# Prefaces

## Preface to the English Edition

An entire generation of mathematicians has grown up during the time between the appearance of the first edition of this textbook and the publication of the fourth edition, a translation of which is before you. The book is familiar to many people, who either attended the lectures on which it is based or studied out of it, and who now teach others in universities all over the world. I am glad that it has become accessible to English-speaking readers.

This textbook consists of two parts. It is aimed primarily at university students and teachers specializing in mathematics and natural sciences, and at all those who wish to see both the rigorous mathematical theory and examples of its effective use in the solution of real problems of natural science.

Note that Archimedes, Newton, Leibniz, Euler, Gauss, Poincaré, who are held in particularly high esteem by us, mathematicians, were more than mere mathematicians. They were scientists, natural philosophers. In mathematics resolving of important specific questions and development of an abstract general theory are processes as inseparable as inhaling and exhaling. Upsetting this balance leads to problems that sometimes become significant both in mathematical education and in science in general.

The textbook exposes classical analysis as it is today, as an integral part of the unified Mathematics, in its interrelations with other modern mathematical courses such as algebra, differential geometry, differential equations, complex and functional analysis.

Rigor of discussion is combined with the development of the habit of working with real problems from natural sciences. The course exhibits the power of concepts and methods of modern mathematics in exploring specific problems. Various examples and numerous carefully chosen problems, including applied ones, form a considerable part of the textbook. Most of the fundamental mathematical notions and results are introduced and discussed along with information, concerning their history, modern state and creators. In accordance with the orientation toward natural sciences, special attention is paid to informal exploration of the essence and roots of the basic concepts and theorems of calculus, and to the demonstration of numerous, sometimes fundamental, applications of the theory.

For instance, the reader will encounter here the Galilean and Lorentz transforms, the formula for rocket motion and the work of nuclear reactor, Euler's theorem on homogeneous functions and the dimensional analysis of physical quantities, the Legendre transform and Hamiltonian equations of classical mechanics, elements of hydrodynamics and the Carnot's theorem from thermodynamics, Maxwell's equations, the Dirac delta-function, distributions and the fundamental solutions, convolution and mathematical models of linear devices, Fourier series and the formula for discrete coding of a continuous signal, the Fourier transform and the Heisenberg uncertainty principle, differential forms, de Rham cohomology and potential fields, the theory of extrema and the optimization of a specific technological process, numerical methods and processing the data of a biological experiment, the asymptotics of the important special functions, and many other subjects.

Within each major topic the exposition is, as a rule, inductive, sometimes proceeding from the statement of a problem and suggestive heuristic considerations concerning its solution, toward fundamental concepts and formalisms. Detailed at first, the exposition becomes more and more compressed as the course progresses. Beginning *ab ovo* the book leads to the most up-to-date state of the subject.

Note also that, at the end of each of the volumes, one can find the list of the main theoretical topics together with the corresponding simple, but nonstandard problems (taken from the midterm exams), which are intended to enable the reader both determine his or her degree of mastery of the material and to apply it creatively in concrete situations.

More complete information on the book and some recommendations for its use in teaching can be found below in the prefaces to the first and second Russian editions.

Moscow, 2003 *V. Zorich*

## Preface to the Fourth Russian Edition

The time elapsed since the publication of the third edition has been too short for me to receive very many new comments from readers. Nevertheless, some errors have been corrected and some local alterations of the text have been made in the fourth edition.

Moscow, 2002 *V. Zorich*

## Preface to the Third Russian edition

This first part of the book is being published after the more advanced Part 2 of the course, which was issued earlier by the same publishing house. For the sake of consistency and continuity, the format of the text follows that adopted in Part 2. The figures have been redrawn. All the misprints that were noticed have been corrected, several exercises have been added, and the list of further readings has been enlarged. More complete information on the subject matter of the book and certain characteristics of the course as a whole are given below in the preface to the first edition.

Moscow, 2001 *V. Zorich*

## Preface to the Second Russian Edition

In this second edition of the book, along with an attempt to remove the misprints that occurred in the first edition,[1] certain alterations in the exposition have been made (mainly in connection with the proofs of individual theorems), and some new problems have been added, of an informal nature as a rule.

The preface to the first edition of this course of analysis (see below) contains a general description of the course. The basic principles and the aim of the exposition are also indicated there. Here I would like to make a few remarks of a practical nature connected with the use of this book in the classroom.

Usually both the student and the teacher make use of a text, each for his own purposes.

At the beginning, both of them want most of all a book that contains, along with the necessary theory, as wide a variety of substantial examples

[1] No need to worry: in place of the misprints that were corrected in the plates of the first edition (which were not preserved), one may be sure that a host of new misprints will appear, which so enliven, as Euler believed, the reading of a mathematical text.

of its applications as possible, and, in addition, explanations, historical and scientific commentary, and descriptions of interconnections and perspectives for further development. But when preparing for an examination, the student mainly hopes to see the material that will be on the examination. The teacher likewise, when preparing a course, selects only the material that can and must be covered in the time alloted for the course.

In this connection, it should be kept in mind that the text of the present book is noticeably more extensive than the lectures on which it is based. What caused this difference? First of all, the lectures have been supplemented by essentially an entire problem book, made up not so much of exercises as substantive problems of science or mathematics proper having a connection with the corresponding parts of the theory and in some cases significantly extending them. Second, the book naturally contains a much larger set of examples illustrating the theory in action than one can incorporate in lectures. Third and finally, a number of chapters, sections, or subsections were consciously written as a supplement to the traditional material. This is explained in the sections "On the introduction" and "On the supplementary material" in the preface to the first edition.

I would also like to recall that in the preface to the first edition I tried to warn both the student and the beginning teacher against an excessively long study of the introductory formal chapters. Such a study would noticeably delay the analysis proper and cause a great shift in emphasis.

To show what in fact can be retained of these formal introductory chapters in a realistic lecture course, and to explain in condensed form the syllabus for such a course as a whole while pointing out possible variants depending on the student audience, at the end of the book I give a list of problems from the midterm exam, along with some recent examination topics for the first two semesters, to which this first part of the book relates. From this list the professional will of course discern the order of exposition, the degree of development of the basic concepts and methods, and the occasional invocation of material from the second part of the textbook when the topic under consideration is already accessible for the audience in a more general form.[2]

In conclusion I would like to thank colleagues and students, both known and unknown to me, for reviews and constructive remarks on the first edition of the course. It was particularly interesting for me to read the reviews of A. N. Kolmogorov and V. I. Arnol'd. Very different in size, form, and style, these two have, on the professional level, so many inspiring things in common.

Moscow, 1997 *V. Zorich*

[2] Some of the transcripts of the corresponding lectures have been published and I give formal reference to the booklets published using them, although I understand that they are now available only with difficulty. (The lectures were given and published for limited circulation in the Mathematical College of the Independent University of Moscow and in the Department of Mechanics and Mathematics of Moscow State University.)

## From the Preface to the First Russian Edition

The creation of the foundations of the differential and integral calculus by Newton and Leibniz three centuries ago appears even by modern standards to be one of the greatest events in the history of science in general and mathematics in particular.

Mathematical analysis (in the broad sense of the word) and algebra have intertwined to form the root system on which the ramified tree of modern mathematics is supported and through which it makes its vital contact with the nonmathematical sphere. It is for this reason that the foundations of analysis are included as a necessary element of even modest descriptions of so-called higher mathematics; and it is probably for that reason that so many books aimed at different groups of readers are devoted to the exposition of the fundamentals of analysis.

This book has been aimed primarily at mathematicians desiring (as is proper) to obtain thorough proofs of the fundamental theorems, but who are at the same time interested in the life of these theorems outside of mathematics itself.

The characteristics of the present course connected with these circumstances reduce basically to the following:

*In the exposition.* Within each major topic the exposition is as a rule inductive, sometimes proceeding from the statement of a problem and suggestive heuristic considerations toward its solution to fundamental concepts and formalisms.

Detailed at first, the exposition becomes more and more compressed as the course progresses.

An emphasis is placed on the efficient machinery of smooth analysis. In the exposition of the theory I have tried (to the extent of my knowledge) to point out the most essential methods and facts and avoid the temptation of a minor strengthening of a theorem at the price of a major complication of its proof.

The exposition is geometric throughout wherever this seemed worthwhile in order to reveal the essence of the matter.

The main text is supplemented with a rather large collection of examples, and nearly every section ends with a set of problems that I hope will significantly complement even the theoretical part of the main text. Following the wonderful precedent of Pólya and Szegő, I have often tried to present a beautiful mathematical result or an important application as a series of problems accessible to the reader.

The arrangement of the material was dictated not only by the architecture of mathematics in the sense of Bourbaki, but also by the position of analysis as a component of a unified mathematical or, one should rather say, natural-science/mathematical education.

*In content.* This course is being published in two books (Part 1 and Part 2).

The present Part 1 contains the differential and integral calculus of functions of one variable and the differential calculus of functions of several variables.

In differential calculus we emphasize the role of the differential as a linear standard for describing the local behavior of the variation of a variable. In addition to numerous examples of the use of differential calculus to study functional relations (monotonicity, extrema) we exhibit the role of the language of analysis in writing simple differential equations – mathematical models of real-world phenomena and the substantive problems connected with them.

We study a number of such problems (for example, the motion of a body of variable mass, a nuclear reactor, atmospheric pressure, motion in a resisting medium) whose solution leads to important elementary functions. Full use is made of the language of complex variables; in particular, Euler's formula is derived and the unity of the fundamental elementary functions is shown.

The integral calculus has consciously been explained as far as possible using intuitive material in the framework of the Riemann integral. For the majority of applications, this is completely adequate.[3] Various applications of the integral are pointed out, including those that lead to an improper integral (for example, the work involved in escaping from a gravitational field, and the escape velocity for the Earth's gravitational field) or to elliptic functions (motion in a gravitational field in the presence of constraints, pendulum motion.)

The differential calculus of functions of several variables is very geometric. In this topic, for example, one studies such important and useful consequences of the implicit function theorem as curvilinear coordinates and local reduction to canonical form for smooth mappings (the rank theorem) and functions (Morse's lemma), and also the theory of extrema with constraint.

Results from the theory of continuous functions and differential calculus are summarized and explained in a general invariant form in two chapters that link up naturally with the differential calculus of real-valued functions of several variables. These two chapters open the second part of the course. The second book, in which we also discuss the integral calculus of functions of several variables up to the general Newton–Leibniz–Stokes formula thus acquires a certain unity.

We shall give more complete information on the second book in its preface. At this point we add only that, in addition to the material already mentioned, it contains information on series of functions (power series and Fourier series included), on integrals depending on a parameter (including the fundamental solution, convolution, and the Fourier transform), and also on asymptotic expansions (which are usually absent or insufficiently presented in textbooks).

We now discuss a few particular problems.

[3] The "stronger" integrals, as is well known, require fussier set-theoretic considerations, outside the mainstream of the textbook, while adding hardly anything to the effective machinery of analysis, mastery of which should be the first priority.

*On the introduction.* I have not written an introductory survey of the subject, since the majority of beginning students already have a preliminary idea of differential and integral calculus and their applications from high school, and I could hardly claim to write an even more introductory survey. Instead, in the first two chapters I bring the former high-school student's understanding of sets, functions, the use of logical symbolism, and the theory of a real number to a certain mathematical completeness.

This material belongs to the formal foundations of analysis and is aimed primarily at the mathematics major, who may at some time wish to trace the logical structure of the basic concepts and principles used in classical analysis. Mathematical analysis proper begins in the third chapter, so that the reader who wishes to get effective machinery in his hands as quickly as possible and see its applications can in general begin a first reading with Chapter 3, turning to the earlier pages whenever something seems nonobvious or raises a question which hopefully I also have thought of and answered in the early chapters.

*On the division of material.* The material of the two books is divided into chapters numbered continuously. The sections are numbered within each chapter separately; subsections of a section are numbered only within that section. Theorems, propositions, lemmas, definitions, and examples are written in italics for greater logical clarity, and numbered for convenience within each section.

*On the supplementary material.* Several chapters of the book are written as a natural extension of classical analysis. These are, on the one hand, Chapters 1 and 2 mentioned above, which are devoted to its formal mathematical foundations, and on the other hand, Chapters 9, 10, and 15 of the second part, which give the modern view of the theory of continuity, differential and integral calculus, and finally Chapter 19, which is devoted to certain effective asymptotic methods of analysis.

The question as to which part of the material of these chapters should be included in a lecture course depends on the audience and can be decided by the lecturer, but certain fundamental concepts introduced here are usually present in any exposition of the subject to mathematicians.

In conclusion, I would like to thank those whose friendly and competent professional aid has been valuable and useful to me during the work on this book.

The proposed course was quite detailed, and in many of its aspects it was coordinated with subsequent modern university mathematics courses – such as, for example, differential equations, differential geometry, the theory of functions of a complex variable, and functional analysis. In this regard my contacts and discussions with V. I. Arnol'd and the especially numerous ones with S. P. Novikov during our joint work with the so-called "experimental student group in natural-science/mathematical education" in the Department of Mathematics at MSU, were very useful to me.

I received much advice from N. V. Efimov, chair of the Section of Mathematical Analysis in the Department of Mechanics and Mathematics at Moscow State University.

I am also grateful to colleagues in the department and the section for remarks on the mimeographed edition of my lectures.

Student transcripts of my recent lectures which were made available to me were valuable during the work on this book, and I am grateful to their owners.

I am deeply grateful to the official reviewers L. D. Kudryavtsev, V. P. Petrenko, and S. B. Stechkin for constructive comments, most of which were taken into account in the book now offered to the reader.

Moscow, 1980 *V. Zorich*

# 2 The Real Numbers

Mathematical theories, as a rule, find uses because they make it possible to transform one set of numbers (the initial data) into another set of numbers constituting the intermediate or final purpose of the computations. For that reason numerical-valued functions occupy a special place in mathematics and its applications. These functions (more precisely, the so-called differentiable functions) constitute the main object of study of classical analysis. But, as you may already have sensed from your school experience, and as will soon be confirmed, any description of the properties of these functions that is at all complete from the point of view of modern mathematics is impossible without a precise definition of the set of real numbers, on which these functions operate.

Numbers in mathematics are like time in physics: everyone knows what they are, and only experts find them hard to understand. This is one of the basic mathematical abstractions, which seems destined to undergo significant further development. A very full separate course could be devoted to this subject. At present we intend only to unify what is basically already known to the reader about real numbers from high school, exhibiting as axioms the fundamental and independent properties of numbers. In doing this, our purpose is to give a precise definition of real numbers suitable for subsequent mathematical use, paying particular attention to their property of completeness or continuity, which contains the germ of the idea of passage to the limit – the basic nonarithmetical operation of analysis.

## 2.1 The Axiom System and some General Properties of the Set of Real Numbers

### 2.1.1 Definition of the Set of Real Numbers

**Definition 1.** A set $\mathbb{R}$ is called the set of *real numbers* and its elements are *real numbers* if the following list of conditions holds, called the axiom system of the real numbers.

### (I) Axioms for addition

*An operation*

$$+ : \mathbb{R} \times \mathbb{R} \to \mathbb{R} ,$$

*(the operation of addition) is defined, assigning to each ordered pair* $(x, y)$ *of elements* $x, y$ *of* $\mathbb{R}$ *a certain element* $x + y \in \mathbb{R}$, *called the* sum *of* $x$ *and* $y$. *This operation satisfies the following conditions:*

$1_+$. *There exists a neutral, or identity element* 0 *(called zero) such that*

$$x + 0 = 0 + x = x$$

*for every* $x \in \mathbb{R}$.

$2_+$. *For every element* $x \in \mathbb{R}$ *there exists an element* $-x \in \mathbb{R}$ *called the* negative *of* $x$ *such that*

$$x + (-x) = (-x) + x = 0 .$$

$3_+$. *The operation* $+$ *is associative, that is, the relation*

$$x + (y + z) = (x + y) + z$$

*holds for any elements* $x, y, z$ *of* $\mathbb{R}$.

$4_+$. *The operation* $+$ *is commutative, that is,*

$$x + y = y + x$$

*for any elements* $x, y$ *of* $\mathbb{R}$.

If an operation is defined on a set $G$ satisfying axioms $1_+$, $2_+$, and $3_+$, we say that a *group structure* is defined on $G$ or that $G$ is a *group*. If the operation is called addition, the group is called an *additive* group. If it is also known that the operation is commutative, that is, condition $4_+$ holds, the group is called *commutative* or *Abelian*.[1]

Thus, Axioms $1_+$–$4_+$ assert that $\mathbb{R}$ is an additive abelian group.

### (II) Axioms for multiplication

*An operation*

$$\bullet : \mathbb{R} \times \mathbb{R} \to \mathbb{R} ,$$

*(the operation of multiplication) is defined, assigning to each ordered pair* $(x, y)$ *of elements* $x, y$ *of* $\mathbb{R}$ *a certain element* $x \cdot y \in \mathbb{R}$, *called the* product *of* $x$ *and* $y$. *This operation satisfies the following conditions:*

[1] N. H. Abel (1802–1829) – outstanding Norwegian mathematician, who proved that the general algebraic equation of degree higher than four cannot be solved by radicals.

$1_\bullet$. *There exists a neutral, or identity element* $1 \in \mathbb{R} \setminus 0$ *(called one) such that*

$$x \cdot 1 = 1 \cdot x = x$$

*for every* $x \in \mathbb{R}$.

$2_\bullet$. *For every element* $x \in \mathbb{R} \setminus 0$ *there exists an element* $x^{-1} \in \mathbb{R}$, *called the* inverse *or* reciprocal *of* $x$, *such that*

$$x \cdot x^{-1} = x^{-1} \cdot x = 1 .$$

$3_\bullet$. *The operation* $\bullet$ *is associative, that is, the relation*

$$x \cdot (y \cdot z) = (x \cdot y) \cdot z$$

*holds for any elements* $x, y, z$ *of* $\mathbb{R}$.

$4_\bullet$. *The operation* $\bullet$ *is commutative, that is,*

$$x \cdot y = y \cdot x$$

*for any elements* $x, y$ *of* $\mathbb{R}$.

We remark that with respect to the operation of multiplication the set $\mathbb{R} \setminus 0$, as one can verify, is a (*multiplicative*) group.

(I, II) The connection between addition and multiplication

*Multiplication is distributive with respect to addition, that is*

$$(x + y)z = xz + yz$$

*for all* $x, y, z \in \mathbb{R}$.

We remark that by the commutativity of multiplication, this equality continues to hold if the order of the factors is reversed on either side.

If two operations satisfying these axioms are defined on a set $G$, then $G$ is called a *field.*

(III) Order axioms

*Between elements of* $\mathbb{R}$ *there is a relation* $\leq$, *that is, for elements* $x, y \in \mathbb{R}$ *one can determine whether* $x \leq y$ *or not. Here the following conditions must hold:*

$0_\leq$. $\forall x \in \mathbb{R}\,(x \leq x)$.

$1_\leq$. $(x \leq y) \wedge (y \leq x) \Rightarrow (x = y)$.

$2_\leq$. $(x \leq y) \wedge (y \leq z) \Rightarrow (x \leq z)$.

$3_\leq$. $\forall x \in \mathbb{R}\, \forall y \in \mathbb{R}\,(x \leq y) \vee (y \leq x)$.

The relation $\leq$ on $\mathbb{R}$ is called *inequality.*

A set on which there is a relation between pairs of elements satisfying axioms $0_{\leq}$, $1_{\leq}$, and $2_{\leq}$, as you know, is said to be *partially ordered.* If in addition axiom $3_{\leq}$ holds, that is, any two elements are comparable, the set is *linearly ordered.* Thus the set of real numbers is linearly ordered by the relation of inequality between elements.

(I, III) THE CONNECTION BETWEEN ADDITION AND ORDER ON $\mathbb{R}$

*If $x, y, z$ are elements of $\mathbb{R}$, then*

$$(x \leq y) \Rightarrow (x + z \leq y + z) .$$

(II, III) THE CONNECTION BETWEEN MULTIPLICATION AND ORDER ON $\mathbb{R}$

*If $x$ and $y$ are elements of $\mathbb{R}$, then*

$$(0 \leq x) \wedge (0 \leq y) \Rightarrow (0 \leq x \cdot y) .$$

(IV) THE AXIOM OF COMPLETENESS (CONTINUITY)

*If $X$ and $Y$ are nonempty subsets of $\mathbb{R}$ having the property that $x \leq y$ for every $x \in X$ and every $y \in Y$, then there exists $c \in \mathbb{R}$ such that $x \leq c \leq y$ for all $x \in X$ and $y \in Y$.*

We now have a complete list of axioms such that any set on which these axioms hold can be considered a concrete realization or *model* of the real numbers.

This definition does not formally require any preliminary knowledge about numbers, and from it "by turning on mathematical thought" we should, again formally, obtain as theorems all the other properties of real numbers. On the subject of this axiomatic formalism we would like to make a few informal remarks.

Imagine that you had not passed from the stage of adding apples, cubes, or other named quantities to the addition of abstract natural numbers; you had not studied the measurement of line segments and arrived at rational numbers; you did not know the great discovery of the ancients that the diagonal of a square is incommensurable with its side, so that its length cannot be a rational number, that is, that irrational numbers are needed; you did not have the concept of "greater" or "smaller" that arises in the process of measurement; you did not picture order to yourself using, for example, the real line. If all these preliminaries had not occurred, the axioms just listed would not be perceived as the outcome of intellectual progress; they would seem at the very least a strange, and in any case arbitrary, fruit of the imagination.

In relation to any abstract system of axioms, at least two questions arise immediately.

First, are these axioms consistent? That is, does there exist a set satisfying all the conditions just listed? This is the problem of *consistency* of the axioms.

Second, does the given system of axioms determine the mathematical object uniquely? That is, as the logicians would say, is the axiom system *categorical*? Here uniqueness must be understood as follows. If two people $A$ and $B$ construct models independently, say of number systems $\mathbb{R}_A$ and $\mathbb{R}_B$, satisfying the axioms, then a bijective correspondence can be established between the systems $\mathbb{R}_A$ and $\mathbb{R}_B$, say $f : \mathbb{R}_A \to \mathbb{R}_B$, preserving the arithmetic operations and the order, that is,

$$\begin{aligned} f(x+y) &= f(x)+f(y)\,, \\ f(x\cdot y) &= f(x)\cdot f(y)\,, \\ x \le y &\Leftrightarrow f(x) \le f(y)\,. \end{aligned}$$

In this case, from the mathematical point of view, $\mathbb{R}_A$ and $\mathbb{R}_B$ are merely distinct but equally valid realizations (models) of the real numbers (for example, $\mathbb{R}_A$ might be the set of infinite decimal fractions and $\mathbb{R}_B$ the set of points on the real line). Such realizations are said to be *isomorphic* and the mapping $f$ is called an *isomorphism*. The result of this mathematical activity is thus not about any particular realization, but about each model in the class of isomorphic models of the given axiom system.

We shall not discuss the questions posed above, but instead confine ourselves to giving informative answers to them.

A positive answer to the question of consistency of an axiom system is always of a hypothetical nature. In relation to numbers it has the following appearance: Starting from the axioms of set theory that we have accepted (see Subsect. 1.4.2), one can construct the set of natural numbers, then the set of rational numbers, and finally the set $\mathbb{R}$ of real numbers satisfying all the properties listed.

The question of the categoricity of the axiom system for the real numbers can be established. Those who wish to do so may obtain it independently by solving Exercises 23 and 24 at the end of this section.

### 2.1.2 Some General Algebraic Properties of Real Numbers

We shall show by examples how the known properties of numbers can be obtained from these axioms.

**a. Consequences of the Addition Axioms** $1^0$. *There is only one zero in the set of real numbers.*

*Proof.* If $0_1$ and $0_2$ are both zeros in $\mathbb{R}$, then by definition of zero,

$$0_1 = 0_1 + 0_2 = 0_2 + 0_1 = 0_2\,. \quad \square$$

$2^0$. *Each element of the set of real numbers has a unique negative.*

*Proof.* If $x_1$ and $x_2$ are both negatives of $x \in \mathbb{R}$, then

$$x_1 = x_1 + 0 = x_1 + (x + x_2) = (x_1 + x) + x_2 = 0 + x_2 = x_2 \,. \quad \square$$

Here we have used successively the definition of zero, the definition of the negative, the associativity of addition, again the definition of the negative, and finally, again the definition of zero.

$3^0$. *In the set of real numbers* $\mathbb{R}$ *the equation*

$$a + x = b$$

*has the unique solution*

$$x = b + (-a) \,.$$

*Proof.* This follows from the existence and uniqueness of the negative of every element $a \in \mathbb{R}$:

$$\begin{aligned}(a + x = b) &\Leftrightarrow \big((x + a) + (-a) = b + (-a)\big) \Leftrightarrow \\ &\Leftrightarrow \big(x + (a + (-a)) = b + (-a)\big) \Leftrightarrow (x + 0 = b + (-a)) \Leftrightarrow \\ &\Leftrightarrow \big(x = b + (-a)\big) \,. \quad \square\end{aligned}$$

The expression $b + (-a)$ can also be written as $b - a$. This is the shorter and more common way of writing it, to which we shall adhere.

**b. Consequences of the Multiplication Axioms** $1^0$. *There is only one multiplicative unit in the real numbers.*

$2^0$. *For each* $x \neq 0$ *there is only one reciprocal* $x^{-1}$.

$3^0$. *For* $a \in \mathbb{R} \setminus 0$, *the equation* $a \cdot x = b$ *has the unique solution* $x = b \cdot a^{-1}$.

The proofs of these propositions, of course, merely repeat the proofs of the corresponding propositions for addition (except for a change in the symbol and the name of the operation); they are therefore omitted.

**c. Consequences of the Axiom Connecting Addition and Multiplication** Applying the additional axiom (I, II) connecting addition and multiplication, we obtain further consequences.

$1^0$. *For any* $x \in \mathbb{R}$

$$x \cdot 0 = 0 \cdot x = 0 \,.$$

*Proof.*

$$(x \cdot 0 = x \cdot (0 + 0) = x \cdot 0 + x \cdot 0) \Rightarrow (x \cdot 0 = x \cdot 0 + (-(x \cdot 0)) = 0) \,. \quad \square$$

From this result, incidentally, one can see that if $x \in \mathbb{R}\backslash 0$, then $x^{-1} \in \mathbb{R}\backslash 0$.

$2^0$. $(x \cdot y = 0) \Rightarrow (x = 0) \vee (y = 0)$.

*Proof.* If, for example, $y \neq 0$, then by the uniqueness of the solution of the equation $x \cdot y = 0$ for $x$, we find $x = 0 \cdot y^{-1} = 0$. □

$3^0$. *For any* $x \in \mathbb{R}$

$$-x = (-1) \cdot x \, .$$

*Proof.* $x + (-1) \cdot x = \big(1 + (-1)\big) \cdot x = 0 \cdot x = x \cdot 0 = 0$, and the assertion now follows from the uniqueness of the negative of a number. □

$4^0$. *For any* $x \in \mathbb{R}$

$$(-1)(-x) = x \, .$$

*Proof.* This follows from $3^0$ and the uniqueness of the negative of $-x$. □

$5^0$. *For any* $x \in \mathbb{R}$

$$(-x) \cdot (-x) = x \cdot x \, .$$

*Proof.*

$$(-x)(-x) = \big((-1) \cdot x\big)(-x) = \big(x \cdot (-1)\big)(-x) = x\big((-1)(-x)\big) = x \cdot x \, .$$

Here we have made successive use of the preceding propositions and the commutativity and associativity of multiplication. □

**d. Consequences of the Order Axioms** We begin by noting that the relation $x \leq y$ (read "$x$ is less than or equal to $y$") can also be written as $y \geq x$ ("$y$ is greater than or equal to $x$"); when $x \neq y$, the relation $x \leq y$ is written $x < y$ (read "$x$ is less than $y$") or $y > x$ (read "$y$ is greater than $x$"), and is called *strict inequality.*

$1^0$. *For any* $x$ *and* $y$ *in* $\mathbb{R}$ *precisely one of the following relations holds:*

$$x < y, \qquad x = y, \qquad x > y \, .$$

*Proof.* This follows from the definition of strict inequality just given and axioms $1_{\leq}$ and $3_{\leq}$. □

$2^0$. *For any* $x, y, z \in \mathbb{R}$

$$(x < y) \wedge (y \leq z) \Rightarrow (x < z) \, ,$$
$$(x \leq y) \wedge (y < z) \Rightarrow (x < z) \, .$$

*Proof.* We prove the first assertion as an example. By Axiom $2_{\leq}$, which asserts that the inequality relation is transitive, we have

$$(x \leq y) \wedge (y < z) \Leftrightarrow (x \leq y) \wedge (y \leq z) \wedge (y \neq z) \Rightarrow (x \leq z) .$$

It remains to be verified that $x \neq z$. But if this were not the case, we would have

$$(x \leq y) \wedge (y < z) \Leftrightarrow (z \leq y) \wedge (y < z) \Leftrightarrow (z \leq y) \wedge (y \leq z) \wedge (y \neq z) .$$

By Axiom $1_{\leq}$ this relation would imply

$$(y = z) \wedge (y \neq z) ,$$

which is a contradiction. □

**e. Consequences of the Axioms Connecting Order with Addition and Multiplication** If in addition to the axioms of addition, multiplication, and order, we use axioms (I,III) and (II, III), which connect the order with the arithmetic operations, we can obtain, for example, the following propositions.

$1^0$. *For any* $x, y, z, w \in \mathbb{R}$

$$\begin{aligned} (x < y) &\Rightarrow (x + z) < (y + z) , \\ (0 < x) &\Rightarrow (-x < 0) , \\ (x \leq y) \wedge (z \leq w) &\Rightarrow (x + z) \leq (y + w) , \\ (x \leq y) \wedge (z < w) &\Rightarrow (x + z < y + w) . \end{aligned}$$

*Proof.* We shall verify the first of these assertions.

By definition of strict inequality and the axiom (I,III) we have

$$(x < y) \Rightarrow (x \leq y) \Rightarrow (x + z) \leq (y + z) .$$

It remains to be verified that $x + z \neq y + z$. Indeed,

$$\big((x + z) = (y + z)\big) \Rightarrow \big(x = (y + z) - z = y + (z - z) = y\big) ,$$

which contradicts the assumption $x < y$. □

$2^0$. *If* $x, y, z \in \mathbb{R}$, *then*

$$\begin{aligned} (0 < x) \wedge (0 < y) &\Rightarrow (0 < xy) , \\ (x < 0) \wedge (y < 0) &\Rightarrow (0 < xy) , \\ (x < 0) \wedge (0 < y) &\Rightarrow (xy < 0) , \\ (x < y) \wedge (0 < z) &\Rightarrow (xz < yz) , \\ (x < y) \wedge (z < 0) &\Rightarrow (yz < xz) . \end{aligned}$$

*Proof.* We shall verify the first of these assertions. By definition of strict inequality and the axiom (II,III) we have

$$(0 < x) \wedge (0 < y) \Rightarrow (0 \leq x) \wedge (0 \leq y) \Rightarrow (0 \leq xy) .$$

Moreover, $0 \neq xy$ since, as already shown,

$$(x \cdot y = 0) \Rightarrow (x = 0) \vee (y = 0) .$$

Let us further verify, for example, the third assertion:

$$\begin{aligned}(x < 0) \wedge (0 < y) &\Rightarrow (0 < -x) \wedge (0 < y) \Rightarrow \\ &\Rightarrow (0 < (-x) \cdot y) \Rightarrow \big(0 < \big((-1) \cdot x\big)y\big) \Rightarrow \\ &\Rightarrow \big(0 < (-1) \cdot (xy)\big) \Rightarrow \big(0 < -(xy)\big) \Rightarrow (xy < 0) .\square\end{aligned}$$

The reader is now invited to prove the remaining relations independently and also to verify that if nonstrict inequality holds in one of the parentheses on the left-hand side, then the inequality on the right-hand side will also be nonstrict.

$3^0$. $0 < 1$.

*Proof.* We know that $1 \in \mathbb{R} \setminus 0$, that is $0 \neq 1$. If we assume $1 < 0$, then by what was just proved,

$$(1 < 0) \wedge (1 < 0) \Rightarrow (0 < 1 \cdot 1) \Rightarrow (0 < 1) .$$

But we know that for any pair of numbers $x, y \in \mathbb{R}$ exactly one of the possibilities $x < y$, $x = y$, $x > y$ actually holds. Since $0 \neq 1$ and the assumption $1 < 0$ implies the relation $0 < 1$, which contradicts it, the only remaining possibility is the one in the statement of the proposition. □

$4^0$. $(0 < x) \Rightarrow (0 < x^{-1})$ and $(0 < x) \wedge (x < y) \Rightarrow (0 < y^{-1}) \wedge (y^{-1} < x^{-1})$.

*Proof.* Let us verify the first of these assertions. First of all, $x^{-1} \neq 0$. Assuming $x^{-1} < 0$, we obtain

$$(x^{-1} < 0) \wedge (0 < x) \Rightarrow (x \cdot x^{-1} < 0) \Rightarrow (1 < 0) .$$

This contradiction completes the proof. □

We recall that numbers larger than zero are called *positive* and those less than zero *negative*.

Thus we have shown, for example, that 1 is a positive number, that the product of a positive and a negative number is a negative number, and that the reciprocal of a positive number is also positive.

### 2.1.3 The Completeness Axiom and the Existence of a Least Upper (or Greatest Lower) Bound of a Set of Numbers

**Definition 2.** A set $X \subset \mathbb{R}$ is said to be *bounded above* (resp. *bounded below*) if there exists a number $c \in \mathbb{R}$ such that $x \leq c$ (resp. $c \leq x$) for all $x \in X$.

The number $c$ in this case is called an *upper bound* (resp. *lower bound*) of the set $X$. It is also called a *majorant* (resp. *minorant*) of $X$.

**Definition 3.** A set that is bounded both above and below is called *bounded.*

**Definition 4.** An element $a \in X$ is called the *largest* or *maximal* (resp. *smallest* or *minimal*) element of $X$ if $x \leq a$ (resp. $a \leq x$) for all $x \in X$.

We now introduce some notation and at the same time give a formal expression to the definition of maximal and minimal elements:

$$(a = \max X) := \big(a \in X \wedge \forall x \in X\,(x \leq a)\big)\,,$$
$$(a = \min X) := \big(a \in X \wedge \forall x \in X\,(a \leq x)\big)\,.$$

Along with the notation $\max X$ (read "the maximum of $X$") and $\min X$ (read "the minimum of $X$") we also use the respective expressions $\max\limits_{x \in X} x$ and $\min\limits_{x \in X} x$.

It follows immediately from the order axiom $1_{\leq}$ that if there is a maximal (resp. minimal) element in a set of numbers, it is the only one.

However, not every set, not even every bounded set, has a maximal or minimal element.

For example, the set $X = \{x \in \mathbb{R} \mid 0 \leq x < 1\}$ has a minimal element. But, as one can easily verify, it has no maximal element.

**Definition 5.** The smallest number that bounds a set $X \subset \mathbb{R}$ from above is called the *least upper bound* (or the *exact upper bound*) of $X$ and denoted $\sup X$ (read "the supremum of $X$") or $\sup\limits_{x \in X} x$.

This is the basic concept of the present subsection. Thus

$$(s = \sup X) := \forall x \in X\,\big((x \leq s) \wedge \big(\forall s' < s\ \exists x' \in X\,(s' < x')\big)\big)\,.$$

The expression in the first set of parentheses on the right-hand side here says that $s$ is an upper bound for $X$; the expression in the second set says that $s$ is the smallest number having this property. More precisely, the expression in the second set of parentheses asserts that any number smaller than $s$ is not an upper bound of $X$.

The concept of the *greatest lower bound* (or *exact lower bound*) of a set $X$ is introduced similarly as the largest of the lower bounds of $X$.

**Definition 6.**

$$(i = \inf X) := \forall x \in X \left((i \leq x) \wedge \left(\forall i' > i \; \exists x' \in X \, (x' < i')\right)\right) .$$

Along with the notation $\inf X$ (read "the infimum of $X$") one also uses the notation $\inf\limits_{x \in X} x$ for the greatest lower bound of $X$.

Thus we have given the following definitions:

$$\sup X := \min \left\{c \in \mathbb{R} | \, \forall x \in X \, (x \leq c)\right\} ,$$
$$\inf X := \max \left\{c \in \mathbb{R} | \, \forall x \in X \, (c \leq x)\right\} .$$

But we said above that not every set has a minimal or maximal element. Therefore the definitions we have adopted for the least upper bound and greatest lower bound require an argument, provided by the following lemma.

**Lemma.** (The least upper bound principle). *Every nonempty set of real numbers that is bounded from above has a unique least upper bound.*

*Proof.* Since we already know that the minimal element of a set of numbers is unique, we need only verify that the least upper bound exists.

Let $X \subset \mathbb{R}$ be a given set and $Y = \left\{y \in \mathbb{R} | \, \forall x \in X \, (x \leq y)\right\}$. By hypothesis, $X \neq \varnothing$ and $Y \neq \varnothing$. Then, by the completeness axiom there exists $c \in \mathbb{R}$ such that $\forall x \in X \, \forall y \in Y \, (x \leq c \leq y)$. The number $c$ is therefore both a majorant of $X$ and a minorant of $Y$. Being a majorant of $X$, $c$ is an element of $Y$. But then, as a minorant of $Y$, it must be the minimal element of $Y$. Thus $c = \min Y = \sup X$. □

Naturally the existence and uniqueness of the greatest lower bound of a set of numbers that is bounded from below is analogous, that is, the following proposition holds.

**Lemma.** ($X$ bounded below) $\Rightarrow$ ($\exists!$ $\inf X$).

We shall not take time to give the proof.

We now return to the set $X = \{x \in \mathbb{R} | \, 0 \leq x < 1\}$. By the lemma just proved it must have a least upper bound. By the very definition of the set $X$ and the definition of the least upper bound, it is obvious that $\sup X \leq 1$.

To prove that $\sup X = 1$ it is thus necessary to verify that for any number $q < 1$ there exists $x \in X$ such that $q < x$; simply put, this means merely that there are numbers between $q$ and 1. This of course, is also easy to prove independently (for example, by showing that $q < 2^{-1}(q+1) < 1$), but we shall not do so at this point, since such questions will be discussed systematically and in detail in the next section.

As for the greatest lower bound, it always coincides with the minimal element of a set, if such an element exists. Thus, from this consideration alone we have $\inf X = 0$ in the present example.

Other, more substantive examples of the use of the concepts introduced here will be encountered in the next section.

## 2.2 The Most Important Classes of Real Numbers and Computational Aspects of Operations with Real Numbers

### 2.2.1 The Natural Numbers and the Principle of Mathematical Induction

**a. Definition of the Set of Natural Numbers** The numbers of the form $1$, $1+1$, $(1+1)+1$, and so forth are denoted respectively by $1, 2, 3, \ldots$ and so forth and are called *natural numbers.*

Such a definition will be meaningful only to one who already has a complete picture of the natural numbers, including the notation for them, for example in the decimal system of computation.

The continuation of such a process is by no means always unique, so that the ubiquitous "and so forth" actually requires a clarification provided by the fundamental principle of mathematical induction.

**Definition 1.** A set $X \subset \mathbb{R}$ is *inductive* if for each number $x \in X$, it also contains $x+1$.

For example, $\mathbb{R}$ is an inductive set; the set of positive numbers is also inductive.

The intersection $X = \bigcap_{\alpha \in A} X_\alpha$ of any family of inductive sets $X_\alpha$, if not empty, is an inductive set.

Indeed,

$$\left(x \in X = \bigcap_{\alpha \in A} X_\alpha\right) \Rightarrow \left(\forall \alpha \in A\,(x \in X_\alpha)\right) \Rightarrow$$

$$\Rightarrow \left(\forall \alpha \in A\,((x+1) \in X_\alpha)\right) \Rightarrow \left((x+1) \in \bigcap_{\alpha \in A} X_\alpha = X\right).$$

We now adopt the following definition.

**Definition 2.** The set of *natural numbers* is the smallest inductive set containing 1, that is, the intersection of all inductive sets that contain 1.

The set of natural numbers is denoted $\mathbb{N}$; its elements are called *natural numbers.*

From the set-theoretic point of view it might be more rational to begin the natural numbers with 0, that is, to introduce the set of natural numbers as the smallest inductive set containing 0; however, it is more convenient for us to begin numbering with 1.

The following fundamental and widely used principle is a direct corollary of the definition of the set of natural numbers.

**b. The Principle of Mathematical Induction** *If a subset $E$ of the set of natural numbers $\mathbb{N}$ is such that $1 \in E$ and together with each number $x \in E$, the number $x+1$ also belongs to $E$, then $E = \mathbb{N}$.*

Thus,

$$(E \subset \mathbb{N}) \wedge (1 \in E) \wedge \big(\forall x \in E \,\big(x \in E \Rightarrow (x+1) \in E\big)\big) \Rightarrow E = \mathbb{N}\,.$$

Let us illustrate this principle in action by using it to prove several useful properties of the natural numbers that we will be using constantly from now on.

$1^0$. *The sum and product of natural numbers are natural numbers.*

*Proof.* Let $m, n \in \mathbb{N}$; we shall show that $(m+n) \in \mathbb{N}$. We denote by $E$ the set of natural numbers $n$ for which $(m+n) \in \mathbb{N}$ for all $m \in \mathbb{N}$. Then $1 \in E$ since $(m \in \mathbb{N}) \Rightarrow \big((m+1) \in \mathbb{N}\big)$ for any $m \in \mathbb{N}$. If $n \in E$, that is, $(m+n) \in \mathbb{N}$, then $(n+1) \in E$ also, since $\big(m+(n+1)\big) = \big((m+n)+1\big) \in \mathbb{N}$. By the principle of induction, $E = \mathbb{N}$, and we have proved that addition does not lead outside of $\mathbb{N}$.

Similarly, taking $E$ to be the set of natural numbers $n$ for which $(m \cdot n) \in \mathbb{N}$ for all $m \in \mathbb{N}$, we find that $1 \in E$, since $m \cdot 1 = m$, and if $n \in E$, that is, $m \cdot n \in \mathbb{N}$, then $m \cdot (n+1) = mn + m$ is the sum of two natural numbers, which belongs to $\mathbb{N}$ by what was just proved above. Thus $(n \in E) \Rightarrow \big((n+1) \in E\big)$, and so by the principle of induction $E = \mathbb{N}$. □

$2^0$. $(n \in \mathbb{N}) \wedge (n \neq 1) \Rightarrow \big((n-1) \in \mathbb{N}\big)$.

*Proof.* Consider the set $E$ consisting of all real numbers of the form $n-1$, where $n$ is a natural number different from 1; we shall show that $E = \mathbb{N}$. Since $1 \in \mathbb{N}$, it follows that $2 := (1+1) \in \mathbb{N}$ and hence $1 = (2-1) \in E$.

If $m \in E$, then $m = n-1$, where $n \in \mathbb{N}$; then $m+1 = (n+1)-1$, and since $n+1 \in \mathbb{N}$, we have $(m+1) \in E$. By the principle of induction we conclude that $E = \mathbb{N}$. □

$3^0$. *For any $n \in \mathbb{N}$ the set $\{x \in \mathbb{N} | \, n < x\}$ contains a minimal element, namely*

$$\min\{x \in \mathbb{N} | \, n < x\} = n+1\,.$$

*Proof.* We shall show that the set $E$ of $n \in \mathbb{N}$ for which the assertion holds coincides with $\mathbb{N}$.

We first verify that $1 \in E$, that is,

$$\min\{x \in \mathbb{N} | \, 1 < x\} = 2\,.$$

We shall also verify this assertion by the principle of induction. Let

$$M = \big\{x \in \mathbb{N} | \, (x = 1) \vee (2 \leq x)\big\}\,.$$

By definition of $M$ we have $1 \in M$. Then if $x \in M$, either $x = 1$, in which case $x + 1 = 2 \in M$, or else $2 \leq x$, and then $2 \leq (x + 1)$, and once again $(x + 1) \in M$. Thus $M = \mathbb{N}$, and hence if $(x \neq 1) \wedge (x \in \mathbb{N})$, then $2 \leq x$, that is, indeed $\min\{x \in \mathbb{N} | \, 1 < x\} = 2$. Hence $1 \in E$.

We now show that if $n \in E$, then $(n + 1) \in E$.

We begin by remarking that if $x \in \{x \in \mathbb{N} | \, n + 1 < x\}$, then

$$(x - 1) = y \in \{y \in \mathbb{N} | \, n < y\} \, .$$

For, by what has already been proved, every natural number is at least as large as 1; therefore $(n + 1 < x) \Rightarrow (1 < x) \Rightarrow (x \neq 1)$, and then by the assertion in $2^0$ we have $(x - 1) = y \in \mathbb{N}$.

Now let $n \in E$, that is, $\min\{y \in \mathbb{N} | \, n < y\} = n+1$. Then $x-1 \geq y \geq n+1$ and $x \geq n + 2$. Hence,

$$\big(x \in \{x \in \mathbb{N} | \, n + 1 < x\}\big) \Rightarrow (x \geq n + 2)$$

and consequently, $\min\{x \in \mathbb{N} | \, n + 1 < x\} = n + 2$, that is, $(n + 1) \in E$.

By the principle of induction $E = \mathbb{N}$, and $3^0$ is now proved. □

As immediate corollaries of $2^0$ and $3^0$ above, we obtain the following properties ($4^0$, $5^0$, and $6^0$) of the natural numbers.

$4^0$. $(m \in \mathbb{N}) \wedge (n \in \mathbb{N}) \wedge (n < m) \Rightarrow (n + 1 \leq m)$.

$5^0$. *The number* $(n+1) \in \mathbb{N}$ *is the immediate successor of the number* $n \in \mathbb{N}$*; that is, if* $n \in \mathbb{N}$*, there are no natural numbers* $x$ *satisfying* $n < x < n + 1$.

$6^0$. *If* $n \in \mathbb{N}$ *and* $n \neq 1$*, then* $(n - 1) \in \mathbb{N}$ *and* $(n - 1)$ *is the immediate predecessor of* $n$ *in* $\mathbb{N}$*; that is, if* $n \in \mathbb{N}$*, there are no natural numbers* $x$ *satisfying* $n - 1 < x < n$.

We now prove one more property of the set of natural numbers.

$7^0$. *In any nonempty subset of the set of natural numbers there is a minimal element.*

*Proof.* Let $M \subset \mathbb{N}$. If $1 \in M$, then $\min M = 1$, since $\forall n \in \mathbb{N} \, (1 \leq n)$.

Now suppose $1 \notin M$, that is, $1 \in E = \mathbb{N} \setminus M$. The set $E$ must contain a natural number $n$ such that all natural numbers not larger than $n$ belong to $E$, but $(n + 1) \in M$. If there were no such $n$, the set $E \subset \mathbb{N}$, which contains 1, would contain along with each of its elements $n$, the number $(n + 1)$ also; by the principle of induction, it would therefore equal $\mathbb{N}$. But the latter is impossible, since $\mathbb{N} \setminus E = M \neq \varnothing$.

The number $(n + 1)$ so found must be the smallest element of $M$, since there are no natural numbers between $n$ and $n + 1$, as we have seen. □

### 2.2.2 Rational and Irrational Numbers

#### a. The Integers

**Definition 3.** The union of the set of natural numbers, the set of negatives of natural numbers, and zero is called the set of *integers* and is denoted $\mathbb{Z}$.

Since, as has already been proved, addition and multiplication of natural numbers do not take us outside $\mathbb{N}$, it follows that these same operations on integers do not lead outside of $\mathbb{Z}$.

*Proof.* Indeed, if $m, n \in \mathbb{Z}$, either one of these numbers is zero, and then the sum $m+n$ equals the other number, so that $(m+n) \in \mathbb{Z}$ and $m \cdot n = 0 \in \mathbb{Z}$, or both numbers are non-zero. In the latter case, either $m, n \in \mathbb{N}$ and then $(m+n) \in \mathbb{N} \subset \mathbb{Z}$ and $(m \cdot n) \in \mathbb{N} \subset \mathbb{Z}$, or $(-m), (-n) \in \mathbb{N}$ and then $m \cdot n = \big((-1)m\big)\big((-1)n\big) \in \mathbb{N}$ or $(-m), n \in \mathbb{N}$ and then $(-m \cdot n) \in \mathbb{N}$, that is, $m \cdot n \in \mathbb{Z}$, or, finally, $m, -n \in \mathbb{N}$ and then $(-m \cdot n) \in \mathbb{N}$ and once again $m \cdot n \in \mathbb{Z}$. □

Thus $\mathbb{Z}$ is an Abelian group with respect to addition. With respect to multiplication $\mathbb{Z}$ is not a group, nor is $\mathbb{Z} \setminus 0$, since the reciprocals of the integers are not in $\mathbb{Z}$ (except the reciprocals of 1 and $-1$).

*Proof.* Indeed, if $m \in \mathbb{Z}$ and $m \neq 0, 1$, then assuming first that $m \in \mathbb{N}$, we have $0 < 1 < m$, and, since $m \cdot m^{-1} = 1 > 0$, we must have $0 < m^{-1} < 1$ (see the consequences of the order axioms in the previous subsection). Thus $m^{-1} \notin \mathbb{Z}$. The case when $m$ is a negative integer different from $-1$ reduces immediately to the one already considered. □

When $k = m \cdot n^{-1} \in \mathbb{Z}$ for two integers $m, n \in \mathbb{Z}$, that is, when $m = k \cdot n$ for some $k \in \mathbb{Z}$, we say that $m$ is *divisible* by $n$ or a *multiple* of $n$, or that $n$ is a *divisor* of $m$.

The divisibility of integers reduces immediately via suitable sign changes, that is, through multiplication by $-1$ when necessary, to the divisibility of the corresponding natural numbers. In this context it is studied in number theory.

We recall without proof the so-called fundamental theorem of arithmetic, which we shall use in studying certain examples.

A number $p \in \mathbb{N}$, $p \neq 1$, is *prime* if it has no divisors in $\mathbb{N}$ except 1 and $p$.

**The fundamental theorem of arithmetic.** *Each natural number admits a representation as a product*

$$n = p_1 \cdots p_k \,,$$

*where $p_1, \ldots, p_k$ are prime numbers. This representation is unique except for the order of the factors.*

Numbers $m, n \in \mathbb{Z}$ are said to be *relatively prime* if they have no common divisors except 1 and $-1$.

It follows in particular from this theorem that if the product $m \cdot n$ of relatively prime numbers $m$ and $n$ is divisible by a prime $p$, then one of the two numbers is also divisible by $p$.

### b. The Rational Numbers

**Definition 4.** Numbers of the form $m \cdot n^{-1}$, where $m, n \in \mathbb{Z}$, are called *rational.*

We denote the set of rational numbers by $\mathbb{Q}$.

Thus, the ordered pair $(m, n)$ of integers defines the rational number $q = m \cdot n^{-1}$ if $n \neq 0$.

The number $q = m \cdot n^{-1}$ can also be written as a quotient[2] of $m$ and $n$, that is, as a so-called rational fraction $\frac{m}{n}$.

The rules you learned in school for operating with rational numbers in terms of their representation as fractions follow immediately from the definition of a rational number and the axioms for real numbers. In particular, "the value of a fraction is unchanged when both numerator and denominator are multiplied by the same non-zero integer", that is, the fractions $\frac{mk}{nk}$ and $\frac{m}{n}$ represent the same rational number. In fact, since $(nk)(k^{-1}n^{-1}) = 1$, that is $(n \cdot k)^{-1} = k^{-1} \cdot n^{-1}$, we have $(mk)(nk)^{-1} = (mk)(k^{-1}n^{-1}) = m \cdot n^{-1}$.

Thus the different ordered pairs $(m, n)$ and $(mk, nk)$ define the same rational number. Consequently, after suitable reductions, any rational number can be presented as an ordered pair of relatively prime integers.

On the other hand, if the pairs $(m_1, n_1)$ and $(m_2, n_2)$ define the same rational number, that is, $m_1 \cdot n_1^{-1} = m_2 \cdot n_2^{-1}$, then $m_1 n_2 = m_2 n_1$, and if, for example, $m_1$ and $n_1$ are relatively prime, it follows from the corollary of the fundamental theorem of arithmetic mentioned above that $n_2 \cdot n_1^{-1} = m_2 \cdot m_1^{-1} = k \in \mathbb{Z}$.

We have thus demonstrated that two ordered pairs $(m_1, n_1)$ and $(m_2, n_2)$ define the same rational number if and only if they are proportional. That is, there exists an integer $k \in \mathbb{Z}$ such that, for example, $m_2 = km_1$ and $n_2 = kn_1$.

### c. The Irrational Numbers

**Definition 5.** The real numbers that are not rational are called *irrational.*

The classical example of an irrational real number is $\sqrt{2}$, that is, the number $s \in \mathbb{R}$ such that $s > 0$ and $s^2 = 2$. By the Pythagorean theorem, the

[2] The notation $\mathbb{Q}$ comes from the first letter of the English word *quotient*, which in turn comes from the Latin *quota*, meaning the unit part of something, and *quot*, meaning *how many.*

irrationality of $\sqrt{2}$ is equivalent to the assertion that the diagonal and side of a square are incommensurable.

Thus we begin by verifying that *there exists a real number* $s \in \mathbb{R}$ *whose square equals* 2, and then that $s \notin \mathbb{Q}$.

*Proof.* Let $X$ and $Y$ be the sets of positive real numbers such that $\forall x \in X\,(x^2 < 2)$, $\forall y \in Y\,(2 < y^2)$. Since $1 \in X$ and $2 \in Y$, it follows that $X$ and $Y$ are nonempty sets.

Further, since $(x < y) \Leftrightarrow (x^2 < y^2)$ for positive numbers $x$ and $y$, every element of $X$ is less than every element of $Y$. By the completeness axiom there exists $s \in \mathbb{R}$ such that $x \leq s \leq y$ for all $x \in X$ and all $y \in Y$.

We shall show that $s^2 = 2$.

If $s^2 < 2$, then, for example, the number $s + \frac{2-s^2}{3s}$, which is larger than $s$, would have a square less than 2. Indeed, we know that $1 \in X$, so that $1^2 \leq s^2 < 2$, and $0 < \Delta := 2 - s^2 < 1$. It follows that

$$\left(s + \frac{\Delta}{3s}\right)^2 = s^2 + 2 \cdot \frac{\Delta}{3s} + \left(\frac{\Delta}{3s}\right)^2 < s^2 + 3 \cdot \frac{\Delta}{3s} < s^2 + 3 \cdot \frac{\Delta}{3s} = s^2 + \Delta = 2\,.$$

Consequently, $\left(s + \frac{\Delta}{3s}\right) \in X$, which is inconsistent with the inequality $x \leq s$ for all $x \in X$.

If $2 < s^2$, then the number $s - \frac{s^2-2}{3s}$, which is smaller than $s$, would have a square larger than 2. Indeed, we know that $2 \in Y$, so that $2 < s^2 \leq 2^2$ or $0 < \Delta := s^2 - 2 < 3$ and $0 < \frac{\Delta}{3} < 1$. Hence,

$$\left(s - \frac{\Delta}{3s}\right)^2 = s^2 - 2 \cdot \frac{\Delta}{3s} + \left(\frac{\Delta}{3s}\right)^2 > s^2 - 3 \cdot \frac{\Delta}{3s} = s^2 - \Delta = 2\,,$$

and we have now contradicted the fact that $s$ is a lower bound of $Y$.

Thus the only remaining possibility is that $s^2 = 2$.

Let us show, finally, that $s \notin Q$. Assume that $s \in \mathbb{Q}$ and let $\frac{m}{n}$ be an irreducible representation of $s$. Then $m^2 = 2 \cdot n^2$, so that $m^2$ is divisible by 2 and therefore $m$ also is divisible by 2. But, if $m = 2k$, then $2k^2 = n^2$, and for the same reason, $n$ must be divisible by 2. But this contradicts the assumed irreducibility of the fraction $\frac{m}{n}$. □

We have worked hard just now to prove that there exist irrational numbers. We shall soon see that in a certain sense nearly all real numbers are irrational. It will be shown that the cardinality of the set of irrational numbers is larger than that of the set of rational numbers and that in fact the former equals the cardinality of the set of real numbers.

Among the irrational numbers we make a further distinction between the so-called algebraic irrational numbers and the transcendental numbers.

A real number is called *algebraic* if it is the root of an algebraic equation

$$a_0 x^n + \cdots + a_{n-1} x + a_n = 0$$

with rational (or equivalently, integer) cofficients.

Otherwise the number is called *transcendental*.

We shall see that the cardinality of the set of algebraic numbers is the same as that of the set of rational numbers, while the cardinality of the set of transcendental numbers is the same as that of the set of real numbers. For that reason the difficulties involved in exhibiting specific transcendental numbers – more precisely, proving that a given number is transcendental – seem at first sight paradoxical and unnatural.

For example, it was not proved until 1882 that the classical geometric number $\pi$ is transcendental,[3] and one of the famous Hilbert[4] problems was to prove the transcendence of the number $\alpha^{\beta}$, where $\alpha$ is algebraic, $(\alpha > 0) \wedge (\alpha \neq 1)$ and $\beta$ is an irrational algebraic number (for example, $\alpha = 2$, $\beta = \sqrt{2}$).

### 2.2.3 The Principle of Archimedes

We now turn to the principle of Archimedes,[5] which is important in both its theoretical aspect and the application of numbers in measurement and computations. We shall prove it using the completeness axiom (more precisely, the least-upper-bound principle, which is equivalent to the completeness axiom). In other axiom systems for the real numbers this fundamental principle is frequently included in the list of axioms.

We remark that the propositions that we have proved up to now about the natural numbers and the integers have made no use at all of the completeness axiom. As will be seen below, the principle of Archimedes essentially reflects the properties of the natural numbers and integers connected with completeness. We begin with these properties.

---

[3] The number $\pi$ equals the ratio of the circumference of a circle to its diameter in Euclidean geometry. That is the reason this number has been conventionally denoted since the eighteenth century, following Euler by $\pi$, which is the initial letter of the Greek word $\pi\varepsilon\rho\iota\varphi\acute{\varepsilon}\rho\iota\alpha$ – *periphery* (circumference). The transcendence of $\pi$ was proved by the German mathematician F. Lindemann (1852–1939). It follows in particular from the transcendence of $\pi$ that it is impossible to construct a line segment of length $\pi$ with compass and straightedge (the problem of rectification of the circle), and also that the ancient problem of squaring the circle cannot be solved with compass and straightedge.

[4] D. Hilbert (1862–1943) – outstanding German mathematician who stated 23 problems from different areas of mathematics at the 1900 International Congress of Mathematicians in Paris. These problems came to be known as the "Hilbert problems". The problem mentioned here (Hilbert's seventh problem) was given an affirmative answer in 1934 by the Soviet mathematician A. O. Gel'fond (1906–1968) and the German mathematician T. Schneider (1911–1989).

[5] Archimedes (287–212 BCE) – brilliant Greek scholar, about whom Leibniz, one of the founders of analysis said, "When you study the works of Archimedes, you cease to be amazed by the achievements of modern mathematicians."

$1^0$. *Any nonempty subset of natural numbers that is bounded from above contains a maximal element.*

*Proof.* If $E \subset \mathbb{N}$ is the subset in question, then by the least-upper-bound lemma, $\exists! \sup E = s \in \mathbb{R}$. By definition of the least upper bound there is a natural number $n \in E$ satisfying the condition $s - 1 < n \leq s$. But then, $n = \max E$, since a natural number that is larger than $n$ must be at least $n + 1$, and $n + 1 > s$. □

**Corollaries** $2^0$. *The set of natural numbers is not bounded above.*

*Proof.* Otherwise there would exist a maximal natural number. But $n < n+1$. □

$3^0$. *Any nonempty subset of the integers that is bounded from above contains a maximal element.*

*Proof.* The proof of $1^0$ can be repeated verbatim, replacing $\mathbb{N}$ with $\mathbb{Z}$. □

$4^0$. *Any nonempty subset of integers that is bounded below contains a minimal element.*

*Proof.* One can, for example, repeat the proof of $1^0$, replacing $\mathbb{N}$ by $\mathbb{Z}$ and using the greatest-lower-bound principle instead of the least-upper-bound principle.

Alternatively, one can pass to the negatives of the numbers ("change signs") and use what has been proved in $3^0$. □

$5^0$. *The set of integers is unbounded above and unbounded below.*

*Proof.* This follows from $3^0$ and $4^0$, or directly from $2^0$. □

We can now state the principle of Archimedes.

$6^0$. (The principle of Archimedes). *For any fixed positive number $h$ and any real number $x$ there exists a unique integer $k$ such that $(k-1)h \leq x < kh$.*

*Proof.* Since $\mathbb{Z}$ is not bounded above, the set $\left\{n \in \mathbb{Z} | \frac{x}{h} < n\right\}$ is a nonempty subset of the integers that is bounded below. Then (see $4^0$) it contains a minimal element $k$, that is $(k-1) \leq x/h < k$. Since $h > 0$, these inequalities are equivalent to those given in the statement of the principle of Archimedes. The uniqueness of $k \in \mathbb{Z}$ satisfying these two inequalities follows from the uniqueness of the minimal element of a set of numbers (see Subsect. 2.1.3). □

And now some corollaries:

$7^0$. *For any positive number $\varepsilon$ there exists a natural number $n$ such that $0 < \frac{1}{n} < \varepsilon$.*

*Proof.* By the principle of Archimedes there exists $n \in \mathbb{Z}$ such that $1 < \varepsilon \cdot n$. Since $0 < 1$ and $0 < \varepsilon$, we have $0 < n$. Thus $n \in \mathbb{N}$ and $0 < \frac{1}{n} < \varepsilon$. □

$8^0$. *If the number* $x \in \mathbb{R}$ *is such that* $0 \leq x$ *and* $x < \frac{1}{n}$ *for all* $n \in \mathbb{N}$, *then* $x = 0$.

*Proof.* The relation $0 < x$ is impossible by virtue of $7^0$. □

$9^0$. *For any numbers* $a, b \in \mathbb{R}$ *such that* $a < b$ *there is a rational number* $r \in \mathbb{Q}$ *such that* $a < r < b$.

*Proof.* Taking account of $7^0$, we choose $n \in \mathbb{N}$ such that $0 < \frac{1}{n} < b - a$. By the principle of Archimedes we can find a number $m \in \mathbb{Z}$ such that $\frac{m-1}{n} \leq a < \frac{m}{n}$. Then $\frac{m}{n} < b$, since otherwise we would have $\frac{m-1}{n} \leq a < b \leq \frac{m}{n}$, from which it would follow that $\frac{1}{n} > b - a$. Thus $r = \frac{m}{n} \in \mathbb{Q}$ and $a < \frac{m}{n} < b$. □

$10^0$. *For any number* $x \in \mathbb{R}$ *there exists a unique integer* $k \in \mathbb{Z}$ *such that* $k \leq x < k + 1$.

*Proof.* This follows immediately from the principle of Archimedes. □

The number $k$ just mentioned is denoted $[x]$ and is called the *integer part* of $x$. The quantity $\{x\} := x - [x]$ is called the *fractional part* of $x$. Thus $x = [x] + \{x\}$, and $\{x\} \geq 0$.

### 2.2.4 The Geometric Interpretation of the Set of Real Numbers and Computational Aspects of Operations with Real Numbers

**a. The Real Line** In relation to real numbers we often use a descriptive geometric language connected with a fact that you know in general terms from school. By the axioms of geometry there is a one-to-one correspondence $f : \mathbb{L} \to \mathbb{R}$ between the points of a line $\mathbb{L}$ and the set $\mathbb{R}$ of real numbers. Moreover this correspondence is connected with the rigid motions of the line. To be specific, if $T$ is a parallel translation of the line $\mathbb{L}$ along itself, there exists a number $t \in \mathbb{R}$ (depending only on $T$) such that $f\big(T(x)\big) = f(x) + t$ for each point $x \in \mathbb{L}$.

The number $f(x)$ corresponding to a point $x \in \mathbb{L}$ is called the *coordinate* of $x$. In view of the one-to-one nature of the mapping $f : \mathbb{L} \to \mathbb{R}$, the coordinate of a point is often called simply a point. For example, instead of the phrase "let us take the point whose coordinate is 1" we say "let us take the point 1". Given the correspondence $f : \mathbb{L} \to \mathbb{R}$, we call the line $\mathbb{L}$ the *coordinate axis* or the *number axis* or the *real line*. Because $f$ is bijective, the set $\mathbb{R}$ itself is also often called the real line and its points are called points of the real line.

As noted above, the bijective mapping $f : \mathbb{L} \to \mathbb{R}$ that defines coordinates on $\mathbb{L}$ has the property that under a parallel translation $T$ the coordinates of the images of points of the line $\mathbb{L}$ differ from the coordinates of the points themselves by a number $t \in \mathbb{R}$, the same for every point. For this reason $f$

is determined completely by specifying the point that is to have coordinate 0 and the point that is to have coordinate 1, or more briefly, by the point 0, called the *origin*, and the point 1. The closed interval determined by these points is called the *unit interval*. The direction determined by the ray with origin at 0 containing 1 is called the positive direction and a motion in that direction (from 0 to 1) is called a motion from left to right. In accordance with this convention, 1 lies to the right of 0 and 0 to the left of 1.

Under a parallel translation $T$ that moves the origin $x_0$ to the point $x_1 = T(x_0)$ with coordinate 1, the coordinates of the images of all points are one unit larger than those of their pre-images, and therefore we locate the point $x_2 = T(x_1)$ with coordinate 2, the point $x_3 = T(x_2)$ with coordinate 3,..., and the point $x_{n+1} = T(x_n)$ with coordinate $n+1$, as well as the point $x_{-1} = T^{-1}(x_0)$ with coordinate $-1$,..., the point $x_{-n-1} = T^{-1}(x_{-n})$ with coordinate $-n-1$. In this way we obtain all points with integer coordinates $m \in \mathbb{Z}$.

Knowing how to double, triple,... the unit interval, we can use Thales' theorem to partition this interval into $n$ congruent subintervals. By taking the subinterval having an endpoint at the origin, we find that the coordinate of its other end, which we denote by $x$, satisfies the equation $n \cdot x = 1$, that is, $x = \frac{1}{n}$. From this we find all points with rational coordinates $\frac{m}{n} \in \mathbb{Q}$.

But there still remain points of $\mathbb{L}$, since we know there are intervals incommensurable with the unit interval. Each such point, like every other point of the line, divides the line into two rays, on each of which there are points with integer or rational coordinates. (This is a consequence of the original geometric principle of Archimedes.) Thus a point produces a partition, or, as it is called, a *cut* of $\mathbb{Q}$ into two nonempty sets $X$ and $Y$ corresponding to the rational points (points with rational coordinates) on the left-hand and right-hand rays. By the axiom of completeness, there is a number $c$ that separates $X$ and $Y$, that is, $x \leq c \leq y$ for all $x \in X$ and all $y \in Y$. Since $X \cup Y = \mathbb{Q}$, it follows that $\sup X = s = i = \inf Y$. For otherwise, $s < i$ and there would be a rational number between $s$ and $i$ lying neither in $X$ nor in $Y$. Thus $s = i = c$. This uniquely determined number $c$ is assigned to the corresponding point of the line.

The assignment of coordinates to points of the line just described provides a visualizable model for both the order relation in $\mathbb{R}$ (hence the term "linear ordering") and for the axiom of completeness or continuity in $\mathbb{R}$, which in geometric language means that there are no "holes" in the line $\mathbb{L}$, which would separate it into two pieces having no points in common. (Such a separation could only come about by use of some point of the line $\mathbb{L}$.)

We shall not go into further detail about the construction of the mapping $f : \mathbb{L} \to \mathbb{R}$, since we shall invoke the geometric interpretation of the set of real numbers only for the sake of visualizability and perhaps to bring into play the reader's very useful geometric intuition. As for the formal proofs,

just as before, they will rely either on the collection of facts we have obtained from the axioms for the real numbers or directly on the axioms themselves.

Geometric language, however, will be used constantly.

We now introduce the following notation and terminology for the number sets listed below:

$]a, b[ := \{x \in \mathbb{R} | \, a < x < b\}$ is the open interval $ab$;
$[a, b] := \{x \in \mathbb{R} | \, a \leq x \leq b\}$ is the closed interval $ab$;
$]a, b] := \{x \in \mathbb{R} | \, a < x \leq b\}$ is the half-open interval $ab$ containing $b$;
$[a, b[ := \{x \in \mathbb{R} | \, a \leq x < b\}$ is the half-open interval $ab$ containing $a$.

**Definition 6.** Open, closed, and half-open intervals are called *numerical intervals* or simply *intervals*. The numbers determining an interval are called its *endpoints*.

The quantity $b - a$ is called the *length* of the interval $ab$. If $I$ is an interval, we shall denote its length by $|I|$. (The origin of this notation will soon become clear.)

The sets

$$]a, +\infty[ := \{x \in \mathbb{R} | \, a < x\}, \qquad ]-\infty, b[ := \{x \in \mathbb{R} | \, x < b\}$$
$$[a, +\infty[ := \{x \in \mathbb{R} | \, a \leq x\}, \qquad ]-\infty, b] := \{x \in \mathbb{R} | \, x \leq b\}$$

and $]-\infty, +\infty[ := \mathbb{R}$ are conventionally called *unbounded intervals* or *infinite intervals*.

In accordance with this use of the symbols $+\infty$ (read "plus infinity") and $-\infty$ (read "minus infinity") it is customary to denote the fact that the numerical set $X$ is not bounded above (resp. below), by writing $\sup X = +\infty$ ($\inf X = -\infty$).

**Definition 7.** An open interval containing the point $x \in \mathbb{R}$ will be called a *neighborhood* of this point.

In particular, when $\delta > 0$, the open interval $]x - \delta, x + \delta[$ is called the *$\delta$-neighborhood* of $x$. Its length is $2\delta$.

The distance between points $x, y \in \mathbb{R}$ is measured by the length of the interval having them as endpoints.

So as not to have to investigate which of the points is "left" and which is "right", that is, whether $x < y$ or $y < x$ and whether the length is $y - x$ or $x - y$, we can use the useful function

$$|x| = \begin{cases} x & \text{when } x > 0\,, \\ 0 & \text{when } x = 0\,, \\ -x & \text{when } x < 0\,, \end{cases}$$

which is called the *modulus* or *absolute value* of the number.

**Definition 8.** The *distance* between $x, y \in \mathbb{R}$ is the quantity $|x - y|$.

The distance is nonnegative and equals zero only when the points $x$ and $y$ are the same. The distance from $x$ to $y$ is the same as the distance from $y$ to $x$, since $|x-y| = |y-x|$. Finally, if $z \in \mathbb{R}$, then $|x-y| \leq |x-z| + |z-y|$. That is, the so-called *triangle inequality* holds.

The triangle inequality follows from a property of the absolute value that is also called the triangle inequality (since it can be obtained from the preceding triangle inequality by setting $z = 0$ and replacing $y$ by $-y$). To be specific, *the inequality*

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

*holds for any numbers $x$ and $y$, and equality holds only when the numbers $x$ and $y$ are both negative or both positive.*

*Proof.* If $0 \leq x$ and $0 \leq y$, then $0 \leq x+y$, $|x+y| = x+y$, $|x| = x$, and $|y| = y$, so that equality holds in this case.

If $x \leq 0$ and $y \leq 0$, then $x+y \leq 0$, $|x+y| = -(x+y) = -x-y$, $|x| = -x$, $|y| = -y$, and again we have equality.

Now suppose one of the numbers is negative and the other positive, for example, $x < 0 < y$. Then either $x < x+y \leq 0$ or $0 \leq x+y < y$. In the first case $|x+y| < |x|$, and in the second case $|x+y| < |y|$, so that in both cases $|x+y| < |x| + |y|$. □

Using the principle of induction, one can verify that

$$|x_1 + \cdots + x_n| \leq |x_1| + \cdots + |x_n| ,$$

and equality holds if and only if the numbers $x_1, \ldots, x_n$ are all nonnegative or all nonpositive.

The number $\frac{a+b}{2}$ is often called the *midpoint* or *center* of the interval with endpoints $a$ and $b$, since it is equidistant from the endpoints of the interval.

In particular, a point $x \in \mathbb{R}$ is the center of its $\delta$-neighborhood $]x-\delta, x+\delta[$ and all points of the $\delta$-neighborhood lie at a distance from $x$ less than $\delta$.

**b. Defining a Number by Successive Approximations** In measuring a real physical quantity, we obtain a number that, as a rule, changes when the measurement is repeated, especially if one changes either the method of making the measurement or the instrument used. Thus the result of measurement is usually an approximate value of the quantity being sought. The quality or precision of a measurement is characterized, for example, by the magnitude of the possible discrepancy between the true value of the quantity and the value obtained for it by measurement. When this is done, it may happen that we can never exhibit the exact value of the quantity (if it exists theoretically). Taking a more constructive position, however, we may (or should) consider that we know the desired quantity completely if we can measure it with any preassigned precision. Taking this position is tantamount to identi-

fying the number with a sequence[6] of more and more precise approximations by numbers obtained from measurement. But every measurement is a finite set of comparisons with some standard or with a part of the standard commensurable with it, so that the result of the measurement will necessarily be expressed in terms of natural numbers, integers, or, more generally, rational numbers. Hence theoretically the whole set of real numbers can be described in terms of sequences of rational numbers by constructing, after due analysis, a mathematical copy or, better expressed, a model of what people do with numbers who have no notion of their axiomatic description. The latter add and multiply the approximate values rather than the values being measured, which are unknown to them. (To be sure, they do not always know how to say what relation the result of these operations has to the result that would be obtained if the computations were carried out with the exact values. We shall discuss this question below.)

Having identified a number with a sequence of approximations to it, we should then, for example, add the sequences of approximate values when we wish to add two numbers. The new sequence thus obtained must be regarded as a new number, called the sum of the first two. But is it a number? The subtlety of the question resides in the fact that not every randomly constructed sequence is the sequence of arbitrarily precise approximations to some quantity. That is, one still has to learn how to determine from the sequence itself whether it represents some number or not. Another question that arises in the attempt to make a mathematical copy of operations with approximate numbers is that different sequences may be approximating sequences for the same quantity. The relation between sequences of approximations defining a number and the numbers themselves is approximately the same as that between a point on a map and an arrow on the map indicating the point. The arrow determines the point, but the point determines only the tip of the arrow, and does not exclude the use of a different arrow that may happen to be more convenient.

A precise description of these problems was given by Cauchy,[7] who carried out the entire program of constructing a model of the real numbers, which we have only sketched. One may hope that after you study the theory of limits you will be able to repeat these constructions independently of Cauchy.

What has been said up to now, of course, makes no claim to mathematical rigor. The purpose of this informal digression has been to direct the reader's attention to the theoretical possibility that more than one natural model of the real numbers may exist. I have also tried to give a picture of the relation

---

[6] If $n$ is the number of the measurement and $x_n$ the result of that measurement, the correspondence $n \mapsto x_n$ is simply a function $f : \mathbb{N} \to \mathbb{R}$ of a natural-number argument, that is, by definition a *sequence* (in this case a sequence of numbers). Section 3.1 is devoted to a detailed study of numerical sequences.

[7] A. Cauchy (1789–1857) – French mathematician, one of the most active creators of the language of mathematics and the machinery of classical analysis.

of numbers to the world around us and to clarify the fundamental role of natural and rational numbers. Finally, I wished to show that approximate computations are both natural and necessary.

The next part of the present section is devoted to simple but important estimates of the errors that arise in arithmetic operations on approximate quantities. These estimates will be used below and are of independent interest.

We now give precise statements.

**Definition 9.** If $x$ is the exact value of a quantity and $\tilde{x}$ a known approximation to the quantity, the numbers

$$\Delta(\tilde{x}) := |x - \tilde{x}|$$

and

$$\delta(\tilde{x}) := \frac{\Delta(\tilde{x})}{|\tilde{x}|}$$

are called respectively the *absolute* and *relative* error of approximation by $\tilde{x}$. The relative error is not defined when $\tilde{x} = 0$.

Since the value $x$ is unknown, the values of $\Delta(\tilde{x})$ and $\delta(\tilde{x})$ are also unknown. However, one usually knows some upper bounds $\Delta(\tilde{x}) < \Delta$ and $\delta(\tilde{x}) < \delta$ for these quantities. In this case we say that the absolute or relative error does not exceed $\Delta$ or $\delta$ respectively. In practice we need to deal only with estimates for the errors, so that the quantities $\Delta$ and $\delta$ themselves are often called the *absolute* and *relative* errors. But we shall not do this.

The notation $x = \tilde{x} \pm \Delta$ means that $\tilde{x} - \Delta \leq x \leq \tilde{x} + \Delta$.

For example,

| | | | |
|---|---|---|---|
| gravitational constant | $G$ | $=$ | $(6.672598 \pm 0.00085) \cdot 10^{-11}\,\mathrm{N} \cdot \mathrm{m}^2/\mathrm{kg}^2$, |
| speed of light *in vacuo* | $c$ | $=$ | $299792458\,\mathrm{m/s}$ (exactly), |
| Planck's constant | $h$ | $=$ | $(6.6260755 \pm 0.0000040) \cdot 10^{-34}\,\mathrm{J} \cdot \mathrm{s}$, |
| charge of an electron | $e$ | $=$ | $(1.60217733 \pm 0.00000049) \cdot 10^{-19}\,\mathrm{Coul}$, |
| rest mass of an electron | $m_e$ | $=$ | $(9.1093897 \pm 0.0000054) \cdot 10^{-31}\,\mathrm{kg}$. |

The main indicator of the precision of a measurement is the relative error in approximation, usually expressed as a percent.

Thus in the examples just given the relative errors are at most (in order):

$$13 \cdot 10^{-5}\,; \quad 0\,; \quad 6 \cdot 10^{-7}\,; \quad 31 \cdot 10^{-8}\,; \quad 6 \cdot 10^{-7}$$

or, as percents of the measured values,

$$13 \cdot 10^{-3}\%\,; \quad 0\%\,; \quad 6 \cdot 10^{-5}\%\,; \quad 31 \cdot 10^{-6}\%\,; \quad 6 \cdot 10^{-5}\%\,.$$

We now estimate the errors that arise in arithmetic operations with approximate quantities.

**Proposition.** *If*

$$|x - \tilde{x}| = \Delta(\tilde{x}) , \quad |y - \tilde{y}| = \Delta(\tilde{y}) ,$$

*then*

$$\Delta(\tilde{x} + \tilde{y}) := |(x + y) - (\tilde{x} + \tilde{y})| \leq \Delta(\tilde{x}) + \Delta(\tilde{y}) , \tag{2.1}$$

$$\Delta(\tilde{x} \cdot \tilde{y}) := |x \cdot y - \tilde{x} \cdot \tilde{y}| \leq |\tilde{x}|\Delta(\tilde{y}) + |\tilde{y}|\Delta(\tilde{x}) + \Delta(\tilde{x}) \cdot \Delta(\tilde{y}) ; \tag{2.2}$$

*if, in addition,*

$$y \neq 0, \quad \tilde{y} \neq 0 \text{ and } \delta(\tilde{y}) = \frac{\Delta(\tilde{y})}{|\tilde{y}|} < 1 ,$$

*then*

$$\Delta\left(\frac{\tilde{x}}{\tilde{y}}\right) := \left|\frac{x}{y} - \frac{\tilde{x}}{\tilde{y}}\right| \leq \frac{|\tilde{x}|\Delta(\tilde{y}) + |\tilde{y}|\Delta(\tilde{x})}{\tilde{y}^2} \cdot \frac{1}{1 - \delta(\tilde{y})} . \tag{2.3}$$

*Proof.* Let $x = \tilde{x} + \alpha$ and $y = \tilde{y} + \beta$. Then

$$\begin{aligned}
\Delta(\tilde{x} + \tilde{y}) &= |(x + y) - (\tilde{x} + \tilde{y})| = |\alpha + \beta| \leq |\alpha| + |\beta| = \Delta(\tilde{x}) + \Delta(\tilde{y}) , \\
\Delta(\tilde{x} \cdot \tilde{y}) &= |xy - \tilde{x} \cdot \tilde{y}| = |(\tilde{x} + \alpha)(\tilde{y} + \beta) - \tilde{x} \cdot \tilde{y}| = \\
&= |\tilde{x}\beta + \tilde{y}\alpha + \alpha\beta| \leq |\tilde{x}|\,|\beta| + |\tilde{y}|\,|\alpha| + |\alpha\beta| = \\
&= |\tilde{x}|\Delta(\tilde{y}) + |\tilde{y}|\Delta(\tilde{x}) + \Delta(\tilde{x}) \cdot \Delta(\tilde{y}) \\
\Delta\left(\frac{\tilde{x}}{\tilde{y}}\right) &= \left|\frac{x}{y} - \frac{\tilde{x}}{\tilde{y}}\right| = \left|\frac{x\tilde{y} - y\tilde{x}}{y\tilde{y}}\right| = \\
&= \left|\frac{(\tilde{x} + \alpha)\tilde{y} - (\tilde{y} + \beta)\tilde{x}}{\tilde{y}^2}\right| \cdot \left|\frac{1}{1 + \beta/\tilde{y}}\right| \leq \frac{|\tilde{x}|\,|\beta| + |\tilde{y}|\,|\alpha|}{\tilde{y}^2} \cdot \frac{1}{1 - \delta(\tilde{y})} = \\
&= \frac{|\tilde{x}|\Delta(\tilde{y}) + |\tilde{y}|\Delta(\tilde{x})}{\tilde{y}^2} \cdot \frac{1}{1 - \delta(\tilde{y})} . \quad \square
\end{aligned}$$

These estimates for the absolute errors imply the following estimates for the relative errors:

$$\delta(\tilde{x} + \tilde{y}) \leq \frac{\Delta(\tilde{x}) + \Delta(\tilde{y})}{|\tilde{x} + \tilde{y}|} , \tag{2.1'}$$

$$\delta(\tilde{x} \cdot \tilde{y}) \leq \delta(\tilde{x}) + \delta(\tilde{y}) + \delta(\tilde{y}) \cdot \delta(\tilde{y}) , \tag{2.2'}$$

$$\delta\left(\frac{\tilde{x}}{\tilde{y}}\right) \leq \frac{\delta(\tilde{x}) + \delta(\tilde{y})}{1 - \delta(\tilde{y})} . \tag{2.3'}$$

In practice, when working with sufficiently good approximations, we have $\Delta(\tilde{x}) \cdot \Delta(\tilde{y}) \approx 0$, $\delta(\tilde{x}) \cdot \delta(\tilde{y}) \approx 0$, and $1 - \delta(\tilde{y}) \approx 1$, so that one can use the following simplified and useful, but formally incorrect, versions of formulas (2.2), (2.3), (2.2′), and (2.3′):

$$\begin{aligned}
\Delta(\tilde{x}\cdot\tilde{y}) &\le |\tilde{x}|\Delta(\tilde{y}) + |\tilde{y}|\Delta(\tilde{x})\,,\\
\Delta\left(\frac{\tilde{x}}{\tilde{y}}\right) &\le \frac{|\tilde{x}|\Delta(\tilde{y}) + \tilde{y}\Delta(\tilde{x})}{\tilde{y}^2}\,,\\
\delta(\tilde{x}\cdot\tilde{y}) &\le \delta(\tilde{x}) + \delta(\tilde{y})\,,\\
\delta\left(\frac{\tilde{x}}{\tilde{y}}\right) &\le \delta(\tilde{x}) + \delta(\tilde{y})\,.
\end{aligned}$$

Formulas (2.3) and (2.3′) show that it is necessary to avoid dividing by a number that is near zero and also to avoid using rather crude approximations in which $\tilde{y}$ or $1 - \delta(\tilde{y})$ is small in absolute value.

Formula (2.1′) warns against adding approximate quantities if they are close to each other in absolute value but opposite in sign, since then $|\tilde{x} + \tilde{y}|$ is close to zero.

In all these cases, the errors may increase sharply.

For example, suppose your height has been measured twice by some device, and the precision of the measurement is $\pm 0.5$ cm. Suppose a sheet of paper was placed under your feet before the second measurement. It may nevertheless happen that the results of the measurement are as follows: $H_1 = (200 \pm 0.5)$ cm and $H_2 = (199.8 \pm 0.5)$ cm respectively.

It does not make sense to try to find the thickness of the paper in the form of the difference $H_2 - H_1$, from which it would follow only that the thickness of the paper is not larger than 0.8 cm. That would of course be a crude reflection (if indeed one could even call it a "reflection") of the true situation.

However, it is worthwhile to consider another more hopeful computational effect through which comparatively precise measurements can be carried out with crude devices. For example, if the device just used for measuring your height was used to measure the thickness of 1000 sheets of the same paper, and the result was $(20 \pm 0.5)$ cm, then the thickness of one sheet of paper is $(0.02 \pm 0.0005)$ cm, which is $(0.2 \pm 0.005)$ mm, as follows from formula (2.1).

That is, with an absolute error not larger than 0.005 mm, the thickness of one sheet is 0.2 mm. The relative error in this measurement is at most 0.025 or 2.5%.

This idea can be developed and has been proposed, for example, as a way of detecting a weak periodic signal amid the larger random static usually called white noise.

**c. The Positional Computation System** It was stated above that every real number can be presented as a sequence of rational approximations. We now recall a method, which is important when it comes to computation, for constructing in a uniform way a sequence of such rational approximations for every real number. This method leads to the positional computation system.

**Lemma.** *If a number $q > 1$ is fixed, then for every positive number $x \in \mathbb{R}$ there exists a unique integer $k \in \mathbb{Z}$ such that*

$$q^{k-1} \leq x < q^k .$$

*Proof.* We first verify that the set of numbers of the form $q^k$, $k \in \mathbb{N}$, is not bounded above. If it were, it would have a least upper bound $s$, and by definition of the least upper bound, there would be a natural number $m \in \mathbb{N}$ such that $\frac{s}{q} < q^m \leq s$. But then $s < q^{m+1}$, so that $s$ could not be an upper bound of the set.

Since $1 < q$, it follows that $q^m < q^n$ when $m < n$ for all $m, n \in \mathbb{Z}$. Hence we have also shown that for every real number $c \in \mathbb{R}$ there exists a natural number $N \in \mathbb{N}$ such that $c < q^n$ for all $n > N$.

It follows that for any $\varepsilon > 0$ there exists $M \in \mathbb{N}$ such that $\frac{1}{q^m} < \varepsilon$ for all natural numbers $m > M$.

Indeed, it suffices to set $c = \frac{1}{\varepsilon}$ and $N = M$; then $\frac{1}{\varepsilon} < q^m$ when $m > M$.

Thus the set of integers $m \in \mathbb{Z}$ satisfying the inequality $x < q^m$ for $x > 0$ is bounded below. It therefore has a minimal element $k$, which obviously will be the one we are seeking, since, for this integer, $q^{k-1} \leq x < q^k$.

The uniqueness of such an integer $k$ follows from the fact that if $m, n \in \mathbb{Z}$ and, for example, $m < n$, then $m \leq n - 1$. Hence if $q > 1$, then $q^m \leq q^{n-1}$.

Indeed, it can be seen from this remark that the inequalities $q^{m-1} \leq x < q^m$ and $q^{n-1} \leq x < q^n$, which imply $q^{n-1} \leq x < q^m$, are incompatible if $m \neq n$. □

We shall use this lemma in the following construction. Fix $q > 1$ and take an arbitrary positive number $x \in \mathbb{R}$. By the lemma we find a unique number $p \in \mathbb{Z}$ such that

$$q^p \leq x < q^{p+1} . \tag{2.4}$$

**Definition 10.** The number $p$ satisfying (2.4) is called the *order of $x$ in the base $q$* or (when $q$ is fixed) simply the *order of $x$*.

By the principle of Archimedes, we find a unique natural number $\alpha_p \in \mathbb{N}$ such that

$$\alpha_p q^p \leq x < \alpha_p q^p + q^p . \tag{2.5}$$

Taking (2.4) into account, one can assert that $\alpha_p \in \{1, \ldots, q-1\}$.

All of the subsequent steps in our construction will repeat the step we are about to take, starting from relation (2.5).

It follows from relation (2.5) and the principle of Archimedes that there exists a unique number $\alpha_{p-1} \in \{0, 1, \ldots, q-1\}$ such that

$$\alpha_p q^p + \alpha_{p-1} q^{p-1} \leq x < \alpha_p q^p + \alpha_{p-1} q^{p-1} + q^{p-1} . \tag{2.6}$$

If we have made $n$ such steps, obtaining the relation

$$\alpha_p q^p + \alpha_{p-1} q^{p-1} + \cdots + \alpha_{p-n} q^{p-n} \leq$$
$$\leq x < \alpha_p q^p + \alpha_{p-1} q^{p-1} + \cdots + \alpha_{p-n} q^{p-n} + q^{p-n},$$

then by the principle of Archimedes there exists a unique number $\alpha_{p-n-1} \in \{0, 1, \ldots, q-1\}$ such that

$$\alpha_p q^p + \cdots + \alpha_{p-n} q^{p-n} + \alpha_{p-n-1} q^{p-n-1} \leq$$
$$x < \alpha_p q^p + \cdots + \alpha_{p-n} q^{p-n} + \alpha_{p-n-1} q^{p-n-1} + q^{p-n-1}.$$

Thus we have exhibited an algorithm by means of which a sequence of numbers $\alpha_p, \alpha_{p-1}, \ldots, \alpha_{p-n}, \ldots$ from the set $\{0, 1, \ldots, q-1\}$ is placed in correspondence with the positive number $x$. Less formally, we have constructed a sequence of rational numbers of the special form

$$r_n = \alpha_p q^p + \cdots + \alpha_{p-n} q^{p-n}, \tag{2.7}$$

and such that

$$r_n \leq x < r_n + \frac{1}{q^{n-p}}. \tag{2.8}$$

In other words, we construct better and better appproximations from below and from above to the number $x$ using the special sequence (2.7). The symbol $\alpha_p \ldots \alpha_{p-n} \ldots$ is a code for the entire sequence $\{r_n\}$. To recover the sequence $\{r_n\}$ from this symbol it is necessary to indicate the value of $p$, the order of $x$.

For $p \geq 0$ it is customary to place a period or comma after $\alpha_0$; for $p < 0$, the convention is to place $|p|$ zeros left of $\alpha_p$ and a period or comma right of the leftmost zero (we recall that $\alpha_p \neq 0$).

For example, when $q = 10$,

$$123.45 := 1 \cdot 10^2 + 2 \cdot 10^1 + 3 \cdot 10^0 + 4 \cdot 10^{-1} + 5 \cdot 10^{-2},$$
$$0.00123 := 1 \cdot 10^{-3} + 2 \cdot 10^{-4} + 3 \cdot 10^{-5};$$

and when $q = 2$,

$$1000.001 := 1 \cdot 2^3 + 1 \cdot 2^{-3}.$$

Thus the value of a digit in the symbol $\alpha_p \ldots \alpha_{p-n} \ldots$ depends on the position it occupies relative to the period or comma.

With this convention, the symbol $\alpha_p \ldots \alpha_0 \ldots$ makes it possible to recover the whole sequence of approximations.

It can be seen by inequalities (2.8) (verify this!) that different sequences $\{r_n\}$ and $\{r'_n\}$, and therefore different symbols $\alpha_p \ldots \alpha_0 \ldots$ and $\alpha'_p \ldots \alpha'_0 \ldots$, correspond to different numbers $x$ and $x'$.

We now answer the question whether some real number $x \in \mathbb{R}$ corresponds to every symbol $\alpha_p \ldots \alpha_0 \ldots$. The answer turns out to be negative.

We remark that by virtue of the algorithm just described for obtaining the numbers $\alpha_{p-n} \in \{0, 1, \ldots, q-1\}$ successively, it cannot happen that all these numbers from some point on are equal to $q-1$.

Indeed, if

$$r_n = \alpha_p q^p + \cdots \alpha_{p-k} q^{p-k} + (q-1) q^{p-k-1} + \cdots + (q-1) q^{p-n}$$

for all $n > k$, that is,

$$r_n = r_k + \frac{1}{q^{k-p}} - \frac{1}{q^{n-p}}\,, \tag{2.9}$$

then by (2.8) we have

$$r_k + \frac{1}{q^{k-p}} - \frac{1}{q^{n-p}} \le x < r_k + \frac{1}{q^{k-p}}\,.$$

Then for any $n > k$

$$0 < r_k + \frac{1}{q^{k-p}} - x < \frac{1}{q^{n-p}}\,,$$

which, as we know from $8^0$ above, is impossible.

It is also useful to note that if at least one of the numbers $\alpha_{p-k-1}, \ldots, \alpha_{p-n}$ is less than $q-1$, then instead of (2.9) we can write

$$r_n < r_k + \frac{1}{q^{k-p}} - \frac{1}{q^{n-p}}$$

or, what is the same

$$r_n + \frac{1}{q^{n-p}} < r_k + \frac{1}{q^{k-p}}\,. \tag{2.10}$$

We can now prove that any symbol $\alpha_n \ldots \alpha_0 \ldots$ composed of the numbers $\alpha_k \in \{0, 1, \ldots, q-1\}$, and in which there are numbers different from $q-1$ with arbitrarily large indices, corresponds to some number $x \ge 0$.

Indeed, from the symbol $\alpha_p \ldots \alpha_{p-n} \ldots$ let us construct the sequence $\{r_n\}$ of the form (2.7). By virtue of the relations $r_0 \le r_1 \le r_n \le \cdots$, taking account of (2.9) and (2.10), we have

$$r_0 \le r_1 \le \cdots \le \cdots < \cdots \le r_n + \frac{1}{q^{n-p}} \le \cdots \le r_1 + \frac{1}{q^{1-p}} \le r_0 + \frac{1}{q^{-p}}\,. \tag{2.11}$$

The strict inequalities in this last relation should be understood as follows: every element of the left-hand sequence is less than every element of the right-hand sequence. This follows from (2.10).

If we now take $x = \sup\limits_{n\in\mathbb{N}} r_n \big( = \inf\limits_{n\in\mathbb{N}} (r_n + q^{-(n-p)}) \big)$, then the sequence $\{r_n\}$ will satisfy conditions (2.7) and (2.8), that is, the symbol $\alpha_p \ldots \alpha_{p-n} \ldots$ corresponds to the number $x \in \mathbb{R}$.

Thus, we have established a one-to-one correspondence between the positive numbers $x \in \mathbb{R}$ and symbols of the form $\alpha_p \ldots \alpha_0, \ldots$ if $p \ge 0$ or $\underbrace{0,0\ldots0}_{|p| \text{ zeros}}\alpha_p \ldots$ if $p < 0$. The symbol assigned to $x$ is called the *q-ary representation of* $x$; the numbers that occur in the symbol are called its *digits*, and the position of a digit relative to the period is called its *rank*.

We agree to assign to a number $x < 0$ the symbol for the positive number $-x$, prefixed by a negative sign. Finally, we assign the symbol $0.0\ldots0\ldots$ to the number 0.

In this way we have constructed the *positional q-ary system of writing real numbers.*

The most useful systems are the decimal system (in common use) and for technical reasons the binary system (in electronic computers). Less common, but also used in some parts of computer engineering are the ternary and octal systems.

Formulas (2.7) and (2.8) show that if only a finite number of digits are retained in the $q$-ary expression of $x$ (or, if we wish, we may say that the others are replaced with zeros), then the absolute error of the resulting approximation (2.7) for $x$ does not exceed one unit in the last rank retained.

This observation makes it possible to use the formulas obtained in Paragraph b to estimate the errors that arise when doing arithmetic operations on numbers as a result of replacing the exact numbers by the corresponding approximate values of the form (2.7).

This last remark also has a certain theoretical value. To be specific, if we identify a real number $x$ with its $q$-ary expression, as was suggested in Paragraph b, once we have learned to perform arithmetic operations directly on the $q$-ary symbols, we will have constructed a new model of the real numbers, seemingly of greater value from the computational point of view.

The main problems that need to be solved in this direction are the following:

To two $q$-ary symbols it is necessary to assign a new symbol representing their sum. It will of course be constructed one step at a time. To be specific, by adding more and more precise rational approximations of the original numbers, we shall obtain rational approximations corresponding to their sum. Using the remark made above, one can show that as the precision of the approximations of the terms increases, we shall obtain more and more $q$-ary digits of the sum, which will then not vary under subsequent improvements in the approximation.

This same problem needs to be solved with respect to multiplication.

Another, less constructive, route for passing from rational numbers to all real numbers is due to Dedekind.

Dedekind identifies a real number with a cut in the set $\mathbb{Q}$ of rational numbers, that is, a partition of $\mathbb{Q}$ into two disjoint sets $A$ and $B$ such that $a < b$ for all $a \in A$ and all $b \in B$. Under this approach to real numbers our axiom of completeness (continuity) becomes a well-known theorem of Dedekind. For that reason the axiom of completeness in the form we have given it is sometimes called Dedekind's axiom.

To summarize, in the present section we have exhibited the most important classes of numbers. We have shown the fundamental role played by the natural and rational numbers. It has been shown how the basic properties of

these numbers follow from the axiom system[8] we have adopted. We have given a picture of various models of the set of real numbers. We have discussed the computational aspects of the theory of real numbers: estimates of the errors arising during arithmetical operations with approximate magnitudes, and the $q$-ary positional computation system.

### 2.2.5 Problems and Exercises

**1.** Using the principle of induction, show that

a) the sum $x_1 + \cdots + x_n$ of real numbers is defined independently of the insertion of parentheses to specify the order of addition;

b) the same is true of the product $x_1 \cdots x_n$;

c) $|x_1 + \cdots + x_n| \leq |x_1| + \cdots + |x_n|$;

d) $|x_1 \cdots x_n| = |x_1| \cdots |x_n|$;

e) $\Big((m, n \in \mathbb{N}) \wedge (m < n)\Big) \Rightarrow \Big((n - m) \in \mathbb{N}\Big)$;

f) $(1 + x)^n \geq 1 + nx$ for $x > -1$ and $n \in \mathbb{N}$, equality holding only when $n = 1$ or $x = 0$ (*Bernoulli's inequality*);

g) $(a+b)^n = a^n + \frac{n}{1!}a^{n-1}b + \frac{n(n-1)}{2!}a^{n-2}b^2 + \cdots + \frac{n(n-1)\cdots 2}{(n-1)!}ab^{n-1} + b^n$ (*Newton's binomial formula*);

**2.** a) Verify that $\mathbb{Z}$ and $\mathbb{Q}$ are inductive sets.

b) Give examples of inductive sets different from $\mathbb{N}$, $\mathbb{Z}$, $\mathbb{Q}$, and $\mathbb{R}$.

**3.** Show that an inductive set is not bounded above.

**4.** a) An inductive set is infinite (that is, equipollent with one of its subsets different from itself).

b) The set $E_n = \{x \in \mathbb{N} | \, x \leq n\}$ is finite. (We denote card $E_n$ by $n$.)

**5.** (The *Euclidean algorithm*) Let $m, n \in \mathbb{N}$ and $m > n$. Their greatest common divisor ($\gcd(m, n) = d \in \mathbb{N}$) can be found in a finite number of steps using the following algorithm of Euclid involving successive divisions with remainder.

$$\begin{aligned} m &= q_1 n + r_1 \qquad (r_1 < n)\,, \\ n &= q_2 r_1 + r_2 \qquad (r_2 < r_1)\,, \\ r_1 &= q_3 r_2 + r_3 \qquad (r_3 < r_2)\,, \\ &\cdots\cdots\cdots\cdots\cdots \\ r_{k-1} &= q_{k+1} r_k + 0\,. \end{aligned}$$

Then $d = r_k$.

b) If $d = \gcd(m, n)$, one can choose numbers $p, q \in \mathbb{Z}$ such that $pm + qn = d$; in particular, if $m$ and $n$ are relatively prime, then $pm + qn = 1$.

---

[8] It was stated by Hilbert in almost the form given above at the turn of the twentieth century. See for example Hilbert, D. *Foundations of Geometry*, Chap. III, § 13. (Translated from the second edition of *Grundlagen der Geometrie*, La Salle, Illinois: Open Court Press, 1971. This section was based on Hilbert's article "Über den Zahlbegriff" in *Jahresbericht der deutschen Mathematikervereinigung* **8** (1900).).

**6.** Try to give your own proof of the fundamental theorem of arithmetic (Paragraph a in Subsect. 2.2.2).

**7.** If the product $m \cdot n$ of natural numbers is divisible by a prime $p$, that is, $m \cdot n = p \cdot k$, where $k \in \mathbb{N}$, then either $m$ or $n$ is divisible by $p$.

**8.** It follows from the fundamental theorem of arithmetic that the set of prime numbers is infinite.

**9.** Show that if the natural number $n$ is not of the form $k^m$, where $k, m \in \mathbb{N}$, then the equation $x^m = n$ has no rational roots.

**10.** Show that the expression of a rational number in any $q$-ary computation system is periodic, that is, starting from some rank it consists of periodically repeating groups of digits.

**11.** Let us call an irrational number $\alpha \in \mathbb{R}$ *well approximated* by rational numbers if for any natural numbers $n, N \in \mathbb{N}$ there exists a rational number $\frac{p}{q}$ such that $\left|\alpha - \frac{p}{q}\right| < \frac{1}{Nq^n}$.

a) Construct an example of a well-approximated irrational number.

b) Prove that a well-approximated irrational number cannot be algebraic, that is, it is transcendental (*Liouville's theorem*).[9]

**12.** Knowing that $\frac{m}{n} := m \cdot n^{-1}$ by definition, where $m \in \mathbb{Z}$ and $n \in \mathbb{N}$, derive the "rules" for addition, multiplication, and division of fractions, and also the condition for two fractions to be equal.

**13.** Verify that the rational numbers $\mathbb{Q}$ satisfy all the axioms for real numbers except the axiom of completeness.

**14.** Adopting the geometric model of the set of real numbers (the real line), show how to construct the numbers $a + b$, $a - b$, $ab$, and $\frac{a}{b}$ in this model.

**15.** a) Illustrate the axiom of completeness on the real line.

b) Prove that the least-upper-bound principle is equivalent to the axiom of completeness.

**16.** a) If $A \subset B \subset \mathbb{R}$, then $\sup A \leq \sup B$ and $\inf A \geq \inf B$.

b) Let $\mathbb{R} \supset X \neq \varnothing$ and $\mathbb{R} \supset Y \neq \varnothing$. If $x \leq y$ for all $x \in X$ and all $y \in Y$, then $X$ is bounded above, $Y$ is bounded below, and $\sup X \leq \inf Y$.

c) If the sets $X, Y$ in b) are such that $X \cup Y = \mathbb{R}$, then $\sup X = \inf Y$.

d) If $X$ and $Y$ are the sets defined in c), then either $X$ has a maximal element or $Y$ has a minimal element. (*Dedekind's theorem.*)

e) (Continuation.) Show that Dedekind's theorem is equivalent to the axiom of completeness.

---

[9] J. Liouville (1809–1882) – French mathematician, who wrote on complex analysis, geometry, differential equations, number theory, and mechanics.

**17.** Let $A+B$ be the set of numbers of the form $a+b$ and $A\cdot B$ the set of numbers of the form $a\cdot b$, where $a\in A\subset\mathbb{R}$ and $b\in B\subset\mathbb{R}$. Determine whether it is always true that

a) $\sup(A+B)=\sup A+\sup B$,

b) $\sup(A\cdot B)=\sup A\cdot\sup B$.

**18.** Let $-A$ be the set of numbers of the form $-a$, where $a\in A\subset\mathbb{R}$. Show that $\sup(-A)=-\inf A$.

**19.** a) Show that for $n\in\mathbb{N}$ and $a>0$ the equation $x^n=a$ has a positive root (denoted $\sqrt[n]{a}$ or $a^{1/n}$).

b) Verify that for $a>0$, $b>0$, and $n,m\in\mathbb{N}$

$$\sqrt[n]{ab}=\sqrt[n]{a}\cdot\sqrt[n]{b}\quad\text{and}\quad\sqrt[n]{\sqrt[m]{a}}=\sqrt[n\cdot m]{a}.$$

c) $(a^{\frac{1}{n}})^m=(a^m)^{\frac{1}{n}}=:a^{m/n}$ and $a^{1/n}\cdot a^{1/m}=a^{1/n+1/m}$.

d) $(a^{m/n})^{-1}=(a^{-1})^{m/n}=:a^{-m/n}$.

e) Show that for all $r_1,r_2\in\mathbb{Q}$

$$a^{r_1}\cdot a^{r_2}=a^{r_1+r_2}\quad\text{and}\quad(a^{r_1})^{r_2}=a^{r_1r_2}\,.$$

**20.** a) Show that the inclusion relation is a partial ordering relation on sets (but not a linear ordering!).

b) Let $A$, $B$, and $C$ be sets such that $A\subset C$, $B\subset C$, $A\setminus B\neq\varnothing$, and $B\setminus A\neq\varnothing$. We introduce a partial ordering into this triple of sets as in a). Exhibit the maximal and minimal elements of the set $\{A,B,C\}$. (Pay attention to the non-uniqueness!)

**21.** a) Show that, just like the set $\mathbb{Q}$ of rational numbers, the set $\mathbb{Q}(\sqrt{n})$ of numbers of the form $a+b\sqrt{n}$, where $a,b\in\mathbb{Q}$ and $n$ is a fixed natural number that is not the square of any integer, is an ordered set satisfying the principle of Archimedes but not the axiom of completeness.

b) Determine which axioms for the real numbers do not hold for $\mathbb{Q}(\sqrt{n})$ if the standard arithmetic operations are retained in $\mathbb{Q}(\sqrt{n})$ but order is defined by the rule $(a+b\sqrt{n}\le a'+b'\sqrt{n}):=\Big((b<b')\vee\Big((b=b')\wedge(a\le a')\Big)\Big)$. Will $\mathbb{Q}(\sqrt{n})$ now satisfy the principle of Archimedes?

c) Order the set $\mathbb{P}[x]$ of polynomials with rational or real coefficients by specifying that

$$P_m(x)=a_0+a_1x+\cdots+a_mx^m\succ 0,\quad\text{if}\quad a_m>0\,.$$

d) Show that the set $\mathbb{Q}(x)$ of rational fractions

$$R_{m,n}=\frac{a_0+a_1x+\cdots+a_mx^m}{b_0+b_1x+\cdots+b_nx^n}$$

with coefficients in $\mathbb{Q}$ or $\mathbb{R}$ becomes an ordered field, but not an Archimedean ordered field, when the order relation $R_{m,n}\succ 0$ is defined to mean $a_mb_n>0$ and the usual arithmetic operations are introduced. This means that the principle of Archimedes cannot be deduced from the other axioms for $\mathbb{R}$ without using the axiom of completeness.

**22.** Let $n \in \mathbb{N}$ and $n > 1$. In the set $E_n = \{0, 1, \ldots, n-1\}$ we define the sum and product of two elements as the remainders when the usual sum and product in $\mathbb{R}$ are divided by $n$. With these operations defined on it, the set $E_n$ is denoted $\mathbb{Z}_n$.

a) Show that if $n$ is not a prime number, then there are nonzero numbers $m, k$ in $\mathbb{Z}_n$ such that $m \cdot k = 0$. (Such numbers are called *zero divisors.*) This means that in $\mathbb{Z}_n$ the equation $a \cdot b = c \cdot b$ does not imply that $a = c$, even when $b \neq 0$.

b) Show that if $p$ is prime, then there are no zero divisors in $\mathbb{Z}_p$ and $\mathbb{Z}_p$ is a field.

c) Show that, no matter what the prime $p$, $\mathbb{Z}_p$ cannot be ordered in a way consistent with the arithmetic operations on it.

**23.** Show that if $\mathbb{R}$ and $\mathbb{R}'$ are two models of the set of real numbers and $f : \mathbb{R} \to \mathbb{R}'$ is a mapping such that $f(x+y) = f(x) + f(y)$ and $f(x \cdot y) = f(x) \cdot f(y)$ for any $x, y \in \mathbb{R}$, then

a) $f(0) = 0'$;

b) $f(1) = 1'$ if $f(x) \not\equiv 0'$, which we shall henceforth assume;

c) $f(m) = m'$ where $m \in \mathbb{Z}$ and $m' \in \mathbb{Z}'$, and the mapping $f : \mathbb{Z} \to \mathbb{Z}'$ is injective and preserves the order.

d) $f\left(\frac{m}{n}\right) = \frac{m'}{n'}$, where $m, n \in \mathbb{Z}$, $n \neq 0$, $m', n' \in \mathbb{Z}'$, $n' \neq 0'$, $f(m) = m'$, $f(n) = n'$. Thus $f : \mathbb{Q} \to \mathbb{Q}'$ is a bijection that preserves order.

e) $f : \mathbb{R} \to \mathbb{R}'$ is a bijective mapping that preserves order.

**24.** On the basis of the preceding exercise and the axiom of completeness, show that the axiom system for the set of real numbers determines it completely up to an isomorphism (method of realizing it), that is, if $\mathbb{R}$ and $\mathbb{R}'$ are two sets satisfying these axioms, then there exists a one-to-one correspondence $f : \mathbb{R} \to \mathbb{R}'$ that preserves the arithmetic operations and the order: $f(x+y) = f(x) + f(y)$, $f(x \cdot y) = f(x) \cdot f(y)$, and $(x \leq y) \Leftrightarrow \big(f(x) \leq f(y)\big)$.

**25.** A number $x$ is represented on a computer as

$$x = \pm q^p \sum_{n=1}^{k} \frac{\alpha_n}{q^n} ,$$

where $p$ is the order of $x$ and $M = \sum_{n=1}^{k} \frac{\alpha_n}{q^n}$ is the mantissa of the number $x$ ($\frac{1}{q} \leq M < 1$).

Now a computer works only with a certain range of numbers: for $q = 2$ usually $|p| \leq 64$, and $k = 35$. Evalute this range in the decimal system.

**26.** a) Write out the $(6 \times 6)$ multiplication table for multiplication in base 6.

b) Using the result of a), multiply "columnwise" in the base-6 system

$$\begin{array}{r} (532)_6 \\ \times\ (145)_6 \\ \hline \end{array}$$

and check your work by repeating the computation in the decimal system.

c) Perform the "long" division

$$(1301)_6 \;\big|\underline{(25)_6}$$

and check your work by repeating the computation in the decimal system.

d) Perform the "columnwise" addition

$$\begin{array}{r} (4052)_6 \\ + \\ \underline{(3125)_6} \end{array}$$

**27.** Write $(100)_{10}$ in the binary and ternary systems.

**28.** a) Show that along with the unique representation of an integer as

$$(\alpha_n \alpha_{n-1} \dots \alpha_0)_3 ,$$

where $\alpha_i \in \{0, 1, 2\}$, it can also be written as

$$(\beta_n \beta_{n-1} \dots \beta_0)_3 ,$$

where $\beta \in \{-1, 0, 1\}$.

b) What is the largest number of coins from which one can detect a counterfeit in three weighings with a pan balance, if it is known in advance only that the counterfeit coin differs in weight from the other coins?

**29.** What is the smallest number of questions to be answered "yes" or "no" that one must pose in order to be sure of determining a 7-digit telephone number?

**30.** a) How many different numbers can one define using 20 decimal digits (for example, two ranks with 10 possible digits in each)? Answer the same question for the binary system. Which system does a comparison of the results favor in terms of efficiency?

b) Evaluate the number of different numbers one can write, having at one's disposal $n$ digits of a $q$-ary system. (Answer: $q^{n/q}$.)

c) Draw the graph of the function $f(x) = x^{n/x}$ over the set of natural-number values of the argument and compare the efficiency of the different systems of computation.

## 2.3 Basic Lemmas Connected with the Completeness of the Real Numbers

In this section we shall establish some simple useful principles, each of which could have been used as the axiom of completeness in our construction of the real numbers.[10]

We have called these principles basic lemmas in view of their extensive application in the proofs of a wide variety of theorems in analysis.

[10] See Problem 4 at the end of this section.

### 2.3.1 The Nested Interval Lemma (Cauchy–Cantor Principle)

**Definition 1.** A function $f : \mathbb{N} \to X$ of a natural-number argument is called a *sequence* or, more fully, a *sequence of elements of* $X$.

The value $f(n)$ of the function $f$ corresponding to the number $n \in \mathbb{N}$ is often denoted $x_n$ and called the $n$th term of the sequence.

**Definition 2.** Let $X_1, X_2, \ldots, X_n, \ldots$ be a sequence of sets. If $X_1 \supset X_2 \supset \cdots \supset X_n \supset \cdots$, that is $X_n \supset X_{n+1}$ for all $n \in \mathbb{N}$, we say the sequence is *nested.*

**Lemma.** (Cauchy–Cantor). *For any nested sequence* $I_1 \supset I_2 \supset \cdots \supset I_n \supset \cdots$ *of closed intervals, there exists a point* $c \in \mathbb{R}$ *belonging to all of these intervals.*

*If in addition it is known that for any* $\varepsilon > 0$ *there is an interval* $I_k$ *whose length* $|I_k|$ *is less than* $\varepsilon$, *then* $c$ *is the unique point common to all the intervals.*

*Proof.* We begin by remarking that for any two closed intervals $I_m = [a_m, b_m]$ and $I_n = [a_n, b_n]$ of the sequence we have $a_m \leq b_n$. For otherwise we would have $a_n \leq b_n < a_m \leq b_m$, that is, the intervals $I_m$ and $I_n$ would be mutually disjoint, while one of them (the one with the larger index) is contained in the other.

Thus the numerical sets $A = \{a_m | \, m \in \mathbb{N}\}$ and $B = \{b_n | \, n \in \mathbb{N}\}$ satisfy the hypotheses of the axiom of completeness, by virtue of which there is a number $c \in \mathbb{R}$ such that $a_m \leq c \leq b_n$ for all $a_m \in A$ and all $b_n \in B$. In particular, $a_n \leq c \leq b_n$ for all $n \in \mathbb{N}$. But that means that the point $c$ belongs to all the intervals $I_n$.

Now let $c_1$ and $c_2$ be two points having this property. If they are different, say $c_1 < c_2$, then for any $n \in \mathbb{N}$ we have $a_n \leq c_1 < c_2 \leq b_n$, and therefore $0 < c_2 - c_1 < b_n - a_n$, so that the length of an interval in the sequence cannot be less than $c_2 - c_1$. Hence if there are intervals of arbitrarily small length in the sequence, their common point is unique. □

### 2.3.2 The Finite Covering Lemma (Borel–Lebesgue Principle, or Heine–Borel Theorem)

**Definition 3.** A system $S = \{X\}$ of sets $X$ is said to *cover* a set $Y$ if $Y \subset \bigcup_{X \in S} X$, (that is, if every element $y \in Y$ belongs to at least one of the sets $X$ in the system $S$).

A subset of a set $S = \{X\}$ that is a system of sets will be called a *subsystem* of $S$. Thus a subsystem of a system of sets is itself a system of sets of the same type.

**Lemma.** (Borel–Lebesgue).[11] *Every system of open intervals covering a closed interval contains a finite subsystem that covers the closed interval.*

*Proof.* Let $S = \{U\}$ be a system of open intervals $U$ that cover the closed interval $[a, b] = I_1$. If the interval $I_1$ could not be covered by a finite set of intervals of the system $S$, then, dividing $I_1$ into two halves, we would find that at least one of the two halves, which we denote by $I_2$, does not admit a finite covering. We now repeat this procedure with the interval $I_2$, and so on.

In this way a nested sequence $I_1 \supset I_2 \supset \cdots \supset I_n \supset \cdots$ of closed intervals arises, none of which admit a covering by a finite subsystem of $S$. Since the length of the interval $I_n$ is $|I_n| = |I_1| \cdot 2^{-n}$, the sequence $\{I_n\}$ contains intervals of arbitrarily small length (see the lemma in Paragraph c of Subsect. 2.2.4). But the nested interval theorem implies that there exists a point $c$ belonging to all of the intervals $I_n$, $n \in \mathbb{N}$. Since $c \in I_1 = [a, b]$ there exists an open interval $]\alpha, \beta[ = U \in S$ containing $c$, that is, $\alpha < c < \beta$. Let $\varepsilon = \min\{c-\alpha, \beta-c\}$. In the sequence just constructed, we find an interval $I_n$ such that $|I_n| < \varepsilon$. Since $c \in I_n$ and $|I_n| < \varepsilon$, we conclude that $I_n \subset U = ]\alpha, \beta[$. But this contradicts the fact that the interval $I_n$ cannot be covered by a finite set of intervals from the system. □

### 2.3.3 The Limit Point Lemma (Bolzano–Weierstrass Principle)

We recall that we have defined a *neighborhood* of a point $x \in \mathbb{R}$ to be an open interval containing the point and the *$\delta$-neighborhood* about $x$ to be the open interval $]x - \delta, x + \delta[$.

**Definition 4.** A point $p \in \mathbb{R}$ is a *limit point* of the set $X \subset \mathbb{R}$ if every neighborhood of the point contains an infinite subset of $X$.

This condition is obviously equivalent to the assertion that every neighborhood of $p$ contains at least one point of $X$ different from $p$ itself. (Verify this!)

We now give some examples.

If $X = \left\{\frac{1}{n} \in \mathbb{R} \mid n \in \mathbb{N}\right\}$, the only limit point of $X$ is the point $0 \in \mathbb{R}$.

For an open interval $]a, b[$ every point of the closed interval $[a, b]$ is a limit point, and there are no others.

For the set $\mathbb{Q}$ of rational numbers every point of $\mathbb{R}$ is a limit point; for, as we know, every open interval of the real numbers contains rational numbers.

[11] É. Borel (1871–1956) and H. Lebesgue (1875–1941) – well-known French mathematicians who worked in the theory of functions.

**Lemma.** (Bolzano–Weierstrass).[12] *Every bounded infinite set of real numbers has at least one limit point.*

*Proof.* Let $X$ be the given subset of $\mathbb{R}$. It follows from the definition of boundedness that $X$ is contained in some closed interval $I \subset \mathbb{R}$. We shall show that at least one point of $I$ is a limit point of $X$.

If such were not the case, then each point $x \in I$ would have a neighborhood $U(x)$ containing either no points of $X$ or at most a finite number. The totality of such neighborhoods $\{U(x)\}$ constructed for the points $x \in I$ forms a covering of $I$ by open intervals $U(x)$. By the finite covering lemma we can extract a system $U(x_1), \ldots, U(x_n)$ of open intervals that cover $I$. But, since $X \subset I$, this same system also covers $X$. However, there are only finitely many points of $X$ in $U(x_i)$, and hence only finitely many in their union. That is, $X$ is a finite set. This contradiction completes the proof. □

### 2.3.4 Problems and Exercises

**1.** Show that

a) if $I$ is any system of nested closed intervals, then

$$\sup\left\{a \in \mathbb{R} \mid [a,b] \in I\right\} = \alpha \leq \beta = \inf\left\{b \in \mathbb{R} \mid [a,b] \in I\right\}$$

and

$$[\alpha, \beta] = \bigcap_{[a,b] \in I} [a,b]\,;$$

b) if $I$ is a system of nested open intervals $]a,b[$ the intersection $\bigcap_{]a,b[ \in I} ]a,b[$ may happen to be empty.

Hint: $]a_n, b_n[ = \left]0, \frac{1}{n}\right[$.

**2.** Show that

a) from a system of closed intervals covering a closed interval it is not always possible to choose a finite subsystem covering the interval;

b) from a system of open intervals covering an open interval it is not always possible to choose a finite subsystem covering the interval;

c) from a system of closed intervals covering an open interval it is not always possible to choose a finite subsystem covering the interval.

**3.** Show that if we take only the set $\mathbb{Q}$ of rational numbers instead of the complete set $\mathbb{R}$ of real numbers, taking a closed interval, open interval, and neighborhood of a point $r \in \mathbb{Q}$ to mean respectively the corresponding subsets of $\mathbb{Q}$, then none of the three lemmas proved above remains true.

---

[12] B. Bolzano (1781–1848) – Czech mathematician and philosopher.
K. Weierstrass (1815–1897) – German mathematician who devoted a great deal of attention to the logical foundations of mathematical analysis.

**4.** Show that we obtain an axiom system equivalent to the one already given if we take as the axiom of completeness

a) the Bolzano–Weierstrass principle
or

b) the Borel–Lebesgue principle (Heine–Borel theorem).

Hint: The principle of Archimedes and the axiom of completeness in the earlier form both follow from a).

c) Replacing the axiom of completeness by the Cauchy–Cantor principle leads to a system of axioms that becomes equivalent to the original system if we also postulate the principle of Archimedes. (See Problem 21 in Subsect. 2.2.2.)

## 2.4 Countable and Uncountable Sets

We now make a small addition to the information about sets that was provided in Chap. 1. This addition will be useful below.

### 2.4.1 Countable Sets

**Definition 1.** A set $X$ is *countable* if it is equipollent with the set $\mathbb{N}$ of natural numbers, that is, $\operatorname{card} X = \operatorname{card} \mathbb{N}$.

**Proposition.** a) *An infinite subset of a countable set is countable.*

b) *The union of the sets of a finite or countable system of countable sets is a countable set.*

*Proof.* a) It suffices to verify that every infinite subset $E$ of $\mathbb{N}$ is equipollent with $\mathbb{N}$. We construct the needed bijective mapping $f : \mathbb{N} \to E$ as follows. There is a minimal element of $E_1 := E$, which we assign to the number $1 \in \mathbb{N}$ and denote $e_1 \in E$. The set $E$ is infinite, and therefore $E_2 := E_1 \setminus e_1$ is nonempty. We assign the minimal element of $E_2$ to the number 2 and call it $e_2 \in E_2$. We then consider $E_3 := E \setminus \{e_1, e_2\}$, and so forth. Since $E$ is an infinite set, this construction cannot terminate at any finite step with index $n \in \mathbb{N}$. As follows from the principle of induction, we assign in this way a certain number $e_n \in E$ to each $n \in \mathbb{N}$. The mapping $f : \mathbb{N} \to E$ is obviously injective.

It remains to verify that it is surjective, that is, $f(\mathbb{N}) = E$. Let $e \in E$. The set $\{n \in \mathbb{N} | \, n \leq e\}$ is finite, and hence the subset of it $\{n \in E | \, n \leq e\}$ is also finite. Let $k$ be the number of elements in the latter set. Then by construction $e = e_k$.

b) If $X_1, \ldots, X_n, \ldots$ is a countable system of sets and each set $X_m = \{x_m^1, \ldots, x_m^n, \ldots\}$ is itself countable, then since the cardinality of the set $X = \bigcup\limits_{n \in \mathbb{N}} X_n$, which consists of the elements $x_m^n$ where $m, n \in \mathbb{N}$, is not less than the cardinality of each of the sets $X_m$, it follows that $X$ is an infinite set.

The element $x_m^n \in X_m$ can be identified with the pair $(m, n)$ of natural numbers that defines it. Then the cardinality of $X$ cannot be greater than the cardinality of the set of all such odered pairs. But the mapping $f : \mathbb{N} \times \mathbb{N} \to \mathbb{N}$ given by the formula $(m, n) \mapsto \frac{(m+n)(m+n+1)}{2} + m$, as one can easily verify, is bijective. (It has a visualizable meaning: we are enumerating the points of the plane with coordinates $(m, n)$ by successively passing from points of one diagonal on which $m + n$ is constant to the points of the next such diagonal, where the sum is one larger.)

Thus the set of ordered pairs $(m, n)$ of natural numbers is countable. But then $\operatorname{card} X \leq \operatorname{card} \mathbb{N}$, and since $X$ is an infinite set we conclude on the basis of a) that $\operatorname{card} X = \operatorname{card} \mathbb{N}$. □

It follows from the proposition just proved that any subset of a countable set is either finite or countable. If it is known that a set is either finite or countable, we say it is *at most countable.* (An equivalent expression is $\operatorname{card} X \leq \operatorname{card} \mathbb{N}$.)

We can now assert, in particular, that *the union of an at most countable family of at most countable sets is at most countable.*

**Corollaries** 1) $\operatorname{card} \mathbb{Z} = \operatorname{card} \mathbb{N}$.

2) $\operatorname{card} \mathbb{N}^2 = \operatorname{card} \mathbb{N}$.

(This result means that the direct product of countable sets is countable.)

3) $\operatorname{card} \mathbb{Q} = \operatorname{card} \mathbb{N}$, that is, *the set of rational numbers is countable.*

*Proof.* A rational number $\frac{m}{n}$ is defined by an ordered pair $(m, n)$ of integers. Two pairs $(m, n)$ and $(m', n')$ define the same rational number if and only if they are proportional. Thus, choosing as the unique pair representing each rational number the pair $(m, n)$ with the smallest possible positive integer denominator $n \in \mathbb{N}$, we find that the set $\mathbb{Q}$ is equipollent to some infinite subset of the set $\mathbb{Z} \times \mathbb{Z}$. But $\operatorname{card} \mathbb{Z}^2 = \operatorname{card} \mathbb{N}$ and hence $\operatorname{card} \mathbb{Q} = \operatorname{card} \mathbb{N}$. □

4) *The set of algebraic numbers is countable.*

*Proof.* We remark first of all that the equality $\mathbb{Q} \times \mathbb{Q} = \operatorname{card} \mathbb{N}$ implies, by induction, that $\operatorname{card} \mathbb{Q}^k = \operatorname{card} \mathbb{N}$ for every $k \in \mathbb{N}$.

An element $r \in \mathbb{Q}^k$ is an ordered set $(r_1, \ldots, r_k)$ of $k$ rational numbers.

An algebraic equation of degree $k$ with rational coefficients can be written in the reduced form $x^k + r_1 x^{k-1} + \cdots + r_k = 0$, where the leading coefficient is 1. Thus there are as many different algebraic equations of degree $k$ as there are different ordered sets $(r_1, \ldots, r_k)$ of rational numbers, that is, a countable set.

The algebraic equations with rational coefficients (of arbitrary degree) also form a countable set, being a countable union (over degrees) of countable sets. Each such equation has only a finite number of roots. Hence the set of algebraic numbers is at most countable. But it is infinite, and hence countable. □

### 2.4.2 The Cardinality of the Continuum

**Definition 2.** The set $\mathbb{R}$ of real numbers is also called the *number continuum*,[13] and its cardinality the *cardinality of the continuum.*

**Theorem.**(Cantor). $\operatorname{card}\mathbb{N} < \operatorname{card}\mathbb{R}$.

This theorem asserts that the infinite set $\mathbb{R}$ has cardinality greater than that of the infinite set $\mathbb{N}$.

*Proof.* We shall show that even the closed interval $[0,1]$ is an uncountable set.

Assume that it is countable, that is, can be written as a sequence $x_1, x_2, \ldots, x_n, \ldots$. Take the point $x_1$ and on the interval $[0,1] = I_0$ fix a closed interval of positive length $I_1$ not containing the point $x_1$. In the interval $I_1$ construct an interval $I_2$ not containing $x_2$. If the interval $I_n$ has been constructed, then, since $|I_n| > 0$, we construct in it an interval $I_{n+1}$ so that $x_{n+1} \notin I_{n+1}$ and $|I_{n+1}| > 0$. By the nested set lemma, there is a point $c$ belonging to all of the intervals $I_0, I_1, \ldots, I_n, \ldots$. But this point of the closed interval $I_0 = [0,1]$ by construction cannot be any point of the sequence $x_1, x_2, \ldots, x_n, \ldots$. □

**Corollaries** 1) $\mathbb{Q} \neq \mathbb{R}$, *and so irrational numbers exist.*

2) *There exist transcendental numbers, since the set of algebraic numbers is countable.*

(After solving Exercise 3 below, the reader will no doubt wish to reinterpret this last proposition, stating it as follows: *Algebraic numbers are occasionally encountered among the real numbers.*)

At the very dawn of set theory the question arose whether there exist sets of cardinality between countable sets and sets having cardinality of the continuum, and the conjecture was made, known as the *continuum hypothesis*, that there are no intermediate cardinalities.

The question turned out to involve the deepest parts of the foundations of mathematics. It was definitively answered in 1963 by the contemporary American mathematician P. Cohen. Cohen proved that the continuum hypothesis is undecidable by showing that neither the hypothesis nor its negation contradicts the standard axiom system of set theory, so that the continuum hypothesis can be neither proved nor disproved within that axiom system. This situation is very similar to the way in which Euclid's fifth postulate on parallel lines is independent of the other axioms of geometry.

### 2.4.3 Problems and Exercises

**1.** Show that the set of real numbers has the same cardinality as the points of the interval $]-1,1[$.

[13] From the Latin *continuum*, meaning *continuous*, or *solid.*

**2.** Give an explicit one-to-one correspondence between

a) the points of two open intervals;

b) the points of two closed intervals;

c) the points of a closed interval and the points of an open interval;

d) the points of the closed interval $[0, 1]$ and the set $\mathbb{R}$.

**3.** Show that

a) every infinite set contains a countable subset;

b) the set of even integers has the same cardinality as the set of all natural numbers.

c) the union of an infinite set and an at most countable set has the same cardinality as the original infinite set;

d) the set of irrational numbers has the cardinality of the continuum;

e) the set of transcendental numbers has the cardinality of the continuum.

**4.** Show that

a) the set of increasing sequences of natural numbers $\{n_1 < n_2 < \cdots\}$ has the same cardinality as the set of fractions of the form $0.\alpha_1\alpha_2\ldots$;

b) the set of all subsets of a countable set has cardinality of the continuum.

**5.** Show that

a) the set $\mathcal{P}(X)$ of subsets of a set $X$ has the same cardinality as the set of all functions on $X$ with values $0, 1$, that is, the set of mappings $f : X \to \{0, 1\}$;

b) for a finite set $X$ of $n$ elements, $\operatorname{card}\mathcal{P}(X) = 2^n$;

c) taking account of the results of Exercises 4b) and 5a), one can write $\operatorname{card}\mathcal{P}(X) = 2^{\operatorname{card} X}$, and, in particular, $\operatorname{card}\mathcal{P}(\mathbb{N}) = 2^{\operatorname{card}\mathbb{N}} = \operatorname{card}\mathbb{R}$;

d) for any set $X$

$$\operatorname{card} X < 2^{\operatorname{card} X}, \text{ in particular, } n < 2^n \text{ for any } n \in \mathbb{N}\,.$$

Hint: See Cantor's theorem in Subsect. 1.4.1.

**6.** Let $X_1, \ldots, X_n$ be a finite system of finite sets. Show that

$$\operatorname{card}\left(\bigcup_{i=1}^{m} X_i\right) = \sum_{i_1} \operatorname{card} X_{i_1} - \sum_{i_1<i_2} \operatorname{card}(X_{i_1} \cap X_{i_2}) + \sum_{i_1<i_2<i_3} \operatorname{card}(X_{i_1} \cap X_{i_2} \cap X_{i_3}) - \cdots + (-1)^{m-1} \operatorname{card}(X_1 \cap \cdots \cap X_m)\,,$$

the summation extending over all sets of indices from 1 to $m$ satisfying the inequalities under the summation signs.

**7.** On the closed interval $[0,1] \subset \mathbb{R}$ describe the sets of numbers $x \in [0,1]$ whose ternary representation $x = 0.\alpha_1\alpha_2\alpha_3\ldots$, $\alpha_i \in \{0,1,2\}$, has the property:

a) $\alpha_1 \neq 1$;

b) $(\alpha_1 \neq 1) \wedge (\alpha_2 \neq 1)$;

c) $\forall i \in \mathbb{N}\,(\alpha_i \neq 1)$ (the *Cantor set*).

**8.** (Continuation of Exercise 7.) Show that

a) the set of numbers $x \in [0,1]$ whose ternary representation does not contain 1 has the same cardinality as the set of all numbers whose binary representation has the form $0.\beta_1\beta_2\ldots$;

b) the Cantor set has the same cardinality as the closed interval $[0,1]$.