb{R}$, we obtain the following discrete version

$$\lim_{n\to\infty}\left(1+\frac{x}{n}\right)^n=e^x.$$

We pause here briefly to derive a significant improvement of the limit in Example 3.2.9 as follows:

Example 10.3.4 Show that

$$\lim_{n\to\infty}\frac{n}{\sqrt[n]{n!}}=e.$$

Letting $a_n=n!/n^n$, $n\in\mathbb{N}$, we calculate

[9] We will treat this is more detail in Section 10.5.

$$\lim_{n\to\infty}\frac{a_{n+1}}{a_n}=\lim_{n\to\infty}\frac{(n+1)!}{(n+1)^{n+1}}\cdot\frac{n^n}{n!}=\lim_{n\to\infty}\frac{n^n}{(n+1)^n}=\lim_{n\to\infty}\frac{1}{(1+1/n)^n}=\frac{1}{e},$$

where we used Euler's limit above. Now, the multiplicative Stolz–Cesàro theorem (Section 3.4) gives

$$\lim_{n\to\infty}a_n^{1/n}=\lim_{n\to\infty}\frac{\sqrt[n]{n!}}{n}=\frac{1}{e}.$$

The example follows.

Remark This limit is usually expressed as the asymptotic relation[10]

$$\sqrt[n]{n!}\sim\frac{n}{e}\quad\text{as}\quad n\to\infty.$$

This can still be improved to give the well-known **Stirling formula**

$$n!\sim\sqrt{2\pi n}\left(\frac{n}{e}\right)^n\quad\text{as}\quad n\to\infty,$$

usually derived in integral calculus.

Returning to the main line, replacing x by $1/n$, $n\in\mathbb{N}$, in our fundamental estimate of the natural logarithm, we get

$$\frac{1}{n+1}<\ln\left(1+\frac{1}{n}\right)<\frac{1}{n},\quad n\in\mathbb{N}.$$

We write the middle term as

$$\frac{1}{n+1}<\ln(n+1)-\ln(n)<\frac{1}{n},\quad n\in\mathbb{N}.$$

Remark The reader versed in elementary calculus will no doubt recognize this inequality as the trivial estimate of the integral

$$\frac{1}{n+1}<\int_n^{n+1}\frac{dx}{x}=\ln(n+1)-\ln(n)<\frac{1}{n},\quad n\in\mathbb{N}.$$

Iterating this estimate over $n=1,2,\dots,n-1$, $2\le n\in\mathbb{N}$, and adding, we obtain

$$\frac{1}{2}+\frac{1}{3}+\cdots+\frac{1}{n}<\ln(n)<1+\frac{1}{2}+\cdots+\frac{1}{n-1}.$$

[10]For two sequences $(a_n)_{n\in\mathbb{N}}$ and $(b_n)_{n\in\mathbb{N}}$ with non-zero terms, we write $a_n\sim b_n$ as $n\to\infty$ if $\lim_{n\to\infty}a_n/b_n=1$.

We rewrite this using the sum of the reciprocals of the first n natural numbers (Example 3.1.6)

$$H_n = 1 + \frac{1}{2} + \cdots + \frac{1}{n}, \quad n \in \mathbb{N},$$

and obtain the following important inequalities

$$H_n - 1 < \ln(n) < H_{n-1}, \quad 2 \leq n \in \mathbb{N}.$$

Of importance is the sequence of differences $(H_n - \ln(n))_{n\in\mathbb{N}}$.
First, this sequence is bounded below, since, by the second inequality above, $0 < H_{n-1} - \ln(n)$, so that we have

$$0 < \frac{1}{n} \leq H_n - \ln(n), \quad n \in \mathbb{N}.$$

(Equality holds only for $n = 1$.)
Second, we claim that this sequence is strictly decreasing. Indeed, using the inequality for the difference $\ln(n + 1) - \ln(n)$ above, we have

$$H_{n+1} - \ln(n+1) = H_n - \ln(n) + \left(\frac{1}{n+1} + \ln(n) - \ln(n+1)\right) < H_n - \ln(n), \quad n \in \mathbb{N}.$$

Finally, by the Monotone Convergence Theorem, this sequence is convergent

$$\lim_{n\to\infty} (H_n - \ln(n)) = \gamma,$$

where the limit γ is called the **Euler–Mascheroni constant**.

Remark It is not known whether γ is rational or irrational. Due to the frequent appearance of γ in various parts of analysis, this is an outstanding problem in mathematics. Using continued fractions one can show that if γ is rational, then in its simple fraction form the denominator must be at least 10^{242080}.

Up to the first 60 digits, we have

$$\gamma = 0.577215664901532860606512090082402431042159335939923598805767\ldots$$

The next example is once again a significant improvement of the limit in Example 3.2.8:

Example 10.3.5 Show that

$$\ln n \leq n\left(\sqrt[n]{n} - 1\right) \leq \ln n \cdot \frac{n}{n + 1 - \sqrt{2n-1}}, \quad n \in \mathbb{N},$$

and consequently

$$\lim_{n\to\infty} \left(n\left(\sqrt[n]{n}-1\right)-\ln n\right)=0.$$

Writing $\sqrt[n]{n}-1=e^{\ln n/n}-1$, since $\ln n < n$, $n\in\mathbb{N}$, the fundamental estimate for natural exponentiation (Section 10.1) gives

$$\frac{\ln n}{n}\le \sqrt[n]{n}-1\le \frac{\ln n}{n}\cdot\frac{1}{1-\ln n/n}.$$

Rearranging and using the sharper upper bound for $\ln n$ derived earlier in this section, we obtain

$$\ln n\le n\left(\sqrt[n]{n}-1\right)\le \ln n\cdot\frac{n}{n-\ln n}\le \ln n\cdot\frac{n}{n-(\sqrt{2n-1}-1)}.$$

The inequality stated above follows.

It remains to derive the associated limit. The estimate just proved gives

$$0\le n\left(\sqrt[n]{n}-1\right)-\ln n\le \ln n\cdot\frac{\sqrt{2n-1}-1}{n+1-\sqrt{2n-1}}.$$

We need to show that the right-hand side is a null-sequence. Simple algebra gives

$$\ln n\cdot\frac{\sqrt{2n-1}-1}{n+1-\sqrt{2n-1}}=2\frac{\ln\sqrt{n}}{\sqrt{n}}\frac{\sqrt{2-1/n}-1/\sqrt{n}}{1+1/n-\sqrt{2/n-1/n^2}}.$$

Since $\lim_{x\to\infty}\ln x/x=0$, the logarithmic factor on the right-hand side has zero limit while the last factor has limit $\sqrt{2}$. The overall limit is therefore zero. The example follows.

For the next example we now return to a previous topic. Recall from Example 3.2.12 the limit formula[11]

$$\lim_{n\to\infty}\frac{s_p(n+1)}{n^{p+1}}=\lim_{n\to\infty}\frac{1^p+2^p+\cdots+n^p}{n^{p+1}}=\frac{1}{p+1},\quad -1<p\in\mathbb{R}.$$

This limit can be interpreted as

$$\lim_{n\to\infty}\frac{\mathcal{A}_p(n)}{n^p}=\frac{1}{p+1},\quad -1<p\in\mathbb{R},$$

[11] Note the extended range of p as a special case of Example 3.4.1, and also the moved up value of n to $n+1$.

where

$$\mathcal{A}_p(n) = \frac{1}{n}\sum_{k=1}^{n} k^p = \frac{1^p + 2^p + \cdots + n^p}{n}$$

is the arithmetic mean of the pth powers of the first n natural numbers.

In view of the AM-GM inequality, it is natural to consider the related problem in which the arithmetic mean is replaced by the geometric mean:

Example 10.3.6 We have

$$\lim_{n\to\infty} \frac{\mathcal{G}_p(n)}{n^p} = e^{-p}, \quad p \in \mathbb{R},$$

where

$$\mathcal{G}_p(n) = \sqrt[n]{\prod_{k=1}^{n} k^p} = \sqrt[n]{1^p \cdot 2^p \cdots n^p}$$

is the geometric mean of the pth powers of the first n natural numbers.

We calculate

$$\lim_{n\to\infty} \frac{\mathcal{G}_p(n)}{n^p} = \lim_{n\to\infty} \frac{\sqrt[n]{1^p \cdot 2^p \cdots n^p}}{n^p} = \lim_{n\to\infty} \left(\frac{\sqrt[n]{n!}}{n}\right)^p = e^{-p},$$

where we used the result of Example 10.3.4.

Combining the limit relations with the arithmetic and geometric means, we obtain[12]

$$(1 \le) \lim_{n\to\infty} \frac{\mathcal{A}_p(n)}{\mathcal{G}_p(n)} = \lim_{n\to\infty} \frac{(1^p + 2^p + \cdots + n^p)/n}{\sqrt[n]{1^p \cdot 2^p \cdots n^p}} = \frac{e^p}{p+1}, \quad -1 < p \in \mathbb{R}.$$

A variation on the theme is the following:

Example 10.3.7 Let $p(x)$ be a polynomial of degree $m \in \mathbb{N}$. We define the arithmetic and geometric means of $p(x)$ by

$$\mathcal{A}_{p(x)}(n) = \frac{1}{n}\sum_{k=1}^{n} p(k) \quad \text{and} \quad \mathcal{G}_{p(x)}(n) = \sqrt[n]{\prod_{k=1}^{n} p(k)}, \quad n \in \mathbb{N}.$$

[12] See also Kubelka, R.P., *Means to an end*, Math. Mag. 74 (2001) 141–142, and Conway Xu, *A GM-AM ratio*, Math. Mag. 83 (2010) 49–50.

Show that

$$(1 \le) \lim_{n\to\infty} \frac{\mathcal{A}_{p(x)}(n)}{\mathcal{G}_{p(x)}(n)} = \frac{e^m}{m+1}.$$

We let

$$p(x) = a_m x^m + a_{m-1}x^{m-1} + \cdots + a_1 x + a_0, \quad a_m \neq 0, \ a_0, a_1, \cdots, a_m \in \mathbb{R}.$$

We calculate

$$\begin{aligned} n \cdot \mathcal{A}_{p(x)}(n) &= \sum_{k=1}^{n} p(k) = \sum_{k=1}^{n} \left(a_m k^m + a_{m-1}k^{m-1} + \cdots + a_1 k + a_0 \right) \\ &= a_m s_m(n+1) + a_{m-1}s_{m-1}(n+1) + \cdots + a_1 s_1(n+1) + a_0 n. \end{aligned}$$

Using the limit in Example 3.2.12, this gives

$$\lim_{n\to\infty} \frac{\mathcal{A}_{p(x)}(n)}{n^m} = \lim_{n\to\infty} \left(a_m \frac{s_m(n+1)}{n^{m+1}} + a_{m-1} \frac{s_{m-1}(n+1)}{n^{m+1}} + \cdots + a_0 \frac{1}{n^m} \right) = \frac{a_m}{m+1}.$$

For the geometric mean we work backwards. We use the multiplicative Stolz–Cesàro limit relation, and calculate

$$\begin{aligned} a_m &= \lim_{n\to\infty} \frac{p(n)}{n^m} = \lim_{n\to\infty} \sqrt[n]{\frac{p(1)}{1^m} \cdot \frac{p(2)}{2^m} \cdots \frac{p(n)}{n^m}} = \lim_{n\to\infty} \frac{\sqrt[n]{p(1)\cdot p(2)\cdots p(n)}}{\sqrt[n]{n!}^m} \\ &= \lim_{n\to\infty} \frac{\sqrt[n]{p(1)\cdot p(2)\cdots p(n)}}{n^m} \lim_{n\to\infty} \left(\frac{n}{\sqrt[n]{n!}} \right)^m = \lim_{n\to\infty} \frac{\mathcal{G}_{p(x)}(n)}{n^m} \cdot e^m, \end{aligned}$$

where we used Example 10.3.4 above.

Putting these together, we obtain

$$\lim_{n\to\infty} \frac{\mathcal{A}_{p(x)}(n)}{\mathcal{G}_{p(x)}(n)} = \lim_{n\to\infty} \frac{\mathcal{A}_{p(x)}(n)}{n^m} \cdot \lim_{n\to\infty} \frac{n^m}{\mathcal{G}_{p(x)}(n)} = \frac{a_m}{m+1} \cdot \frac{e^m}{a_m} = \frac{e^m}{m+1}.$$

The example follows.

We finish this section by a cadre of interesting limits.

Example 10.3.8 [13] Derive the limit

$$\lim_{n\to\infty} \left(\sqrt[n+1]{(n+1)!} - \sqrt[n]{n!} \right) = \frac{1}{e}.$$

[13] This is due to the Roumanian mathematician Traian Lalescu (1882–1929).

This is an easy application of the Stolz–Cesàro theorem as follows:

$$\lim_{n\to\infty}\left(\sqrt[n+1]{(n+1)!}-\sqrt[n]{n!}\right)=\lim_{n\to\infty}\frac{\sqrt[n+1]{(n+1)!}-\sqrt[n]{n!}}{(n+1)-n}=\lim_{n\to\infty}\frac{\sqrt[n]{n!}}{n}=\frac{1}{e},$$

where we used the limit in Example 10.3.4 again.

Example 10.3.9 [14] Show that

$$\lim_{n\to\infty}\left(\frac{(n+1)^2}{\sqrt[n+1]{(n+1)!}}-\frac{n^2}{\sqrt[n]{n!}}\right)=e.$$

We will derive a generalization of this as follows:[15] Let $(a_n)_{n\in\mathbb{N}}$ be a real sequence with positive terms. Then we have the implication

$$\lim_{n\to\infty}\frac{a_{n+1}-a_n}{n}=L>0\quad\Rightarrow\quad\lim_{n\to\infty}\left(\frac{a_{n+1}}{\sqrt[n+1]{(n+1)!}}-\frac{a_n}{\sqrt[n]{n!}}\right)=e\cdot\frac{L}{2}.$$

To show this, we first use Example 10.3.4, and calculate

$$\lim_{n\to\infty}\left(\frac{a_{n+1}}{\sqrt[n+1]{(n+1)!}}-\frac{a_n}{\sqrt[n]{n!}}\right)=\lim_{n\to\infty}\frac{a_n}{\sqrt[n]{n!}}\left(\frac{a_{n+1}}{a_n}\frac{\sqrt[n]{n!}}{\sqrt[n+1]{(n+1)!}}-1\right)$$

$$=\lim_{n\to\infty}\frac{n}{\sqrt[n]{n!}}\frac{a_n}{n^2}n\left(\frac{a_{n+1}}{a_n}\frac{\sqrt[n]{n!}}{\sqrt[n+1]{(n+1)!}}-1\right)=\frac{eL}{2}\lim_{n\to\infty}n\left(\frac{a_{n+1}}{a_n}\frac{\sqrt[n]{n!}}{\sqrt[n+1]{(n+1)!}}-1\right),$$

where we used the limit in the previous example and the Stolz–Cesàro Theorem to the effect that

$$\lim_{n\to\infty}\frac{a_n}{n^2}=\lim_{n\to\infty}\frac{a_{n+1}-a_n}{(n+1)^2-n^2}=\lim_{n\to\infty}\frac{a_{n+1}-a_n}{n}\frac{n}{2n+1}=\frac{L}{2}.$$

As a byproduct of the last limit to be used below, we also have

$$\lim_{n\to\infty}\frac{a_{n+1}}{a_n}=\lim_{n\to\infty}\frac{a_{n+1}}{(n+1)^2}\cdot\frac{n^2}{a_n}\cdot\frac{(n+1)^2}{n^2}=\frac{L}{2}\cdot\frac{2}{L}=1.$$

Returning to our main computation, it remains to show that

$$\lim_{n\to\infty}n\,(b_n-1)=1,$$

[14] This is due to the Roumanian mathematician D.M. Bătineţu-Giurgiu (1936-).

[15] This generalization and the next example are due to Virgil Nicula.

where

$$b_n = \frac{a_{n+1}}{a_n} \cdot \frac{\sqrt[n]{n!}}{\sqrt[n+1]{(n+1)!}}.$$

First note that

$$\begin{aligned}\lim_{n\to\infty} b_n &= \lim_{n\to\infty} \frac{a_{n+1}}{a_n} \cdot \frac{\sqrt[n]{n!}}{\sqrt[n+1]{(n+1)!}} \\ &= \lim_{n\to\infty} \frac{a_{n+1}}{a_n} \cdot \frac{\sqrt[n]{n!}}{n} \cdot \frac{n+1}{\sqrt[n+1]{(n+1)!}} \cdot \frac{n+1}{n} \\ &= 1 \cdot \frac{1}{e} \cdot e = 1,\end{aligned}$$

where we used Example 10.3.4 again. By continuity of the natural logarithm function, we obtain $\lim_{n\to\infty} \ln b_n = 0$.

Once again, returning to the main line, we have

$$\lim_{n\to\infty} n\,(b_n - 1) = \lim_{n\to\infty} n\left(e^{\ln b_n} - 1\right) = \lim_{n\to\infty} n \ln b_n,$$

where the last equality follows from the fundamental estimate of the natural exponential function in Section 10.1 applied to the null-sequence $(\ln b_n)_{n\in\mathbb{N}}$ (provided that the last limit exists). We now use the explicit formula for b_n, $n \in \mathbb{N}$ and obtain

$$\begin{aligned}\lim_{n\to\infty} n \ln b_n &= \lim_{n\to\infty} n \ln\left(\frac{a_{n+1}}{a_n} \cdot \frac{\sqrt[n]{n!}}{\sqrt[n+1]{(n+1)!}}\right) \\ &= \lim_{n\to\infty} n\left(\ln\frac{a_{n+1}}{a_n} + \frac{\ln n!}{n} - \frac{\ln(n+1)!}{n+1}\right).\end{aligned}$$

For the first term in the parentheses, we calculate

$$\begin{aligned}\lim_{n\to\infty} n \ln\frac{a_{n+1}}{a_n} &= \lim_{n\to\infty} n \ln\left(\left(\frac{a_{n+1}}{a_n} - 1\right) + 1\right) \\ &= \lim_{n\to\infty} n\left(\frac{a_{n+1}}{a_n} - 1\right) = \lim_{n\to\infty} \frac{n^2}{a_n} \cdot \frac{a_{n+1} - a_n}{n} = \frac{2}{L} \cdot L = 2,\end{aligned}$$

where we applied the fundamental estimate for the natural logarithm to the null-sequence $(a_{n+1}/a_n - 1)_{n\in\mathbb{N}}$, and the previous limits.

For the remaining terms in the parentheses, we use the Stolz–Cesàro theorem again, and calculate

$$\lim_{n\to\infty} n\left(\frac{\ln n!}{n} - \frac{\ln(n+1)!}{n+1}\right) = \lim_{n\to\infty} \frac{(n+1)\ln n! - n\ln(n+1)!}{n+1}$$

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

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

$$= \lim_{n\to\infty} (n+1)\ln\frac{n+1}{n+2} = -\ln\left(\lim_{n\to\infty}\left(\frac{n+2}{n+1}\right)^{n+1}\right)$$

$$= -\ln\left(\lim_{n\to\infty}\left(1+\frac{1}{n+1}\right)^{n+1}\right) = -\ln e = -1,$$

where we also used Euler's limit.

Putting everything together, we obtain

$$\lim_{n\to\infty} n\,(b_n - 1) = 2 - 1 = 1.$$

The example follows.

Example 10.3.10 Let $(a_n)_{n\in\mathbb{N}_0}$ be a sequence such that $0 < a_0 < 1$, and $a_{n+1} = a_n - a_n^2$, $n \in \mathbb{N}_0$. Show that

$$(1)\ \lim_{n\to\infty} a_n = 0;\quad (2)\ \lim_{n\to\infty} na_n = 1;\quad (3)\ \lim_{n\to\infty} \frac{n(1-na_n)}{\ln n} = 1.$$

We first claim that $a_n \in (0, 1)$, $n \in \mathbb{N}$. By Peano's Principle of Induction, we need to perform only the general induction step $n \Rightarrow n+1$. But this is clear since $a_{n+1} = a_n(1 - a_n) \in (0, 1)$.

Next, $a_{n+1} = a_n - a_n^2 < a_n$, $n \in \mathbb{N}_0$, so that the sequence $(a_n)_{n\in\mathbb{N}_0}$ is strictly decreasing. By the Monotone Convergence Theorem, this sequence is convergent. Let $\lim_{n\to\infty} a_n = L \in [0, 1)$. By the recurrence relation, we have $L = L - L^2$. We obtain $L = 0$. Thus, (1) follows.

To show (2), we write

$$\lim_{n\to\infty} na_n = \lim_{n\to\infty} \frac{n}{1/a_n},$$

and make use of the Stolz–Cesàro theorem (with n moved up to $n+1$) as follows:

$$\lim_{n\to\infty} \frac{(n+1) - n}{1/a_{n+1} - 1/a_n} = \lim_{n\to\infty} \frac{a_n \cdot a_{n+1}}{a_n - a_{n+1}} = \lim_{n\to\infty} \frac{a_n(a_n - a_n^2)}{a_n^2} = \lim_{n\to\infty} (1 - a_n) = 1,$$

where the last equality is because of (1). Hence (2) follows.

For (3), we write the limit as

$$\lim_{n\to\infty} \frac{n(1-na_n)}{\ln n} = \lim_{n\to\infty} \frac{na_n}{\ln n}\left(\frac{1}{a_n} - n\right) = \lim_{n\to\infty} \frac{1/a_n - n}{\ln n},$$

where we used (2). Again, making use of the Stolz–Cesàro theorem, we calculate

$$\lim_{n\to\infty} \frac{(1/a_{n+1} - (n+1)) - (1/a_n - n)}{\ln(n+1) - \ln(n)} = \lim_{n\to\infty} \frac{1/a_{n+1} - 1/a_n - 1}{\ln(1+1/n)}$$
$$= \lim_{n\to\infty} n\left(\frac{1}{a_{n+1}} - \frac{1}{a_n} - 1\right) = \lim_{n\to\infty} n\left(\frac{1}{a_n(1-a_n)} - \frac{1}{a_n} - 1\right) = \lim_{n\to\infty} \frac{n\cdot a_n}{1-a_n} = 1,$$

where we used Euler's limit $\lim_{n\to\infty} n\ln(1+1/n) = \ln\lim_{n\to\infty}(1+1/n)^n = 1$ as well as (1) and (2). Now (3) follows.

Exercises

10.3.1. Use the exponential and logarithmic functions to show $\lim_{n\to\infty} \sqrt[n]{a^n + b^n} = \max(a, b)$, $a, b > 0$.

10.3.2. Derive the inequality

$$\frac{\ln(x) + \ln(y)}{2} \le \ln\left(\frac{x+y}{2}\right), \quad x, y > 0.$$

10.3.3. Calculate the derivatives of the general exponential and logarithmic functions.

For the next set of exercises, define the **cosine and sine hyperbolic functions** $\cosh : \mathbb{R} \to \mathbb{R}$ and $\sinh : \mathbb{R} \to \mathbb{R}$ as

$$\cosh(x) = \frac{e^x + e^{-x}}{2} \quad \text{and} \quad \sinh(x) = \frac{e^x - e^{-x}}{2}, \quad x \in \mathbb{R}.$$

10.3.4. Derive the identity $\cosh^2(x) - \sinh^2(x) = 1$, $x \in \mathbb{R}$.

10.3.5. For $x, y \in \mathbb{R}$, derive the addition formulas

$$\cosh(x+y) = \cosh(x)\cosh(y) + \sinh(x)\sinh(y);$$
$$\sinh(x+y) = \sinh(x)\cosh(y) + \cosh(x)\sinh(y).$$

10.3.6. Show that $\cosh'(c) = \sinh(c)$ and $\sinh'(c) = \cosh(c)$, $c \in \mathbb{R}$.

10.3.7. Prove that, for $q \in \mathbb{Q}$, the numbers $\sinh(\ln q)$ and $\cosh(\ln q)$ are rational numbers. Calculate $\sinh(\ln 2)$ and $\cosh(\ln 2)$.

10.3.8. For $n \in \mathbb{N}_0$, derive the lower estimates

$$1 + \frac{x^2}{2!} + \frac{x^4}{4!} + \cdots + \frac{x^{2n}}{(2n)!} < \cosh(x), \quad 0 < x \in \mathbb{R}$$

and

$$\frac{x}{1!} + \frac{x^3}{3!} + \cdots + \frac{x^{2n+1}}{(2n+1)!} < \sinh(x), \quad 0 < x \in \mathbb{R}.$$

10.3.9. Show that

$$\cosh(x) = \sum_{n=0}^{\infty} \frac{x^{2n}}{(2n)!} \quad \text{and} \quad \sinh(x) = \sum_{n=0}^{\infty} \frac{x^{2n+1}}{(2n+1)!}, \quad x \in \mathbb{R}.$$

10.3.10. Use Exercise 10.3.8 to show the following:

$$\ln(x) < \frac{1}{2}\left(x - \frac{1}{x}\right), \ x > 1 \text{ and } \ln(x) > \frac{1}{2}\left(x - \frac{1}{x}\right), \quad 0 < x < 1;$$

$$|\ln(x)| \leq \sqrt{x + \frac{1}{x} - 2}, \quad x > 0;$$

$$|\ln(x)| \leq \sqrt{2\sqrt{3\left(x + \frac{1}{x} + 1\right)} - 6}, \quad x > 0,$$

and in all the estimates equalities hold if and only if $x = 1$.

10.3.11. Determine the horizontal and vertical asymptotes of the functions

$$f(x) = \frac{1}{\cosh(x)} \quad \text{and} \quad g(x) = \frac{1}{\sinh(x)}, \quad x \in \mathbb{R}.$$

10.4 The General Exponential and Logarithmic Functions

For a given positive real base $0 < a \in \mathbb{R}$, we **define**

$$a^x = e^{x \cdot \ln(a)}, \quad x \in \mathbb{R}.$$

Since the natural exponential function is differentiable, we see that the function $y = a^x$, $x \in \mathbb{R}$, is differentiable (and hence continuous).

We claim that, for x integral or rational number, this new definition of the exponentiation reverts back to our earlier definition in Section 3.2.

First, for $n \in \mathbb{N}$, we have

$$a^n = e^{n \cdot \ln(a)} = e^{\overbrace{\ln(a) + \cdots + \ln(a)}^{n}} = \overbrace{e^{\ln(a)} \cdots e^{\ln(a)}}^{n} = \overbrace{a \cdots a}^{n}.$$

Clearly, $a^0 = e^{0 \cdot \ln(a)} = e^0 = 1$. For $n \in \mathbb{N}$, we also have

$$a^{-n} = e^{-n \ln(a)} = \frac{1}{e^{n \ln(a)}} = \frac{1}{a^n}.$$

Hence, for integral exponents, the two definitions are the same.

Second, recall from Section 3.2 that, for $m/n \in \mathbb{Q}$ with $m, n \in \mathbb{Z}$ and $n \neq 0$, the exponential $a^{m/n}$ is defined as the unique positive real number for which

$$\left(a^{\frac{m}{n}}\right)^n = \left(\sqrt[n]{a^m}\right)^n = a^m.$$

We now calculate

$$\left(e^{\frac{m}{n} \cdot \ln(a)}\right)^n = e^{\frac{m}{n} \cdot \ln(a)} \cdot e^{\frac{m}{n} \cdot \ln(a)} \cdots e^{\frac{m}{n} \cdot \ln(a)} = e^{n \cdot \frac{m}{n} \cdot \ln(a)} = e^{m \cdot \ln(a)} = a^m.$$

Setting $q = m/n \in \mathbb{Q}$, we obtain

$$a^q = e^{q \cdot \ln(a)}, \quad q \in \mathbb{Q}.$$

The claim follows.

Finally, (sequential) continuity of our new exponentiation implies that it coincides with the old definition for **real** exponents.

Since the exponential and logarithmic functions (of the same base) are inverses of each other, we have $\ln(x) = \log_e(x)$, $0 < x \in \mathbb{R}$. Moreover, the change of base formula implies that the general logarithmic function is differentiable (and hence continuous).

Example 10.4.1 [16] Let $2 \leq n \in \mathbb{N}$. For what value of $0 < a \in \mathbb{R}$ do we have

$$\sum_{k=2}^{n} \frac{1}{\log_k a} = \frac{1}{\log_2 a} + \frac{1}{\log_3 a} + \cdots + \frac{1}{\log_n a} = 1?$$

[16] A special case was a problem in the American Mathematics Competition, 2015.

Using the change of base formula and the logarithmic identities (Section 3.3), we obtain

$$\sum_{k=2}^{n} \frac{1}{\log_k a} = \sum_{k=2}^{n} \log_a k = \log_a \left(\prod_{k=2}^{n} k \right) = \log_a(n!) = 1.$$

Hence, $a = n!$.

In addition to the natural base e, the base 10 logarithmic function, the so-called **common logarithm**, is particularly well suited in computations when using the decimal system. The base 10 of the common logarithm is often suppressed from the notation, and we write $\log_{10}(x) = \log(x)$, $x > 0$.

As a simple illustration, we claim that, for $n \in \mathbb{N}$, the greatest integer of the common logarithm, $[\log(n)]$, is one less than the number of decimal digits in n.

Indeed, write n using decimal digits as

$$n = d_k d_{k-1} \dots d_1 d_0$$

with the digits $d_0, d_1, \dots, d_{k-1}, d_k$ ranging from 0 to 9 and $d_k \neq 0$. We thus have

$$n = 10^{k+1} \cdot 0.d_k d_{k-1} \dots d_1 d_0.$$

Taking the common logarithm of both sides and using the logarithmic identities, we obtain

$$\begin{aligned} \log(n) &= \log(10^{k+1} \cdot 0.d_k d_{k-1} \dots d_1 d_0) \\ &= \log(10^{k+1}) + \log(0.d_k d_{k-1} \dots d_1 d_0) \\ &= k + 1 + \log(0.d_k d_{k-1} \dots d_1 d_0). \end{aligned}$$

Since $d_k \geq 1$, we have $1/10 \leq 0.d_k d_{k-1} \dots d_1 d_0 < 1$. Thus $-1 \leq \log(1/10) \leq \log(0.d_k d_{k-1} \dots d_1 d_0) < \log(1) = 0$. With this, we have $k \leq \log(n) < k + 1$, and the claim follows.

Example 10.4.2 To express 2^{100} in decimal notation, how many decimal digits are needed?

We have $\log(2) \approx 0.3010299957$ so that

$$\log(2^{100}) \approx 100 \cdot 0.3010299957 = 30.10299957.$$

By the above, the decimal representation of 2^{100} has 31 digits. By the way, the number itself is

$$2^{100} = 1267650600228229401496703205376.$$

Exercise

10.4.1. As a generalization of the Bernoulli inequality in Section 3.2, show that the exponential function $y = a^x$ with domain variable $x \in \mathbb{R}$ is **convex**, that is, for $x_0 < x_1$ we have

$$a^{(1-x)x_0+xx_1} \leq (1-x)a^{x_0} + xa^{x_1}, \quad 0 \leq x \leq 1.$$

What is the geometric meaning of this inequality?

10.5 The Natural Exponential Function According to Euler

In this section we start anew, and discuss Euler's approach to the natural exponential function. Recall from Section 3.2 our notation

$$e_n^*(x) = \left(1 + \frac{x}{n}\right)^n, \quad x \in \mathbb{R},\ n \in \mathbb{N}.$$

Note that we showed there the monotonicity property

$$e_n^*(x) < e_{n+1}^*(x), \quad 0 \neq x > -n,\ n \in \mathbb{N}.$$

One of our purposes in the present section is to give a direct proof (without the use of the natural logarithm function) of the limit formula

$$\lim_{n\to\infty} e_n^*(x) = \lim_{n\to\infty} \left(1 + \frac{x}{n}\right)^n = e^x, \quad x \in \mathbb{R}.$$

(Following Newton, we derived this in Section 10.3 in a rather circuitous way.)

Remark 1 For completeness, we note here that, using the natural logarithm function and its properties, a quick proof can be given as follows.

We may assume $x \neq 0$. We have

$$\ln\left(\lim_{n\to\infty} e_n^*(x)\right) = \lim_{n\to\infty} \ln e_n^*(x) = \lim_{n\to\infty} n \cdot \ln\left(1 + \frac{x}{n}\right)$$
$$= \lim_{h\to 0} \frac{x}{h} \cdot \ln(1+h) = x \cdot \lim_{h\to 0} \frac{\ln(1+h) - \ln 1}{h} = x \cdot \ln'(1) = x.$$

Remark 2 Recall the Compound Interest Formula in Example 6.1.4. It gives the future (compound) value V of a principal deposit P with x interest rate after t years as

$$V = P\left(1 + \frac{x}{n}\right)^{n\cdot t},$$

assuming n compounding periods per year. Using our notations, this can be written as $V = P \cdot e_n^*(x)^t$. Taking the limit, and using the limit relation above (yet to be discussed) along with continuity of exponentiation, we have

$$\lim_{n\to\infty} P \cdot e_n^*(x)^t = P \cdot \left(\lim_{n\to\infty} e_n^*(x)\right)^t = P \cdot (e^x)^t = P \cdot e^{x\cdot t}.$$

This is called the **Continuous Compound Interest Formula**.

Although it is physically unrealistic, it reflects the future value, assuming continuous compounding. For t relatively large (retirement accounts), it shows a near exponential growth of an initial investment assuming stable average market conditions for a long stretch of time.

History
As noted in Section 6.1, it was Jacob Bernoulli who, in studying compound interest, first tried to find the actual value of e considering $(1 + 1/n)^n$ for large values of $n \in \mathbb{N}$. Subsequently, Johann Bernoulli (1667–1748) in 1697 studied the analytic properties of $(1 + x/n)^n$ for large n. The number e was first used by Leibniz, but as the base of the natural logarithm, the inverse of the natural exponentiation, it appeared first in the works of Euler. In particular, he noted the limits $\lim_{n\to\infty}(1 + x/n)^n = e^x$ and $\lim_{n\to\infty} n(\sqrt[n]{x} - 1) = \ln(x)$.

Returning to the main line, recall the Bernoulli inequalities from Section 3.2: For $a > 0$, we have

$$a^x \le 1 + x(a - 1) \text{ for } 0 < x < 1, \text{ and } a^x \ge 1 + x(a - 1) \text{ for } x > 1, \ x \in \mathbb{R}.$$

Remark Having the exponential function in place, we can now give the simple geometric interpretation of these inequalities. The line given by $y = 1 + x(a - 1)$ is a secant that cuts the graph of the exponential function $y = a^x$ at the points $(0, 1)$ and $(1, a)$. The graph itself is "convex" in the sense that it is below the (finite) secant segment cut out from the graph by the secant line, and above beyond.

As the first task, and as in the case of rational exponents, we claim that equality holds in either of the inequalities above if and only if $a = 1$.

We show this for the first inequality; the argument for the second is analogous. Given $0 < x < 1$, $x \in \mathbb{R}$, choose $n \in \mathbb{N}$ large enough such that

$$0 < x - \frac{1}{n} < x < x + \frac{1}{n} < 1.$$

By the first Bernoulli inequality applied to these modified values, we have

$$a^{x\pm 1/n} \le 1 + \left(x \pm \frac{1}{n}\right)(a - 1).$$

Assume now that $1 \ne a \in \mathbb{R}$. Then, we have $a^{x-1/n} \ne a^{x+1/n}$ since the exponential function is strictly monotonic. We now use the (strict) AM-GM inequality (in two indeterminates) with the previous inequality and calculate

$$a^x = \sqrt{a^{x-1/n} \cdot a^{x+1/n}} < \frac{a^{x-1/n} + a^{x+1/n}}{2}$$
$$\leq \frac{1 + (x - 1/n)(a-1) + 1 + (x + 1/n)(a-1)}{2} = 1 + x(a-1).$$

Our claim of strict inequality follows.

Of paramount importance in Euler's study of the exponential function are the two functions $f, g : (0, \infty) \to \mathbb{R}$ defined by

$$f(x) = \left(1 + \frac{1}{x}\right)^x \quad \text{and} \quad g(x) = \left(1 + \frac{1}{x}\right)^{x+1}, \quad 0 < x \in \mathbb{R}.$$

Since $x > 0$, we have

$$f(x) < g(x), \quad 0 < x \in \mathbb{R}.$$

We now claim that f is (strictly) **increasing** and g is (strictly) **decreasing**. These are consequences of the Bernoulli inequalities above.

Indeed, letting $x' < x''$, and substituting $a = 1 + 1/x'$ and $x = x'/x''$ into the (first) Bernoulli inequality (since $0 < x'/x'' < 1$), we have

$$\left(1 + \frac{1}{x'}\right)^{x'/x''} < 1 + \frac{x'}{x''} \cdot \frac{1}{x'} = 1 + \frac{1}{x''}.$$

Raising both sides to the exponent x'', we obtain

$$\left(1 + \frac{1}{x'}\right)^{x'} < \left(1 + \frac{1}{x''}\right)^{x''}.$$

This gives $f(x') < f(x'')$, $0 < x' < x''$; and monotonicity of f follows. Similarly, using the substitution $a = 1 - 1/(x'+1)$ and $x = (x'+1)/(x''+1)$, we have

$$\left(1 - \frac{1}{x'+1}\right)^{(x'+1)/(x''+1)} < 1 - \frac{1}{x''+1}.$$

Raising both sides to the exponent $x'' + 1$ and taking reciprocals, we obtain

$$\left(1 + \frac{1}{x'}\right)^{x'+1} > \left(1 + \frac{1}{x''}\right)^{x''+1}.$$

This gives $g(x') > g(x'')$, $0 < x' < x''$; and monotonicity of g follows.

Finally, we consider the difference

$$g(x) - f(x) = \left(1 + \frac{1}{x}\right)^{x+1} - \left(1 + \frac{1}{x}\right)^{x} = \left(1 + \frac{1}{x}\right)^{x} \cdot \frac{1}{x}, \quad x > 0.$$

This gives

$$\lim_{x\to\infty} (g(x) - f(x)) = \lim_{x\to\infty} \left(1 + \frac{1}{x}\right)^{x} \cdot \frac{1}{x} = 0$$

since the first factor in the last product stays bounded while the second factor $(1/x)$ has limit zero.

By monotonicity just proved, we conclude

$$\lim_{x\to\infty} f(x) = \lim_{x\to\infty} g(x).$$

We now **define** the real number e^* as the **common value** of these two limits.

By construction, for all $0 < x \in \mathbb{R}$, we have

$$f(x) = \left(1 + \frac{1}{x}\right)^{x} < e^* < \left(1 + \frac{1}{x}\right)^{x+1} = g(x),$$

where we inserted the definitions of f and g.

We now claim that $e^* = e$, where

$$e = \lim_{n\to\infty} e_n(1) = \lim_{n\to\infty} \left(1 + \frac{1}{1!} + \cdots + \frac{1}{n!}\right),$$

as in Section 10.1.

To show this we choose the simplest sequence $\mathbb{N} = (n)_{n\in\mathbb{N}}$. We have

$$f(n) = e_n^*(1) = \left(1 + \frac{1}{n}\right)^{n} < e^*, \quad n \in \mathbb{N},$$

with limit

$$\lim_{n\to\infty} f(n) = \lim_{n\to\infty} e_n^*(1) = \lim_{n\to\infty} \left(1 + \frac{1}{n}\right)^{n} = e^*.$$

We expand the power in the last limit by the Binomial Formula. For $n \in \mathbb{N}$, we have

$$e_n^*(1) = \left(1 + \frac{1}{n}\right)^{n} = \sum_{k=1}^{n} \binom{n}{k} \frac{1}{n^k} = \sum_{k=0}^{n} \frac{1}{k!} \cdot \frac{n(n-1)\cdots(n-k+1)}{n^k}$$

$$= \sum_{k=0}^{n} \frac{1}{k!} \left(1 - \frac{1}{n}\right) \cdots \left(1 - \frac{k-1}{n}\right) \le \sum_{k=0}^{n} \frac{1}{k!} = e_n(1).$$

Taking limits, we obtain $e^* = \lim_{n\to\infty} e_n^*(1) \le \lim_{n\to\infty} e_n(1) = e$.

For the reverse inequality, we fix $m \in \mathbb{N}$, let $m < n \in \mathbb{N}$, and add up only the first $m+1$ binomial terms

$$1+\frac{1}{1!}+\cdots+\frac{1}{k!}\left(1-\frac{1}{n}\right)\cdots\left(1-\frac{k-1}{n}\right)+\cdots+\frac{1}{m!}\left(1-\frac{1}{n}\right)\cdots\left(1-\frac{m-1}{n}\right) \leq \left(1+\frac{1}{n}\right)^n < e_*.$$

Keeping m fixed and letting $n \to \infty$, the left-hand side approaches $e_m(1)$ (since there are only $m+1$ terms). We obtain $e_m(1) \leq e_*$. Finally, if we now take the limit as $m \to \infty$, the left-hand side approaches e while the right-hand side stays fixed. We finally arrive at $e = \lim_{m\to\infty} e_m(1) \leq e_*$. The reverse inequality, and hence the claim follows.

Returning to the main line, using $e_* = e$, we obtain

$$f(x) = \left(1+\frac{1}{x}\right)^x < e < \left(1+\frac{1}{x}\right)^{x+1} = g(x), \quad 0 < x \in \mathbb{R},$$

and we recover Euler's limit

$$\lim_{x\to\infty} f(x) = \lim_{x\to\infty} g(x) = \lim_{x\to\infty}\left(1+\frac{1}{x}\right)^x = e$$

already obtained in Section 10.3.

Example 10.5.1 For the functions $f, g : (0,\infty) \to \mathbb{R}$, we have the identity

$$f(x)^{g(x)} = g(x)^{f(x)}, \quad 0 < x \in \mathbb{R}.$$

Indeed, using $g(x) = (1+1/x)\cdot f(x)$, $0 < x \in \mathbb{R}$, we calculate

$$f(x)^{g(x)} = f(x)^{(1+1/x)\cdot f(x)} = f(x)^{f(x)} \cdot f(x)^{f(x)/x}$$
$$= f(x)^{f(x)} \cdot \left(1+\frac{1}{x}\right)^{f(x)} = \left(f(x)\cdot\left(1+\frac{1}{x}\right)\right)^{f(x)} = g(x)^{f(x)},$$

where, in the third equality, we also used the definition of f.
The identity just proved shows that $y = f(x)$ and $z = g(x)$, $0 < x \in \mathbb{R}$, are (real) solutions of the equation

$$y^z = z^y, \quad 1 < y < z, \ \ y, z \in \mathbb{R}.$$

(For an interesting contrast, see Remark 1 after Example 10.3.2.) Observe now that this equation is equivalent to

$$e^{\ln(y)/y} = e^{\ln(z)/z}, \quad 1 < y < z, \ \ y, z \in \mathbb{R}.$$

We know from Example 10.3.2 that the function $x \mapsto \ln(x)/x$, $1 < x \in \mathbb{R}$, is strictly increasing on $(1, e]$ (actually on $(0, e]$) and strictly decreasing on $[e, \infty)$ with absolute maximum at e and $\lim_{x\to\infty} \ln(x)/x = 0$. It follows that, the (real) solutions of the equation above are pairs (y, z) with $1 < y < e < z$, $y, z \in \mathbb{R}$, where each component of (y, z) uniquely determines the other.
We are interested in finding the **rational** solutions of this equation; that is, pairs (y, z) with $y, z \in \mathbb{Q}$. Since, for $n \in \mathbb{N}$, the numbers $f(n)$ and $g(n)$ are both rational, by the above, we have an infinite sequence of pairs $(f(n), g(n))$, $n \in \mathbb{N}$, which are rational solutions of our equation.
As a final task, we now show that these are the only rational solutions. To do this, assume that the pair (y, z), $0 < y < e < z$, is a rational solution; that is, $y, z \in \mathbb{Q}$. We let $1 < w = z/y \in \mathbb{Q}$. We substitute $z = w \cdot y$ into the equation and express y in terms of w. We have

$$y^{w\cdot y} = \left(y^w\right)^y = (w \cdot y)^y.$$

This gives $y^w = w \cdot y$, and hence

$$y = w^{\frac{1}{w-1}}.$$

Letting $w = m/n$, $m > n$, $\gcd(m, n) = 1$, $m, n \in \mathbb{N}$, this gives

$$y = \left(\frac{m}{n}\right)^{\frac{n}{m-n}} = \left(\frac{m}{n}\right)^{\frac{n}{k}},$$

where $k = m - n \in \mathbb{N}$.
If $k = 1$, then $m = n + 1$, and we have

$$y = \left(\frac{n+1}{n}\right)^n = \left(1 + \frac{1}{n}\right)^n = f(n),$$

and we arrive at the pair $(f(n), g(n))$, $n \in \mathbb{N}$.
It remains to show that $k > 1$ cannot happen. Assuming the contrary, we let $y = a/b$, $\gcd(a, b) = 1$, $a, b \in \mathbb{N}$. With this the general expression of y above can be written as

$$\frac{a}{b} = \left(\frac{m^n}{n^n}\right)^{\frac{1}{k}}.$$

We claim that m and n are k**th powers**; that is, we have $m = u^k$ and $n = v^k$ for some $u, v \in \mathbb{N}$. Indeed, eliminating the denominators, the equation above takes the form

$$a^k \cdot n^n = b^k \cdot m^n.$$

By $\gcd(a, b) = \gcd(m, n) = 1$, this splits into two equations

$$m^n = a^k \quad \text{and} \quad n^n = b^k.$$

Since $\gcd(n, k) = \gcd(n, m - n) = \gcd(n, m) = 1$, by Proposition 1.3.1, there exists $c, d \in \mathbb{Z}$ such that $c \cdot n + d \cdot k = 1$. Using this, we obtain

$$m = m^{c\cdot n+d\cdot k} = (m^n)^c \cdot (m^d)^k = (a^k)^c \cdot (m^d)^k = (a^c \cdot m^d)^k.$$

Now, the rational number $a^c \cdot m^d$ is actually a positive integer, $u \in \mathbb{N}$, say, since its kth power is $m \in \mathbb{N}$. Thus, we have $m = u^k$. The proof that $n = v^k$ for some $v \in \mathbb{N}$ is analogous. Note that $u > v$ as $m > n$.
With these, we have

$$k = m - n = u^k - v^k = (v + 1)^k - v^k \geq v^k + k \cdot v + 1 - v^k = k \cdot v + 1 \geq k + 1,$$

where, in the first inequality, we used the Binomial Formula (Section 6.3) keeping only the first two terms and the last. This is a contradiction. Our claim follows.
A final note. Taking reciprocals, our equation can be put into the equivalent form

$$\left(\frac{1}{y}\right)^{1/y} = \left(\frac{1}{z}\right)^{1/z}, \quad 0 < 1/z < 1/e < 1/y < 1.$$

Now, replacing the variables by their reciprocals and retaining the original notation, this means that the equation

$$y^y = z^z, \quad 0 < z < 1/e < y < 1,$$

has the pair $(1/g(x), 1/f(x))$ as real solution for all $0 < x \in \mathbb{R}$; and the only rational solutions[17] are $(1/f(n), 1/g(n)), n \in \mathbb{N}$.

Remark There are many inequalities amongst the general powers $y^z, 0 < y, z \in \mathbb{R}$, in various forms, and, although they are well-known, they abound in mathematical contests. Using the Bernoulli inequality, in Example 3.2.10, we showed

$$y^z + z^y > 1, \quad 0 < y, z \in \mathbb{R}.$$

As another example, we have

$$(y \cdot z)^{\frac{y+z}{2}} \leq \left(\frac{y + z}{2}\right)^{y+z} \leq y^y \cdot z^z, \quad 0 < y, z \in \mathbb{R}.$$

[17] This was Problem 5 in Round 1 and Year 32 of the USA Mathematical Talent Search; Academic Year 2020/2021.

The first inequality here is a simple application of the AM-GM inequality. The second is equivalent to

$$\frac{y+z}{2}\cdot\ln\left(\frac{y+z}{2}\right)\le\frac{y\ln y+z\ln z}{2},\quad 0<y,z\in\mathbb{R},$$

and this, in turn, follows from convexity of the function $x\mapsto x\cdot\ln x$, $0<x\in\mathbb{R}$ (usually derived by basic calculus).

Example 10.5.2 We have

$$\lim_{n\to\infty}\left(\frac{1}{n+1}+\cdots+\frac{1}{2n}\right)=\lim_{n\to\infty}\sum_{k=n+1}^{2n}\frac{1}{k}=\ln 2.$$

(Note the lower bound $1/2$ of the expression in parentheses in Exercise 1.4.3 at the end of Section 1.4.)

To derive this limit we use the discrete version of the estimates just obtained above as follows

$$\left(1+\frac{1}{k}\right)^k<e<\left(1+\frac{1}{k-1}\right)^k,\quad 2\le k\in\mathbb{N}.$$

We use monotonicity of the natural logarithm to rewrite this in the equivalent form

$$\ln\left(1+\frac{1}{k}\right)^k<1<\ln\left(1+\frac{1}{k-1}\right)^k,\quad 2\le k\in\mathbb{N}.$$

We now divide by k and sum up for $k=n+1,\dots,2n$, $n\in\mathbb{N}$ and obtain

$$\sum_{k=n+1}^{2n}\ln\left(1+\frac{1}{k}\right)<\sum_{k=n+1}^{2n}\frac{1}{k}<\sum_{k=n+1}^{2n}\ln\left(1+\frac{1}{k-1}\right),\quad n\in\mathbb{N}.$$

We calculate the lower bound as follows

$$\begin{aligned}\sum_{k=n+1}^{2n}\ln\left(1+\frac{1}{k}\right)&=\ln\prod_{k=n+1}^{2n}\left(1+\frac{1}{k}\right)=\ln\prod_{k=n+1}^{2n}\left(\frac{k+1}{k}\right)\\&=\ln\left(\frac{2n+1}{n+1}\right)=\ln\left(2-\frac{1}{n+1}\right).\end{aligned}$$

The calculation for the upper bound is similar

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

Putting everything together, we obtain

$$\ln\left(2-\frac{1}{n+1}\right) < \sum_{k=n+1}^{2n} \frac{1}{k} < \ln 2, \quad n \in \mathbb{N}.$$

Letting $n \to \infty$, the limit relation follows.

Returning to the main line, as in Section 10.3, we now replace the variable x by n/x in Euler's limit, where $0 < x \in \mathbb{R}$ is **fixed**, and $n \in \mathbb{N}$. We have

$$\left(1+\frac{x}{n}\right)^{n/x} < e, \quad n \in \mathbb{N},$$

and

$$\lim_{n\to\infty}\left(1+\frac{x}{n}\right)^{n/x} = e, \quad 0 < x \in \mathbb{R}.$$

Finally, raising the expressions to the exponent $0 < x \in \mathbb{R}$, we obtain

$$e_n^*(x) = \left(1+\frac{x}{n}\right)^{n} < e^x, \quad n \in \mathbb{N},$$

and

$$\lim_{n\to\infty} e_n^*(x) = \lim_{n\to\infty}\left(1+\frac{x}{n}\right)^{n} = e^x, \quad 0 < x \in \mathbb{R}.$$

This is Euler's representation of the natural exponential function for positive exponents.

To extend this to negative exponents, using again $e^* = e$, we recall

$$e < \left(1+\frac{1}{x}\right)^{x+1} = g(x), \quad x > 0,$$

and

$$\lim_{x\to\infty} g(x) = e.$$

We rework $g(x)$ as

$$g(x) = \left(1+\frac{1}{x}\right)^{x+1} = \left(\frac{x+1}{x}\right)^{x+1} = \left(\frac{x}{x+1}\right)^{-(x+1)} = \left(1-\frac{1}{x+1}\right)^{-(x+1)}.$$

Taking reciprocals, we arrive at

$$\left(1-\frac{1}{x+1}\right)^{x+1} < e^{-1}, \quad 0 < x \in \mathbb{R}.$$

As before, replacing the variable $x+1>1$ by $n/x>1$ with a **fixed** $x>0$ and $x<n$, $n\in\mathbb{N}$, and raising the expressions to the exponent $0<x\in\mathbb{R}$, we obtain

$$e_n^*(-x)=\left(1-\frac{x}{n}\right)^n<e^{-x},\quad 0<x<n,$$

and

$$\lim_{n\to\infty}e_n^*(-x)=\lim_{n\to\infty}\left(1-\frac{x}{n}\right)^n=e^{-x},\quad 0<x\in\mathbb{R}.$$

We conclude that Euler's representation of the natural exponential function also holds for negative exponents. Combining the two representations, we obtain

$$e_n^*(x)=\left(1+\frac{x}{n}\right)^n\le e^x,\quad x>-n,\ n\in\mathbb{N},$$

with equality if and only if $x=0$, and

$$\lim_{n\to\infty}e_n^*(x)=\lim_{n\to\infty}\left(1+\frac{x}{n}\right)^n=e^x,\quad x\in\mathbb{R}.$$

Example 10.5.3 Show that

$$\lim_{n\to\infty}\prod_{k=1}^{n}\left(1-\frac{k}{n^2}\right)=\frac{1}{\sqrt{e}}.$$

Using the fundamental estimate for the natural exponentiation function, for $n\in\mathbb{N}$, we calculate

$$\prod_{k=1}^{n}\left(1-\frac{k}{n^2}\right)\le\prod_{k=1}^{n}e^{-k/n^2}=e^{-(1+2+\cdots+n)/n^2}=e^{-(n+1)/(2n)},$$

where we used $1+2+\cdots+n=n(n+1)/2$.

For the lower bound, we have

$$\begin{aligned}\left(\prod_{k=1}^{n}\left(1-\frac{k}{n^2}\right)\right)^2&=\prod_{k=1}^{n}\left(1-\frac{k}{n^2}\right)\left(1-\frac{n-k+1}{n^2}\right)\\&=\prod_{k=1}^{n}\left(1-\frac{n+1}{n^2}+\frac{k}{n^2}\frac{n-k+1}{n^2}\right)\\&\ge\prod_{k=1}^{n}\left(1-\frac{n+1}{n^2}\right)=\left(1-\frac{n+1}{n^2}\right)^n\\&\ge\left(1-\frac{n+1}{n^2-1}\right)^n=\left(1-\frac{1}{n-1}\right)^n,\end{aligned}$$

where, after the last inequality, we assumed $2\le n\in\mathbb{N}$.

Putting these together, we obtain

$$\left(1-\frac{1}{n-1}\right)^{n/2} \le \prod_{k=1}^{n}\left(1-\frac{k}{n^2}\right) \le e^{-(n+1)/(2n)}, \quad 2 \le n \in \mathbb{N}.$$

Finally, we have

$$\lim_{n\to\infty}\left(1-\frac{1}{n-1}\right)^{n/2} = \lim_{n\to\infty}\left(1-\frac{1}{n-1}\right)^{(n-1)/2}\left(1-\frac{1}{n-1}\right)^{1/2} = \frac{1}{\sqrt{e}},$$

and

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

The example follows.

Remark For $n \in \mathbb{N}$, the expression

$$e_n^*(x) = \left(1+\frac{x}{n}\right)^n$$

is a degree n polynomial. It has a single root at $x = -n$. The change of variable $x \mapsto -x-2n$ results in the $(-1)^n$ multiple of the polynomial. We obtain that, for n odd, the graph of the polynomial is symmetric with respect to the point $(-n, 0)$, and, for n even, it is symmetric with respect to the vertical line $x = -n$. In particular, for n odd, the restriction $x > -n$ of the lower bound for e^x can be removed as the polynomial is negative for $x < -n$.

As a byproduct, the estimates above give polynomial lower bounds for e^x. For example, for $n = 1$, we recover our earlier estimate $1 + x \le e^x$, and, for $n = 2$, we have the new (extended) lower bound

$$1 + x + \frac{x^2}{4} \le e^x, \quad x > -2.$$

Exercises

10.5.1. Derive the limit

$$\lim_{x\to\infty}\left(1+\frac{1}{x^2}\right)^x = 1.$$

10.5.2. Show that

$$\sqrt[n]{e}\frac{n}{e} < \sqrt[n]{n!} < n, \quad n \in \mathbb{N}.$$

10.5.3. Let $n \in \mathbb{N}$. The equation

$$\left(1 + \frac{x}{n}\right)^n = e^x$$

has the trivial solution $x_0 = 0$. Show that, for n odd, this is the only solution, and for n even, there is another solution $x_1 < -n$.

10.5.4. Show directly that the only rational solutions of the equation $y^y = z^z$, $0 < z < 1/e < y < 1$ are the pairs $(1/g(n), 1/f(n))$, $n \in \mathbb{N}$.

10.5.5. Find all positive solutions $0 < x, y \in \mathbb{R}$ of the equation $e^{x+y} = x/y$.

Chapter 11
Trigonometry

"If a pyramid is 250 cubits high and the side of its base 360 cubits long, what is its seqed?"
by Ahmes (c. 1680–1620 BCE)
The Rhind Mathematical Papyrus.

Note: The cubit is an ancient measurement of length; 1 cubit is approximately 18 inches or 457 mm. (The Bible says that Noah's Ark was 300 cubits in length, 50 cubits in width, and 30 cubits in height.) The seqed is an ancient Egyptian term to express the inclination of the triangular face of a pyramid; it is proportional to our reciprocal of the slope or the cotangent of the angle of inclination.

In this chapter we develop trigonometry, the circular analog of arithmetic on the real line. Our treatment has many novel features: explicit algebraic formulas for a large number of special angles using Archimedes' duplication formula discussed in Section 5.9; the Chebyshev polynomials that are used to derive trigonometric identities involving multiple angles; and a thorough discussion on the geometry of triangles, including the concepts of incircle, circumcircle, and Heron's formula (with an extremal property through the AM-GM inequality). One of the highlights of this chapter is Newton's lesser known elementary approach (using means) to derive the power series of the sine and cosine functions well before the advent of the Taylor series. Another highlight is an optional section that contains a complete and (the only) elementary proof of Euler's famous result solving the Basel problem introduced in Section 3.1. Finally, Ptolemy's theorem on cyclic quadrilaterals and its applications finish this chapter.

11.1 The Unit Circle $\mathbb{S}$ vs. the Real Line $\mathbb{R}$

In Section 5.5, we introduced the unit circle $\mathbb{S}$ in the Birkhoff plane $\mathbb{R}^2$ as the set of points that are at unit distance from the origin 0. Recall that, as a simple application of the Cartesian distance formula, a point $P = (x, y)$ is on $\mathbb{S}$ if and only if the

G. Toth, *Elements of Mathematics*, Undergraduate Texts in Mathematics,
https://doi.org/10.1007/978-3-030-75051-0_11

coordinates x and y of P satisfy the equation of the unit circle

$$x^2 + y^2 = 1, \quad (x, y) \in \mathbb{R}^2.$$

A point $P = (x, y)$ on $\mathbb{S}$ determines and is uniquely determined by the **angle measure** $\theta = \mu(\angle E0P)$ of the angle $\angle E0P$, where $E = (1, 0) \in \mathbb{S}$ is on the positive first axis, the initial half-line of the angle, and P is on the half-line with end-point at the origin 0, the terminal half-line of the angle. We say that this angle is in **standard position**.

Remark We define the **positive orientation** of the Birkhoff plane $\mathbb{R}^2$ by setting the positive right angle from the positive first axis to the positive second axis. This corresponds to $\omega(0, E, E') = 1 > 0$, $E = (1, 0)$ and $E' = (0, 1)$, as in Section 5.3.

As discussed in Section 5.7, the angle measure θ of an angle $\angle E0P$ is the arc length of the circular arc in $\mathbb{S}$ with end-points $E = (1, 0)$ and $P = (x, y)$. It is customary to call this angle measure **radian**.

History

Another (classical) measurement of angle is the **degree**, denoted by °. It is defined by the agreement that the full angle of 2π radians is 360°. Thus, to convert degrees to radians amounts to multiplication by $\pi/180°$; in particular, we have $30° = \pi/6, 45° = \pi/4, 60° = \pi/3, 90° = \pi/2, 180° = \pi$, etc.

The origins of the use of degree as a measurement of angle go back to antiquity. It must relate to the early astronomical discovery that the Sun advances every day approximately 1°, giving a rough approximation of the days of the year as 360. With some rare but notable exceptions, as the Persian calendar, most ancient calendars realized that the number of days of the year is actually 365. For example, the ancient Egyptian calendar consisted of 360 regular days (30 days in a month and 10 days in a week) **plus** five Epagomenal days.[1]

The oldest extant Vedic Sanskrit text, the **Rigveda** (c. 1500–c. 1200 BCE), provides a clear evidence that the Indian mathematicians during the Vedic period used the 360 division of the circle: *"... one wheel... On it are placed together three hundred and sixty like pegs."*

The use of the degree may also be related to the Sumerian and Babylonian sexagesimal (base 60) arithmetic, in that a **chord** of length equal to the radius of a circle is also the side length of an equilateral triangle, six of which make up a hexagon inscribed into the circle. Dividing the central angle of a participating triangle into 60 equal parts, one arrives at 1°.

Starting with the early works of Aristarchus of Samos (c. 310–c. 230 BCE) and Hipparchus, the first extant records of the use of degree appear in the works of Timocharis of Alexandria (c. 320–c. 360 BCE), Aristillus (c. 261 BCE) of Timocharis' School, and Archimedes.

We quickly note that an angle measure θ associated with a point $P \in \mathbb{S}$ is determined only up to an additive integer multiple of 2π, that is, $\theta + 2n\pi$, $n \in \mathbb{Z}$, correspond to the same point P. These angles are called **coterminal angles**. This non-uniqueness of the angle is also clear from the non-uniqueness of the circular arc in $\mathbb{S}$ connecting E and P. In fact, depending on how many times and in what direction we wind around $\mathbb{S}$, there are infinitely many such circular arcs parametrized by the set of integers.

Note that choosing the shortest among all these arcs does not solve the problem of non-unicity for several reasons. For example, the shortest arc for $P = (-1, 0)$

[1] Also in the Coptic, pre-Columbian, etc. calendars.

is not unique, and, in the third and fourth quadrants, the shortest circular arcs are usually given by negative angle measures.

Some special or common angle measures recur in various fields of mathematics and sciences. A few of these are $\pi/6$, $\pi/4$, $\pi/3$, $\pi/2$ and their multiples. These angles in standard position intercept a circular arc on the unit circle from $E = (1, 0)$ to particular points $P = (x, y)$. Simple geometry can be used to find the points P associated with these angles.

For the angle $\pi/3$ in standard position, we first note that the triangle $\triangle[0, E, P]$ is equilateral. Thus, the perpendicular bisector through P bisects the opposite side $[0, E]$ at a point M. This immediately gives the first coordinate of P as $1/2$. The second can be obtained by the Pythagorean theorem applied to the right triangle $\triangle[0, M, P]$. We obtain that the second coordinate of P is $\sqrt{1 - (1/2)^2} = \sqrt{3}/2$. With these, we have $P = (1/2, \sqrt{3}/2)$. Finally, since the triangle $\triangle[O, M, P]$ has unit hypotenuse, as a byproduct, we also obtain that the point for the angle $\pi/6$ in standard position is $(\sqrt{3}/2, 1/2)$.

The terminal side of the angle $\pi/4$ in standard position is given by the equation $y = x$. Therefore, we have $2x^2 = 2y^2 = 1$ so that the associated point is $(\sqrt{2}/2, \sqrt{2}/2)$. Moreover, since the terminal side of the angle $\pi/2$ is the positive second axis, the point corresponding to the right angle is $(0, 1)$.

Exercise

11.1.1. Given a rectangle $[A, B, C, D]$ with $d(A, B)/d(B, C) = 2$, we let $E \in [A, B]$ such that $\mu(\angle BCE) = \pi/12$. Show that the triangle $\triangle[C, D, E]$ is isosceles.

11.2 The Sine and Cosine Functions

As in the previous section, let θ be an angle measure associated with the point $P \in \mathbb{S}$ on the unit circle. The coordinates x and y of P are functions of θ. We use these to define the **cosine** and **sine** functions $\cos : \mathbb{R} \to \mathbb{R}$ and $\sin : \mathbb{R} \to \mathbb{R}$ by

$$x = \cos(\theta) \quad \text{and} \quad y = \sin(\theta).$$

This definition immediately implies that the range of both the cosine and sine functions is the closed interval $[-1, 1]$.

History

The earliest possible attestation of a trigonometric table is in the Babylonian clay tablet, Plimpton 322, already noted in Section 5.7 for its relation to Pythagorean triples. The tablet itself is a matrix of four columns and fifteen rows filled with numeral entries in Babylonian sexagesimal notation (see Figure 11.1). The numbers in the second column can be interpreted as the shortest sides of

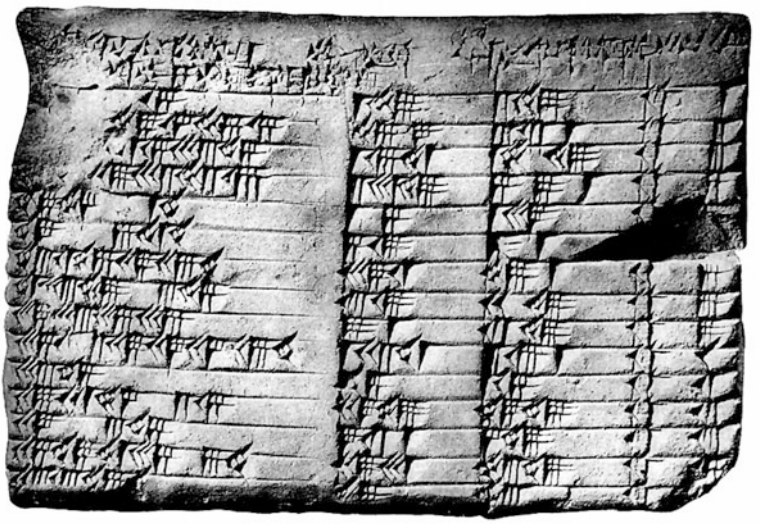


Fig. 11.1 The Plimpton 322 Babylonian clay tablet. G.A. Plimpton Collection of Columbia University.

a right triangle, and the numbers in the third column may be the hypotenuses of the respective triangles. A possible trigonometric explanation of the numbers in the first column is that they are the squared secants or tangents of the respective angles opposite to the shortest sides. If this is a valid interpretation, then the entries in the fifteen rows roughly correspond in one degree increments 15 secants or tangents between 35° and 45°.
Hipparchus is believed to be the first mathematician who had a "chord table," a trigonometric table of chords of a circle subtended by central angles. He used this table to calculate the eccentricity of the orbits of the Moon and the Sun.
Aryabhatta (476–550 CE) was an Indian mathematician who composed what is known as the Āryabhaṭīya Sine Table. Actually, this is not a table arranged in a matrix form, rather a set of 24 numbers that represent the first differences of the values of trigonometric sines expressed in arcminutes. A lesser known fact is that about a century later (in 629), in his commentary Āryabhaṭīyabhāṣya to the Āryabhaṭīya, Bhāskara I gave very accurate rational approximations to the sine function. (This latter work is also significant because it is one of the oldest extant works in Sanskrit on mathematics and astronomy. Compare this with the historical note on the much older Rigveda in the Vedic period above.)

Example 11.2.1 Find the sine and cosine of the angle measures $\pi/6$, $\pi/4$, $\pi/3$, $\pi/2$.

Using the points found in the previous section, we have $\sin(\pi/6) = \cos(\pi/3) = 1/2$, $\sin(\pi/3) = \cos(\pi/6) = \sqrt{3}/2$, $\sin(\pi/4) = \cos(\pi/4) = \sqrt{2}/2$, and $\sin(\pi/2) = 1, \cos(\pi/2) = 0$.

Since an angle measure is determined only up to an additive integer multiple of 2π, it follows from our definition that the cosine and sine functions are **periodic** with period 2π:

$$\cos(\theta + 2n\pi) = \cos(\theta) \quad \text{and} \quad \sin(\theta + 2n\pi) = \sin(\theta), \quad n \in \mathbb{Z}.$$

Reflecting the point $P = (x, y)$ to the first axis, we obtain the point $P' = (x, -y)$ with the angle measure θ changing to its negative $-\theta$. We thus obtain the so-called **even-odd identities**

$$\cos(-\theta) = \cos(\theta) \quad \text{and} \quad \sin(-\theta) = -\sin(\theta), \quad \theta \in \mathbb{R}.$$

The name comes from the fact that these identities assert that cosine is an even function and sine is an odd function.

Note finally that the cosine and sine functions are not independent since the equation of the unit circle $x^2 + y^2 = 1$ gives

$$\cos^2(\theta) + \sin^2(\theta) = 1, \quad \theta \in \mathbb{R}.$$

This is called the **Pythagorean Identity** for cosine and sine, since the equation of the unit circle and, more generally, the Cartesian distance formula are equivalent to the Pythagorean theorem.

An equivalent and more geometric definition of the cosine and sine functions is to consider a right triangle $\triangle[A, B, C]$ with (acute) angle measure θ at the vertex A and right angle at the vertex C.[2] Since the sum of the angle measures in any triangle is π, the angle at B has complementary angle measure $\pi/2 - \theta$. Letting a, b, c be the side lengths of the sides opposite to A, B, C, we define[3]

$$\cos(\theta) = \frac{b}{c} \quad \text{and} \quad \sin(\theta) = \frac{a}{c}.$$

Now the crux is that these ratios depend only on θ (and not on the specific triangle chosen) since any other triangle with the same angles is similar to this, and, by the Birkhoff Postulate of Similarity, the ratios of the corresponding side lengths are equal.

If the right triangle is constructed within the unit circle (with the length of the hypotenuse equal to the radius) and with θ in standard position, then these definitions reduce to the previous definition of sine and cosine.

Swapping the roles of A and B, we immediately obtain the identities for complementary angles

$$\cos(\theta) = \sin\left(\frac{\pi}{2} - \theta\right) \quad \text{and} \quad \sin(\theta) = \cos\left(\frac{\pi}{2} - \theta\right), \quad \theta \in \mathbb{R}.$$

A slight drawback of the geometric definition above is that it defines the cosine and sine functions only for an acute angle $0 < \theta < \pi/2$. There are several (analytic

[2]Here, by standard practice, we briefly abandon our convention to list the vertices of a (non-degenerate) triangle $\triangle[A, B, C]$ such that (A, B, C) is positively oriented.

[3]Recall our convention $a = d(B, C), b = d(C, A), c = d(A, B)$.

and geometric) ways to extend these definitions for all $\theta \in \mathbb{R}$. We have chosen our initial definition to avoid this problem.

We now introduce a more powerful way of evaluating sine and cosine on specific angles. Let $\mathcal{P}_n$, $n \geq 3$, be a regular n-sided polygon inscribed in $\mathbb{S}$. Denote by l_n the half of the side length of $\mathcal{P}_n$. Recall from Section 5.9 Archimedes' duplication formula:

$$l_{2n} = \sqrt{\frac{1 - \sqrt{1 - l_n^2}}{2}}.$$

The central angle with vertex at the origin 0 subtended by a side of $\mathcal{P}_n$ is $2\pi/n$. We obtain that $l_n = \sin(\pi/n)$. These formulas allow us to calculate the sine (and cosine) of many special angles.

First, for $n = 4$, we have the square inscribed in $\mathbb{S}$ with diagonal length 2. The Pythagorean theorem gives $l_4 = \sin(\pi/4) = \sqrt{2}/2$.

For $n = 8$, using the duplication formula, we calculate

$$l_8 = \sin\left(\frac{\pi}{8}\right) = \sqrt{\frac{1 - \sqrt{1 - l_4^2}}{2}} = \sqrt{\frac{1 - \sqrt{1 - \frac{1}{2}}}{2}} = \sqrt{\frac{1 - \frac{1}{\sqrt{2}}}{2}} = \frac{\sqrt{2 - \sqrt{2}}}{2}.$$

Continuing, a similar computation gives

$$l_{16} = \sin\left(\frac{\pi}{16}\right) = \sqrt{\frac{1 - \sqrt{1 - l_8^2}}{2}} = \frac{\sqrt{2 - \sqrt{2 + \sqrt{2}}}}{2}.$$

We now see the general pattern as

$$l_{2^n} = \sin\left(\frac{\pi}{2^n}\right) = \frac{\sqrt{2 - \sqrt{2 + \sqrt{2 + \cdots + \sqrt{2}}}}}{2}$$

with $n - 1$ nested square roots.[4]

The Pythagorean identity gives the respective values of the cosine function

$$\cos\left(\frac{\pi}{2^n}\right) = \sqrt{1 - \sin^2\left(\frac{\pi}{2^n}\right)} = \frac{\sqrt{2 + \sqrt{2 + \sqrt{2 + \cdots + \sqrt{2}}}}}{2}$$

[4] It is a standard problem for the use of the ratio test (Section 3.4) to sum the infinite series $\sqrt{2} + \sqrt{2 - \sqrt{2}} + \sqrt{2 - \sqrt{2 + \sqrt{2}}} + \cdots$ (without mentioning the trigonometric formula above). This series can then be written as $2\sum_{n=2}^{\infty} \sin(\pi/2^n) < 2\sum_{n=2}^{\infty} \pi/2^n = \pi$, where we used the standard inequality for the sine function (Section 11.5) along with the Infinite Geometric Series Formula.

with $n-1$ nested square roots.[5]

Another sequence starts with $n=6$. Since a hexagon is made up of six equilateral triangles, we have

$$l_6 = \sin\left(\frac{\pi}{6}\right) = \cos\left(\frac{\pi}{3}\right) = \frac{1}{2},$$

where $\pi/6$ and $\pi/3$ are complementary angles. The duplication formula now gives

$$l_{12} = \sin\left(\frac{\pi}{12}\right) = \frac{\sqrt{2-\sqrt{3}}}{2}.$$

Continuing this, we arrive at the general formula

$$l_{3\cdot 2^n} = \sin\left(\frac{\pi}{3\cdot 2^n}\right) = \frac{\sqrt{2-\sqrt{2+\sqrt{2+\cdots+\sqrt{2+\sqrt{3}}}}}}{2}$$

with n nested square roots.

Once again the Pythagorean identity gives the respective values of the cosine

$$\cos\left(\frac{\pi}{3\cdot 2^n}\right) = \frac{\sqrt{2+\sqrt{2+\sqrt{2+\cdots+\sqrt{2+\sqrt{3}}}}}}{2}$$

with n nested square roots.

We now turn to a geometric description of the graphs of the sine and cosine functions.

The natural space for the graphs of the sine and cosine functions is the Cartesian product $\mathbb{S}\times\mathbb{R}$. This is a (vertical) cylinder in the 3-dimensional space $\mathbb{R}^3$ since $\mathbb{S}\times\mathbb{R}\subset\mathbb{R}^2\times\mathbb{R}=\mathbb{R}^3$. In terms of the Cartesian coordinates $(x,y,z)\in\mathbb{R}^3$, the equation $z=x$ is the plane that subtends $\pi/4$ angle with the third axis and intersects the coordinate plane spanned by the first two axes (given by $z=0$) in the second coordinate axis. This plane further intersects the cylinder in an ellipse. For $P=(x,y)=(\cos(\theta),\sin(\theta))\in\mathbb{S}$, the point on this ellipse above P has elevation $z=x=\cos(\theta)$, so this ellipse is the **graph of the cosine function** in the cylinder $\mathbb{S}\times\mathbb{R}$.

In a similar vein, the plane $z=y$ subtends $\pi/4$ angle with the third axis and cuts an ellipse out of the cylinder $\mathbb{S}\times\mathbb{R}$. This is the graph of the sine function.

Since these ellipses subtend $\pi/4$ angle with the third axis, we now see that the cosine and sine functions play the same roles in trigonometry as the identity function $y=x$

[5] See this also in the Kettering University Mathematics Olympiad, 2009.

for arithmetic on the real line $\mathbb{R}$. In particular, akin to forming polynomials in the indeterminate x, we can also form **trigonometric polynomials** in the indeterminates $\cos(\theta)$ and $\sin(\theta)$.

A byproduct of these constructions is the pair of identities

$$\sin\left(\theta + \frac{\pi}{2}\right) = \cos(\theta) \quad \text{and} \quad \cos\left(\theta + \frac{\pi}{2}\right) = -\sin(\theta), \quad \theta \in \mathbb{R},$$

since the two ellipses can be rotated into each other by a quarter turn about the third axis. (Note that replacing θ by $-\theta$ in our earlier "swapping" identities and appealing to the even-odd properties of cosine and sine, the identities above also follow.)

Finally, note that rolling out the cylinder to the plane $\mathbb{R}^2$ corresponds to the transformation $(\cos(\theta), \sin(\theta), z) \mapsto (\theta, z)$, and the two ellipses are mapped to the graphs of the cosine and sine functions on the plane $\mathbb{R}^2$.

While the sine and cosine functions are not one-to-one on $\mathbb{R}$, we can restrict them to suitable domains to find **inverses**.

First, the cosine function is strictly decreasing on the closed interval $[0, \pi]$ and gives a one-to-one correspondence $\cos : [0, \pi] \to [-1, 1]$. We use this branch of the cosine function to define the inverse $\cos^{-1} : [-1, 1] \to [0, \pi]$. This inverse is traditionally called the **arccosine** function and denoted by arccos.

Second, we can restrict the sine function to the interval $[-\pi/2, \pi/2]$ on which it is strictly increasing. We then obtain the inverse sine or **arcsine** function $\sin^{-1} = \arcsin : [-1, 1] \to [-\pi/2, \pi/2]$.

Remark The names "arccosine" and "arcsine" come from the fact that the domain variable for the cosine and sine functions is an angle, the length of the respective circular arc on the unit circle. Inverting, this arc length becomes the range variable.

Example 11.2.2 Calculate arcsin and arccos of $1/2$.

To determine $\arcsin(1/2)$, we need to find the angle $\theta \in [-\pi/2, \pi/2]$ such that $\arcsin(1/2) = \theta$, that is, $\sin(\theta) = 1/2$. From our earlier computations, we find that this angle is $\theta = \pi/6$. Therefore, we have $\arcsin(1/2) = \pi/6$. Similarly, solving $\cos(\theta) = 1/2$ with $\theta \in [0, \pi]$, we obtain $\arccos(1/2) = \pi/3$.

Remark The domains and ranges of the inverse functions are restricted. For example, while $\sin(5\pi/6) = 1/2$, this does not mean that $\arcsin(1/2) = 5\pi/6$ since $5\pi/6$ is not in the range of the arcsine function.

Example 11.2.3 Determine the domains and algebraic representations of the compositions $\cos \circ \arcsin$ and $\sin \circ \arccos$.

For the first composition $\cos \circ \arcsin$, the domain of the cosine function is $\mathbb{R}$, but the arcsine function has a domain of $[-1, 1]$. Therefore the domain of the composition $\cos \circ \arcsin$ is $[-1, 1]$. This is also the case for the composition $\sin \circ \arccos$.

Turning to the algebraic representation of $\cos \circ \arcsin$, we let $\arcsin(x) = \theta$, or equivalently, $\sin(\theta) = x$, with $x \in [-1, 1]$ and $\theta \in [-\pi/2, \pi/2]$. Since cosine is an even function, we may assume that $\theta > 0$, and therefore θ is an acute angle. We

then construct a right triangle with angle θ, side length opposite to θ equal to x, and hypotenuse equal to 1. The Pythagorean theorem gives the third side as $\sqrt{1-x^2}$. Thus, we have $\cos(\theta) = \sqrt{1-x^2}$. We conclude that $(\cos \circ \arcsin)(x) = \cos(\theta) = \sqrt{1-x^2}$.

A similar procedure gives $(\sin \circ \arccos)(x) = \sqrt{1-x^2}$ with $x \in [-1, 1]$.

Exercises

11.2.1. Let $S = \mathbb{S}_{(a,b)}$ be a unit circle with center $(a, b) \in \mathbb{R}^2$ such that $a^2 + b^2 > 1$ (that is, the origin $(0, 0)$ is exterior to S). What is the shortest path from $(0, 0)$ to the point $(2a, 2b)$ avoiding S?

11.2.2. Determine how to split the unit square into two rectangles such that one can be inscribed into the other in a tilted position (with its vertices on the respective sides of the other).

11.3 Principal Identities for Sine and Cosine

The pair of identities in the previous section raises the following question: Are there general identities expressing the cosine and sine of the sum of two angles in terms of the cosine and sine of the angles themselves? The answer is "yes," and we now proceed to derive these so-called **trigonometric addition formulas**.

Let $\alpha, \beta \in \mathbb{R}$. We denote $P = (\cos(\alpha), \sin(\alpha))$, $Q = (\cos(\beta), \sin(\beta))$, and $R = (\cos(\alpha-\beta), \sin(\alpha-\beta))$, three points on $\mathbb{S}$ with respective angle measures α, β, and $\alpha-\beta$. By the Birkhoff Postulate of Similarity, the isosceles triangles $\triangle[0, P, Q]$ and $\triangle[0, E, R]$ with $E = (1, 0)$ are congruent since their angle measures at 0 are the same $(\alpha - \beta)$. Thus, we have $d(P, Q)^2 = d(E, R)^2$. The Cartesian distance formula gives

$$(\cos(\alpha) - \cos(\beta))^2 + (\sin(\alpha) - \sin(\beta))^2 = (\cos(\alpha - \beta) - 1)^2 + \sin(\alpha - \beta)^2.$$

Expanding, and using the Pythagorean identity three times, we obtain

$$2 - 2\cos(\alpha)\cos(\beta) - 2\sin(\alpha)\sin(\beta) = 2 - 2\cos(\alpha - \beta).$$

Simplifying, we arrive at the identity

$$\cos(\alpha - \beta) = \cos(\alpha)\cos(\beta) + \sin(\alpha)\sin(\beta) \quad \alpha, \beta \in \mathbb{R}.$$

Replacing β by its negative and using the even-odd identities, the identity above immediately gives

$$\cos(\alpha+\beta)=\cos(\alpha)\cos(\beta)-\sin(\alpha)\sin(\beta)\quad \alpha,\beta\in\mathbb{R}.$$

Finally, translating by $\pi/2$, we calculate

$$\begin{aligned}\sin(\alpha+\beta)&=-\cos\left(\alpha+\beta+\frac{\pi}{2}\right)\\&=-\cos\left(\alpha+\frac{\pi}{2}\right)\cos(\beta)+\sin\left(\alpha+\frac{\pi}{2}\right)\cos(\beta)\\&=\sin(\alpha)\cos(\beta)+\cos(\alpha)\sin(\beta).\end{aligned}$$

Once again replacing β by its negative, we obtain

$$\sin(\alpha-\beta)=\sin(\alpha)\cos(\beta)-\cos(\alpha)\sin(\beta).$$

We summarize that the **addition formulas** for sine and cosine are as follows:

$$\begin{aligned}\cos(\alpha+\beta)&=\cos(\alpha)\cos(\beta)-\sin(\alpha)\sin(\beta)\\\cos(\alpha-\beta)&=\cos(\alpha)\cos(\beta)+\sin(\alpha)\sin(\beta)\\\sin(\alpha+\beta)&=\sin(\alpha)\cos(\beta)+\cos(\alpha)\sin(\beta)\\\sin(\alpha-\beta)&=\sin(\alpha)\cos(\beta)-\cos(\alpha)\sin(\beta).\end{aligned}$$

The following example is a simple application of the addition formula for sine:

Example 11.3.1 Let $a,b\in\mathbb{R}$, $a^2+b^2>0$. Write[6] $a\sin\alpha+b\cos\alpha$ as an expression involving a single sine.

Let $P=(a/\sqrt{a^2+b^2},b/\sqrt{a^2+b^2})\in\mathbb{R}^2$. Then $P\in\mathbb{S}$ so that $P=(\cos\beta,\sin\beta)$ for some $\beta\in\mathbb{R}$. With this, we obtain

$$a\sin\alpha+b\cos\alpha=\sqrt{a^2+b^2}\,(\sin\alpha\cos\beta+\cos\alpha\sin\beta)=\sqrt{a^2+b^2}\sin(\alpha+\beta).$$

The Cauchy–Schwarz inequality can be combined with trigonometric identities to obtain new trigonometric inequalities. The following is a simple example of this.

Example 11.3.2 For $0<\alpha,\beta<\pi/2$, $\alpha,\beta\in\mathbb{R}$, show that

$$\frac{\cos^3\alpha}{\cos\beta}+\frac{\sin^3\alpha}{\sin\beta}\geq\frac{1}{\cos(\alpha-\beta)}.$$

To show this, we first note that on the domain $(0,\pi/2)$ both cosine and sine are positive. The Cauchy–Schwarz inequality gives

[6] To simplify the notation, whenever convenient, we suppress the parentheses in $\sin(\alpha)$ and $\cos(\alpha)$, etc.

$$\left(\frac{\cos^3\alpha}{\cos\beta}+\frac{\sin^3\alpha}{\sin\beta}\right)\cdot(\cos\alpha\cos\beta+\sin\alpha\sin\beta)\geq\left(\cos^2\alpha+\sin^2\alpha\right)^2=1.$$

Using the addition formula for cosine, the inequality now follows.

Example 11.3.3 We have

$$\arcsin x\pm\arcsin y=\arcsin\left(x\sqrt{1-y^2}\pm y\sqrt{1-x^2}\right)$$
$$\arccos x\pm\arccos y=\arccos\left(xy\mp\sqrt{(1-x^2)(1-y^2)}\right).$$

These formulas are direct consequences of the addition formulas for the sine and cosine functions. For the first formula, we calculate

$$\begin{aligned}\sin(\arcsin x\pm\arcsin y)&=\sin(\arcsin x)\cos(\arcsin y)\pm\cos(\arcsin x)\sin(\arcsin y)\\&=x\sqrt{1-y^2}\pm y\sqrt{1-x^2},\end{aligned}$$

where we used the results of Example 11.2.3.

The second formula can be derived in a similar way.

Setting $\alpha=\beta$ in the addition formulas, we obtain the so-called **double angle formulas**

$$\begin{aligned}\cos(2\alpha)&=\cos^2(\alpha)-\sin^2(\alpha)=1-2\sin^2(\alpha)=2\cos^2(\alpha)-1\\\sin(2\alpha)&=2\cos(\alpha)\sin(\alpha),\end{aligned}$$

where in the first equality we used the Pythagorean identity and gave two alternatives.

The first equality gives the **power reducing formulas**

$$\cos^2(\alpha)=\frac{1+\cos(2\alpha)}{2}\quad\text{and}\quad\sin^2(\alpha)=\frac{1-\cos(2\alpha)}{2}.$$

Replacing α by its half, we arrive at the **half angle formulas**

$$\cos^2\left(\frac{\alpha}{2}\right)=\frac{1+\cos(\alpha)}{2}\quad\text{and}\quad\sin^2\left(\frac{\alpha}{2}\right)=\frac{1-\cos(\alpha)}{2}.$$

(We did not take the square roots of both sides on purpose as they depend on the sign of the cosine and sine of the half angle. Note that these half angle formulas can also be used instead of Archimedes' duplication formula to obtain the root formulas for the sine and cosine of the special angles in Section 11.2.)

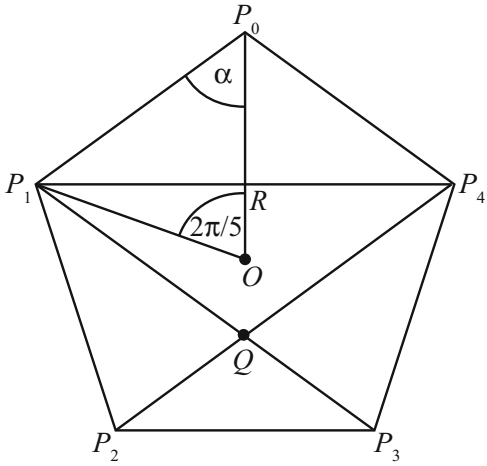


Fig. 11.2 The regular pentagon and the golden number.

History
The addition formulas for sine and cosine were discovered by the Persian mathematician Abū al-Wafā' Būzjānī (940–997/998 CE).

Example 11.3.4 (Regular Pentagon and the Golden Number) Consider a regular pentagon with vertices P_0, P_1, P_2, P_3, P_4 (see Figure 11.2). By scaling, we may assume that the side length of the pentagon is unity. Let Q be the intersection of the diagonal line segments $[P_1, P_3]$ and $[P_2, P_4]$. Since a diagonal of a regular pentagon is always parallel to one of its sides, we see that the quadrilateral with vertices P_0, P_1, Q, P_4 is a rhombus. Thus, we have $d(P_1, Q) = d(P_4, Q) = 1$. Define $\tau = d(P_1, P_4)$. (We will see shortly that this is the golden number (Example 3.1.2), so that this notation will be justified.) Clearly, the isosceles triangles $\triangle[P_1, Q, P_4]$ and $\triangle[P_2, Q, P_3]$ are similar. By Birkhoff's Postulate of Similarity, we have

$$\frac{d(P_1, Q)}{d(P_3, Q)} = \frac{d(P_1, P_4)}{d(P_2, P_3)}.$$

Substituting the known quantities, we obtain $d(P_3, Q) = 1/\tau$. On the other hand, $d(P_1, P_3) = d(P_1, Q) + d(Q, P_3)$ so that $\tau = 1 + 1/\tau$. We see that τ is the golden number $\tau = (1 + \sqrt{5})/2$.

We now change the settings, and let O be the center of the pentagon. The central angle $\angle P_0 O P_1$ has measure $2\pi/5$. Since the sum of the (interior) angles in a triangle is equal to π, we obtain $\alpha = \mu(\angle P_1 P_0 O) = \pi/2 - \pi/5$. Let the radial segment $[O, P_0]$ intersect with the diagonal $[P_1, P_4]$ at the point R. Then the triangle $\triangle[P_0, P_1, R]$ has right angle at R and we obtain $\sin(\alpha) = \tau/2$. Substituting the value of α, we obtain

$$\sin\left(\frac{\pi}{2} - \frac{\pi}{5}\right) = \cos\left(\frac{\pi}{5}\right) = \frac{\tau}{2}.$$

The root formula for the golden number now gives

$$\cos\left(\frac{\pi}{5}\right) = \frac{1+\sqrt{5}}{4}.$$

Using the Pythagorean identity, we also obtain

$$\sin\left(\frac{\pi}{5}\right) = \frac{\sqrt{10-2\sqrt{5}}}{4}.$$

As an application, we derive a root formula for $\sin(\pi/10)$. Using the half angle formula for sine, we calculate

$$\sin^2\left(\frac{\pi}{10}\right) = \frac{1-\cos\left(\frac{\pi}{5}\right)}{2} = \frac{1-\frac{1+\sqrt{5}}{4}}{2} = \frac{3-\sqrt{5}}{8} = \left(\frac{\sqrt{5}-1}{4}\right)^2.$$

Thus, we have

$$\sin\left(\frac{\pi}{10}\right) = \frac{\sqrt{5}-1}{4}.$$

A somewhat more advanced exercise using trigonometry developed so far is the following:

Example 11.3.5 Let $n \in \mathbb{N}$. Show that

$$\prod_{j=0}^{n-1} \cos(2^j\alpha) = \frac{\sin(2^n\alpha)}{2^n \sin\alpha}.$$

We proceed with Peano's Principle of Induction with respect to $n \in \mathbb{N}$. For $n = 1$, we have

$$\cos(\alpha) = \frac{\sin(2\alpha)}{2\sin\alpha}.$$

This is the double angle formula for sine. For the general induction step $n \Rightarrow n+1$, we use the induction hypothesis and calculate

$$\prod_{j=0}^{n} \cos(2^j\alpha) = \prod_{j=0}^{n-1} \cos(2^j\alpha)\cdot\cos(2^n\alpha) = \frac{\sin(2^n\alpha)}{2^n\sin\alpha}\cdot\cos(2^n\alpha)$$

$$= \frac{2\sin(2^n\alpha)\cos(2^n\alpha)}{2^{n+1}\sin\alpha} = \frac{\sin(2^{n+1}\alpha)}{2^{n+1}\sin\alpha},$$

where we used the double angle formula for the sine. The induction is complete, and the example follows.

Another direct consequence of the addition formulas for cosine and sine is the set of so-called **product to sum** formulas:

$$\begin{aligned} 2\cos\alpha\cos\beta &= \cos(\alpha-\beta)+\cos(\alpha+\beta)\\ 2\sin\alpha\cos\beta &= \sin(\alpha+\beta)+\sin(\alpha-\beta)\\ 2\sin\alpha\sin\beta &= \cos(\alpha-\beta)-\cos(\alpha+\beta). \end{aligned}$$

These come handy at times in computations as the following example shows:

Example 11.3.6 Let $n \in \mathbb{N}_0$. Show that

$$\sum_{k=0}^{n}\sin(\alpha+k\beta) = \sin\alpha+\sin(\alpha+\beta)+\cdots+\sin(\alpha+n\beta) = \frac{\sin\left(\alpha+\frac{n\beta}{2}\right)\sin\frac{(n+1)\beta}{2}}{\sin\frac{\beta}{2}}$$

$$\sum_{k=0}^{n}\cos(\alpha+k\beta) = \cos\alpha+\cos(\alpha+\beta)+\cdots+\cos(\alpha+n\beta) = \frac{\cos\left(\alpha+\frac{n\beta}{2}\right)\sin\frac{(n+1)\beta}{2}}{\sin\frac{\beta}{2}}.$$

We derive only the first formula; the proof of the second formula is analogous. We proceed by induction with respect to $n \in \mathbb{N}$. The initial case $n = 0$ is a tautology. Using the induction hypothesis in the general induction step $n-1 \Rightarrow n$, we need to show

$$\sin\left(\alpha+\frac{(n-1)\beta}{2}\right)\sin\frac{n\beta}{2}+\sin(\alpha+n\beta)\sin\frac{\beta}{2} = \sin\left(\alpha+\frac{n\beta}{2}\right)\sin\frac{(n+1)\beta}{2}.$$

Using the last product to sum formula for each of the three products, all terms cancel, so that equality holds. The example follows.

Since the cosine and sine functions play dual roles, we define a **trigonometric polynomial** as an expression $p(\cos(\theta), \sin(\theta))$, where $p(x, y)$ is a polynomial in the indeterminates x and y.

For example, the right-hand sides in the double angle formulas are trigonometric polynomials: $2x^2-1$ and $2xy$ with $x = \cos(\alpha)$ and $y = \sin(\alpha)$.
Using these, we derive the triple angle formulas for cosine and sine as follows:

$$\begin{aligned} \cos(3\alpha) &= \cos(2\alpha+\alpha) = \cos(2\alpha)\cos(\alpha)-\sin(2\alpha)\sin(\alpha)\\ &= (2\cos^2(\alpha)-1)\cos(\alpha)-2\cos(\alpha)\sin^2(\alpha)\\ &= 2\cos^3(\alpha)-\cos(\alpha)-2\cos(\alpha)(1-\cos^2(\alpha)) = 4\cos^3(\alpha)-3\cos(\alpha). \end{aligned}$$

In a similar vein, we calculate

$$\begin{aligned}\sin(3\alpha) &= \sin(2\alpha+\alpha) = \sin(2\alpha)\cos(\alpha) + \cos(2\alpha)\sin(\alpha)\\ &= 2\cos^2(\alpha)\sin(\alpha) + (2\cos^2(\alpha) - 1)\sin(\alpha)\\ &= (4\cos^2(\alpha) - 1)\sin(\alpha) = -4\sin^3(\alpha) + 3\sin(\alpha),\end{aligned}$$

where in the last equality we used the Pythagorean identity.
Summarizing, we have the triple angle formulas

$$\begin{aligned}\cos(3\alpha) &= 4\cos^3(\alpha) - 3\cos(\alpha) = 4x^3 - 3x\\ \sin(3\alpha) &= -4\sin^3(\alpha) + 3\sin(\alpha) = -4y^3 + 3y.\end{aligned}$$

Example 11.3.7 Derive the identity

$$\frac{\sin(3\alpha)}{\sin(\alpha)} - \frac{\cos(3\alpha)}{\cos(\alpha)} = 2, \quad \alpha \neq k\frac{\pi}{2},\ k \in \mathbb{Z}.$$

Indeed, by the triple angle formulas, we have

$$\begin{aligned}\frac{\sin(3\alpha)}{\sin(\alpha)} - \frac{\cos(3\alpha)}{\cos(\alpha)} &= -4\sin^2(\alpha) + 3 - 4\cos^2(\alpha) + 3\\ &= -4(\sin^2(\alpha) + \cos^2(\alpha)) + 6 = -4 + 6 = 2,\end{aligned}$$

where we used the Pythagorean identity.

Note that another way of solving this problem is to represent the trigonometric expressions in α as polynomial expressions in x and y and use $x^2 + y^2 = 1$.

We now digress from the main line and show yet another application of the triple angle formula for cosine, to find the roots of a cubic polynomial (Section 7.2). More specifically, recall that if, for the critical expression, we have

$$\left(\frac{q}{2}\right)^2 + \left(\frac{p}{3}\right)^3 < 0,$$

then the reduced cubic equation

$$x^3 + px + q = 0$$

has three real roots, but our cubic formula gives them only in complex form.

The novel idea here, due to François Viète, is to compare the reduced cubic equation with the triple angle formula written in the following form:

$$4\cos^3(\theta) - 3\cos(\theta) - \cos(3\theta) = 0.$$

Letting $x = u\cos(\theta)$ with $u = 2\sqrt{-p/3}$ (note that, due to our assumption on the critical expression, $p < 0$), our reduced cubic equation takes the form

$$4\cos^3(\theta) - 3\cos(\theta) + 4q/u^3 = 0,$$

since $4p/u^2 = -3$. For the constant term, we have

$$\frac{4q}{u^3} = \frac{q}{2\sqrt{-\left(\frac{p}{3}\right)^3}} = \frac{3q}{2p}\sqrt{-\frac{3}{p}}.$$

Matching this with the triple angle formula, we obtain

$$3\theta = \arccos\left(\frac{3q}{2p}\sqrt{-\frac{3}{p}}\right).$$

(Note that our assumption on the critical expression implies that the argument in arccos is in $[-1, 1]$, so that it is well-defined.) Since $= u\cos(\theta)$, this gives the three **real** solutions of our reduced cubic as follows:

$$x = 2\sqrt{-\frac{p}{3}}\cos\left(\frac{1}{3}\arccos\left(\frac{3q}{2p}\sqrt{-\frac{3}{p}}\right) + \frac{2k\pi}{3}\right) \quad k = 0, 1, 2,$$

where we incorporated the periodicity with an integer multiple of 2π.

Remark As yet another application of the triple angle formula for cosine, letting $\alpha = \pi/9$, we have

$$\cos\left(\frac{\pi}{3}\right) = 4\cos^3\left(\frac{\pi}{9}\right) - 3\cos\left(\frac{\pi}{9}\right).$$

Since $\cos(\pi/3) = 1/2$, we obtain that $\cos(\pi/9)$ is a root of the cubic equation

$$8x^3 - 6x - 1 = 0.$$

We encountered this in Example 7.4.5. Recall that, according to the Rational Root Theorem (Section 7.4), the possible rational roots are $\pm 1, \pm 1/2, \pm 1/4, \pm 1/8$. Upon substituting, none of these solve the cubic equation. (In particular, this cubic is irreducible over $\mathbb{Q}$.) We conclude that $\cos(\pi/9)$ is an **irrational** number. Since it is a root of an **irreducible cubic** polynomial, it follows by a somewhat advanced algebraic reasoning that $\pi/9$ is not constructible (as the length of a line segment) by straightedge and compass. Since this is $1/3$ of the constructible $\pi/6$, we see that there is no geometric construction by straightedge and compass to **trisect an arbitrary angle**.

Example 11.3.8 Show that

$$8 \cdot \cos\left(\frac{\pi}{9}\right) \cdot \cos\left(\frac{2\pi}{9}\right) \cdot \cos\left(\frac{4\pi}{9}\right) = 1.$$

We have seen in the previous example that $\cos(\pi/9)$ is an irrational root of the polynomial equation $8x^3 = 6x + 1$. Letting $x = \cos(\pi/9)$ and using the double angle formula for the cosine function, we have

$$\cos\left(\frac{2\pi}{9}\right) = 2\cos\left(\frac{\pi}{9}\right)^2 - 1 = 2x^2 - 1$$

$$\cos\left(\frac{4\pi}{9}\right) = 2\cos\left(\frac{2\pi}{9}\right)^2 - 1 = 2(2x^2 - 1)^2 - 1.$$

The triple product in question can be written as $8x(2x^2 - 1)(2(2x^2 - 1)^2 - 1)$. We expand this product while systematically reducing its degree using the cubic equation for x above. We calculate

$$\begin{aligned} 8x(2x^2 - 1)(2(2x^2 - 1)^2 - 1) &= 8(2x^3 - x)(8x^4 - 8x^2 + 1) \\ &= 8\left(\frac{6x + 1}{4} - x\right)\Big(x(6x + 1) - 8x^2 + 1\Big) = 2(2x + 1)(-2x^2 + x + 1) \\ &= 2(-4x^3 + 3x + 1) = -8x^3 + 6x + 2 = 1. \end{aligned}$$

The example follows.

The following formulas, still due to François Viète, give the expansion of $\sin(n\alpha)$ and $\cos(n\alpha)$, $n \in \mathbb{N}$, as trigonometric polynomials in the indeterminates $\cos\alpha$ and $\sin\alpha$:

$$\cos(n\alpha) = \sum_{k=0}^{n} \cos\left(\frac{k\pi}{2}\right) \binom{n}{k} \sin^k \alpha \cos^{n-k} \alpha$$

$$\sin(n\alpha) = \sum_{k=0}^{n} \sin\left(\frac{k\pi}{2}\right) \binom{n}{k} \sin^k \alpha \cos^{n-k} \alpha.$$

Note that the coefficients $\cos(k\pi/2)$ and $\sin(k\pi/2)$ take values ± 1 and 0, and half of the terms in each sum above are zero.

These formulas can be derived simultaneously by Peano's Principle of Induction. The initial case $n = 1$ for both formulas is a tautology. We perform the general induction step $n \Rightarrow n = 1$ for the second formula (for a change); the computations for the first formula are analogous. We have

$$\sin((n + 1)\alpha) = \sin(n\alpha)\cos\alpha + \cos(n\alpha)\sin\alpha$$

$$
\begin{aligned}
&= \cos\alpha \left(\sum_{k=0}^{n} \sin\left(\frac{k\pi}{2}\right) \binom{n}{k} \sin^k\alpha \cos^{n-k}\alpha \right) \\
&\quad + \sin\alpha \left(\sum_{k=0}^{n} \cos\left(\frac{k\pi}{2}\right) \binom{n}{k} \sin^k\alpha \cos^{n-k}\alpha \right) \\
&= \sum_{k=0}^{n} \sin\left(\frac{k\pi}{2}\right) \binom{n}{k} \sin^k\alpha \cos^{(n+1)-k}\alpha \\
&\quad + \sum_{k=0}^{n} \cos\left(\frac{k\pi}{2}\right) \binom{n}{k} \sin^{k+1}\alpha \cos^{(n+1)-(k+1)}\alpha.
\end{aligned}
$$

Shifting the index in the second sum, it becomes

$$
\begin{aligned}
&\sum_{k=1}^{n+1} \cos\left(\frac{(k-1)\pi}{2}\right) \binom{n}{k-1} \sin^k\alpha \cos^{(n+1)-k}\alpha \\
&\quad = \sum_{k=1}^{n+1} \sin\left(\frac{k\pi}{2}\right) \binom{n}{k-1} \sin^k\alpha \cos^{(n+1)-k}\alpha.
\end{aligned}
$$

Substituting this, noticing the vanishing of the initial term ($k = 0$), splitting off the final term ($k = n + 1$), and joining the two sums, we calculate

$$
\begin{aligned}
\sin((n+1)\alpha) &= \sin\left(\frac{(n+1)\pi}{2}\right) \sin^{n+1}\alpha \\
&\quad + \sum_{k=1}^{n} \sin\left(\frac{k\pi}{2}\right) \left(\binom{n}{k-1} + \binom{n}{k} \right) \sin^k\alpha \cos^{(n+1)-k}\alpha \\
&= \sin\left(\frac{(n+1)\pi}{2}\right) \sin^{n+1}\alpha \\
&\quad + \sum_{k=1}^{n} \sin\left(\frac{k\pi}{2}\right) \binom{n+1}{k} \sin^k\alpha \cos^{(n+1)-k}\alpha,
\end{aligned}
$$

where we used the inductive binomial identity in Section 6.3. Finally, putting back the initial (vanishing) term and the final term, we arrive at

$$
\sin((n+1)\alpha) = \sum_{k=0}^{n+1} \sin\left(\frac{k\pi}{2}\right) \binom{n+1}{k} \sin^k\alpha \cos^{(n+1)-k}\alpha.
$$

The general induction step is complete, and the formula follows.

Despite their compact appearance, the general multiple angle formulas for cosine and sine are not very convenient to work with.

About three centuries later, another approach was put forward by Chebyshev. To motivate this, we return to our double and triple angle formulas and observe that, for $n = 1, 2, 3$, the expressions $\cos(n\alpha)$ and $\sin(n\alpha)/\sin(\alpha)$ are polynomials in the indeterminate $\cos(\alpha)$.

This is true in general; in fact, we have

$$\cos(n\alpha) = T_n(\cos(\alpha)) \quad \text{and} \quad \frac{\sin(n\alpha)}{\sin(\alpha)} = U_{n-1}(\cos(\alpha)), \quad n \in \mathbb{N},$$

where T_n and U_n are polynomials of degree n.
According to our computations, we have

$$\begin{aligned} T_1(x) &= x & U_0(x) &= 1, \\ T_2(x) &= 2x^2 - 1 & U_1(x) &= 2x \\ T_3(x) &= 4x^3 - 3x & U_2(x) &= 4x^2 - 1. \end{aligned}$$

In general, T_n and U_n satisfy the following inductive relations:

$$\begin{aligned} T_{n+1}(x) &= xT_n(x) - (1 - x^2)U_{n-1}(x) \\ U_{n+1}(x) &= xU_n(x) + T_{n+1}(x). \end{aligned}$$

Indeed, these relations are direct consequences of the addition formulas

$$\begin{aligned} \cos((n+1)\alpha) &= \cos(n\alpha)\cos(\alpha) - \sin(n\alpha)\sin(\alpha) \\ &= \cos(\alpha)\cos(n\alpha) - (1 - \cos^2(\alpha))\frac{\sin(n\alpha)}{\sin(\alpha)} \\ \sin((n+2)\alpha) &= \sin((n+1)\alpha)\cos(\alpha) + \cos((n+1)\alpha)\sin(\alpha) \\ &= \left(\cos(\alpha)\frac{\sin((n+1)\alpha)}{\sin(\alpha)} + \cos((n+1)\alpha)\right)\sin(\alpha). \end{aligned}$$

Now, a simple induction in the use of these recurrence formulas shows that T_n and U_n are polynomials of degree n. These are called **Chebyshev polynomials**.

Example 11.3.9 Use the Chebyshev inductive relations to derive the quadruple angle formulas.

We calculate

$$T_4(x) = xT_3(x) - (1-x^2)U_2(x) = x(4x^3 - 3x) - (1-x^2)(4x^2 - 1) = 8x^4 - 8x^2 + 1,$$

and

$$U_3(x) = xU_2(x) + T_3(x) = x(4x^2 - 1) + 4x^3 - 3x = 8x^3 - 4x.$$

Thus, we have

$$\cos(4\alpha) = 8\cos^4(\alpha) - 8\cos^2(\alpha) + 1$$
$$\sin(4\alpha) = 8\cos^3(\alpha)\sin(\alpha) - 4\cos(\alpha)\sin(\alpha).$$

(Note that these formulas can also be obtained by applying twice the double angle formulas.)

We close this section by deriving several important formulas pertaining to the side lengths and angles of a general (non-degenerate) triangle $\triangle[A, B, C]$ with (non-collinear) vertices A, B, C. As usual, we denote the angle measures at the vertices A, B, C by α, β, γ and the side lengths by $a = d(B, C)$, $b = d(C, A)$, $c = d(A, B)$. The metric quantities $\alpha, \beta, \gamma, a, b, c$ are not independent. We have $\alpha + \beta + \gamma = \pi$. In particular, we see that the angles have the following restrictions: $\alpha + \beta < \pi$, $\beta + \gamma < \pi$, $\gamma + \alpha < \pi$. In addition, by the triangle inequality, we have $a < b + c$, $b < c + a$, $c < a + b$. Apart from these, we claim that the choice of any three **independent** quantities from a, b, c and α, β, γ (that is, with the exception of choosing the three angles) determines the triangle $\triangle[A, B, C]$ (up to congruence), and thereby the rest of the quantities can be computed.

This can be shown by the **Laws of Cosines and Sines**, which we now proceed to discuss.

We recall the following formula from Section 5.6:

$$d(A_0, B_0)^2 = 2 + \frac{c^2 - a^2 - b^2}{ab},$$

where A_0, and respectively B_0, is the point at unit distance from the vertex C on the half-line with end-point C and containing A, and respectively B. The triangle $\triangle[A_0, B_0, C]$ is isosceles (since $d(A_0, C) = d(B_0, C) = 1$). The altitude through C splits this triangle into two congruent right triangles. We thus have $\sin(\gamma/2) = d(A_0, B_0)/2$. The half angle formula gives $\sin^2(\gamma/2) = d(A_0, B_0)^2/4 = (1 - \cos\gamma)/2$. Using this to eliminate $d(A_0, B_0)$ in the formula above, after rearranging, we arrive at the **Law of Cosines**

$$c^2 = a^2 + b^2 - 2ab\cos\gamma.$$

Remark Note that $\gamma = \pi/2$ gives the Pythagorean Theorem $c^2 = a^2 + b^2$.

History
Trigonometric functions were not known to the ancient Greeks mainly because of the notion of function had not been developed at that time. On the other hand, they certainly knew that the ratios of the respective side lengths of two similar triangles are equal. This, applied to right triangles, immediately gives that the ratios of side lengths depend only on the (two acute) angles of the right triangle. This implicitly leads to the fact that these ratios are functions depending only on these angles. With this in mind, Propositions 12 and 13 in Book II of Euclid's *Elements* give an essentially equivalent formulation of the Law of Cosines.

In the following example, we return to origami, this time performed on an equilateral triangle.

Example 11.3.10 Fold an equilateral triangle $\triangle[A, B, C]$ of side length 1 along a crease line segment $[P, Q]$ with $P \in [A, C]$ and $Q \in [B, C]$ such that the vertex C folds over to a point $C' \in [A, B]$. Assume that C' splits the side $[A, B]$ in the ratio $p \div q$, $p + q = 1$. Show that the length of the crease is

$$d(P, Q) = \sqrt{\frac{(p^2 - p + 1)^2}{(2 - p)^2} - \frac{(p^2 - p + 1)(q^2 - q + 1)}{(2 - p)(2 - q)} + \frac{(q^2 - q + 1)^2}{(2 - q)^2}}.$$

(Note the special case $p = 1$ and $q = 0$ ($C' = B$) giving $d(P, Q) = \sqrt{3}/2$, the height of the equilateral triangle.)

Let $x = d(C, P) = d(C', P)$ and $y = d(C, Q) = d(C', Q)$. The Law of Cosines applied to the triangles $\triangle[A, P, C']$ and $\triangle[B, Q, C']$ gives $x^2 = (1-x)^2 + p^2 - p(1-x)$ and $y^2 = (1-y)^2 + q^2 - q(1-y)$, where we used that the side length of our original triangle is unity and $\cos(\pi/3) = 1/2$. Simplifying, and solving for x and y, we obtain $x = (p^2-p+1)/(2-p)$ and $y = (q^2-q+1)/(2-q)$. Finally, we apply the Law of Cosines to the triangle $\triangle[P, Q, C]$ to get $d(P, Q)^2 = x^2+y^2-xy$. Substituting, the claimed formula follows.

Example 11.3.11 We briefly revisit Example 8.3.2 here and give a more illuminating solution to the problem: In an ellipse, the product of the distances of the two foci from any tangent line to the ellipse is equal to the square of the semiminor axis.

We let $F_\pm$ be the foci, P_0 the point of tangency of the tangent line to the ellipse, $d(F_-, F_+) = 2c$, $d(F_\pm, P_0) = d_\pm$, $d_- + d_+ = 2a$, and, finally, $G_\pm$ the perpendicular projections of $F_\pm$ to the tangent line ℓ, $d(F_\pm, \ell) = d(F_\pm, G_\pm)$. By the reflective property of the ellipse, we have $\alpha = \mu(\angle G_- P_0 F_-) = \mu(\angle F_+ P_0 G_+)$.

The Law of Cosines applied to the triangle $\triangle[F_-, P_0, F_+]$ can be written as

$$(2c)^2 = d_-^2 + d_+^2 - 2d_-d_+ \cos(\pi - 2\alpha).$$

This, combined with $(2a)^2 = (d_- + d_+)^2 = d_-^2 + d_+^2 + 2d_-d_+$, gives

$$4b^2 = 4(a^2 - c^2) = 2d_-d_+(1 - \cos(2\alpha)) = 4d_-d_+ \sin^2\alpha.$$

We arrive at

$$d(F_-, \ell)d(F_+, \ell) = d_- \sin\alpha \cdot d_+ \sin\alpha = b^2.$$

The example follows.

The Law of Cosines relates the three side lengths of a triangle to one of the angle measures. Another formula, the so-called Law of Sines, relates two side lengths to two angle measures. We now proceed to derive this.

Let $\mathcal{C}$ be the **circumcircle** of the triangle $\triangle[A, B, C]$ with **circumradius** R. (Recall from Section 5.5 that the circumcircle is the unique circle through the three vertices A, B, C whose circumcenter O is the common meeting point of the perpendicular bisectors of the three sides $[A, B]$, $[B, C]$, and $[C, A]$.) We claim

$$\sin\gamma = \frac{c}{2R}.$$

We consider the side $[A, B]$ as a chord of $\mathcal{C}$. Let $\mathcal{C}_0 \subset \mathcal{C}$ be the circular arc with end-points A and B containing C. We distinguish three cases.

I. The chord $[A, B]$ is a diameter of the circumcircle $\mathcal{C}$, and hence $c = d(A, B) = 2R$. By Thales' Theorem,[7] $\triangle[A, B, C]$ is a right triangle with right angle at C. We thus have $\gamma = \pi/2$ and hence $\sin\gamma = 1$. The claim follows in this case.

II. $\mathcal{C}_0$ is the longer circular arc of $\mathcal{C}$ with end-points A and B. In this case we move the vertex $C \in \mathcal{C}_0$ to another point $C' \in \mathcal{C}_0$ such that $O \in [B, C']$. By the Central Angle Theorem, the angle measure at the vertex C' of the triangle $\triangle[A, B, C']$ stays γ. Since $[B, C']$ is a diameter of $\mathcal{C}$, again by Thales' Theorem, $\triangle[A, B, C']$ is a right triangle (with right angle at the vertex A). The claim follows in this case from the definition of sine.

III. $\mathcal{C}_0$ is the shorter circular arc of $\mathcal{C}$ with end-points A and B. In this case we move $C \in \mathcal{C}_0$ to a point $C' \in \mathcal{C} \setminus \mathcal{C}_0$. By the Central Angle Theorem again, the angle measure γ changes to $\pi - \gamma$. But $\sin(\pi - \gamma) = \sin\gamma$, and the previous case applies. The claim follows.

Applying the formula to all sides of the triangle, we arrive at the **Law of Sines**[8]

$$\frac{\sin\alpha}{a} = \frac{\sin\beta}{b} = \frac{\sin\gamma}{c} = \frac{1}{2R}.$$

In addition to their side lengths and angles, triangles have many more metric characteristics such as perimeter, inradius, circumradius, etc. (For the last two, see

[7] Here and in what follows, we use Thales' Theorem and its generalization, the Central Angle Theorem. These can be derived as straightforward applications of the pons asinorum; see Exercise 5.5.2 at the end of Section 5.5.

[8] Note that another very simple proof of the first two equalities can be obtained by writing down the definition of sine for the three angles with respect to the three lengths of the altitude lines of the triangle.

Section 5.5.) To close this section, we derive a few classical formulas for these in terms of the sides a, b, c and angles α, β, γ of our triangle $\triangle[A, B, C]$.

First, the Law of Cosines can be written as $\cos\alpha = (b^2 + c^2 - a^2)/(2bc)$. Using the Pythagorean identity, we replace the cosine by sine and calculate

$$\begin{aligned}\sin\alpha &= \sqrt{1-\cos^2(\alpha)} = \sqrt{1-\left(\frac{b^2+c^2-a^2}{2bc}\right)^2} = \frac{\sqrt{(2bc)^2-(b^2+c^2-a^2)}}{2bc}\\ &= \frac{\sqrt{(2bc-b^2-c^2+a^2)(2bc+b^2+c^2-a^2)}}{2bc} = \frac{\sqrt{(a^2-(b-c)^2)((b+c)^2-a^2)}}{2bc}\\ &= \frac{\sqrt{(a+b+c)(-a+b+c)(a-b+c)(a+b-c)}}{2bc} = \frac{2\sqrt{s(s-a)(s-b)(s-c)}}{bc},\end{aligned}$$

where, in the last equality, we used the **semiperimeter** (half of the perimeter) $s = (a+b+c)/2$ of the triangle $\triangle[A, B, C]$. Using the Law of Sines, we write the formula above in a more symmetric form as

$$\frac{\sin\alpha}{a} = \frac{\sin\beta}{b} = \frac{\sin\gamma}{c} = \frac{2\sqrt{s(s-a)(s-b)(s-c)}}{abc} = \frac{1}{2R},$$

where we inserted the expression in the circumradius at the end.

Remark Although in this book we systematically avoided discussing areas (and integrals), we see no harm noting that, taking the side $[A, B]$ with length c as the base, the height of the triangle $\triangle[A, B, C]$ is $b\sin\alpha$. Thus, the area of our triangle is $\mathcal{A} = (1/2)bc\sin\alpha$. Using our formula for $\sin\alpha$ above, we finally arrive at **Heron's formula**

$$\mathcal{A} = \sqrt{s(s-a)(s-b)(s-c)}.$$

As a beautiful application, we show that, among the triangles of a given perimeter, the equilateral triangle has the largest area.

Let $s > 0$ be the given semiperimeter. For a triangle with side lengths a, b, c, the AM-GM inequality in the three variables $s-a, s-b, s-c$ gives

$$(s-a)(s-b)(s-c) \le \left(\frac{s-a+s-b+s-c}{3}\right)^3 = \left(\frac{s}{3}\right)^3.$$

Moreover, equality holds if and only if $s-a = s-b = s-c$, that is, if and only if $a = b = c$. Now, by Heron's formula, we have

$$\mathcal{A} = \sqrt{s(s-a)(s-b)(s-c)} \le \sqrt{s\left(\frac{s}{3}\right)^3} = \frac{s^2}{3\sqrt{3}}.$$

Equality, maximum of $\mathcal{A}$, holds if and only if $a = b = c$. Notice that, as a byproduct, we also obtained the area of an equilateral triangle in terms of its semiperimeter.

As a second application of Heron's formula, recall that the incircle is the largest circle inscribed in the triangle. As such, it touches each side at a point of tangency. By the characteristic property of the circle discussed in Section 5.5, the line segments connecting the incenter (the center of the incircle) and these points of tangency are perpendicular to the respective sides. Thus, the inradius $r > 0$ is the height of the three sub-triangles that the original triangle is split by the three line segments from the incenter to the vertices. The areas of these three sub-triangles add up to the area $\mathcal{A}$ of our triangle. We have

$$\mathcal{A} = \frac{ar}{2} + \frac{br}{2} + \frac{cr}{2} = r\frac{a+b+c}{2} = rs.$$

Using Heron's formula, we obtain

$$r = \frac{\mathcal{A}}{s} = \sqrt{\frac{(s-a)(s-b)(s-c)}{s}}.$$

Combining our formulas for the inradius and circumradius, we obtain

$$rR = \frac{\mathcal{A}}{s} \cdot \frac{abc}{4\mathcal{A}} = \frac{abc}{4s} = \frac{abc}{2(a+b+c)}.$$

Exercises

11.3.1. Derive the addition formulas for cosine and sine in the following geometric way (for $0 < \alpha, \beta, \alpha + \beta < \pi/2$) (see Figure 11.3). Let T_1 be a right triangle with an acute angle α and hypotenuse $\cos\beta$ and T_2 a right triangle with an acute angle β and hypotenuse 1. Paste T_1 and T_2 together along the common length sides such that the acute angles α and β share a common vertex. Finally, insert this configuration into a rectangle and calculate each side length of the rectangle in two ways.

11.3.2. Let $a^2 + b^2 = c^2 + d^2 = 1$, $a, b, c, d \in \mathbb{R}$. Show that $|ac + bd| \leq 1$.

11.3.3. Let $0 < a, b \in \mathbb{R}$ such that $a^2 + b^2 = 1$. Define the real sequence $(c_n)_{n=0}^{\infty}$ inductively by $c_0 = 0$ and $c_{n+1} = a \cdot c_n + b \cdot \sqrt{1 - c_n^2}$, $n \in \mathbb{N}_0$. Show that $0 < c_n \leq 1$, $n \in \mathbb{N}$; in particular, the sequence $(c_n)_{n=0}^{\infty}$ is well-defined. Prove that, for $a \leq \sqrt{2}/2$, we have $c_{n+2} = c_n$, $n \in \mathbb{N}$; that is, the sequence $(c_n)_{n=1}^{\infty}$ is periodic with period 2.

11.3.4. Derive the following identities:

$$\sin^3(\alpha) = \frac{3\sin(\alpha) - \sin(3\alpha)}{4} \quad \text{and} \quad \sin^4(\alpha) = \frac{3 - 4\cos(2\alpha) + \cos(4\alpha)}{8}.$$

Derive the similar identities for the powers of cosine.

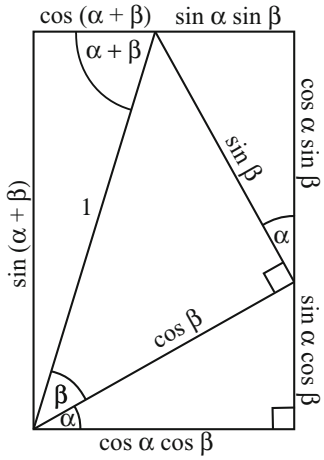


Fig. 11.3 Geometric proof of the addition formulas for sine and cosine.

11.3.5. Derive the following identities:

$$\sin^2(\alpha)\cos^2(\alpha) = \frac{1-\cos(4\alpha)}{8};$$
$$\sin^3(\alpha)\cos^3(\alpha) = \frac{3\sin(2\alpha)-\sin(6\alpha)}{32};$$
$$\sin^4(\alpha)\cos^4(\alpha) = \frac{3-4\cos(4\alpha)+\cos(8\alpha)}{128}.$$

11.3.6. Show that $\arcsin(x) + \arccos(x) = \pi/2$

11.3.7. Given $\alpha + \beta + \gamma = \pi$, show that

$$\sin(2\alpha) + \sin(2\beta) + \sin(2\gamma) = 4\sin\alpha\sin\beta\sin\gamma.$$

11.3.8. Given $\alpha + \beta + \gamma = \pi$, show that

$$\tan(\alpha) + \tan(\beta) + \tan(\gamma) = \tan(\alpha)\tan(\beta)\tan(\gamma).$$

11.3.9. Show that if $n \in \mathbb{N}$ is **not** divisible by 3, then a (given) angle with angle measure π/n can be trisected by straightedge and compass.[9] (Note the contrast with the remark following Example 11.3.7.)

11.3.10. Show that the Chebyshev polynomials $T_n(x)$ and U_{n-1} with $n \in \mathbb{N}$ satisfy **"Pell's Equation"**

$$T_n^2(x) - (x^2-1)U_{n-1}^2(x) = 1.$$

11.3.11. Calculate $T_n(\pm 1)$ and $U_{n-1}(\pm 1)$ for $n \in \mathbb{N}$.

[9] Inspired by a problem in the USA Mathematical Olympiad, 1981.

11.3.12. Derive the identity

$$2T_m(x)T_n(x) = T_{m+n}(x) + T_{m-n}(x), \quad m > n,\ m, n \in \mathbb{N}.$$

11.3.13. Show that the Chebyshev polynomial $T_n(x)$ restricted to the interval $[-1, 1]$ has n roots and has range $[-1, 1]$.

11.3.14. Use induction with respect to $n \in \mathbb{N}$ to show

$$T_n'(c) = nU_{n-1}(c) \quad \text{and} \quad (c^2-1)U_n'(c)+cU_n(c)=(n+1)T_{n+1}(c), \quad c \in \mathbb{R}.$$

11.3.15. Derive the **sum to product** formulas:

$$\cos\alpha + \cos\beta = 2\sin\frac{\alpha+\beta}{2}\sin\frac{\alpha-\beta}{2}$$
$$\cos\alpha - \cos\beta = -2\sin\frac{\alpha+\beta}{2}\sin\frac{\alpha-\beta}{2}$$
$$\sin\alpha + \sin\beta = 2\sin\frac{\alpha+\beta}{2}\cos\frac{\alpha-\beta}{2}.$$

11.3.16. Use Example 11.3.6 to derive the **Lagrange identities**:

$$\sum_{k=1}^{n}\sin(k\alpha) = \frac{\cos(\alpha/2) - \cos((n+1/2)\alpha)}{2\sin(\alpha/2)}$$
$$\sum_{k=1}^{n}\cos(k\alpha) = \frac{-\sin(\alpha/2) + \sin((n+1/2)\alpha)}{2\sin(\alpha/2)}.$$

11.4 Trigonometric Rational Functions

Just as rational functions can be constructed from polynomials by allowing divisions, we can form trigonometric rational functions from trigonometric polynomials.

The most basic trigonometric rational functions are the **tangent** and **cotangent** functions $\tan : \mathbb{R} \to \mathbb{R}$ and $\cot : \mathbb{R} \to \mathbb{R}$ defined by

$$\tan(\theta) = \frac{y}{x} = \frac{\sin(\theta)}{\cos(\theta)} \quad \text{and} \quad \cot(\theta) = \frac{x}{y} = \frac{\cos(\theta)}{\sin(\theta)}.$$

The domain of the tangent function is the set of real numbers $\theta \in \mathbb{R}$ for which $\cos(\theta) \neq 0$. Since the cosine function vanishes on the odd multiples of $\pi/2$, we obtain that the tangent function is defined on the domain $\{\theta \in \mathbb{R} \,|\, \theta \neq (2n+1)\pi/2,\ n \in \mathbb{Z}\}$.

Similarly, the cotangent function is defined away from the zero-set of the sine function, the integer multiples of π. Therefore the domain of definition of the cotangent function is $\{\theta \in \mathbb{R} \,|\, \theta \neq n\pi,\ n \in \mathbb{Z}\}$.

By definition, the tangent and cotangent functions are connected through the identity

$$\tan(\theta) \cdot \cot(\theta) = 1, \quad \theta \in \mathbb{R}.$$

(We use here our convention that the variable θ is unrestricted in $\mathbb{R}$ with the tacit understanding that the respective functions may not be defined on the whole $\mathbb{R}$.)

Since they are fractions of the cosine and sine functions, the tangent and cotangent functions are automatically periodic with period 2π. In fact, their shorter period is π. It is enough to show this for the tangent function:

$$\tan(\theta + n\pi) = \frac{\sin(\theta + n\pi)}{\cos(\theta + n\pi)} = \frac{\sin(\theta)\cos(n\pi)}{\cos(\theta)\cos(n\pi)} = \frac{\sin(\theta)}{\cos(\theta)} = \tan(\theta), \quad n \in \mathbb{Z}.$$

Since the cosine function is even and the sine function is odd, both the tangent and cotangent functions are odd: $\tan(-\theta) = -\tan(\theta)$ and $\cot(-\theta) = -\cot(\theta)$.

Of lesser importance but sometimes useful are the **secant** and **cosecant** functions $\sec : \mathbb{R} \to \mathbb{R}$ and $\csc : \mathbb{R} \to \mathbb{R}$ defined by

$$\sec(\theta) = \frac{1}{x} = \frac{1}{\cos(\theta)} \quad \text{and} \quad \csc(\theta) = \frac{1}{y} = \frac{1}{\sin(\theta)}.$$

The properties of the secant and cosecant functions are readily derived from those of the cosine and sine functions.

Dividing the Pythagorean identity by the squares of cosine and sine functions, we obtain the Pythagorean identities for the tangent and cotangent functions:

$$\tan^2(\theta) + 1 = \sec^2(\theta) \quad \text{and} \quad \cot^2(\theta) + 1 = \csc^2(\theta).$$

Returning to our **right** triangle $\triangle[A, B, C]$ above, with angle $\theta \in (0, \pi/2)$ at A and right angle at C, we have

$$\tan(\theta) = \frac{a}{b}, \quad \cot(\theta) = \frac{b}{a}, \quad \sec(\theta) = \frac{c}{b}, \quad \csc(\theta) = \frac{c}{a}.$$

With these, we exhausted all possible ratios of the side lengths a, b, c.

Remark Note that the tangent of the angle measure θ that a line makes with the positive first axis is the slope m of the line: $m = \tan(\theta)$.

Swapping the roles of A and B above, we obtain the identities

$$\cot(\theta) = \tan\left(\frac{\pi}{2} - \theta\right) \quad \text{and} \quad \csc(\theta) = \sec\left(\frac{\pi}{2} - \theta\right), \quad \theta \in \mathbb{R}.$$

Due to periodicity, the trigonometric rational functions above are not one-to-one on their entire domains. Just like in the case of the sine and cosine functions, we need to restrict them to obtain suitable inverses. To begin with, it follows directly from the definition that the tangent function is strictly increasing on the interval $(-\pi/2, \pi/2)$ and its range is the whole $\mathbb{R}$. The inverse $\tan^{-1} = \arctan : \mathbb{R} \to (\pi/2, \pi/2)$ is therefore defined on this branch. Similarly, the cotangent function is strictly decreasing on the interval $(0, \pi)$ with range $\mathbb{R}$, and we obtain the inverse $\cot^{-1} = \operatorname{arccot} : \mathbb{R} \to (0, \pi)$.

Using the same reasoning, we define $\sec^{-1} = \operatorname{arcsec} : (-\infty, -1] \cup [1, \infty) \to [0, \pi/2) \cup (\pi/2, \pi]$ and $\csc^{-1} = \operatorname{arccsc} : (-\infty, -1] \cup [1, \infty) \to [-\pi/2, 0) \cup (0, \pi/2]$.

Example 11.4.1 Determine the domain and the algebraic representation of the composition $\cos \circ \arctan$.

Both functions arctan and cos are defined on $\mathbb{R}$; therefore, the domain of the composition is also $\mathbb{R}$. We let $\arctan(x) = \theta$, that is, $\tan(\theta) = x$ with $\theta \in (-\pi/2, \pi/2)$. Since the cosine function is even and the tangent function is odd, we may assume that $\theta > 0$, an acute angle. We now construct a right triangle of angle θ with side length opposite to θ equal to x and adjacent side length equal to 1. The Pythagorean Theorem gives the length of the hypotenuse as $\sqrt{1 + x^2}$. Moreover, from this triangle, we have $\cos(\theta) = 1/\sqrt{1 + x^2}$. Therefore $(\cos \circ \arctan)(x) = 1/\sqrt{1 + x^2}$, $x \in \mathbb{R}$.

Addition formulas for our new trigonometric functions are readily obtained. We give some details only for the tangent and cotangent functions. Using the addition formulas for sine and cosine, we calculate

$$\tan(\alpha + \beta) = \frac{\sin(\alpha + \beta)}{\cos(\alpha + \beta)} = \frac{\sin(\alpha)\cos(\beta) + \cos(\alpha)\sin(\beta)}{\cos(\alpha)\cos(\beta) - \sin(\alpha)\sin(\beta)} = \frac{\tan(\alpha) + \tan(\beta)}{1 - \tan(\alpha)\tan(\beta)}.$$

With this we obtain the addition formulas for the tangent function

$$\tan(\alpha \pm \beta) = \frac{\tan(\alpha) \pm \tan(\beta)}{1 \mp \tan(\alpha)\tan(\beta)}.$$

Similarly (taking reciprocals), we have

$$\cot(\alpha \pm \beta) = \frac{\cot(\alpha)\cot(\beta) \mp 1}{\cot(\alpha) \pm \cot(\beta)}.$$

Example 11.4.2 Let ℓ_1 and ℓ_2 be two intersecting non-perpendicular (non-vertical) lines in the plane forming a (positive) angle θ. Show that

$$\tan(\theta) = \frac{m_2 - m_1}{1 + m_1 m_2},$$

where m_1 and m_2 are the slopes of ℓ_1 and ℓ_2.

Letting $m_1 = \tan(\alpha_1)$ and $m_2 = \tan(\alpha_2)$ with $-\pi/2 < \alpha_1 < \alpha_2 < \pi/2$, we have $\theta = \alpha_2 - \alpha_1$. The addition formula for the tangent gives

$$\tan(\theta) = \tan(\alpha_2 - \alpha_1) = \frac{\tan(\alpha_2) - \tan(\alpha_2)}{1 + \tan(\alpha_1)\tan(\alpha_2)} = \frac{m_2 - m_1}{1 + m_1 m_2}.$$

Example 11.4.3 Show that

$$\arctan\frac{1}{x} = \arctan\frac{1}{x+y} + \arctan\frac{y}{x^2+xy+1}.$$

Using the addition formula for tangent, we calculate

$$\tan\left(\arctan\frac{1}{x+y} + \arctan\frac{y}{x^2+xy+1}\right) = \frac{\frac{1}{x+y} + \frac{y}{x^2+xy+1}}{1 - \frac{1}{x+y}\cdot\frac{y}{x^2+xy+1}} = \frac{(x+y)^2+1}{x((x+y)^2+1)} = \frac{1}{x}.$$

The example follows.

This example can be readily generalized. In fact, just as in the case of the sine and cosine functions, we have the following;

$$\arctan x + \arctan y = \arctan\left(\frac{x+y}{1-xy}\right)$$
$$\operatorname{arccot} x + \operatorname{arccot} y = \operatorname{arccot}\left(\frac{xy-1}{x+y}\right).$$

Returning to the main line, setting $\alpha = \beta$ in the addition formulas above, we obtain the **double angle formulas** for tangent and cotangent

$$\tan(2\alpha) = \frac{2\tan(\alpha)}{1 - \tan^2(\alpha)} \quad \text{and} \quad \cot(2\alpha) = \frac{\cot^2(\alpha) - 1}{2\cot(\alpha)}.$$

Similarly, we have

$$\sec(2\alpha) = \frac{\sec^2(\alpha)}{2 - \sec^2(\alpha)} \quad \text{and} \quad \csc(2\alpha) = \frac{\sec(\alpha)\csc(\alpha)}{2}.$$

An interesting consequence of the double angle formula for the tangent function is that all the six trigonometric functions can be expressed as rational functions of the tangent of the respective half angle. These formulas are as follows:

$$\sin(\alpha) = \frac{2\tan\left(\frac{\alpha}{2}\right)}{1 + \tan^2\left(\frac{\alpha}{2}\right)} \quad \cos(\alpha) = \frac{1 - \tan^2\left(\frac{\alpha}{2}\right)}{1 + \tan^2\left(\frac{\alpha}{2}\right)}$$

$$\tan(\alpha) = \frac{2\tan\left(\frac{\alpha}{2}\right)}{1-\tan^2\left(\frac{\alpha}{2}\right)} \quad \cot(\alpha) = \frac{1-\tan^2\left(\frac{\alpha}{2}\right)}{2\tan\left(\frac{\alpha}{2}\right)}$$
$$\sec(\alpha) = \frac{1+\tan^2\left(\frac{\alpha}{2}\right)}{1-\tan^2\left(\frac{\alpha}{2}\right)} \quad \csc(\alpha) = \frac{1+\tan^2\left(\frac{\alpha}{2}\right)}{2\tan\left(\frac{\alpha}{2}\right)}.$$

(Strictly speaking, these identities hold for angles that are not odd multiples of π, that is, $\alpha \neq (2k+1)\pi$, $k \in \mathbb{Z}$.)

To derive these formulas is straightforward. For example, we have

$$\sin(\alpha) = 2\cos\left(\frac{\alpha}{2}\right)\sin\left(\frac{\alpha}{2}\right) = \frac{2\cos\left(\frac{\alpha}{2}\right)\sin\left(\frac{\alpha}{2}\right)}{\cos^2\left(\frac{\alpha}{2}\right)+\sin^2\left(\frac{\alpha}{2}\right)} = \frac{2\tan\left(\frac{\alpha}{2}\right)}{1+\tan^2\left(\frac{\alpha}{2}\right)},$$

where, in the last step, we divided both the numerator and the denominator by $\cos^2(\alpha/2)$.

Given a polynomial $p(x, y)$ in the indeterminates x and y, the corresponding trigonometric polynomial can be written as

$$p(\cos(\alpha), \sin(\alpha)) = p\left(\frac{1-\tan^2\left(\frac{\alpha}{2}\right)}{1+\tan^2\left(\frac{\alpha}{2}\right)}, \frac{2\tan\left(\frac{\alpha}{2}\right)}{1+\tan^2\left(\frac{\alpha}{2}\right)}\right).$$

The right-hand side is the rational function

$$p\left(\frac{1-z^2}{1+z^2}, \frac{2z}{1+z^2}\right)$$

in the indeterminate z evaluated at $\tan(\alpha/2)$. Thus, at the expense of getting a rational function from a polynomial, the two indeterminates x and y are reduced to the single indeterminate z. In deriving trigonometric identities, this is not as useful as it may seem since the resulting rational function is often too complex.

Remark The substitution $z = \tan(\alpha/2)$ and the formula above are used in integral calculus to reduce a trigonometric (rational) integral to the integral of a rational function (which can then be integrated by using the method of partial fractions).

Another application is also noteworthy. The Pythagorean identity for cosine and sine gives $p(x, y) = x^2 + y^2 = 1$, for $x = \cos(\alpha)$ and $y = \sin(\alpha)$. Using this substitution, we have

$$p\left(\frac{1-z^2}{1+z^2}, \frac{2z}{1+z^2}\right) = \left(\frac{1-z^2}{1+z^2}\right)^2 + \left(\frac{2z}{1+z^2}\right)^2 = 1,$$

where $z = \tan(\alpha/2)$. Multiplying out, we obtain

$$(1-z^2)^2+(2z)^2=(1+z^2)^2.$$

Finally, substituting $z=\tan(\alpha/2)=s/t$, $t>s>0$, $s,t\in\mathbb{N}$, and simplifying, we arrive at

$$(t^2-s^2)^2+(2st)^2=(t^2+s^2)^2.$$

This gives all Pythagorean triples $(a,b,c)=(t^2-s^2,2st,t^2+s^2)$ as in Section 5.7.

As a last note, multiple angle formulas can be easily obtained from those of the sine and cosine. The Viète formula for the tangent function is

$$\tan(n\alpha)=\frac{\sum_{k=0}^{n}\sin\left(\frac{k\pi}{2}\right)\binom{n}{k}\tan^k\alpha}{\sum_{k=0}^{n}\cos\left(\frac{k\pi}{2}\right)\binom{n}{k}\tan^k\alpha},\quad n\in\mathbb{N}.$$

Exercises

11.4.1. Given $\alpha+\beta+\gamma=\pi/2$, show that

$$\cot(\alpha)+\cot(\beta)+\cot(\gamma)=\cot(\alpha)\cot(\beta)\cot(\gamma).$$

11.4.2. Let $x=\tan(\alpha/2)$. Show that

$$\sin(\alpha)=\frac{2x}{1+x^2}\quad\text{and}\quad\cos(\alpha)=\frac{1-x^2}{1+x^2}.$$

11.4.3. Derive the following triple angle formulas for the tangent and cotangent functions:

$$\tan(3\alpha)=\frac{3\tan(\alpha)-\tan^3(\alpha)}{1-3\tan^2(\alpha)}\quad\text{and}\quad\cot(3\alpha)=\frac{3\cot(\alpha)-\cot^3(\alpha)}{1-3\cot^2(\alpha)}.$$

11.4.4. Use the identity $\cot(\theta)-\cot(2\theta)=1/\sin(2\theta)$, $\theta\neq m\pi/2$, $m\in\mathbb{Z}$, to derive the formula[10]

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

11.4.5. Derive root formulas for the following: (a) $\cos(2\pi/3)$ and $\sin(2\pi/3)$, (b) $\cos(3\pi/4)$ and $\sin(3\pi/4)$, and (c) $\cos(5\pi/12)$ and $\sin(5\pi/12)$.

[10] For a special numerical example using the idea of this exercise, see Problem 13 in the American High School Mathematics Examination, 1988.

11.4.6. Using the notations in Section 11.3, for the triangle $\triangle[A, B, C]$, derive the Law of Cotangents

$$\frac{\cot(\alpha/2)}{s-a} = \frac{\cot(\beta/2)}{s-b} = \frac{\cot(\gamma/2)}{s-c} = \frac{1}{r}.$$

11.4.7. Use the Law of Cotangents and the triple angle formula for the cotangent function to derive Heron's formula.

11.4.8. In a triangle $\triangle[A, B, C]$, the sequence $\cot\alpha, \cot\beta, \cot\gamma$ is arithmetic. Show that $a^2 + c^2 = 2b^2$.

11.4.9. The first three terms of a geometric sequence are $\sin(\alpha)$, $\cos(\alpha)$, and $\tan(\alpha)$, for some $\alpha \in \mathbb{R}$. Find $\cos(\alpha)$.

11.5 Trigonometric Limits

Although trigonometric functions are radically different from polynomials, there are many inequalities among them. To incorporate trigonometry into our study, these inequalities are of crucial importance.

To begin with, we recall the basic construction in Section 5.6, specified to our case of the unit circle $\mathbb{S}$ with center at the origin 0. Let $P_0, P_1 \in \mathbb{S}$ with $0 < d(P_0, P_1) < 2$, and denote by $\mathcal{C} \subset \mathbb{S}$ the shorter circular arc with end-points P_0 and P_1. Let m_0, and respectively m_1, be the tangent line to $\mathcal{C}$ through the point P_0, and respectively P_1. Finally, let M be the intersection of m_0 and m_1. The main result of Section 5.6 is

$$(d(P_0, P_1) <)\ \mathcal{L}_{\mathcal{C}} < d(P_0, M) + d(P_1, M),$$

where we inserted the first (trivial) inequality. Let $0 < x < \pi$ be the angle measure of the angle $\angle P_0 0 P_1$. (Due to our present purpose to compare trigonometric functions with polynomials, we use x as a variable for an angle measure.) Then, by definition of the Birkhoff angle measure, we have $x = \mathcal{L}_{\mathcal{C}}$. In addition, the triangle $\triangle[0, P_0, M]$ has right angle at the vertex P_0 (by tangency), and the angle measure at the origin (as a vertex) is $x/2$. Since $d(0, P_0) = 1$, we obtain $d(P_0, M) = \tan(x/2)$. Since the triangles $\triangle[0, P_0, M]$ and $\triangle[0, P_1, M]$ are congruent, we also have $d(P_1, M) = \tan(x/2)$. Finally, splitting the triangle $\triangle[0, P_0, P_1]$ into two congruent right triangles by the line segment $[0, M]$, we obtain $d(P_0, P_1) = 2\sin(x/2)$. Substituting these into the inequality above, we obtain

$$2\sin\left(\frac{x}{2}\right) < x < 2\tan\left(\frac{x}{2}\right), \quad 0 < x < \pi.$$

This fundamental inequality has several applications. First, squaring and using the half angle formulas, a simple computation gives

$$\frac{x^2}{4}\frac{1+\cos(x)}{2} < \frac{1-\cos(x)}{2} < \frac{x^2}{4}, \quad 0 < x < \pi.$$

Rearranging, we obtain

$$1 - \frac{x^2}{2} < \cos(x) < \frac{1-(x/2)^2}{1+(x/2)^2} = \frac{2}{1+(x/2)^2} - 1, \quad 0 < x < \pi.$$

Notice that this also holds for $-\pi < x < 0$ (since the ingredients are even functions) and that, at $x = 0$, equality holds throughout. (Note that the usual upper bound is the constant 1 function, but here we preferred to give a much better approximation of the cosine.)

Monotonicity of the limit now gives

$$1 = \lim_{x\to 0}\left(1 - \frac{x^2}{2}\right) \le \lim_{x\to 0}\cos x \le \lim_{x\to 0}\left(\frac{2}{1+(x/2)^2} - 1\right) = 1.$$

This gives $\lim_{x\to 0}\cos x = 1$.

The estimate for cosine above is refined enough to get an estimate for the **difference quotient of cosine** at $x = 0$:

$$-\frac{x}{2} < \mathfrak{m}_{\cos}(x, 0) = \frac{\cos x - 1}{x} < -\frac{x/2}{1+(x/2)^2}, \quad 0 < |x| < \pi.$$

This gives the derivative

$$\cos'(0) = \lim_{x\to 0}\frac{\cos x - 1}{x} = 0.$$

To obtain an estimate for the difference quotient for the sine function, we return to our fundamental inequality. Doubling x, we obtain

$$\sin x < x < \tan x, \quad 0 < x < \frac{\pi}{2}.$$

Replacing $\tan(x)$ by $\sin(x)/\cos(x)$, and rearranging, we arrive at the following:

$$x\cos x < \sin x < x, \quad 0 < x < \frac{\pi}{2}.$$

Notice that the opposite chain of inequalities holds for $-\pi/2 < x < 0$ since the functions involved are here odd.

By monotonicity of the limit, we obtain

$$0 = \lim_{x\to 0^+} x\cos x \le \lim_{x\to 0^+}\sin x \le \lim_{x\to 0^+} x = 0,$$

and hence $\lim_{x\to 0} \sin x = 0$.

Remark By the Pythagorean identity, we have

$$\lim_{x\to 0} \sin^2 x = \lim_{x\to 0} \left(1 - \cos^2 x\right) = 0,$$

and this also gives the last limit formula above.

The estimate for sine above is refined enough to get an estimate for the **difference quotient of sine** at $x = 0$:

$$\cos x < \mathfrak{m}_{\sin}(x, 0) = \frac{\sin x - \sin 0}{x} = \frac{\sin x}{x} < 1, \quad 0 < |x| < \pi/2.$$

(Notice that this also holds for $-\pi/2 < x < 0$ since the functions involved are even.) This gives

$$1 = \lim_{x\to 0} \cos x \le \lim_{x\to 0} \frac{\sin x}{x} \le 1,$$

and we obtain the derivative

$$\sin'(0) = \lim_{x\to 0} \frac{\sin x}{x} = 1.$$

We now calculate the derivative of the cosine and sine functions at an arbitrary $c \in \mathbb{R}$. We claim that, for the difference quotients, we have the following:

$$\begin{aligned}
\mathfrak{m}_{\cos}(x, c) &= \cos c \cdot \mathfrak{m}_{\cos}(x - c, 0) - \sin c \cdot \mathfrak{m}_{\sin}(x - c, 0) \\
\mathfrak{m}_{\sin}(x, c) &= \cos c \cdot \mathfrak{m}_{\sin}(x - c, 0) + \sin c \cdot \mathfrak{m}_{\cos}(x - c, 0).
\end{aligned}$$

Indeed, we calculate

$$\begin{aligned}
\mathfrak{m}_{\cos}(x, c) &= \frac{\cos x - \cos c}{x - c} = \frac{\cos((x - c) + c) - \cos c}{x - c} \\
&= \cos c \cdot \frac{\cos(x - c) - 1}{x - c} - \sin c \cdot \frac{\sin(x - c)}{x - c} \\
&= \cos c \cdot \mathfrak{m}_{\cos}(x - c, 0) - \sin c \cdot \mathfrak{m}_{\sin}(x - c, 0).
\end{aligned}$$

The first formula for cosine follows. The proof of the second formula for sine is analogous.

Taking the limit $x \to c$ (or $x - c \to 0$), $c \in \mathbb{R}$, we obtain

$$\begin{aligned}
\cos'(c) &= \cos c \cdot \cos'(0) - \sin c \cdot \sin'(0) = -\sin c \\
\sin'(c) &= \cos c \cdot \sin'(0) + \sin c \cdot \cos'(0) = \cos c.
\end{aligned}$$

Finally, note that, since differentiable functions are continuous, as a byproduct, we obtain that the sine and cosine functions are **continuous everywhere**.

History
In his work *Siddhānta Shiromani* (Section III entitled Grahagaṇita) Bhāskara II arrived at the following approximation:[11] $\sin(x) - \sin(c) \approx (x - c)\cos(c)$, where $x \approx c$. This is essentially the differentiation formula $\sin'(c) = \cos(c)$ obtained above. As noted previously, he used this for astronomical calculations.

For the tangent function, using our inequalities above, we have

$$x < \tan(x) = \frac{\sin(x)}{\cos(x)} < \frac{x}{1 - x^2/2}, \quad 0 < x < \sqrt{2}.$$

The lower bound here is a direct consequence of the inequality $x\cos(x) < \sin(x)$ via dividing by the cosine function which is positive for $0 < x < \pi/2$. For the upper bound, we use $\sin(x) < x$ and $1 - x^2/2 < \cos(x)$. For the latter, we need to restrict the variable to the shorter range $0 < x < \sqrt{2}\,(< \pi/2)$ to make sure that $1 - x^2/2 > 0$.
We now rearrange and calculate

$$0 < \tan(x) - x < \frac{x}{1 - x^2/2} - x = x\left(\frac{1}{1 - x^2/2} - 1\right) = \frac{x^3/2}{1 - x^2/2}, \quad 0 < x < \sqrt{2}.$$

Dividing by x, we obtain

$$0 < \frac{\tan(x)}{x} - 1 < \frac{x^2/2}{1 - x^2/2}, \quad 0 < |x| < \sqrt{2}.$$

Notice that $\mathfrak{m}_{\tan}(x, 0) = \tan(x)/x$ is the difference quotient for the tangent function at 0. As a byproduct, taking limits, we obtain $\tan'(0) = 1$. Next, the derivative of the tangent function at an arbitrary $c \in \mathbb{R}$, $c \neq \pi/2 + k\pi$, $k \in \mathbb{Z}$, can be calculated by first deriving the following formula for the difference quotient:

$$\mathfrak{m}_{\tan}(x, c) = \mathfrak{m}_{\tan}(x - c, 0) \cdot \frac{1 + \tan^2 c}{1 - \mathfrak{m}_{\tan}(x - c, 0)\tan c \cdot (x - c)}.$$

This can be shown using the addition formula for the tangent function along the same lines as the analogous formulas for the cosine and sine functions. Letting $x \to c$ (or $x - c \to 0$), we then obtain

$$\tan'(c) = 1 + \tan^2 c = \sec^2 c, \quad c \neq \pi/2 + k\pi, k \in \mathbb{Z}.$$

Remark Alternatively, using the quotient rule of differentiation (Section 4.3), we calculate

[11] Using modern notation.

$$\tan'(c)=\left(\frac{\sin}{\cos}\right)'(c)=\frac{\sin'(c)\cdot\cos(c)-\sin(c)\cdot\cos'(c)}{\cos^2(c)}=\frac{\cos^2(c)+\sin^2(c)}{\cos^2(c)}=1+\tan^2(c).$$

The tangent function has vertical asymptotes $x=\pi/2+k\pi$, $k\in\mathbb{Z}$. We have

$$\lim_{x\to\pi/2^{\pm}}\tan x=\lim_{x\to\pi/2^{\pm}}\frac{\sin x}{\cos x}=\mp\infty,$$

and, by periodicity, this also holds when any integer multiple of π is added.

We wish to obtain a more precise description of the "asymptotic behavior" of the tangent function near the asymptotes.

Since $\cot(x)=\tan(\pi/2-x)$, $x\neq k\pi$, $k\in\mathbb{Z}$, it is more convenient to do this for the cotangent function at 0.

Once again, for $0<x<\pi/2$, our earlier estimates give

$$\frac{1}{x}-\frac{x}{2}=\frac{1-x^2/2}{x}<\frac{\cos x}{\sin x}=\cot x<\frac{1}{x}.$$

Rearranging, we find

$$-\frac{x}{2}<\cot x-\frac{1}{x}<0,\quad 0<x<\frac{\pi}{2}.$$

As before, for $-\pi<x<0$, the inequality signs are reversed since the functions are odd.

This gives

$$\left|\cot x-\frac{1}{x}\right|<\frac{|x|}{2},\quad 0<|x|<\frac{\pi}{2},$$

showing that the cotangent function near 0 behaves like the rectangular hyperbola given by $y=1/x$.

Finally, since $\cot(x)=\tan(\pi/2-x)$, for the asymptotic behavior of the tangent function at $\pi/2$, we have

$$\left|\tan x+\frac{1}{x-\pi/2}\right|<\frac{|x-\pi/2|}{2},\quad 0<x<\pi,\ x\neq\pi/2.$$

We finish this section with a set of examples that shed light on continuity, differentiability, monotonicity, and critical points (Section 4.3) involving trigonometric functions.

We begin with the simplest one.

Example 11.5.1 Show that $\lim_{x\to 0}\sin(1/x)$ does not exist.

We define two null-sequences $(a_n)_{n\in\mathbb{N}_0}$ and $(b_n)_{n\in\mathbb{N}_0}$, which will also be useful in the sequel. We let

$$a_n = \frac{2}{(4n+1)\pi} \quad \text{and} \quad b_n = \frac{2}{(4n+3)\pi}, \quad n \in \mathbb{N}_0.$$

Clearly, we have $0 < \cdots < b_{n+1} < a_{n+1} < b_n < a_n < \cdots < b_0 < a_0$ and $\lim_{n\to\infty} a_n = \lim_{n\to\infty} b_n = 0$. On the other hand, we have $\sin(1/a_n) = 1$ and $\sin(1/b_n) = -1$, $n \in \mathbb{N}_0$. By Corollary to Proposition 4.1.1, the example follows.

Example 11.5.2 We have $\lim_{x\to 0}(x \cdot \sin(1/x)) = 0$. In particular, the function

$$f(x) = \begin{cases} x \cdot \sin(1/x), & \text{if} \quad x \neq 0 \\ 0, & \text{if} \quad x = 0 \end{cases}$$

is continuous everywhere.

Since the range of the sine function is $[-1, 1]$, for $0 \neq x \in \mathbb{R}$, we have

$$-|x| \leq x \cdot \sin\frac{1}{x} \leq |x|.$$

By monotonicity of the limit, we obtain

$$0 = -\lim_{x\to 0} |x| \leq \lim_{x\to 0} \left(x \cdot \sin\frac{1}{x} \right) \leq \lim_{x\to 0} |x| = 0.$$

Since continuity away from 0 is clear, the example follows.

Example 11.5.3 Prove that the function $f : \mathbb{R} \to \mathbb{R}$ defined by

$$f(x) = \begin{cases} x^2 \cdot \sin(1/x), & \text{if} \quad x \neq 0 \\ 0, & \text{if} \quad x = 0 \end{cases}$$

is differentiable everywhere.

Differentiability away from 0 is clear. Therefore, we only need to consider the difference quotient at 0 as follows:

$$\mathfrak{m}_f(x, 0) = \frac{x^2 \cdot \sin\frac{1}{x}}{x} = x \cdot \sin\frac{1}{x}, \quad 0 \neq x \in \mathbb{R}.$$

By the previous example, we have

$$f'(0) = \lim_{x\to 0} \mathfrak{m}_f(x, 0) = \lim_{x\to 0} x \cdot \sin\frac{1}{x} = 0.$$

Differentiability at 0 follows.

Example 11.5.4 Let the function $f : \mathbb{R} \to \mathbb{R}$ be defined by

$$f(x) = \begin{cases} x + 2x^2 \cdot \sin \frac{1}{x}, & \text{if} \quad x \neq 0 \\ 0, & \text{if} \quad x = 0. \end{cases}$$

Show that $f'(0) = 1$, but, for any $0 < \delta \in \mathbb{R}$, the function f is **not** monotonic on the interval $(-\delta, \delta)$. In particular, f has infinitely many critical points on $(-\delta, \delta)$.

As the previous example shows, we have $f'(0) = 1$. Note also that the second statement follows from the first since a continuous function with no critical points must be strictly monotonic (Section 4.3).

We now make use of the sequences $(a_n)_{n\in\mathbb{N}_0}$ and $(b_n)_{n\in\mathbb{N}_0}$ defined in Example 11.5.1 above. To show non-monotonicity, we claim

$$f(b_n) < f(a_n) \quad \text{and} \quad f(b_n) < f(a_{n+1}), \quad n \in \mathbb{N}_0.$$

The first inequality is clear since

$$f(a_n) - f(b_n) = a_n + 2a_n^2 - (b_n - 2b_n^2) = a_n - b_n + 2(a_n^2 + b_n^2) > 0, \quad n \in \mathbb{N}_0.$$

For the second, using $\sin(1/a_n) = 1$ and $\sin(1/b_n) = -1$, we calculate

$$\begin{aligned} f(b_n) - f(a_{n+1}) &= b_n - 2b_n^2 - (a_{n+1} + 2a_{n+1}^2) = b_n - a_{n+1} - 2(b_n^2 + a_{n+1}^2) \\ &= \frac{2}{(4n+3)\pi} - \frac{2}{(4n+5)\pi} - 2\left(\frac{4}{(4n+3)^2\pi^2} + \frac{4}{(4n+5)^2\pi^2}\right) \\ &= \frac{4}{(4n+3)(4n+5)\pi} - 8\frac{(4n+3)^2 + (4n+5)^2}{(4n+3)^2(4n+5)^2\pi^2}. \end{aligned}$$

This is negative if and only if

$$(4n+3)(4n+5)\frac{\pi}{2} < (4n+3)^2 + (4n+5)^2, \quad n \in \mathbb{N}_0.$$

This, however, holds by the AM-GM inequality (since $\pi/2 < 2$).
The example follows.

Remark The reader versed in differential calculus will no doubt realize that the derivative f', as a function, is

$$f'(x) = 1 + 4x \sin \frac{1}{x} + 2x^2 \cos \frac{1}{x}\left(-\frac{1}{x^2}\right) = 1 + 4x \sin \frac{1}{x} - 2\cos \frac{1}{x}.$$

This has no limit at 0 since $\lim_{x\to 0} \cos(1/x)$ does not exist (even though $f'(0) = 1$). In particular, the derivative f' is not continuous at 0.

Exercises

11.5.1. Let $\mathbb{S}$ be the unit circle with center at the origin. Given an angle $0 < x < \pi$, let $d(x)$, and respectively $\ell(x)$, denote the length of a chord, and respectively the length of a circular arc, of $\mathbb{S}$, both subtended by x as a central angle at the origin. Calculate the limit

$$\lim_{x\to 0^+} \frac{\ell(x)}{d(x)}.$$

11.5.2. Let $\alpha \in (0, 2\pi)$ such that α/π is irrational. Use the Equidistribution Theorem (Section 2.4) to derive the following: $\limsup_{n\to\infty} \sin(n\alpha) = 1$ and $\liminf_{n\to\infty} \sin(n\alpha) = -1$.

11.6 Cosine and Sine Series According to Newton

The series expansions of the cosine and sine functions can be obtained using limits of their arithmetic means over equidistant subdivisions of the domain interval.

Recall from Section 3.2 the concept of **arithmetic mean** of a real function $f : [a, b] \to \mathbb{R}$, $a < b$, $a, b \in \mathbb{R}$:

$$\mathcal{A}_f(n, a, b) = \frac{1}{n} \sum_{k=1}^{n} f\left(a + k\frac{b-a}{n}\right), \quad n \in \mathbb{N},$$

and the **mean** of f:

$$\mathcal{A}_f(a, b) = \lim_{n\to\infty} \mathcal{A}_f(n, a, b).$$

As noted there, the mean is linear and monotonic. Finally, we calculated the mean of the power function $\mathfrak{p}_p(x) = x^p$, $0 < x \in \mathbb{R}$, $0 < p \in \mathbb{R}$, as $\mathcal{A}_{\mathfrak{p}_p}(x) = x^p/(p+1)$, $0 < x \in \mathbb{R}$, where the mean is taken over the interval $[0, x]$ (with 0 suppressed).

We now calculate the mean of the cosine and sine functions on an interval $[0, x]$, where $0 < x \in \mathbb{R}$ is a fixed positive real number.

For the cosine function, we use the summation formula in Example 11.3.6 (with $\alpha = 0$ and $\beta = x/n$). For $n \in \mathbb{N}$, we calculate

$$\begin{aligned}\mathcal{A}_{\cos}(n, x) &= \frac{1}{n} \sum_{k=1}^{n} \cos\left(k\frac{x}{n}\right) = \frac{\cos\left(\frac{x}{2}\right) \sin\left(\frac{(n+1)x}{2n}\right)}{n \cdot \sin\left(\frac{x}{2n}\right)} - \frac{1}{n} \\ &= \cos\left(\frac{x}{2}\right) \cdot \frac{\frac{x}{2n}}{\sin\left(\frac{x}{2n}\right)} \cdot \frac{2}{x} \cdot \sin\left(\frac{x}{2} + \frac{x}{2n}\right) - \frac{1}{n},\end{aligned}$$

where the term $1/n$ corresponds to $k = 0$ in the summation. Taking the limit, we obtain

$$\mathcal{A}_{\cos}(x) = \frac{2\sin\left(\frac{x}{2}\right)\cos\left(\frac{x}{2}\right)}{x} = \frac{\sin x}{x}, \quad 0 < x \in \mathbb{R},$$

where we used

$$\lim_{n\to\infty} \frac{\frac{x}{2n}}{\sin\left(\frac{x}{2n}\right)} = \lim_{u\to 0} \frac{u}{\sin u} = 1.$$

The calculations for the mean of the sine function are entirely analogous in the use of the summation formula for the sine in Example 11.3.6. We obtain

$$\mathcal{A}_{\sin}(x) = \frac{1-\cos x}{x}, \quad 0 < x \in \mathbb{R}.$$

Armed with these explicit formulas for the means, we are now ready to start with the series expansions of cosine and sine. Throughout, we set $0 < x \in \mathbb{R}$.

We start with the inequality $\cos x \leq 1$. We take the means of both sides and have

$$\frac{\sin x}{x} = \mathcal{A}_{\cos}(x) \leq \mathcal{A}_{\mathfrak{p}_0}(x) = 1,$$

or equivalently, $\sin x \leq x$. Taking the means of both sides of this, we obtain

$$\frac{1-\cos x}{x} = \mathcal{A}_{\sin}(x) \leq \mathcal{A}_{\mathfrak{p}_1}(x) = \frac{x}{2}.$$

Equivalently, $1 - x^2/2 \leq \cos x$. Once again, taking the means of both sides, we get $1 - x^2/3! \leq \sin x/x$, or equivalently, $x - x^3/3! \leq \sin x$. Taking the means again, we obtain $x/2 - x^3/4! \leq (1-\cos x)/x$, or equivalently, $\cos x \leq 1 - x^2/2! + x^4/4!$.

The patterns emerging here can be readily generalized. We now claim that, for $n \in \mathbb{N}_0$, we have

$$1 - \frac{x^2}{2!} + \frac{x^4}{4!} - \cdots - \frac{x^{4n+2}}{(4n+2)!} \leq \cos x \leq 1 - \frac{x^2}{2!} + \cdots + \frac{x^{4n}}{(4n)!}$$

and

$$x - \frac{x^3}{3!} + \frac{x^5}{5!} - \cdots - \frac{x^{4n+3}}{(4n+3)!} \leq \sin x \leq x - \frac{x^3}{3!} + \cdots + \frac{x^{4n+1}}{(4n+1)!}.$$

We show these simultaneously by Peano's Principle of Induction with respect to $n \in \mathbb{N}$.

In view of the above, only the general induction step $n \Rightarrow n+1$ needs to be performed.

We take the first chain of inequalities and calculate the means of all terms. We obtain

$$1 - \frac{x^2}{3!} + \frac{x^4}{5!} - \cdots - \frac{x^{4n+2}}{(4n+3)!} \le \frac{\sin x}{x} \le 1 - \frac{x^2}{3!} + \cdots + \frac{x^{4n}}{(4n+1)!}$$

Multiplying through by x, we obtain the second inequality.

We now take the second chain of inequalities and calculate the means of all terms. We obtain

$$\frac{x}{2!} - \frac{x^3}{4!} + \frac{x^5}{6!} - \cdots - \frac{x^{4n+3}}{(4n+4)!} \le \frac{1-\cos x}{x} \le \frac{x}{2!} - \frac{x^3}{4!} + \cdots + \frac{x^{4n+1}}{(4n+2)!}.$$

Rearranging, we arrive at the first chain of inequalities with n moved up to $n+1$. The general induction step is complete, and the formulas follow.

As direct consequences of the formulas above, we have the following estimates:

$$\left|\cos x - \left(1 - \frac{x^2}{2!} + \cdots + \frac{x^{4n}}{(4n)!}\right)\right| \le \frac{|x|^{4n+2}}{(4n+2)!}, \quad x \in \mathbb{R},$$

and

$$\left|\sin x - \left(x - \frac{x^3}{3!} + \cdots + \frac{x^{4n+1}}{(4n+1)!}\right)\right| \le \frac{|x|^{4n+3}}{(4n+3)!}, \quad x \in \mathbb{R}.$$

We now recall that, for fixed $x \in \mathbb{R}$, we have $\lim_{n\to\infty} x^n/n! = 0$. This means that the general final term in each sum converges to zero as $n \to \infty$. This gives the convergent **infinite series expansion** of cosine and sine as follows:

$$\cos x = \sum_{n=0}^{\infty} (-1)^n \frac{x^{2n}}{(2n)!} \quad \text{and} \quad \sin x = \sum_{n=0}^{\infty} (-1)^n \frac{x^{2n+1}}{(2n+1)!}.$$

Finally, note that these hold for negative $x < 0$ as well since the functions in either side are even and, respectively, odd.

Example 11.6.1 Find a rational number that approximates $\cos(1/2)$ up to 15-decimal digit precision.[12]

In view of the estimate for cosine above, we need to find $n \in \mathbb{N}$ such that

$$\frac{(1/2)^{4n+2}}{(4n+2)!} \le 10^{-15}.$$

[12]This problem needs a computer algebra system.

A check of the first few values of $n \in \mathbb{N}$ shows that $n = 3$ satisfies the inequality; that is, we have

$$1,000,000,000,000,000 = 10^{15} \leq 2^{14}14! = 1,428,329,123,020,800$$

The approximating fraction can be obtained by substituting $x = 1/2$ into the finite series

$$1 - \frac{x^2}{2!} + \frac{x^4}{4!} - \frac{x^6}{6!} + \frac{x^8}{8!} - \frac{x^{10}}{(10)!} + \frac{x^{12}}{12!}.$$

The approximating fraction is

$$\frac{245,972,670,919}{280,284,364,800}.$$

History
The power series expansions of the sine, cosine, and the inverse tangent functions can be traced back to the Indian mathematician Mādhava (c. 1340–c. 1425), the founder of the Kerala School of Astronomy and Mathematics. Most of his writings have been lost, but later Kerala scholars refer to his results, among others, notably Nilakantha Somayaji (1444–1544) in his *Tantrasanghara* (c. 1500).
In the West, first the Scottish mathematician and astronomer James Gregory (1638–1675) published several power series expansions. The general method of constructing these series (including the series expansions of sine and cosine) at an arbitrary point was developed by Brook Taylor (1685–1731). Finally, the Scottish mathematician Colin Maclaurin (1698–1746) also developed and extensively used power series expansions centered at zero; consequently, this special case of Taylor series is often named after him as Maclaurin series.
In his work *Tractatus de Methodis Serierum et Fluxionum* dated in 1671 (but unpublished) Newton calculated the power series expansion of the sine function (as well as the binomial expansion and the series expansion of $\ln(1 + x)$ and the inverse sine function). We essentially followed his calculations here; he considered calculus as the algebraic counterpart of arithmetic for infinite decimals.

Exercise

11.6.1. Use the Cauchy Product Rule to find an infinite series expansion of the function $e^x / \cos(x)$.

11.7 The Basel Problem of Euler*

Recall the Basel problem from Section 3.1:

$$\sum_{n=1}^{\infty} \frac{1}{n^2} = 1 + \frac{1}{2^2} + \frac{1}{3^2} + \cdots = \frac{\pi^2}{6}.$$

In this section we will present an elementary proof of this formula using identities involving the cotangent and cosecant functions.

History
All proofs of the Basel problem use advanced mathematical tools except the one given below. This elementary proof goes back to Cauchy's *Course D'Analyse* (Note VIII) published in 1821. This proof also appeared in the twin Yaglom brother's work *Nonelementary Problems in an Elementary Exposition* published in 1954.

We begin by developing trigonometric formulas for the ratios $\cos(n\alpha)/\sin^n(\alpha)$ and $\sin(n\alpha)/\sin^n(\alpha)$, $n \in \mathbb{N}$.

Using multiple angle formulas (Section 11.3), the conversions $\cos(\alpha)/\sin(\alpha) = \cot(\alpha)$ and $1/\sin(\alpha) = \csc(\alpha)$, and the Pythagorean identity $\csc^2(\alpha) = 1 + \cot^2(\alpha)$, $\alpha \in \mathbb{R}$ (Section 11.4), for $n = 1, 2, 3, 4$, we easily obtain

$$\begin{aligned}
\frac{\cos(\alpha)}{\sin(\alpha)} &= \cot(\alpha), \\
\frac{\cos(2\alpha)}{\sin^2(\alpha)} &= \cot^2(\alpha) - 1, \\
\frac{\cos(3\alpha)}{\sin^3(\alpha)} &= 4\cot^3(\alpha) - 3\cot(\alpha)\csc^2(\alpha) = \cot^3(\alpha) - 3\cot(\alpha), \\
\frac{\cos(4\alpha)}{\sin^4(\alpha)} &= 8\cot^4(\alpha) - 8\cot^2(\alpha)\csc^2(\alpha) + \csc^2(\alpha) = \cot^4(\alpha) - 6\cot^2(\alpha) + 1.
\end{aligned}$$

Similarly, we have

$$\begin{aligned}
\frac{\sin(\alpha)}{\sin(\alpha)} &= 1, \\
\frac{\sin(2\alpha)}{\sin^2(\alpha)} &= 2\cot(\alpha), \\
\frac{\sin(3\alpha)}{\sin^3(\alpha)} &= -4 + 3\csc^2(\alpha) = 3\cot^2(\alpha) - 1, \\
\frac{\sin(4\alpha)}{\sin^4(\alpha)} &= 8\cot^4(\alpha) - 8\cot^2(\alpha)\csc^2(\alpha) + \csc^4(\alpha) = \cot^4(\alpha) - 6\cot^2(\alpha) + 1.
\end{aligned}$$

The pattern of the coefficients is binomial, and it is not hard to guess the general formulas. For $n \in \mathbb{N}$, we have

$$\frac{\cos(n\alpha)}{\sin^n(\alpha)} = \sum_{k=0}^{[n/2]} (-1)^k \binom{n}{2k} \cot^{n-2k}(\alpha),$$

and

$$\frac{\sin(n\alpha)}{\sin^n(\alpha)} = \sum_{k=0}^{[(n-1)/2]} (-1)^k \binom{n}{2k+1} \cot^{n-2k-1}(\alpha),$$

where $[\cdot]$ is the greatest integer function. We call these the **cotangent expansion formulas**.

We now prove these **simultaneously** using induction with respect to $n \in \mathbb{N}$. By the above, we need only to perform the general induction step $n \Rightarrow n+1$. We use the Chebyshev inductive formulas (Section 11.3) as

$$\begin{aligned}\frac{\cos((n+1)\alpha)}{\sin^{n+1}(\alpha)} &= \frac{T_{n+1}(\cos(\alpha))}{\sin^{n+1}(\alpha)} = \frac{\cos(\alpha)T_n(\cos(\alpha)) - (1-\cos^2(\alpha))U_{n-1}(\cos(\alpha))}{\sin^{n+1}(\alpha)} \\ &= \cot(\alpha)\frac{T_n(\cos(\alpha))}{\sin^n(\alpha)} - \frac{U_{n-1}(\cos(\alpha))}{\sin^{n-1}(\alpha)},\end{aligned}$$

and

$$\begin{aligned}\frac{\sin((n+1)\alpha)}{\sin^{n+1}(\alpha)} &= \frac{U_n(\cos(\alpha))}{\sin^n(\alpha)} = \frac{\cos(\alpha)U_{n-1}(\cos(\alpha)) + T_n(\cos(\alpha))}{\sin^n(\alpha)} \\ &= \cot(\alpha)\frac{U_{n-1}(\cos(\alpha))}{\sin^{n-1}(\alpha)} + \frac{T_n(\cos(\alpha))}{\sin^n(\alpha)}.\end{aligned}$$

By the induction hypothesis, we have

$$\frac{T_n(\cos(\alpha))}{\sin^n(\alpha)} = \frac{\cos(n\alpha)}{\sin^n(\alpha)} = \sum_{k=0}^{[n/2]} (-1)^k \binom{n}{2k} \cot^{n-2k}(\alpha)$$

and

$$\frac{U_{n-1}(\cos(\alpha))}{\sin^{n-1}(\alpha)} = \frac{\sin(n\alpha)}{\sin^n(\alpha)} = \sum_{k=0}^{[(n-1)/2]} (-1)^k \binom{n}{2k+1} \cot^{n-2k-1}(\alpha).$$

Substituting these into the formulas above, and using the binomial identity

$$\binom{n+1}{m} = \binom{n}{m} + \binom{n}{m-1}, \quad 0 \le m \le n, \quad m, n \in \mathbb{N}$$

(for $m = 2k$ and $m = 2k + 1$), and shifting the index within summations, the cotangent expansion formulas follow for $n + 1$. The induction is complete.

Remark The cotangent expansion formulas above are usually derived using the **de Moivre formula**. Since this involves some basic arithmetic in complex numbers, we preferred to stay in the real field $\mathbb{R}$ and used induction instead.

We make use of the second cotangent expansion formula for $n = 2m + 1$ odd. We write this in the expanded form

$$\frac{\sin((2m+1)\alpha)}{\sin^{2m+1}(\alpha)} = \binom{2m+1}{1}\cot^{2m}(\alpha) - \binom{2m+1}{3}\cot^{2m-2}(\alpha) + \cdots + (-1)^m\binom{2m+1}{2m+1}.$$

We substitute for α the m numbers

$$\alpha_k = \frac{k\pi}{2m+1}, \quad k = 1, 2, \ldots, m,$$

which are all zeros of the numerator $\sin((2m + 1)\alpha)$. Letting $t_k = \cot^2(\alpha_k)$, $k = 1, 2, \ldots, m$, we obtain

$$0 = \binom{2m+1}{1}t_k^m - \binom{2m+1}{3}t_k^{m-1} + \cdots + (-1)^m\binom{2m+1}{2m+1}.$$

We rephrase this by saying that t_k, $k = 1, 2, \ldots, m$, are **roots** of the polynomial

$$p(t) = \binom{2m+1}{1}t^m - \binom{2m+1}{3}t^{m-1} + \cdots + (-1)^m\binom{2m+1}{2m+1}.$$

Now the crux is that the m numbers α_k, $k = 1, 2, \ldots, m$, are distinct. Moreover, they are all contained in the interval $(0, \pi/2)$ on which the cotangent (square) is strictly decreasing. Hence, the m roots t_k, $k = 1, 2, \ldots, m$, of $p(t)$ are also distinct. Since the polynomial $p(t)$ has degree m, these are all the roots. The Factor Theorem gives the factorization

$$p(t) = \binom{2m+1}{1}(t - t_1)(t - t_2)\cdots(t - t_m).$$

Using the first Viète formula to extract the coefficient of the t^{m-1} term, we obtain

$$t_1 + t_2 + \cdots + t_m = \frac{\binom{2m+1}{3}}{\binom{2m+1}{1}} = \frac{2m(2m-1)}{6}.$$

Returning to α_k, $k = 1, 2, \ldots, m$, this gives

$$\cot^2(\alpha_1) + \cot^2(\alpha_2) + \cdots + \cot^2(\alpha_m) = \frac{2m(2m-1)}{6}.$$

We now use the Pythagorean identity to change the cotangents into cosecants:

$$\csc^2(\alpha_1) + \csc^2(\alpha_2) + \cdots + \csc^2(\alpha_m) = \frac{2m(2m-1)}{6} + m = \frac{2m(2m+2)}{6}.$$

For the final step, we use the estimates

$$\cot^2 \alpha < \frac{1}{\alpha^2} < \csc^2 \alpha, \quad 0 < \alpha < \frac{\pi}{2},$$

which can be obtained from the estimate $\sin(\alpha) < \alpha < \tan(\alpha)$, $0 < \alpha < \pi/2$, in Section 11.5, by taking reciprocals.

Combining these, we have

$$\frac{2m(2m-1)}{6} < \left(\frac{2m+1}{\pi}\right)^2 + \left(\frac{2m+1}{2\pi}\right)^2 + \cdots + \left(\frac{2m+1}{m\pi}\right)^2 < \frac{2m(2m+2)}{6}.$$

Rearranging, we obtain

$$\frac{\pi^2}{6}\frac{2m(2m-1)}{(2m+1)^2} < 1 + \frac{1}{2^2} + \cdots + \frac{1}{m^2} < \frac{\pi^2}{6}\frac{2m(2m+2)}{(2m+1)^2}.$$

By the monotonicity of the limit, we have

$$\frac{\pi^2}{6} = \frac{\pi^2}{6}\lim_{m\to\infty}\frac{2m(2m-1)}{(2m+1)^2} \le \lim_{m\to\infty}\left(1 + \frac{1}{2^2} + \cdots + \frac{1}{m^2}\right) \le \frac{\pi^2}{6}\lim_{m\to\infty}\frac{2m(2m+2)}{(2m+1)^2} = \frac{\pi^2}{6}.$$

Thus, we obtain

$$\sum_{n=1}^{\infty}\frac{1}{n^2} = \lim_{m\to\infty}\left(1 + \frac{1}{2^2} + \cdots + \frac{1}{m^2}\right) = \frac{\pi^2}{6}.$$

The Basel problem follows.

Exercise

11.7.1. Show that

$$\sum_{n=1}^{\infty}\frac{(-1)^{n+1}}{n^2} = \frac{\pi^2}{12}.$$

11.8 Ptolemy's Theorem

We have seen that any triangle has a unique circumscribed circle (Section 5.5). This clearly fails in general for quadrilaterals. The question arises: What condition guarantees that the quadrilateral is cyclic; that is, it possesses a circumcircle? The following beautiful result is due to the Greek mathematician and astronomer Claudius Ptolemy (c. 100–170 CE): If a quadrilateral is cyclic, then the product of its two diagonal lengths is equal to the sum of the products of its opposite side lengths.

It this section we derive a somewhat more extended version of Ptolemy's Theorem and its converse.

Let A, B, C, D be the four vertices of the quadrilateral in positively oriented cyclic order, and let $\alpha, \beta, \gamma, \delta$ be the angle measures at the respective vertices. We denote the side lengths as $a = d(A, B)$, $b = d(B, C)$, $c = d(C, D)$, $d = d(D, A)$, and the two diagonals as $u = d(A, C)$, $v = d(B, D)$.

(Extended) Ptolemy Theorem *A quadrilateral is cyclic if and only if*

$$uv = ac + bd \quad \text{and} \quad u(ab + cd) = v(ad + bc).$$

Proof The Law of Cosines applied to the sub-triangles $\triangle[A, B, C]$ and $\triangle[C, D, A]$ gives

$$2\cos\beta = \frac{a^2 + b^2 - u^2}{ab} \quad \text{and} \quad 2\cos\delta = \frac{c^2 + d^2 - u^2}{cd}.$$

The quadrilateral is cyclic if and only if $\beta + \delta = \pi$, or equivalently, if and only if $\cos\beta + \cos\delta = 0$. Adding the two equations above, we obtain that the quadrilateral is cyclic if and only if

$$2(\cos\beta + \cos\delta) = \frac{a^2 + b^2}{ab} + \frac{c^2 + d^2}{cd} - u^2\left(\frac{1}{ab} + \frac{1}{cd}\right) = 0,$$

or equivalently, if and only if

$$u^2 = \frac{(a^2 + b^2)cd + (c^2 + d^2)ab}{ab + cd} = \frac{(ac + bd)(ad + bc)}{ab + cd},$$

where in the last equality we performed a simple factoring.

We perform the same procedure for the sub-triangles $\triangle[B, C, D]$ and $\triangle[D, A, B]$ and obtain that the quadrilateral is cyclic if and only if

$$2(\cos\alpha + \cos\gamma) = \frac{a^2 + d^2}{ad} + \frac{b^2 + c^2}{bc} - v^2\left(\frac{1}{ad} + \frac{1}{bc}\right) = 0,$$

or equivalently, if and only if

$$v^2 = \frac{(a^2+d^2)bc + (b^2+c^2)ad}{ad+bc} = \frac{(ab+cd)(ac+bd)}{ad+bc}.$$

The two equations for u^2 and v^2 above are clearly equivalent to the two conditions given in the theorem. The proof is complete.

Remark The Law of Cosines was used in the proof above to derive Ptolemy's Theorem. Conversely, Ptolemy's Theorem **implies** the Law of Cosines as a special case.

In fact, any triangle $\triangle[A, B, C]$ with its circumcircle $\mathcal{C}$ can be extended to a cyclic (symmetric) trapezoid inscribed into the same circle by adding an extra vertex $D \in \mathcal{C}$ such that the "base" $[B, C]$ is parallel to the "top" $[A, D]$. Using the notations above, by symmetry, we have $d(A, B) = a = c = d(C, D)$. Again by symmetry, the diagonal lengths are equal. Ptolemy's Theorem gives $u^2 = a^2 + bd$. On the other hand, the base b and top d lengths are related by $b = d + 2a\cos\beta$ (by projecting the top line segment $[A, D]$ perpendicularly to the base $[B, C]$ and applying the definition of cosine to the two sub-triangles thus obtained). Eliminating d, we obtain $u^2 = a^2 + b(b - 2a\cos\beta) = a^2 + b^2 - 2ab\cos\beta$. This is the Law of Cosines for the triangle $\triangle[A, B, C]$.

Ptolemy's Theorem has many beautiful applications. We mention here a few as follows:

Example 11.8.1 Consider an **equilateral triangle** inscribed in a circle. Then any point of the circle has the following property: The distance of the point from the farthest vertex of the triangle is equal to the sum of the distances from the other two nearer vertices.

Indeed, if $\triangle[A, B, C]$ is the equilateral triangle with circumcircle $\mathcal{C}$ and $D \in \mathcal{C}$ is the additional point, then Ptolemy's Theorem gives $sd(D, B) = sd(D, A) + sd(D, C)$, where s is the side length of the triangle. Canceling s, we obtain $d(D, B) = d(D, A) + d(D, C)$.

Example 11.8.2 The ratio of a diagonal to the side length of a regular **pentagon** is the golden number τ (see Examples 3.1.2 and 11.3.4).

Inscribe the pentagon into a circle. Let a be the side length and b the length of a diagonal. Ptolemy's Theorem (applied to a quadrilateral with omitting one vertex of the pentagon) gives $b^2 = a^2 + ab$. Dividing, we obtain $(b/a)^2 = 1 + (b/a)$. This gives the golden number τ (since it satisfies $\tau^2 = 1 + \tau$).

Example 11.8.3 The side length of a regular **decagon** inscribed in a circle of radius R is equal to R/τ, where τ is the golden number.

We construct the regular decagon by the Archimedean duplication from a regular pentagon by taking perpendicular bisectors for each side. We apply Ptolemy's Theorem to the quadrilateral one of whose diagonals is a perpendicular bisector of a side of the pentagon (as well as the diagonal of the circle), and two other vertices

are the end-points of this side. Letting l denote the side length of the decagon, and using the notations of the previous example, Ptolemy's Theorem gives $2Ra = 2bl$. Hence, $l = R(a/b) = R/\tau$ as claimed.

History

Ptolemy's *Almagest* was the most important and influential text on the motion of the planets and stars in a geocentric model of the universe, until the introduction of the heliocentric model by Copernicus (1473–1543). In the *Almagest* (Book I, chapter 11), Ptolemy compiled a "Table of Chords," which, using our modern notations, is essentially equivalent to a sine table. In creating this table, Ptolemy used several geometric propositions of Euclid and the theorem on quadrilaterals inscribed in a circle, the result that came down to us as Ptolemy's Theorem.

Exercise

11.8.1. Prove Ptolemy's Theorem

$$d(A, B) \cdot d(C, D) + d(B, C) \cdot d(A, D) = d(A, C) \cdot d(B, D),$$

by converting the side lengths to angles using the Law of Sines with the diameter of the circumscribed circle.

Further Reading

1. Apostol, T.M., *Mathematical Analysis*, Addison-Wesley, Reading, MA, 1974.
2. Binmore, K.G., *Mathematical Analysis: A Straightforward Approach*, Cambridge University Press, Cambridge, 1977.
3. Boyer, C.B., *A History of Mathematics*, John Wiley, New York, 1968.
4. Clark, C., *Elementary Mathematical Analysis*, Wadsworth, Belmont, CA, 1982.
5. Cohen, L., and Ehrlich, L., *The Structure of the Real Number System*, Van Nostrand, Princeton, NJ, 1963.
6. de Souza, P.N., and Silva, J.-N., *Berkeley Problems in Mathematics*, Springer, New York, 1998.
7. Engel, A., *Problem solving strategies*, Springer, Berlin, 1997.
8. Gelfand, I.M. and Gelfand, A., *Geometry*, Birkhäuser, Basel, 2020.
9. Hardy, G.H., Littlewood, J.E., and Pólya, G., *Inequalities*, 2nd ed. Cambridge University Press, 1988.
10. Hardy, G.H., Wright, E.M., *An Introduction to the Theory of Numbers*, 5th ed. Oxford: Clarendon Press, New York, 1979.
11. Halmos, P., *Naive Set Theory*, Van Nostrand, Princeton, NJ, 1960.
12. Jech, T., *Set Theory*, 3rd ed. Springer, New York, 2002.
13. Kline, M., *Mathematical Thought from Ancient to Modern Times*, Oxford University Press, Oxford, 1972.
14. Landau, E., *Foundations of Analysis*, Chelsea House, New York, 1951.
15. Lang, S., *Analysis I-II*, Addison-Wesley, Reading, MA, 1968.
16. Marsden, J.E. *Elementary Classical Analysis*, W.H. Freeman, San Francisco, 1974.
17. Monk, J.D., *Introduction to Set Theory*, McGraw-Hill, New York, 1969.
18. Royden, H.L., *Real Analysis*, The Macmillan Company, New York, 1963.
19. Rudin, W., *Principles of Mathematical Analysis*, McGraw-Hill, New York, 1976.
20. Sierpiński, W., *Elementary Theory of Numbers*, 2nd ed. North Holland, 1985.
21. Suppes, P., *Axiomatic Set Theory*, Van Nostrand, Pprinceton, NJ, 1960.
22. Toth, G., *Glimpses of Algebra and Geometry*, 2nd ed. Springer, New York, 2002.
23. Wilder, R.L., *Introduction to Foundations of Mathematics*, 2nd ed. John Wiley, New York, 1965.

G. Toth, *Elements of Mathematics*, Undergraduate Texts in Mathematics,
https://doi.org/10.1007/978-3-030-75051-0

Index

G. Toth, *Elements of Mathematics*, Undergraduate Texts in Mathematics,
https://doi.org/10.1007/978-3-030-75051-0