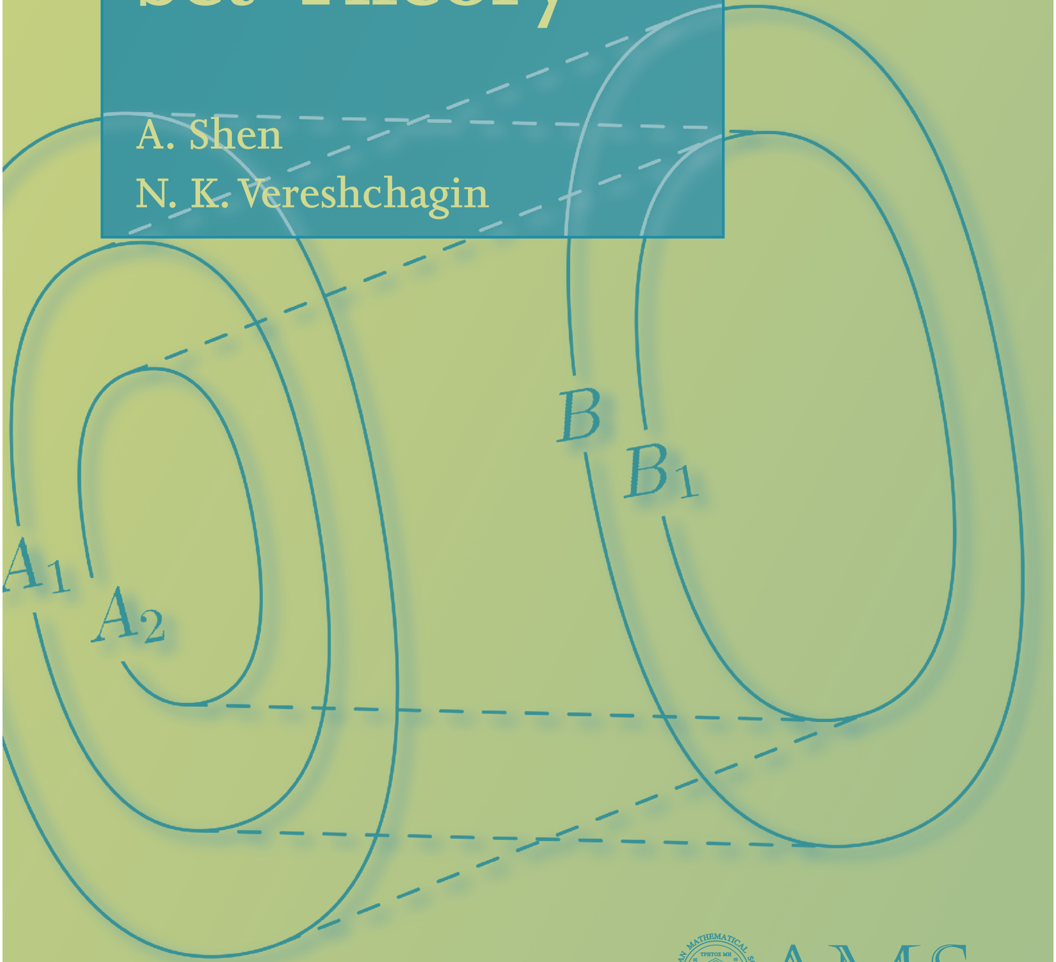
STUDENT MATHEMATICAL LIBRARY
Volume 17

Basic Set Theory

A. Shen
N. K. Vereshchagin



AMS
AMERICAN MATHEMATICAL SOCIETY

Basic
Set Theory

STUDENT MATHEMATICAL LIBRARY
Volume 17

Basic Set Theory

A. Shen

N. K. Vereshchagin

AMS
American Mathematical Society

Н. К. Верещагин, А. Шень

ОСНОВЫ ТЕОРИИ МНОЖЕСТВ

МЦНМО, Москва, 1999

Translated from the Russian by A. Shen

2000 *Mathematics Subject Classification*. Primary 03–01, 03Exx.

Abstract. The book is based on lectures given by the authors to undergraduate students at Moscow State University. It explains basic notions of "naive" set theory (cardinalities, ordered sets, transfinite induction, ordinals). The book can be read by undergraduate and graduate students and all those interested in basic notions of set theory. The book contains more than 100 problems of various degrees of difficulty.

Library of Congress Cataloging-in-Publication Data

Vereshchagin, Nikolai Konstantinovich, 1958–
[Osnovy teorii mnozhestv. English]
Basic set theory / A. Shen, N. K. Vereshchagin.
p. cm. — (Student mathematical library, ISSN 1520-9121 ; v. 17)
Authors' names on t.p. of translation reversed from original.
Includes bibliographical references and index.
ISBN 0-8218-2731-6 (acid-free paper)
1. Set theory. I. Shen, A. (Alexander), 1958– II. Title. III. Series.

QA248 .V4613 2002
511.3′22—dc21 2002066533

∞ The paper used in this book is acid-free and falls within the guidelines
established to ensure permanence and durability.
Visit the AMS home page at http://www.ams.org/

10 9 8 7 6 5 4 3 2 1 07 06 05 04 03 02

Contents

Preface

This book is based on notes from several undergraduate courses the authors offered for a number of years at the Department of Mathematics and Mechanics of Moscow State University. (We hope to extend this series: the books "Calculi and Languages" and "Computable Functions" are in preparation.)

The main notions of set theory (cardinals, ordinals, transfinite induction) are among those any professional mathematician should know (even if (s)he is not a specialist in mathematical logic or set-theoretic topology). Usually these notions are briefly discussed in the opening chapters of textbooks on analysis, algebra, or topology, before passing to the main topic of the book. This is, however, unfortunate—the subject is sufficiently interesting, important, and simple to deserve a leisurely treatment.

It is such a leisurely exposition that we are trying to present here, having in mind a diversified audience: from an advanced high school student to a professional mathematician (who, on his/her way to vacations, wants to finally find out what is this transfinite indiction which is always replaced by Zorn's Lemma). For deeper insight into set theory the reader can turn to other books (some of which are listed in references).

We would like to use this opportunity to express deep gratitude to our teacher Vladimir Andreevich Uspensky, whose lectures, books, and comments influenced us (and this book) perhaps even more than we realize.

We are grateful to the AMS and Sergei Gelfand (who suggested to translate this book into English) for patience. We also thank Yuri Burman who helped a lot with the translation.

Finally, we wish to thank all participants of our lectures and seminars and all readers of preliminary versions of this book.

We would appreciate learning about all errors and typos in the book found by the readers (and sent by e-mail to `ver@mccme.ru` or `shen@mccme.ru`).

A. Shen, N. K. Vereshchagin

Chapter 1

Sets and Their Cardinalities

1. Sets

Let us recall several operations on sets and notation for them:

- A *set* consists of *elements*. Notation: $x \in M$ means that x is an element of a set M (belongs to M).
- A set A is a *subset* of a set B ($A \subset B$) if each element of A is also an element of B. In this case B is called a *superset* of A.
- Two sets A and B are *equal* ($A = B$) if they consist of the same elements (i.e., if $A \subset B$ and $B \subset A$).
- If A is a subset of B and $A \neq B$, then A is called a *proper* subset of B (notation: $A \subsetneq B$).
- The *empty* set $\varnothing$ (called also the *null* set) contains no elements. It is a subset of any set.
- The *intersection* $A \cap B$ of two sets A and B consists of all elements that belong both to A and to B:

$$A \cap B = \{x \mid x \in A \text{ and } x \in B\}.$$

- The *union* $A \cup B$ consists of all elements of A and all elements of B (and no other elements):

$$A \cup B = \{x \mid x \in A \text{ or } x \in B\}.$$

- The *set difference* $A \setminus B$ consists of elements of A that are *not* elements of B:

$$A \setminus B = \{x \mid x \in A \text{ and } x \notin B\}.$$

  There is a special case: if B is a subset of A, the difference $A \setminus B$ is also called a *complement* of B in A.

- The *symmetric difference* $A \mathbin{\triangle} B$ consists of all elements that belong to exactly one of the sets A and B:

$$A \mathbin{\triangle} B = (A \setminus B) \cup (B \setminus A) = (A \cup B) \setminus (A \cap B).$$

- By $\{a, b, c\}$ we denote the set that contains a, b, c and no other elements. Some of the elements a, b, c may coincide; it this case $\{a, b, c\}$ consists of one or two elements. This notation is also used in a less formal way. For example, the set of all elements of a sequence $a_0, a_1, \ldots$ is denoted by $\{a_0, a_1, \ldots\}$ (and sometimes even $\{a_i\}$). More pedantic notation would be $\{a_i \mid i \in \mathbb{N}\}$, where $\mathbb{N}$ is the set of all natural numbers ($\mathbb{N} = \{0, 1, 2, \ldots\}$).

The notion of a set is relatively new. It appeared at the end of the 19th century when Cantor started comparing cardinalities of sets; see Section 3 of this chapter. The notion of a set turned out to be convenient and even found its way into high school mathematics. Instead of saying that equation $x^2 + 1 = 0$ has no solutions, teachers started explaining that the set of all solutions of this equation is empty, etc. Some teachers even tried to explain the difference between the empty set $\varnothing$ and the set $\{\varnothing\}$, whose only element is the empty set, but with very limited success. The idea to modernize the high school curriculum by using set-theoretic language from the very beginning created a lot of problems.

We assume, however, that the reader is familiar with the set-theoretic language, and we will use it freely. Here are some problems for self-assessment; we hope that most of them will be easy for you.

Problem 1. Consider the oldest mathematician among chess players and the oldest chess player among mathematicians. Could they be two different people?

Problem 2. The same question for the best mathematician among chess players and the best chess player among mathematicians.

Problem 3. One tenth of mathematicians are chess players, and one sixth of chess players are mathematicians. Which group (mathematicians or chess players) is bigger? What is the ratio of sizes of these two groups?

Problem 4. Do there exist sets A, B and C such that $A \cap B \neq \varnothing$, $A \cap C = \varnothing$ and $(A \cap B) \setminus C = \varnothing$?

Problem 5. Which of the following formulas (a)–(f) are true for arbitrary sets A, B, C: (**a**) $(A \cap B) \cup C = (A \cup C) \cap (B \cup C)$; (**b**) $(A \cup B) \cap C = (A \cap C) \cup (B \cap C)$; (**c**) $(A \cup B) \setminus C = (A \setminus C) \cup B$; (**d**) $(A \cap B) \setminus C = (A \setminus C) \cap B$; (**e**) $A \setminus (B \cup C) = (A \setminus B) \cap (A \setminus C)$; (**f**) $A \setminus (B \cap C) = (A \setminus B) \cup (A \setminus C)$?

Problem 6. Give formal proofs of all valid formulas from the previous problem, starting from definitions. (Your proof should go like this: "We have to prove that the left-hand side equals the right-hand side. Let x be any element of the left-hand side set. Then … . Therefore, x belongs to the right-hand side set. On the other hand, let … ".)

Give counterexamples to the formulas which are not always true.

Problem 7. Prove that the symmetric difference operation is associative: $A \triangle (B \triangle C) = (A \triangle B) \triangle C$ for any A, B and C. (*Hint*: Addition modulo 2 is associative.)

Problem 8. Prove that $(A_1 \cap \cdots \cap A_n) \triangle (B_1 \cap \cdots \cap B_n) \subset (A_1 \triangle B_1) \cup \cdots \cup (A_n \triangle B_n)$ for arbitrary sets $A_1, \ldots, A_n$ and $B_1, \ldots, B_n$.

Problem 9. Consider an equality whose left-hand side and right-hand side contain set variables and operations $\cap$, $\cup$, $\setminus$. Prove that if this equality is false for some sets, then it is false for some sets that contain at most one element.

Problem 10. How many different expressions can be formed from set variables A and B by using union, intersection and set difference? (Variables and operations can be used more than once. Two expressions are considered identical if they assume the same value for each set of values of the variables involved.) Solve the same problem for three sets and for n sets. (*Answer* in the general case: 2^{2^n-1}.)

Problem 11. Solve the same problem if only $\cup$ and $\cap$ are allowed. (For $n = 2$ and $n = 3$ this problem is easy to solve; however, no general formula for any n is known. This problem is also called "counting monotone Boolean functions in n variables".)

Problem 12. How many subsets does an n-element set have?

Problem 13. Assume that A consists of n elements and $B \subset A$ consists of k elements. Find the number of different sets C such that $B \subset C \subset A$.

Problem 14. A set U contains $2n$ elements. We select k subsets of A in such a way that none of them is a subset of another one. What is the maximum possible value of k? (*Hint*: Maximal k is achieved when all subsets have n elements. Indeed, imagine the following process: We start with an empty set and add random elements one by one until we get U. At most one selected set can appear during this process. On the other hand, the expected number of selected sets that appear during this process can be computed using the linearity of expectation. Take into account that the probability to come across some set $Z \subset U$ is minimal when Z contains n elements, since all the sets of a given size are equiprobable.)

2. Cardinality

Cardinality of a finite set A is defined as the number of its elements. Cardinality of a set A is denoted by $\#A$ or $|A|$. This notation will be extended to infinite sets (see below). The following formula gives the cardinality of the union of several sets in terms of their cardinalities and the cardinalities of all their intersections.

Theorem 1 (Inclusion-Exclusion Principle)**.**

$$\begin{aligned}|A \cup B| &= |A| + |B| - |A \cap B|;\\ |A \cup B \cup C| &= |A| + |B| + |C|\\ &\quad - |A \cap B| - |A \cap C| - |B \cap C|\\ &\quad + |A \cap B \cap C|.\end{aligned}$$

In general, $|A_1 \cup \cdots \cup A_n|$ *equals*

$$\sum_i |A_i| - \sum_{i<j} |A_i \cap A_j| + \sum_{i<j<k} |A_i \cap A_j \cap A_k| - \cdots .$$

Proof. This formula can be proved by induction on n, but we provide another, more interesting, proof. Let U be an arbitrary superset of $A_1, \ldots, A_n$. For a set $X \subset U$ consider its *characteristic function* χ_X that is defined on U as follows: $\chi_X(x) = 1$ for $x \in X$ and $\chi_X(x) = 0$ for $x \notin X$. Set-theoretic operations can be expressed in terms of characteristic functions. For example, the characteristic function of the intersection of two sets A and B is the product of characteristic functions of A and B: $\chi_{A\cap B}(u) = \chi_A(u)\chi_B(u)$. If B is a complement of A in U, then $\chi_B(x) = 1 - \chi_A(x)$ for all $x \in U$.

The cardinality of a set $X \subset U$ is the sum of all values of χ_X:

$$|X| = \sum_u \chi_X(u).$$

The union $A_1 \cup \cdots \cup A_N$ is a complement of the intersection of complements; therefore,

$$\chi_{A_1\cup\cdots\cup A_n} = 1 - (1 - \chi_{A_1}) \cdots (1 - \chi_{A_n}).$$

The right-hand side equals

$$\sum_i \chi_{A_i} - \sum_{i<j} \chi_{A_i}\chi_{A_j} + \sum_{i<j<k} \chi_{A_i}\chi_{A_j}\chi_{A_k} - \cdots .$$

Summation over all elements of U (both sides of the equality are integer-valued functions defined on U) gives the Inclusion-Exclusion Principle. □

Problem 15. Prove that $|A_1 \triangle \cdots \triangle A_n|$ equals

$$\sum_i |A_i| - 2\sum_{i<j} |A_i \cap A_j| + 4 \sum_{i<j<k} |A_i \cap A_j \cap A_k| - \cdots$$

(the coefficients are powers of 2).

Some other theorems of elementary combinatorics are stated below as exercises. We are mostly interested in the following principle:

> *if there is a one-to-one correspondence between two sets A and B, then $|A| = |B|$.*

One-to-one correspondence between two sets means that each element of the first set corresponds to precisely one element of the second set (and vice versa).

Here are some problems that use this principle.

Problem 16. Consider 1000 white points and one black point on a circle. Count all triangles whose vertices are white points. Count all convex quadrangles formed by three white points and one black point. Which number is larger? (*Solution*: The numbers are equal because each quadrangle corresponds to precisely one triangle formed by three white vertices of the quadrangle.)

Problem 17. Fix a set of cardinality 100. Count all its 57-subsets, i.e., subsets of cardinality 57. Count all its subsets of cardinality 43. Which number is larger? (*Hint*: $57 + 43 = 100$.)

Problem 18. Prove that the number of all binary strings of length n equals the number of all subsets of the set $\{1, 2, \ldots, n\}$. (*Hint*: Each subset $X \subset \{1, 2, \ldots, n\}$ corresponds to its characteristic sequence; ith element of this sequence equals 1 if and only if $i \in X$.)

Problem 19. Prove that the number of binary sequences of length n that consist of k ones and $n - k$ zeros equals the number of k-subsets of an n-set.

The number of k-subsets of an n-set is denoted by $\binom{n}{k}$ and is called the number of *k-combinations* of an n-set. It is also called a *binomial coefficient* and appears in the binomial expansion (see below).

Problem 20. Prove that

$$\binom{n}{k} = \binom{n}{n-k}.$$

Problem 21. Prove that

$$\binom{n}{0} + \binom{n}{1} + \binom{n}{2} + \cdots + \binom{n}{n} = 2^n.$$

Problem 22. Let U be any finite set. Prove that the number of subsets $X \subset U$ having even cardinality equals the number of subsets $X \subset U$ having odd cardinality. (*Hint*: Fix some $u \in U$ and consider pairs formed by subsets that differ only at u.)

Problem 23. Prove that

$$\binom{n}{0} - \binom{n}{1} + \binom{n}{2} + \cdots + (-1)^n \binom{n}{n} = 0.$$

(*Hint*: Use the preceding problem.)

Problem 24. Prove the *Newton binomial expansion* formula:

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

Problem 25. Consider a nonassociative product of n terms. There are several ways to insert parentheses that indicate the order of operations. Prove that the number of different ways equals the number of triangulations of a convex $(n+1)$-gon by diagonals. (For example, the product abc can be either $(ab)c$ or $a(bc)$; on the other hand, there are two ways of cutting a quadrangle into two triangles by a diagonal. For the product $abcd$ and a pentagon there are 5 possibilities.) These numbers are called *Catalan numbers*.

3. Equal cardinalities

We say that two sets A and B *have the same cardinality* if there exists a one-to-one correspondence between A and B (each element of A corresponds to exactly one element of B and vice versa). Notation: $A \simeq B$.

Evidently, two finite sets A and B have the same cardinality if and only if $|A| = |B|$ (the number of elements in A equals the number of elements in B). However, the definition makes sense for infinite sets as well. For example, let us prove that closed intervals $[0, 1]$ and $[0, 2]$ have the same cardinality. Indeed, the mapping $x \mapsto 2x$ is a one-to-one correspondence between $[0, 1]$ and $[0, 2]$.

Problem 26. Prove that any two intervals (a, b) and (c, d) have the same cardinality.

Problem 27. Prove that any two circles have the same cardinality. Prove that any two disks have the same cardinality.

Problem 28. Prove that $[0,1) \simeq (0,1]$.

The following problem is somewhat more difficult: Prove that $(0,1) \simeq (0,+\infty)$. It can be done as follows. Note that the mapping $x \mapsto 1/x$ is a one-to-one correspondence between $(0,1)$ and $(1,+\infty)$. Note also that the mapping $x \mapsto (x-1)$ is a one-to-one correspondence between $(1,+\infty)$ and $(0,+\infty)$. Therefore their composition, i.e., the mapping $x \mapsto (1/x)-1$ is a one-to-one correspondence between $(0,1)$ and $(0,+\infty)$. Q.e.d.

More generally, one can say that the relation "to have the same cardinality" is an *equivalence relation*. This means that this relation is *reflexive* ($A \simeq A$ for any set A), *symmetric* (if $A \simeq B$ then $B \simeq A$) and *transitive* (if $A \simeq B$ and $B \simeq C$ then $A \simeq C$). We have applied the transitivity property using $B = (1,+\infty)$ as an intermediate set.

Other examples:

- The set of all infinite sequences of zeros and ones has the same cardinality as the set of all subsets of the set $\mathbb{N}$ of natural numbers. (Indeed, for each sequence $a_0a_1a_2\ldots$ consider a set of all $i \in \mathbb{N}$ such that $a_i = 1$. For example, the sequence $00000\ldots$ corresponds to the empty set, the sequence $11111\ldots$ corresponds to the set $\mathbb{N}$, and the sequence $10101010\ldots$ corresponds to the set of all even numbers.)
- The set of all infinite sequences of digits 0, 1, 2, 3 has the same cardinality as the set of all infinite sequences of zeros and ones. (Indeed, one can encode 0, 1, 2, 3 by blocks 00, 01, 10, 11. The inverse mapping splits sequences of zeros and ones into blocks of length 2. Then each block is replaced by one of the four digits $0,\ldots,3$.)
- The set of all infinite sequences of digits 0, 1, 2 has the same cardinality as the set of all infinite sequences of zeros and ones. (A naïve approach: the set of all sequences of digits $0,1,2$ lies in between two sets of the same cardinality (i.e., the set of all sequences of digits $0,1$ and all sequences of digits

$0, 1, 2, 3$) and therefore has the same cardinality. This argument in indeed valid; see the Cantor–Bernstein Theorem in Section 5 of this chapter. However, we can construct a one-to-one correspondence explicitly if we encode the digits 0, 1, and 2 by the blocks 0, 10, and 11. It is easy to see that each infinite sequence of zeros and ones can be split into these blocks uniquely from left to right. These three blocks form the so-called "prefix code".)

- Generalizing the example above, one can prove that the set $P(U)$ of all subsets of any set U (it is called the *power set* of U) has the same cardinality as the set 2^U of all (everywhere defined) functions of type $U \to \{0, 1\}$. (Indeed, each subset $X \subset U$ corresponds to its characteristic function χ_X.)

To continue this list, we need to prove some properties of countable sets.

4. Countable sets

A set X is called *countable* if X has the same cardinality as the set $\mathbb{N}$ of natural numbers. Reformulation: X is countable if $X = \{x_0, x_1, x_2, \dots\}$ (here x_i corresponds to natural number i; we need a one-to-one correspondence, so all x_i should be different).

For example, the set $\mathbb{Z}$ of all integers is countable since $\mathbb{Z} = \{0, 1, -1, 2, -2, 3, -3, \dots\}$.

Theorem 2. (a) *Any subset of a countable set is finite or countable.*

(b) *Any infinite set has a countable subset.*

(c) *The union of a finite or countable family of finite or countable sets is finite or countable.*

Proof. (a) Let B be a subset of a countable set $A = \{a_0, a_1, a_2, \dots\}$. Delete all elements of the sequence $a_0, a_1, \dots$ that do not belong to B. The remaining elements form either a finite sequence (B is finite) or an infinite sequence (B is countable).

(b) Let A be an infinite set. Then A is not empty; let b_0 be some element of A. Since A is infinite, it contains other elements. Let b_1

be one of them (i.e., $b_1 \in A$, $b_1 \neq b_0$). Since A has more than two elements, we can choose element b_2 that differs from b_0 and b_1, etc.

We get a sequence $b_0, b_1, \ldots$ (since A if infinite, for each i we can find a new element $b_i \in A$). Then the set $B = \{b_0, b_1, \ldots\}$ is a countable subset of A. (Note that B may be a proper subset of A even if A is countable.)

(c) Consider a countable family of countable sets $A_0, A_1, A_2, \ldots$. Since A_i is countable, the elements of A can be arranged in a sequence $A_i = \{a_{i0}, a_{i1}, \ldots\}$. Let us write down all these sequences; we get a table

$$\begin{array}{ccccc}
a_{00} & a_{01} & a_{02} & a_{03} & \ldots \\
a_{10} & a_{11} & a_{12} & a_{13} & \ldots \\
a_{20} & a_{21} & a_{22} & a_{23} & \ldots \\
a_{30} & a_{31} & a_{32} & a_{33} & \ldots \\
\ldots & \ldots & \ldots & \ldots & \ldots
\end{array}$$

Now this table can be converted into a sequence. For example, we can walk along diagonals

$$a_{00}, \quad a_{01}, a_{10}, \quad a_{02}, a_{11}, a_{20}, \quad a_{03}, a_{12}, a_{21}, a_{30}, \ldots.$$

If all A_i are disjoint, this sequence provides a one-to-one correspondence between $\mathbb{N}$ and the union of all A_i. If A_i are not disjoint, we must delete repetitions.

If we have only finitely many sets (or some sets are finite), some elements of our sequence disappear and remaining elements form a finite or countable set. □

Problem 29. We have described a one-to-one correspondence between the set of all ordered pairs of natural numbers (denoted by $\mathbb{N} \times \mathbb{N}$) and $\mathbb{N}$: a pair $\langle x, y\rangle$ corresponds to some natural number $p(x, y)$. For example, $p(0,0) = 0$, $p(0,1) = 1$, $p(1,0) = 2$, $p(0,2) = 3$, $p(1,1) = 4$, etc. It turns out that p is a polynomial with rational coefficients. Find this polynomial.

Remark. There is a subtle point in the proof of Theorem 2(b). We have selected elements of A one by one. We know that (at each step) some unused element of A does exist. However, there is no rule that determines which element of A should be selected. More rigorous approach uses a special axiom, called the *axiom of choice*. This axiom was considered doubtful (and harmful) at the beginning of the 20th century. However, now people are accustomed to it, and the majority of contemporary mathematicians use it and do not worry about that. In the middle of the 20th century Kurt Gödel, perhaps the greatest logician of the century, proved that the axiom of choice cannot be refuted (its negation does not follow from the remaining axioms of set theory, assuming the remaining axioms are consistent). In 1963 Paul Cohen proved that the axiom of choice cannot be derived from the remaining axioms (if they are consistent). Of course, to explain the Gödel and Cohen theorems (not to mention their proofs), we would need to develop axiomatic set theory, and this goes far beyond the scope of our book.

Problem 30. The axiom of choice is also used in the proof of part (c). Can you see where? (*Answer*: We know that the sets A_i are countable. This means that for each i there *exists* a one-to-one correspondence between $\mathbb{N}$ and A_i. But the mere existence is not enough; we have to fix these mappings, and only after that we can construct a one-to-one correspondence between the union of all A_i and $\mathbb{N}$.)

Some other examples of countable sets:

- The set $\mathbb{Q}$ of rational numbers is countable. Indeed, rational numbers are fractions of two integers. The set of fractions with a given denominator is countable. Therefore, $\mathbb{Q}$ is a union of a countable family of countable sets. (As we shall see in Section 6 of this chapter, the set $\mathbb{R}$ of all real numbers is uncountable.)
- The set $\mathbb{N}^k$ formed by k-tuples of natural numbers, is countable. Let us prove this using induction on k. For $k = 2$ the set $\mathbb{N}^2 = \mathbb{N} \times \mathbb{N}$ (whose elements are pairs of natural numbers) is a countable union of countable sets $\{0\} \times \mathbb{N}, \{1\} \times \mathbb{N}, \ldots$ (elements of an ith set are pairs $\langle i, \text{something}\rangle$). Therefore $\mathbb{N}^2$ is countable.

  Similarly, $\mathbb{N}^3$ is a countable union of sets $\{i\} \times \mathbb{N} \times \mathbb{N}$. Each of these sets is a set of triples with fixed first element. Therefore, $\{i\} \times \mathbb{N} \times \mathbb{N}$ has the same cardinality as $\mathbb{N} \times \mathbb{N}$ and is countable (induction assumption).

The same argument works for $\mathbb{N}^4$, $\mathbb{N}^5$, etc. (we prove that $\mathbb{N}^{k+1}$ is countable using that $\mathbb{N}^k$ is countable).

- The set of all finite sequences of natural numbers is countable. Indeed, as we have seen, the set of all sequences of a given length k (i.e., $\mathbb{N}^k$) is countable, so the set of all finite sequences of natural numbers is a countable union of countable sets.

- In the previous example we can replace natural numbers by elements of any countable (or finite) set. For example, we can consider the set of all English texts. This set is countable (text is a finite sequence of letters, digits, punctuation marks and other ASCII characters). The same is true for the set of all (possible) computer programs, etc.

- A real number x is said to be *algebraic* if x is a root of a nonzero polynomial with integer coefficients. (For example, any rational number is algebraic since it is a root of a polynomial of degree 1 with integer coefficients; $\sqrt{2}$ and $\sqrt{2}+\sqrt{3}$ are also algebraic numbers because $x^2-2=0$ for $x=\sqrt{2}$ and $(x^2-5)^2-24=0$ for $x=\sqrt{2}+\sqrt{3}$.)

  The set of all algebraic numbers is countable. Indeed, the set $\mathbb{Z}[x]$ of all polynomials with integer coefficients is countable (each polynomial is determined by a finite sequence of integer coefficients), and each (nonzero) polynomial has finitely many roots (at most n for a polynomial of degree n).

- The set of all periodic decimal fractions is countable. (Indeed, each periodic fraction can be represented by a finite string that includes digits, decimal period and parentheses. For example, $1/6 = 0.16666\ldots$ can be written as $0.1(6)$. Now recall that the set of all finite strings is countable.)

Problem 31. Prove that any family of disjoint open intervals (p,q) (where p and q are any real numbers such that $p<q$) is finite or countable. (*Hint*: Any interval contains a rational point.)

Problem 32. (**a**) Prove that any family of disjoint 8-signs on the plane is countable. (By an 8-sign we mean a union of two tangent circles of any size; the interior part of the circles is not included.)

(**b**) Prove a similar statement for letters T or E on the plane (but not for M or O!).

Problem 33. A point $x \in \mathbb{R}$ is called a *maximum point* for a function $f\colon \mathbb{R} \to \mathbb{R}$ if there exists some $\varepsilon > 0$ such that $f(x) > f(x+h)$ for any h such that $|h| < \varepsilon$ and $h \neq 0$. Prove that the set of all maximum points (for any function f) is either finite or countable.

Problem 34. Let $f\colon \mathbb{R} \to \mathbb{R}$ be a nondecreasing function. Prove that f is continuous everywhere except for some countable set.

Theorem 3. *If A is infinite and B is countable (or finite), the union $A \cup B$ has the same cardinality as A.*

Proof. Without loss of generality we may assume that A and B are disjoint: $A \cap B = \varnothing$. Indeed, the intersection $A \cap B$ can be deleted from B; the remaining set $B' = B \setminus A$ is still countable (or finite).

Let P be a countable subset of A; let Q be the rest: $Q = A \setminus P$. We have to prove that $B+P+Q$ has the same cardinality as $P+Q$ (we use $+$ instead of $\cup$ to emphasize that the sets are disjoint). Both $B+P$ and P are countable. Consider a one-to-one correspondence between them and extend it to the one-to-one correspondence between $B+P+Q$ and $P+Q$ (which is the identity on Q: each element $q \in Q$ corresponds to itself). □

Problem 35. Using this approach, construct a one-to-one correspondence between the closed interval $[0,1]$ and the half-open interval $[0,1)$. (*Hint*: Take $B = \{1\}$.)

Problem 36. Theorem 3 guarantees that adding a finite or countable set to an infinite set does not change its cardinality. Prove a similar result for subtraction: If A is infinite and uncountable, and B is finite or countable, then $A \setminus B$ has the same cardinality as A.

Problem 37. R. Dedekind suggested the following definition of an infinite set: A set A is infinite if there exists a one-to-one correspondence between A and a proper subset $B \subsetneq A$. Show that this property does characterize infinite sets.

Using Theorem 3, one can easily prove that the real line $\mathbb{R}$ and each of the intervals ($[a, b], (a, b), [a, b), [a, +\infty)$, etc.) have the same cardinality.

Problem 38. Construct a one-to-one correspondence between the set $[0, 1] \cup [2, 3] \cup [4, 5] \cup \cdots$ and $[0, 1]$.

Problem 39. Prove that the set of all points on the plane has the same cardinality as the set of all lines. (*Hint*: The line $y = ax + b$ corresponds to the pair (a, b); do not forget about vertical lines.)

Problem 40. Prove that a half-plane (the set of all points on one side of a line) has the same cardinality as the entire plane. (It does not matter whether we include the boundary line in the half-plane or not.)

Theorem 4. *The interval* $[0, 1]$ *has the same cardinality as the set* $2^{\mathbb{N}}$ *of all infinite sequences of zeros and ones.*

Proof. Indeed, any real number $x \in [0, 1]$ can be represented by an infinite binary fraction. The first digit (after binary point) is 0 if x belongs to the left half of [0,1] and is 1 if x belongs to the right half. To find out the next digit, we divide the selected part in two halves and see which half contains x, etc.

The same correspondence can be defined "from right to left": a sequence $x_0x_1x_2\ldots$ corresponds to the real number

$$\frac{x_0}{2} + \frac{x_1}{4} + \frac{x_2}{8} + \cdots.$$

(We assume that the reader is familiar with calculus; this assumption is unavoidable since we speak about real numbers!)

A careful reader will notice that our description ignores an important problem: fractions with denominator 2^n (for integer n) have two representations. For example, the fraction 3/8 can be written either as .011000... or as .010111... . To get a one-to-one correspondence we must eliminate sequences of zeros and ones that have only finitely many zeros (i.e., periodic fractions with period 1). But we know that periodic fractions form a countable set, so it does not matter whether we eliminate them or not. □

Problem 41. Write down a binary fraction that corresponds to 1/3.

We have used binary fractions, but we can use ordinary decimal fractions as well and prove that the set $[0,1]$ has the same cardinality as the set of all infinite sequences of digits $0,1,\ldots,9$. (Therefore, the set of all infinite binary fractions has the same cardinality as the set of all infinite decimal fractions. This statement can be proved directly using the trick described on p. 8.)

Now we are ready to prove the following remarkable theorem:

Theorem 5. *The unit square* (*with interior*) *has the same cardinality as the closed unit interval.*

Proof. Points of a square are determined by their coordinates, so the unit square has the same cardinality as the set $[0,1] \times [0,1]$ formed by pairs $\langle x, y\rangle$ (where $x, y \in [0,1]$). We know already that elements of $[0,1]$ can be replaced by sequences of zeros and ones. It remains to note that a pair of sequences

$$\langle x_0x_1x_2\ldots, y_0y_1y_2\ldots\rangle$$

can be mapped to a "mixed" sequence

$$x_0y_0x_1y_1x_2y_2\ldots$$

and this mapping provides one-to-one correspondence between sequences and pairs of sequences. □

The German mathematician Georg Cantor, who invented set theory, proved this result in 1877 and was very surprised by it. Indeed, this result contradicts our intuitive perception of "dimension" (a square has dimension 2, whereas a line segment has dimension 1; therefore, a square should have "more points"). Cantor wrote to Dedekind that he was interested to know whether spaces of different dimension have the same number of points; he remarked: "it seems that this question should be answered affirmatively, though I had a different opinion for several years" (June 20, 1877).

Dedekind answered that Cantor's result had not made the notion of dimension meaningless; it only showed that we have to restrict our attention to one-to-one correspondences that are continuous (in both directions), and then we can distinguish between spaces of different dimensions. This conjecture turned out to be true; however, it is quite nontrivial. The first attempts to prove it (including the proof in one of Cantor's papers) contained errors. Only thirty years later Brouwer gave a correct proof. (One should note that for a line segment and a square the proof is simple; problems arise in higher dimensions. Let us note also that there exists a continuous mapping $\pi\colon [0,1] \to [0,1] \times [0,1]$ whose range is $[0,1] \times [0,1]$. This strange mapping is called "Peano's curve".)

Theorem 5 has many corollaries: it is easy to prove now that a disk has the same cardinality as its boundary, the line has the same cardinality as the plane, etc.

A one-to-one correspondence between the pairs of reals and the reals can be extended to a one-to-one correspondence between the triples of reals and the pairs of reals (if a pair $\langle x, y\rangle$ corresponds to u, the triple $\langle x, y, z\rangle$ corresponds to $\langle u, z\rangle$). Therefore, three-dimensional space has the same cardinality as the two-dimensional plane (and therefore the same cardinality as a one-dimensional line). Similar argument shows that the spaces $\mathbb{R}^n$ (= set of all n-tuples of reals) for all n have the same cardinality.

Problem 42. Prove that the set of all *finite* sequences of real numbers has the same cardinality as the set $\mathbb{R}$ (the set of reals).

Problem 43. Prove that the set of all *infinite* sequences of real numbers has the same cardinality as $\mathbb{R}$.

Note that we are still unable to show that the set $\mathbb{R}$ (or the set of infinite sequences of zeros and ones) is uncountable. See Section 6 of this chapter.

One says that the set $\mathbb{R}$ has the *cardinality of the continuum* (because a point can move *continuously* along a line).

5. Cantor–Bernstein Theorem

We have given a formal definition for the intuitive notion of "having the same size" (requiring the existence of a one-to-one correspondence). Now we want to give a formal definition that reflects the intuitive idea of "one set being larger than another".

We say that *the cardinality of a set A does not exceed the cardinality of a set B* if there exists a one-to-one correspondence between A and a subset of B (which may be equal to the entire B).

Problem 44. Professor X suggests the following definition: the cardinality of a set A is strictly less than the cardinality of a set B if there exists a one-to-one correspondence between A and a proper subset $B' \subsetneq B$. Explain why this definition is not really good. (*Hint*: Popular expositions of set theory often start with the following paradox that goes back to Galilei. Are there as many squares of integers $(0, 1, 4, 9, 16, \ldots)$ as all nonnegative integers? Squares are rare: most nonnegative integers are not squares. On the other hand, there are as many squares as all natural numbers; i^2 corresponds to i.)

The relation "the cardinality of A does not exceed the cardinality of B" between two sets A and B has the following natural properties:

- If A and B have the same cardinality, then the cardinality of A does not exceed the cardinality of B. Indeed, if there exists a one-to-one correspondence between A and B, then there exists a one-to-one correspondence between A and a subset B' of B. (Evident: let $B' = B$.)
- If the cardinality of A does not exceed the cardinality of B and the cardinality of B does not exceed the cardinality of C, then the cardinality of A does not exceed the cardinality of C. (Indeed, consider a one-to-one correspondence between A and some $B' \subset B$, and another one-to-one correspondence between B and some $C' \subset C$. The latter maps B' onto some $C'' \subset C' \subset C$ (see Figure 1), and C'' has the same cardinality as A.)

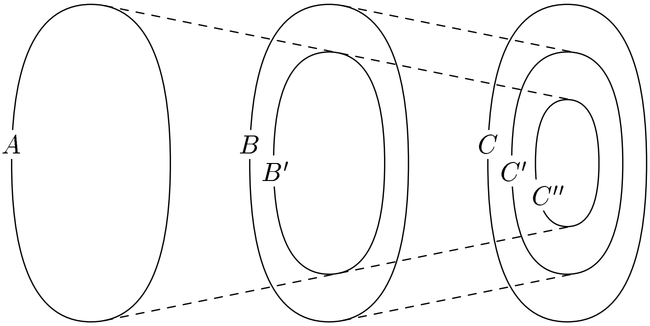


Figure 1. Transitivity.

- If the cardinality of A does not exceed the cardinality of B, and the cardinality of B does not exceed the cardinality of A, then A and B have the same cardinality. (This innocently looking statement is quite nontrivial. It is called the Cantor–Bernstein Theorem and will be proved later in this section.)
- For any two sets A and B either the cardinality of A does not exceed the cardinality of B or the cardinality of B does not exceed the cardinality of A. (To prove this statement, we need to use the so-called "transfinite induction"; see Theorem 25 in Section 6 of Chapter 2.

Theorem 6 (Cantor–Bernstein)**.** *Let A and B be two sets. Assume that there exists a one-to-one correspondence between A and a subset $B_1 \subset B$ and a one-to-one correspondence between B and some $A_1 \subset A$. Then the sets A and B have the same cardinality* (*i.e., there exists a one-to-one correspondence between A and B*)*.*

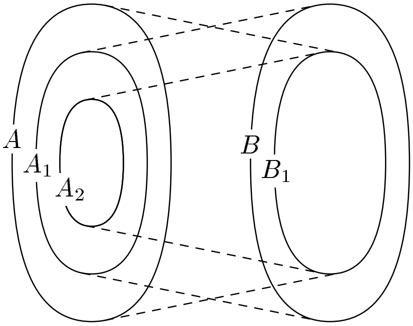


Figure 2. Cantor–Bernstein Theorem and its special case.

Proof. Consider a one-to-one correspondence between B and A_1. The subset $B_1 \subset B$ then corresponds to some $A_2 \subset A_1$ (see Figure 2). All three sets A, B_1 and A_2 have the same cardinality. We have to prove that they have the same cardinality as B (or A_1).

Now we may forget about B and prove the following special case of the Cantor–Bernstein Theorem:

If $A_2 \subset A_1 \subset A_0$ and A_2 has the same cardinality as A_0, then all three sets have the same cardinality.

(To make our notation consistent, we denote A by A_0.)

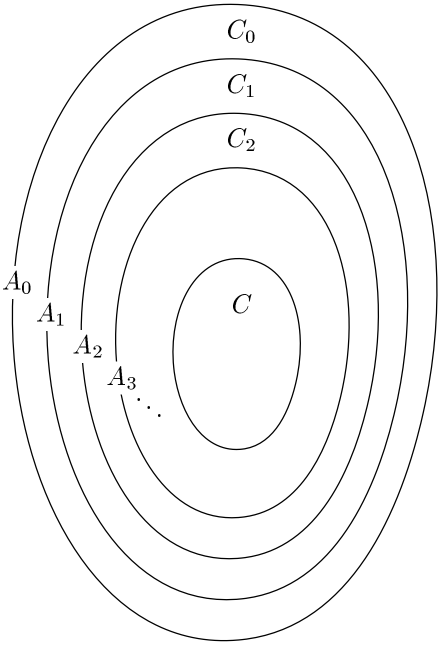


Figure 3. Sets A_i and C_i.

Let f be a function that provides a one-to-one correspondence $A_0 \to A_2$ (an element $x \in A_0$ corresponds to $f(x) \in A_2$). Function f maps A_0 onto A_2. Since A_1 is a subset of A_0, function f maps it onto some $A_3 \subset A_2$ (see Figure 3). In a similar way f maps A_2 onto some set A_4, and $A_4 \subset A_3$ because $A_2 \subset A_1$.

Repeating this procedure, we get a nonincreasing sequence of sets,

$$A_0 \supset A_1 \supset A_2 \supset A_3 \supset A_4 \supset \cdots .$$

A one-to-one function $f\colon A_0 \to A_2$ maps A_i onto A_{i+2} (notation: $f(A_i) = A_{i+2}$). The set A_{2n} is the set of all elements that are obtained from some element of A_0 by n applications of f. Similarly, A_{2n+1} is the set of elements that are obtained from some element of A_1 by n applications of f.

Important remark: the intersection of all A_i may be nonempty. Indeed, it consists of all elements x such that f^{-1} can be applied to x once, twice, … , k times for any k.

We have split A_0 into disjoint layers $C_i = A_i \setminus A_{i+1}$ and the "core" $C = \bigcap_i A_i$.

All layers C_0, C_2, $C_4, \ldots$ have the same cardinality because f provides a one-to-one correspondence between C_0 and C_2, between C_2 and C_4, etc.:

$$C_0 \xrightarrow{f} C_2 \xrightarrow{f} C_4 \xrightarrow{f} \cdots .$$

The same is true for layers with odd indices:

$$C_1 \xrightarrow{f} C_3 \xrightarrow{f} C_5 \xrightarrow{f} \cdots .$$

We may also note (though this fact is not used in the sequel) that f is a permutation on C (i.e., is a one-to-one correspondence between C and C).

Now it is easy to describe a one-to-one correspondence g between A_0 and A_1. Let $x \in A_0$. Then $g(x)$ is equal to $f(x)$ for $x \in C_{2k}$ (for any k) and $g(x) = x$ for $x \in C_{2k+1}$ (for any k) and for $x \in C$ (Figure 4). □

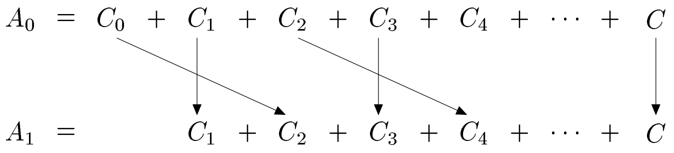

$A_0 = C_0 + C_1 + C_2 + C_3 + C_4 + \cdots + C$
$A_1 = C_1 + C_2 + C_3 + C_4 + \cdots + C$

Figure 4. Vertical arrows are the identity mappings; diagonal arrows are parts of f.

This theorem (sometimes called also the Schröder–Bernstein Theorem) was stated (without proof) by Cantor in 1883; he promised to give the proof in subsequent papers. However, he had not kept his promise, and first proofs were given by Schröder (1896) and Bernstein (1897). It is clear from Cantor's writings that he planned to prove this theorem together with another result mentioned above (for any two sets there exists a one-to-one correspondence between one of them and a subset of the other (see Theorem 25 in Section 6 of Chapter 2). But it is not clear what argument Cantor had in mind.

The Cantor–Bernstein Theorem is useful when we want to prove that two sets have the same cardinality. Here is an example. We want to prove that a ball and a torus (regarded, with their interiors, as subsets of three-dimensional space), have the same cardinality. A small torus inside the ball has the same cardinality as the big torus; a small ball inside the torus has the same cardinality as the big one. It remains to apply the Cantor–Bernstein Theorem. (Direct construction is also possible: both sets can be dissected into circles. However, our argument is much more general and can be applied to any two bounded sets with interior points.)

Problem 45. Look again at the problems above and find out which of them can be easily solved by using the Cantor–Bernstein Theorem.

Problem 46. Prove that any two sets (planar or three-dimensional) that contain a piece of a line (or a curve), have the same cardinality.

Problem 47. A square is represented as the union of two sets: $[0,1]\times[0,1] = A \cup B$. Prove that at least one of the sets A and B has the cardinality of the continuum. (*Hint*: If either A or B contains a line segment, one can apply the Cantor–Bernstein Theorem. If (say) A does not contain a line segment, then each subset $\{x\}\times[0,1]$ intersects B, and we can apply the Cantor–Bernstein Theorem again.)

Problem 48. Prove that if $[0,1] = A \cup B$, then either A or B has the cardinality of the continuum.

The proof of the Cantor–Bernstein Theorem given above can be explained in more abstract terms (without explicit use of natural numbers). Recall that $f\colon A \to A_2$ is a one-to-one correspondence between the set A and its subset A_2, while A_1 is some intermediate set.

We say that a set $X \subset A$ is “good” (just in this proof) if it contains $A \setminus A_1$ and is closed under f, i.e., if

$$X \supset (A \setminus A_1) + f(X).$$

(Here we use "+" instead of "$\cup$" to emphasize that the sets are disjoint.) It is easy to see that the intersection of any family of good sets is a good set. Therefore, the intersection of all good sets is the least good set (the good set that is a subset of any good set). Let M be this intersection. It is easy to check that $(A \setminus A_1) + f(M)$ is a good set. Therefore (since M is a subset of any good set) the inclusion (see the definition of a good set) becomes the equality for M:

$$M = (A \setminus A_1) + f(M).$$

Now we can construct a bijection (one-to-one correspondence) $g\colon A \to A_1$. This bijection g coincides with f on M and coincides with the identity mapping outside M.

Problem 49. Give a detailed proof for the Cantor–Bernstein Theorem following this scheme.

This argument is useful when we develop axiomatic set theory. The advantage is that it does not use natural numbers (the notion of a natural number should be defined in axiomatic set theory, and this definition is not straightforward). But in fact the arguments remain the same, because the least good set M equals $C_0 \cup C_2 \cup \cdots$.

Let us revisit the four possibilities for two sets A and B mentioned above:

- A has the same cardinality as some subset of B, and B has the same cardinality as some subset of A. (In this case A and B have the same cardinality, as the Cantor–Bernstein Theorem says.)
- A has the same cardinality as some subset of B, but B differs from all subsets of A (there is no one-to-one correspondence between B and a subset of A). In this case we say that *A has smaller cardinality than B*.
- B has the same cardinality as some subset of A, but not vice versa (as in the previous case, but with A and B changed places). In this case *A has bigger cardinality than B*.
- There is no one-to-one correspondence between A and a subset of B and there is no one-to-one correspondence between

B and a subset of A. This case is in fact impossible, but we cannot prove this yet (see Section 6 of Chapter 2).

Problem 50. Prove that a countable set has smaller cardinality that any (infinite) uncountable one.

Problem 51. Give a detailed proof of the following statement: if A has smaller cardinality than B, and B has smaller cardinality than C, then A has smaller cardinality than C (transitivity). (*Hint*: Use the Cantor–Bernstein Theorem.)

For a long time we have used the word “cardinality” in a context like “sets A and B have the same cardinality” or “set A has smaller cardinality than B”. But what is the cardinality of a given set A?

One may try to define the cardinality of A as the class of all sets that have the same cardinality as A:

$$|A| = \{X \mid A \simeq X\}.$$

It is easy to see that $|A| = |B|$ (according to this definition) if and only if the sets A and B have the same cardinality. Therefore, our expression “to have the same cardinality” can be understood literally.

The problem is that there are too many sets that have the same cardinality as A, since we allow anything in the world to be elements of these sets. There are so many of them that it is difficult to create a set of all these sets; it may lead to set-theoretical paradoxes (see Section 6 of this chapter, page 28).

How can we overcome this difficulty? The simplest approach is to use the word “cardinality” only in phrases like “has the same cardinality” and “has smaller cardinality” but never speak about cardinalities as objects. Another approach is to introduce a notion of a “class”; classes may contain more elements than sets but cannot be elements of other sets (and classes). This leads to another version of axiomatic set theory that speaks not only about sets but also about classes. Then we define the cardinality of A as the class of all sets that have the same cardinality as A.

Finally, there is a third approach. We can choose for each set A a “standard” set that has the same cardinality as A. Usually this standard set is a minimal *ordinal* that has the same cardinality as A.

We will not go into details because the construction of ordinals as sets is beyond the scope of this book. Let us give one example, however: the cardinality of a set $\{a, b, c\}$ (the number 3) is the ordinal

$$\{\varnothing, \{\varnothing\}, \{\varnothing, \{\varnothing\}\}\}.$$

Here is what Cantor have said about cardinalities in 1895: "We will call by the name 'power' or 'cardinal number' of M the general concept which, by means of our active faculty of thought, arises from the aggregate M when we make abstraction of the nature of its various elements m and of the order in which they are given ... Since every single element m, if we abstract from its nature, becomes a 'unit', the cardinal number $\overline{\overline{M}}$ is a definite aggregate composed of units, and this number has existence in our mind as an intellectual image or projection of the given aggregate M." [Originally in German; this translation is due to Philip E. B. Jourdain.]

Anyway, we use the notation $|A|$ for the cardinality of A just as a shortcut: $|A| = |B|$ means that there exists a one-to-one correspondence between A and B; $|A| \leq |B|$ means that there exists a one-to-one correspondence between A and some subset of B; $|A| < |B|$ means that $A \leq B$ but not $B \leq A$ (see page 22).

6. Cantor's Theorem

We still do not have any example of an infinite uncountable set. A classical "diagonal construction" invented by Cantor provides such an example.

Theorem 7 (Cantor). *The set* $2^{\mathbb{N}}$ *of all infinite sequences of zeros and ones is uncountable.*

Proof. Assume that this set is countable. Then all infinite sequences of zeros and ones can be numbered: $\alpha_0, \alpha_1, \ldots$.

Let $\alpha_0, \alpha_1, \ldots$ be the rows of the infinite two-dimensional table:

$$\begin{array}{ccllll}
\alpha_0 & = & \underline{\alpha_{00}} & \alpha_{01} & \alpha_{02} & \ldots \\
\alpha_1 & = & \alpha_{10} & \underline{\alpha_{11}} & \alpha_{12} & \ldots \\
\alpha_2 & = & \alpha_{20} & \alpha_{21} & \underline{\alpha_{22}} & \ldots \\
\multicolumn{6}{c}{\dotfill}
\end{array}$$

Here α_{ij} is the jth element of the ith sequence (α_i). Now consider the sequence on the diagonal: its elements are α_{00}, α_{11}, α_{22}, ...

(the ith element of the diagonal sequence is α_{ii}, and it equals the ith element of the ith sequence). Inverting the diagonal sequence, we get a sequence

$$\beta_i = 1 - \alpha_{ii}.$$

This sequence β differs from α_i (has a different ith term). Therefore β does not appear in the table: $\beta \neq \alpha_i$ for any i. But our assumption was that *all* infinite sequences of zeros and ones appear among α_i. We get a contradiction that shows that the set $2^{\mathbb{N}}$ of all sequences of zeros and ones is not countable. □

This theorem implies that the set $\mathbb{R}$ of real numbers is uncountable. (Indeed, we know that $\mathbb{R}$ has the same cardinality as the set of infinite binary fractions.) Therefore, $\mathbb{R}$ cannot coincide with the countable set of algebraic numbers. Therefore, there exists a real number which is not algebraic (is not a root of any nonzero polynomial with integer coefficients). The nonalgebraic numbers are also called *transcendental* numbers.

When Cantor created set theory, it was already known that transcendental numbers do exist. The first example of a transcendental number was constructed in 1844 by the French mathematician J. Liouville. He proved that if a real number α can be approximated by rational numbers with high precision (there exist approximations with denominator m and error that is very small compared to $1/m$), then α is not algebraic. For example, the number $\sum(1/10^{n!})$ satisfies Liouville's condition. Liouville's theorem is not very hard to prove; however, it requires some estimates of approximation errors, and Cantor's proof compared with Liouville's one looks like magic. Cantor's proof was published in 1874. This was the first paper devoted to set theory. Its first section proves that the set of algebraic numbers is countable, while the second proves that the set of real numbers is not countable. The general definition of the notion "have the same cardinality" was introduced three years later (together with the proof that spaces of different dimension have the same cardinality; see above).

Let us also note that in 1873 the French mathematician Charles Hermite has proved that the number e (base of natural logarithms) is not algebraic, and in 1882 the German mathematician F. Lindemann proved that π is transcendental and therefore cannot be constructed using a straight edge and a compass.

Our next few problems assume that the reader is familiar with the basic notions of calculus.

Problem 52. Prove that for any uncountable set $A \subset \mathbb{R}$ there exists a *condensation* point a such that any neighborhood of a has an uncountable intersection with A. (The statement remains true if we replace "uncountable set" by "set of the cardinality of the continuum".)

Problem 53. Prove that a closed set $A \subset \mathbb{R}$ that has no isolated points has the cardinality of the continuum.

Problem 54. Prove that any closed set $A \subset \mathbb{R}$ is either countable (or finite) or has the cardinality of the continuum. (*Hint*: Consider the subset $B \subset A$ whose elements are condensation points of A, i.e., points b such that every neighborhood of b has an uncountable intersection with A. If B is empty, then A is finite or countable. If B is not empty, then B is a closed set without isolated points and therefore has the cardinality of the continuum.)

This problem shows that the statement of the Continuum Hypothesis (CH) is true for all closed subsets. This hypothesis says that every subset of $\mathbb{R}$ is either countable (or finite) or has the cardinality of the continuum. Cantor proved CH for closed sets and regarded this proof as the first step towards the proof of CH in the general case, but this idea failed.

Problem 55. Let A be a countable set of points on the plane. Prove that the remaining part of the plane is connected: any two points outside A can be connected by a polygonal line (with two segments) that does not intersect A.

Let us revisit Cantor's diagonal construction. We know that the set of infinite sequences of zeros and ones has the same cardinality as the set of all subsets of $\mathbb{N}$. (Each subset X corresponds to its "characteristic sequence" that has ones at places that belong to X.) Therefore Cantor's Theorem can be reformulated as follows:

> *The set $\mathbb{N}$ cannot be put into one-to-one correspondence with the set of all its subsets.*

The proof can be translated into this language, too. Imagine that there exists a one-to-one correspondence $i \mapsto A_i$ between natural numbers and sets of natural numbers. The diagonal sequence now corresponds to the set D of all $i \in \mathbb{N}$ such that $i \in A_i$. The sequence

β that is not covered is the complement of D: $B = \{i \mid i \notin A_i\}$. It is clear that B differs from any A_i (at place i).

Now we are ready to generalize this argument and prove the following general form of Cantor's Theorem:

Theorem 8 (General form of Cantor's Theorem)**.** *For an arbitrary set X there is no one-to-one correspondence between X and the power set $P(X)$ (the set of all subsets of X).*

Proof. Let φ be a one-to-one correspondence between X and $P(X)$ (and $\varphi(x)$ is a set that corresponds to $x \in X$). Consider the set Z of all elements $x \in X$ that do not belong to the corresponding subset $\varphi(x)$:

$$Z = \{x \in X \mid x \notin \varphi(x)\}.$$

Let us prove that Z does not correspond to any element of X, i.e., that $Z \neq \varphi(z)$ for any $z \in X$. Indeed, assume that $Z = \varphi(z)$ for some z. Then

$$z \in Z \Leftrightarrow z \notin \varphi(z) \Leftrightarrow z \notin Z$$

(according to the definition of Z; recall that $\varphi(z) = Z$). This contradiction shows that the set Z does not correspond to any element z and φ is not a one-to-one correspondence. □

Note that Theorem 8 and its proof are still valid for the empty X (in this case $P(X)$ has one element).

Note also that any set X can be put into one-to-one correspondence with some subset of the set $P(X)$. Indeed, each element $x \in X$ corresponds to a singleton $\{x\}$. Therefore we can say that X has smaller cardinality than $P(X)$ (according to the definition on p. 22).

Problem 56. Prove that $n < 2^n$ for any natural $n = 0, 1, 2, \ldots$.

Theorem 8 has appeared in Cantor's paper dated 1890/91. Cantor considers functions with values 0 and 1 instead of subsets.

Now we have come close to the dangerous point where our intuition about sets becomes self-contradictory. Consider the "universal" set U that consists of all sets. Then all subsets of U are elements of U and therefore $P(U) \subset U$. This contradicts Cantor's Theorem.

We can unfold this argument and get the so-called *Russell's paradox*. Traditionally, Russell's paradox is explained as follows.

Normally a set is not its own element. For example, the set $\mathbb{N}$ of all natural numbers is not a natural number itself and therefore $\mathbb{N} \notin \mathbb{N}$. On the other hand, one may imagine a set that is its own element. For example, the set U of all sets is a set and therefore $U \in U$. Let us say that a set X is "normal" if X is not its own element, i.e., $X \notin X$. Now consider the set N of all normal sets. Is N normal or not? If N is normal, then N belongs to the set of all normal sets; $N \in N$ and therefore N is not normal. On the other hand, if N is not normal, then it does not belong to the set N of all normal sets and therefore is normal. How is that possible?

Another version of this paradox: an adjective is called "self-referential" if it has the property it describes. For example, the adjective "English" is self-referential while "Russian" is not. Now the question: is the adjective "non-self-referential" self-referential or not? (Any answer immediately leads to a contradiction.)

This reminds of the famous liar's paradox ("This statement is false") or a story about a barber who shaves every man in the village who does not shave himself. (Question: does the barber shave himself or not?) [*Note for American readers*: If the barber story looks like a paradox to you, this is only because you are a sexist. Of course, the barber is a woman.]

Trying to develop set theory, we need to isolate the problem and find out what was illegal in our arguments that have led to Russell's paradox. This is quite a nontrivial question, and was widely discussed during the first half of the 20th century. Here are some conclusions:

- The notion of a set is not intuitively clear. Different people (and different scientific traditions) may give different meanings to the word "set".
- Sets are too abstract objects to make the question "What is in fact true?" meaningful. Recall the Continuum Hypothesis (stated in Cantor's 1878 paper) saying that any infinite subset of $\mathbb{R}$ is either countable or has the cardinality of the continuum. Cantor claimed in his paper that CH can be

proven by some "new approach that uses induction and will be explained elsewhere", but in fact he never succeeded in proving CH. Later people started realizing that CH can be regarded both as true or false depending on our viewpoint. Accepting CH or its negation, we come to different theories, but neither of them is clearly preferable.

The situation is somewhat similar to non-Euclidean geometries. We may consider Euclid's fifth postulate (only one line parallel to a given line can pass through a given point) as true statement. The resulting geometry is called Euclidean. On the other hand, we may declare the negation of the fifth postulate to be an axiom. Then we get non-Euclidean geometry where there exist a point and two lines that both pass through this point and are parallel to a third line. This geometry was developed (among others) by Lobachevsky and is sometimes called Lobachevsky geometry.

Which geometry is the "right" one: Euclidean or non-Euclidean? This is a not a mathematical question, and we should ask physicists, not mathematicians. (And modern physicists will not answer this question either. Instead, they will explain to you that they use both geometries to construct models of the real world.)

The same is true for set theory: it is quite clear that neither mathematicians nor physicists should be asked whether CH is true or not, and only theology may provide the ultimate answer. (By the way, Cantor discussed questions related to set theory with professional theologists.)

- To avoid troubles while reasoning about sets, we have to be cautious. Some kinds of arguments are especially dangerous. The safety rules (that work, at least for now) are formulated in axiomatic set theory. There are several versions of axiomatic set theory. The most well known is ZF (named after Zermelo and Fraenkel). Adding the axiom of choice AC to ZF, we get a theory called ZFC.

We will not develop axiomatic set theory in this book. Instead, we give an informal description of restrictions that are useful to avoid paradoxes. It is not allowed to consider the set of all elements with some property because there are too many "potential candidates". The sets can be constructed only stepwise, using the sets already constructed. For example, for any set X it is allowed to consider the power set $P(X)$ that consists of all subsets of X (power set axiom). It is allowed to consider a subset of a given set that is formed by the elements that have some property (axiom of separation). It is allowed to consider the union of a given set of sets, i.e., for a given set X it is allowed to consider the set that consists of all elements of all elements of X (axiom of union). There are some other axioms.

We will continue our informal exposition but will try to warn the reader when we come close to a dangerous place, as we have done when we tried to define the cardinality of a set as the class of all sets having the same cardinality.

7. Functions

Up to now we have avoided formal definitions when speaking about functions, their arguments, values, composition of functions, etc. Now we give more formal definitions.

Let A and B be two sets. Consider the set of all ordered pairs $\langle a, b\rangle$ for all $a \in A$ and $b \in B$. This set is called the *Cartesian product* of A and B and is denoted by $A \times B$. (But what is an "ordered pair"? We return to this question later; see p. 34.)

Any subset $R \subset A \times B$ is called a *binary relation* between elements of A and B. Sometimes A coincides with B and we get a binary relation on A. For example, there is a binary relation "to be a divisor of" on the set $\mathbb{N}$ that can be denoted by "$|$". One may write that "$\langle 2, 6\rangle \in |$" (2 is a divisor of 6) and "$\langle 2, 7\rangle \notin |$" (2 is not a divisor of 7). However, the "infix" notation (like "$2|6$") is traditionally used in this case.

Problem 57. Are the relations "to be a divisor of" and "to be a multiple of" the same relation or are they different? (*Answer*: Of course, they are different, since an ordered pair $\langle a, b\rangle$ differs from the ordered pair $\langle b, a\rangle$.)

If a function f can be applied to elements of a set A and its values are elements of a set B, we may consider the relation between elements of A and B that consists of all pairs $\langle x, f(x)\rangle$ (for all $x \in X$). This relation can be called the *graph* of the function f. (This terminology is used in calculus where each function $f\colon \mathbb{R} \to \mathbb{R}$ has the graph that consists of all points on the plane having coordinates $\langle x, f(x)\rangle$.)

However, from the formal viewpoint it is easier to identify a function with its graph. We come to the following definition:

A relation $F \subset A \times B$ is called a (partial) *function from A to B* if F does not contain two pairs $\langle a, b_1\rangle$ and $\langle a, b_2\rangle$ with $b_1 \ne b_2$. In other terms, F is a (partial) function from A to B if for any $a \in A$ there exists at most one element $b \in B$ such that $\langle a, b\rangle \in F$.

The *domain* $\operatorname{Dom} F$ of the function F is the set of all $a \in A$ for which such b exists. For any $a \in \operatorname{Dom} F$ we may regard the *value of F at a* as the (only) element $b \in B$ such that $\langle a, b\rangle \in F$. This element is denoted by $F(a)$.

All values $F(a)$ for all $a \in \operatorname{Dom} F$ form a set that is called the *range* of the function F and is denoted by $\operatorname{Val} F$.

We say that F is *undefined* on a if $a \notin \operatorname{Dom} F$. Note that our definition does not require a function from A to B be defined on all elements of A; its domain may be any subset of A. (And its range $\operatorname{Val} F$ may be any subset of B.)

If F is defined on all elements of A, we write $f\colon A \to B$ and say that f is a *total* function defined on A.

Here is an example. An identity function $\mathrm{id}_A\colon A \to A$ has domain A and range A; it is a set of pairs $\langle a, a\rangle$ (for all $a \in A$), and $\mathrm{id}_A(a) = a$ for any $a \in A$. (The subscript A in id_A is sometimes omitted when A is clear from the context.)

The *composition* of two functions $f\colon A \to B$ and $g\colon B \to C$ is the function $h\colon A \to C$ such that $h(x) = g(f(x))$ for any $x \in A$. In other terms, h is a set of pairs

$$\{\langle a, c\rangle \mid \langle a, b\rangle \in f \text{ and } \langle b, c\rangle \in g \text{ for some } b \in B\}.$$

We denote the composition by $g \circ f$ (note that the function on the right of the $\circ$ sign is applied first).

The composition is an associative operation on functions, i.e., $h \circ (f \circ g) = (h \circ f) \circ g$; therefore we may omit parentheses when several functions are composed.

Let $f\colon A \to B$, and let B' be a subset of B. Consider the set of all $x \in A$ such that $f(x) \in B'$. This set is called the *preimage* of B' under f and is denoted by $f^{-1}(B')$:

$$f^{-1}(B') = \{x \in A \mid f(x) \in B'\}.$$

The *image* of a set $A' \subset A$ under f is the set of all values $f(a)$ for all $a \in A'$. Notation: $f(A')$. In other terms,

$$\begin{aligned} f(A') &= \{f(a) \mid a \in A'\} \\ &= \{b \in B \mid \langle a, b\rangle \in f \text{ for some } a \in A'\}. \end{aligned}$$

A pedantic reader will note that the notation $f(\dots)$ is used both for the image of a set and for the value of the function. However, the risk of misunderstanding is minimal and the meaning is clear from the context.

Problem 58. Which of the following equalities are true for any $f\colon A \to B$, $g\colon B \to C$, $A', A'' \subset A$, $B', B'' \subset B$, $C' \subset C$?

$$\begin{gathered} f(A' \cap A'') = f(A') \cap f(A''); \\ f(A' \cup A'') = f(A') \cup f(A''); \\ f(A' \setminus A'') = f(A') \setminus f(A''); \\ f^{-1}(B' \cap B'') = f^{-1}(B') \cap f^{-1}(B''); \\ f^{-1}(B' \cup B'') = f^{-1}(B') \cup f^{-1}(B''); \\ f^{-1}(B' \setminus B'') = f^{-1}(B') \setminus f^{-1}(B''); \\ f^{-1}(f(A')) \subset A'; \\ f^{-1}(f(A')) \supset A'; \\ f(f^{-1}(B')) \subset B'; \\ f(f^{-1}(B')) \supset B'; \\ (g \circ f)(A) = g(f(A)); \\ (g \circ f)^{-1}(C') = f^{-1}(g^{-1}(C')). \end{gathered}$$

Functions are also called *mappings*.

A function $f\colon A \to B$ is an *injection* if $f(a) \neq f(a')$ for any $a, a' \in A$ such that $a' \neq a$.

A function $f\colon A \to B$ is a *surjection* if its range coincides with B.

These two definitions are more similar than one may think, as the following problems show:

Problem 59. Prove that a function $f\colon A \to B$ is an injection if and only if f has a *left inverse*, i.e., there exists a function $g\colon B \to A$ such that $g \circ f = \mathrm{id}_A$. Prove that a function $f\colon A \to B$ is a surjection if and only if f has a *right inverse*, i.e., there exists a function $g\colon B \to A$ such that $f \circ g = \mathrm{id}_B$.

Problem 60. Prove that a function $f\colon A \to B$ is an injection if and only if the left cancellation property holds: $f \circ g_1 = f \circ g_2$ implies $g_1 = g_2$ for any two functions g_1, g_2 whose ranges are subsets of A. Prove that a function $f\colon A \to B$ is a surjection if and only if the right cancellation property holds: $g_1 \circ f = g_2 \circ f$ implies $g_1 = g_2$ for any two functions g_1, g_2 whose domains equal B.

A function $f\colon A \to B$ is a *bijection* or *one-to-one correspondence* if it is injective and surjective.

For each bijection $f\colon A \to B$ there exists an inverse function f^{-1} such that $f^{-1}(y) = x \Leftrightarrow f(x) = y$.

Problem 61. Let a function $f\colon A \to B$ has both the left inverse function $g_1\colon B \to A$ and the right inverse function $g_2\colon B \to A$. Is it possible that $g_1 \neq g_2$?

Recall that sets A and B have the same cardinality if there exists a bijection $f\colon A \to B$. What can be said about A and B if there is an injection $f\colon A \to B$? It is easy to see that f is a one-to-one correspondence between A and $f(A)$. Therefore, such an injection exists if and only if B has a subset that has the same cardinality as A, i.e., if the cardinality of A does not exceed the cardinality of B; see the definition given in Section 5 of this chapter.

A “dual” result is also true: a surjection $f\colon A \to B$ exists if and only if the cardinality of B does not exceed the cardinality of A.

Indeed, let $f\colon A \to B$ be a surjection. Then for each $b \in B$ there exists at least one element $a \in A$ such that $f(a) = b$. Selecting one element for each $b \in B$, we form a set $A' \subset A$ and get a one-to-one correspondence between A' and B. (Note that we again use the axiom of choice; see p. 10.)

On the other hand, if a subset $A' \subset A$ has the same cardinality as B and $g\colon A' \to B$ is a bijection, we can get a surjection $f\colon A \to B$ by extending g (so that $f(x) = g(x)$ for $x \in A'$; the values $f(x)$ for $x \notin A'$ may be chosen arbitrarily).

Problem 62. Before moving further, find an error in the argument presented in the previous paragraph.

In fact such an extension is possible only if B is not empty, and so the correct statement reads as follows: a surjection $f\colon A \to B$ exists if and only if B is not empty and the cardinality of B does not exceed the cardinality of A, or if both sets A and B are empty.

There is one more question that we have to discuss: what is an "ordered pair"? Informally speaking, we need a tool that combines two objects x and y into one composite object $\langle x, y\rangle$ in such a way that

$$\langle x_1, y_1\rangle = \langle x_2, y_2\rangle \;\Leftrightarrow\; x_1 = x_2 \text{ and } y_1 = y_2.$$

One may take the notion of an ordered pair as a basic notion and regard the equivalence above as an axiom about ordered pairs. However, more traditional approach uses a trick invented by the Polish mathematician Kuratowski. Recall that $\{x\}$ is a set whose only element is x (a singleton), and $\{x, y\}$ is a set that consists of x and y. (So $\{x, y\} = \{x\} = \{y\}$ if $x = y$.)

Theorem 9 (Ordered pairs)**.** *Let us define an ordered pair $\langle x, y\rangle$ as $\{\{x\}, \{x, y\}\}$. Then*

$$\langle x_1, y_1\rangle = \langle x_2, y_2\rangle \;\Leftrightarrow\; x_1 = x_2 \textit{ and } y_1 = y_2.$$

Proof. Assume that $\langle x_1, y_1\rangle = \langle x_2, y_2\rangle$. By definition this means that

$$\{\{x_1\}, \{x_1, y_1\}\} = \{\{x_2\}, \{x_2, y_2\}\}.$$

Now we have to consider all possibilities case by case (and be careful not to mix x and $\{x\}$).

A. Assume that $x_1 \neq y_1$. Then the set $\{x_1, y_1\}$ consists of two elements. This set belongs to the left-hand side; therefore it belongs to the right-hand side. Thus, it must be equal to either $\{x_2\}$ or

$\{x_2, y_2\}$. The first case is impossible (a two-element set cannot be equal to a singleton). Therefore, $\{x_1, y_1\} = \{x_2, y_2\}$. On the other hand, the singleton $\{x_1\}$ belongs to the left-hand side; therefore it belongs to the right-hand side and is equal to $\{x_2\}$ (since it cannot be equal to a two-element set). Hence $x_1 = x_2$ and $y_1 = y_2$.

B. The case $x_2 \neq y_2$ is similar.

C. Let $x_1 = y_1$ and $x_2 = y_2$. Then $\{x_1, y_1\} = \{x_1\}$, and so the left-hand side is $\{\{x_1\}\}$. For the same reason the right-hand side is $\{\{x_2\}\}$. Therefore, $x_1 = x_2$, and all four elements x_1, x_2, y_1, y_2 coincide. □

Note that a similar theorem is true for other definitions of the ordered pair, and there is nothing special about Kuratowski's definition: it is just a convenient trick.

Problem 63. Prove a statement similar to Theorem 9 for another definition of the ordered pair (suggested by Norbert Wiener):

$$\langle x, y\rangle = \{\{\varnothing, \{x\}\}, \{\{y\}\}\}.$$

8. Operations on cardinals

Let A and B be two finite sets; A consists of a elements and B consists of b elements. Then $A \times B$ consists of ab elements. This statement can be taken as a definition of multiplication: ab is the cardinality of the set $A \times B$, where $|A| = a$, $|B| = b$.

We can define addition of natural numbers in a similar way: $a+b$ is the cardinality of the set $A \cup B$, where $|A| = a$, $|B| = b$, and the sets A and B are disjoint.

These definitions allow us to extend addition and multiplication from natural numbers to any cardinal numbers (cardinalities of arbitrary sets). We define the *product* of cardinal numbers a and b as the cardinality of $A \times B$, where A has cardinality a and B has cardinality b. We define the *sum* of two cardinal numbers a and b as the cardinality of $A \cup B$, where A and B are disjoint sets of cardinalities a and b.

Remarks. 1. As we have explained, the use of cardinals (cardinal numbers) as separate entities requires some caution. A safer approach allows only statements like "the cardinality of the set X is the product of the cardinalities of the sets A and B" (which means that there exists a one-to-one correspondence between X and $A \times B$) but does not allow us to talk about cardinal numbers.

A careful reader can verify that all statements about cardinal numbers that appear in our book can be translated into this language in a straightforward way.

2. A pedantic reader would also mention that we have to prove that the cardinality of $A \times B$ does not depend on the choice of specific sets A and B: if $A \simeq A'$ and $B \simeq B'$, then $A \times B \simeq A' \times B'$. (This is indeed evident.) A similar remark can be made about addition.

3. An even more pedantic reader would also mention that we have to prove that for given A and B we can always find disjoint sets A' and B' that have the same cardinalities as A and B. This is also evident: let (say) $A' = A \times \{0\}$ and $B' = B \times \{1\}$.

Now we define *exponentiation* for cardinals. Let A and B be finite sets that have a and b elements. We have to find a set that can be naturally defined in terms of A and B and consists of a^b elements.

The natural choice is the set of all functions $f\colon B \to A$ (functions that have domain B and whose values are elements of A). This set is denoted by A^B.

Let us explain why $|A^B| = a^b$ (for finite sets A and B). Indeed, to specify a function $f \in A^B$ we have to specify its value at each element $x \in B$. For each $x \in B$ we have a different possibilities, and choices for different x are independent; therefore we have $a \cdot a \cdots a$ (b times) possibilities.

Therefore, we can define a^b as the cardinality of the set A^B, where A and B are sets of cardinalities a and b. This definition is consistent with the standard definition of a^b for natural numbers a and b.

Problem 64. What is 0^0 according to our definition? (*Answer*: 1.)

Example. Let 2 denote a set that consists of two elements, e.g., $\{0,1\}$. What is $2^{\mathbb{N}}$? According to our definition, this is a set of all functions $f\colon \mathbb{N} \to \{0,1\}$. These functions are infinite sequences of zeros and ones (a function is a sequence $f(0)f(1)f(2)\ldots$). There exists a natural one-to-one correspondence between 2^X and $P(X)$ (we have seen it for the special case $X = \mathbb{N}$, but the same construction works for any X).

Standard properties of addition and multiplication (commutative, associative and distributive laws) are true for the operations on cardinals:

$$\begin{aligned} a+b &= b+a; \\ a+(b+c) &= (a+b)+c; \\ a\times b &= b\times a; \\ a\times(b\times c) &= (a\times b)\times c; \\ (a+b)\times c &= (a\times c)+(b\times c). \end{aligned}$$

These laws can be stated without using cardinals as separate entities. For example, $a\times b = b\times a$ means that there exists a one-to-one correspondence between $A\times B$ and $B\times A$ (indeed, $\langle x,y\rangle \mapsto \langle y,x\rangle$ can be used). Other properties are also easy to prove.

Somewhat more work is needed for laws that involve exponentiation:

$$\begin{aligned} a^{b+c} &= a^b \times a^c; \\ (ab)^c &= a^c \times b^c; \\ (a^b)^c &= a^{b\times c}. \end{aligned}$$

Let us prove the first one. What is A^{B+C} for disjoint B and C? Elements of A^{B+C} are functions of type $B\cup C \to A$. Such a function consists of two parts: its restriction to B (that is obtained if we forget about arguments in C) and its restriction on C. Thus, for each element of A^{B+C} we get a pair of functions, the first belonging to A^B, the second belonging to A^C. This mapping is a one-to-one correspondence between A^{B+C} and $A^B \times A^C$.

A correspondence between $(A\times B)^C$ and $A^C\times B^C$ is also often used. For example, an element of $(\mathbb{R}\times\mathbb{R})^{\mathbb{R}}$ is a mapping of type $\mathbb{R}\to\mathbb{R}\times\mathbb{R}$, i.e., a curve $t\mapsto z(t) = \langle x(t), y(t)\rangle$ on the coordinate plane. Such a curve is determined by a pair of functions $x, y\colon \mathbb{R}\to\mathbb{R}$.

A correspondence between $(A^B)^C$ and $A^{B\times C}$ is used less frequently. An element $f\in A^{B\times C}$ is a mapping of type $B\times C\to A$, i.e., a function of two arguments. The first argument belongs to B, and the second to C. If we fix the second argument, we get a function $f_c\colon B \to A$, defined as $f_c(b) = f(b,c)$ (to be formal, we should write $f(\langle b,c\rangle)$ instead of $f(b,c)$). The mapping $c \mapsto f_c$ belongs to $(A^B)^C$ and corresponds to an element $f \in A^{B\times C}$. (A similar construction is used in algebra when we regard a polynomial in two variables x, y as a polynomial in one variable x whose coefficients are polynomials in y; the ring $\mathbb{Z}[x,y]$ is isomorphic to $(\mathbb{Z}[y])[x]$.)

Cardinality of countable sets is denoted by $\aleph_0$. The continuum cardinality (the cardinality of $\mathbb{R}$ or the set of infinite sequences of zeros and ones) is denoted by $\mathfrak{c}$. By definition, $\mathfrak{c} = 2^{\aleph_0}$.

A curious reader would ask: what does subscript 0 in $\aleph_0$ mean? What is, say, $\aleph_1$? Usually $\aleph_1$ means the minimal uncountable (infinite) cardinal. (As we will see later, it does exist.) The continuum hypothesis (see p. 28) says that $\mathfrak{c} = \aleph_1$.

Now we can write known properties of countable sets as identities:

- $\aleph_0 + n = \aleph_0$ for finite n (the union of a finite set and a countable set is countable);
- $\aleph_0 + \aleph_0 = \aleph_0$ (the union of two countable sets is countable);
- $\aleph_0 \times \aleph_0 = \aleph_0$ (the union of a countable family of countable sets is countable).

These identities can be combined to get other theorems about cardinalities. For example,

$$\mathfrak{c} \times \mathfrak{c} = 2^{\aleph_0} \times 2^{\aleph_0} = 2^{\aleph_0+\aleph_0} = 2^{\aleph_0} = \mathfrak{c}$$

means that real line and coordinate plane have the same cardinality.

In a similar way,

$$\mathfrak{c}^{\aleph_0} = (2^{\aleph_0})^{\aleph_0} = 2^{\aleph_0\times\aleph_0} = 2^{\aleph_0} = \mathfrak{c}.$$

(The set of infinite sequences of real numbers has the same cardinality as the set of real numbers.)

Problem 65. Explain the following computation:

$$\mathfrak{c} + \mathfrak{c} = 1 \times \mathfrak{c} + 1 \times \mathfrak{c} = 2 \times \mathfrak{c} = 2^1 \times 2^{\aleph_0} = 2^{1+\aleph_0} = 2^{\aleph_0} = \mathfrak{c}.$$

Problem 66. Prove that $\aleph_0 \times \mathfrak{c} = \mathfrak{c}$.

These properties of cardinals are useful together with the Cantor–Bernstein Theorem. For example, note that

$$\mathfrak{c} = 2^{\aleph_0} \le {\aleph_0}^{\aleph_0} \le \mathfrak{c}^{\aleph_0} = \mathfrak{c};$$

therefore ${\aleph_0}^{\aleph_0} = \mathfrak{c}$ (the set of all infinite sequences of natural numbers has the cardinality of the continuum).

Problem 67. The latter computation has implicitly used the monotonicity of exponentiation ($a_1 \le a_2$ implies $a_1^b \le a_2^b$). Prove this property and similar properties for other operations (they are quite evident, though).

Problem 68. Construct an explicit one-to-one correspondence between the infinite sequences of natural numbers and the irrational numbers in the interval $(0, 1)$ using continuous fractions, i.e., fractions of type $1/(n_0 + 1/(n_1 + 1/(n_2 + \cdots)))$.

Problem 69. Prove that $\aleph_0^{\mathfrak{c}} = 2^{\mathfrak{c}}$. (Cantor's Theorem says that this cardinal is greater than $\mathfrak{c}$.)

Problem 70. Find the cardinality of the set of all continuous functions of type $\mathbb{R} \to \mathbb{R}$. Does it change if we omit the continuity requirement?

Problem 71. Find the cardinality of the set of all monotone functions of type $\mathbb{R} \to \mathbb{R}$.

Problem 72. A family of subsets of $\mathbb{N}$ has the following property: any two elements of the family have finite intersection. Can this family have the cardinality of the continuum? The same question if any two elements of this family have finite symmetric difference.

We shall see later that $a \times b = a + b = \max(a, b)$ for any infinite cardinals a and b, but the proof requires transfinite induction. Therefore, solving Problems 47 and 48 we had to use a special construction to prove that $a + b = \mathfrak{c}$ implies $a = \mathfrak{c}$ or $b = \mathfrak{c}$. The following theorem generalizes this trick:

Theorem 10. *Assume that*

$$A_1 \times A_2 \times \cdots \times A_n = B_1 \cup B_2 \cup \cdots \cup B_n.$$

Then there exists i such that the cardinality of A_i does not exceed the cardinality of B_i.

Proof. Consider the projection of $B_i \subset A_1 \times \cdots \times A_n$ onto A_i for $i = 1, \ldots, n$, i.e., the set of all elements of A_i that appear as ith elements of tuples in B_i. If this projection coincides with A_i (for some i), the theorem is proved. If not, let $x_i \in A_i$ be a point that is not covered. The n-tuple $\langle x_1, \ldots, x_n \rangle$ does not belong to any B_i, but this is impossible because of our assumption. □

The statement of Theorem 10 (it is sometimes called *Koenig's Theorem*) involves the Cartesian product of n sets. It can be defined inductively (e.g., $A \times B \times C$ consists of triples $\langle a, b, c \rangle$ that can be identified with pairs $\langle \langle a, b \rangle, c \rangle$). This approach does not allow us to define the Cartesian product of a countable family of sets. However, we can overcome this difficulty and define $A_0 \times A_1 \times A_2 \times \cdots$ (countably many factors) as the set of all sequences $a_0, a_1, a_2, \ldots$ such that $a_i \in A_i$ (the set of all functions of type $\mathbb{N} \to A_0 \cup A_1 \cup A_2 \cup \cdots$ such that $a(i) \in A_i$ for all $i \in \mathbb{N}$). This definition allows us to extend Koenig's Theorem to countable (or even uncountable) products.

Reformulation of Koenig's Theorem: If $b_i < a_i$ for all $i = 0, 1, \ldots$ (here a_i and b_i are cardinal numbers), then

$$b_0 + b_1 + b_2 + \cdots < a_0 \times a_1 \times a_2 \times \cdots .$$

Recalling that $\mathfrak{c} \times \mathfrak{c} \times \cdots$ (countably many factors) is $\mathfrak{c}^{\aleph_0} = \mathfrak{c}$, we get the following corollary of Koenig's Theorem: If a set of continuum cardinality is a union of a countable family of sets, then at least one of these sets has continuum cardinality.

Problem 73. Give a detailed proof of the last statement.

Problem 74. Let $a_0, a_1, a_2, \ldots$ be cardinals such that $a_i \geq 2$ for all i. Prove that

$$a_0 + a_1 + a_2 + \cdots \leq a_0 \times a_1 \times a_2 \times \cdots .$$

Problem 75. Let $m_0 < m_1 < m_2 < \cdots$ be an increasing sequence of cardinals. Prove that the sum $m_0 + m_1 + m_2 + \cdots$ differs from $a^{\aleph_0}$ for any cardinal a.

Chapter 2

Ordered Sets

1. Equivalence relations and orderings

Recall that a *binary relation* on a set X is defined simply as a subset $R \subset X \times X$; usually one writes xRy instead of $\langle x, y\rangle \in R$.

A binary relation R on X is called an *equivalence relation* if it possesses the following properties:

- (reflexivity) xRx for all $x \in X$.
- (symmetry) $xRy \Rightarrow yRx$ for all $x, y \in X$.
- (transitivity) xRy and $yRz \Rightarrow xRz$ for all $x, y, z \in X$.

The following obvious statement is used frequently:

Theorem 11. **(a)** *If a set X is split into a union of disjoint subsets, then the relation "to be in the same subset" is an equivalence relation.*

(b) *Any equivalence relation can be obtained in such a way.*

Proof. The first statement is absolutely obvious. We prove the second to show how to make use of all the properties mentioned in the definition. Indeed, let R be an equivalence relation. For any element $x \in X$ consider its *equivalence class*, the set of all $y \in X$ such that xRy.

Let us prove that for any two elements x_1, x_2 their equivalence classes are either disjoint or coincide. Let z be a common element of the two classes. Then x_1Rz and x_2Rz, and therefore zRx_2 (by symmetry) and x_1Rx_2 (by transitivity), and also x_2Rx_1 (by symmetry again). Thus for any z the statement x_1Rz implies x_2Rz (by transitivity), and vice versa.

It remains to note that by reflexivity any element x belongs to its own class, which means that the set X is indeed split into disjoint equivalence classes. □

Problem 76. Show that the symmetry and transitivity requirements can be replaced by the following single requirement: xRz and $yRz \Rightarrow xRy$ (reflexivity is still needed).

Problem 77. How many equivalence relations are there on the set $\{1, 2, 3, 4, 5\}$?

Problem 78. There are two equivalence relations $\sim_1$ and $\sim_2$ defined on the set M, having n_1 and n_2 equivalence classes, respectively. Is their intersection $x \sim y \Leftrightarrow [(x \sim_1 y)$ and $(x \sim_2 y)]$ an equivalence relation? How many classes can it have? What can be said about the union of the relations?

Problem 79. (Ramsey Theorem.) The set of all k-element subsets of some infinite set A is split into l classes (k, l are positive integers). Prove that there exists an infinite set $B \subset A$ such that all its k-element subsets belong to the same class.

(It is evident for $k = 1$: if an infinite set is split into a finite number of classes, then one of these classes is infinite. For $k = 2$ and $l = 2$ the statement can be reformulated as follows: given an infinite set of people, one can choose either infinitely many people such that any two of them are acquainted or infinitely many people such that no two of them are acquainted. A finite version of this statement claims that of every six people there are either three acquaintances or three people unacquainted with one another; this is a well-known puzzle.)

The set of all equivalence classes is called the *quotient set* of the set X modulo the equivalence relation R. (If the relation respects an additional structure on X, then one obtains quotient groups, quotient rings, etc.)

We will come across equivalence relations many more times, but our main topic now is orderings (order relations).

A binary relation $\le$ on a set X is called a *partial order* if it has the following properties:

- (reflexivity) $x \le x$ for all $x \in X$;
- (antisymmetry) $x \le y$ and $y \le x \Rightarrow x = y$ for all $x, y \in X$;
- (transitivity) $x \le y$ and $y \le z \Rightarrow x \le z$ for all $x, y, z \in X$.

(Following the tradition, we use the symbol "$\le$" rather than a letter, to denote an ordering relation.) The set with a partial order defined on it is called a *partially ordered set*, or just *poset.*

Two elements x, y of a partially ordered set are called *comparable* if either $x \le y$ or $y \le x$. Notice that the definition of a partially ordered set does not require that any two elements are comparable. Such a requirement added, we obtain the definition of the *linear* order (called also *total* order). A set with a linear (= total) order is called *linearly* (or *totally*) *ordered set.*

Here are some examples of posets:

- Any subset of $\mathbb{R}$ with the usual $\le$-relation is a linearly ordered set.
- Consider the following (*coordinatewise*) order on the set $\mathbb{R}\times\mathbb{R}$ of pairs of real numbers: $(x_1, x_2) \le (y_1, y_2)$ if $x_1 \le y_1$ and $x_2 \le y_2$. This order is not linear (for example, the pairs $(1, 0)$ and $(0, 1)$ are not comparable).
- Consider the following (*pointwise*) order on the set of all functions of type $\mathbb{R} \to \mathbb{R}$: $f \le g$ if $f(x) \le g(x)$ for all $x \in \mathbb{R}$. This order is not linear.
- Consider the following order on the set of positive integers: $x \le y$ if x is a divisor of y. This order is not linear either.
- A relation "any prime factor of the number x is a factor of the number y" is not an order on the set of positive integers: it is reflexive and transitive, but not antisymmetric.
- Let U be an arbitrary set. Then the subset relation $\subset$ is a partial order on the set $P(U)$ of all subsets of U.

- There exists a traditional alphabetical order of letters of the English alphabet: (a $\le$ b $\le$ c $\le \cdots \le$ z). This order is linear — for any two letters it is known which one comes first (consult a dictionary if necessary).
- There is a *lexicographical order* (the one used in dictionaries) on the set of all English words. Its formal definition looks like that: if a word x is a prefix of a word y, then $x \le y$ (e.g., fact $\le$ factor). If neither word is a prefix of the other, then find the leftmost position in which the letters in two words are different and look which letter goes first in the alphabetical order. This ordering is also linear (otherwise dictionary makers and users would have been in trouble).
- The equality relation $((x \le y) \Leftrightarrow (x = y))$ is also a partial order, such that no two different elements are comparable.
- An almost real-life example: suppose X is a set of boxes. We say that $x \le y$ if the box x can be put inside the box y (or if x and y are the same box). Depending on the set of boxes available, this order may be linear or not.

Let x, y be elements of a partially ordered set X. One says that $x < y$ if $x \le y$ and $x \ne y$. The <-relation has the following properties:

$$x \not< x;$$
$$(x < y) \text{ and } (y < z) \Rightarrow x < z.$$

(The first property is evident. Prove the second: assume that $x < y$ and $y < z$, that is, $x \le y$, $x \ne y$, $y \le z$ and $y \ne z$. Then $x \le z$ by transitivity. If $x = z$ then $x \le y \le x$, and therefore $x = y$ by antisymmetry, contrary to our assumptions.)

Note the usual word usage: we read the sign "$\le$" as "less or equal", and the sign "<" as "less (than)", assuming implicitly that $x \le y$ if and only if either $x < y$ or $x = y$. Luckily, this is indeed true. Another note: expression $x > y$ (x is greater than y) means that $y < x$, and expression $x \ge y$ (x is greater than or equal to y) means that $y \le x$.

Problem 80. Explain why it is a bad idea to read $x \le y$ as "x is not greater than y".

Some authors define a partial ordering as a relation $<$ having the two properties mentioned above. In this case the relation $x \le y \Leftrightarrow [(x < y) \text{ or } (x = y)]$ is a partial ordering according to our definition.

Problem 81. Check the latter assertion.

To avoid confusion, a relation $<$ is sometimes called a *strict ordering*, while the relation $\le$ is called a *nonstrict ordering*. There are different ways to define a partially ordered set: it is possible to start with a definition of a nonstrict ordering $\le$ (reflexive, antisymmetric and transitive) and to derive a strict ordering $<$ from it (as we have done), and it is possible to go in the other direction.

Problem 82. Omitting the property of antisymmetry in the definition of a partial order, one obtains a definition of a *preorder*. Prove that any preorder can be described as follows: the set is split into disjoint classes such that $x \le y$ for any two elements x, y of the same class, and the quotient set bears a partial order that defines the result of comparison for two elements of different classes.

Here are some operations on posets.

- Let Y be a subset of a partially ordered set $(X, \le)$. Then a partial order on the set Y is defined as follows:

  $$(\le_Y) = (\le) \cap (Y \times Y).$$

  This order is called the *induced* order. If X is linearly ordered, then the induced order on Y is also linear.

- Let X and Y be two disjoint posets. Then one can define a partial order on their union as follows: an element of X is (by definition) less than any element of Y, and two elements of the same set are compared as before. This partially ordered set is denoted by $X + Y$ and called the *sum* of posets X and Y. The ordering on $X + Y$ is linear if both X and Y are linearly ordered.

  The same notation is used for nondisjoint (or even equal) sets. For example, we can define an ordered set $\mathbb{N} + \mathbb{N}$ as follows. Take two disjoint copies of the set of natural numbers, $\{0, 1, \dots\}$ and $\{\bar{0}, \bar{1}, \dots\}$, and consider the set $\{0, 1, \dots, \bar{0}, \bar{1}, \dots\}$, where $k \le \bar{l}$ for any k and l, and inside either copy the ordering is the usual one.

- Let $(X, \leq_X)$ and $(Y, \leq_Y)$ be two partially ordered sets. There are several ways to define an ordering on their product $X \times Y$. One can assume that $(x_1, y_1) \leq (x_2, y_2)$ if $x_1 \leq_X x_2$ and $y_1 \leq_Y y_2$ (*componentwise* ordering). This ordering is not linear even if both the original orderings were: if the first component is greater for one pair and the second for the other pair, the pairs are not comparable. To obtain a linear ordering, we may agree that some coordinate is "principal", and compare first these coordinates, and only afterwards (as a tie-break) the second ones. If the X-coordinate is "principal" then $(x_1, y_1) \leq (x_2, y_2)$ if either $x_1 <_X x_2$ or $x_1 = x_2$ and $y_1 \leq_Y y_2$. For technical reasons, though, it is more convenient to take the second coordinate as "principal". Speaking about a linear ordering on the product of two linearly ordered sets, we will always assume this latter ordering (the second coordinates are compared first).

Problem 83. Prove that a partially ordered set $\mathbb{N} \times \mathbb{N}$ (with componentwise ordering) does not contain an infinite subset X such that any two distinct elements of X are noncomparable. Is a similar statement true for $\mathbb{Z} \times \mathbb{Z}$?

Problem 84. Prove a similar statement for $\mathbb{N}^k$ with the componentwise ordering.

Problem 85. Let U be a finite set of n elements. Consider the set $P(U)$ of all the subsets of U partially ordered by inclusion. What is the maximum possible cardinality of a set $S \subset P(U)$ such that the ordering induced on S is linear? What is the maximum possible cardinality if no two distinct elements of S are comparable? (*Hint*: See Problem 14.)

Problem 86. How many linear orderings are there on a set of n elements?

Problem 87. Prove that any partial ordering on a finite set can be extended to a linear ordering ("extension" means that if $x \leq y$ in the original ordering, then it is true also in the new ordering).

Problem 88. Let X be an infinite partially ordered set. Prove that it either contains an infinite subset whose elements are pairwise incomparable or it contains an infinite subset such that the induced ordering on it is linear (or both).

Problem 89. (A finite version of the previous problem.) Let m and n be positive integers. Prove that any partially ordered set of cardinality $mn+1$ contains either $m+1$ pairwise incomparable elements or $n+1$ pairwise comparable elements.

Problem 90. A sequence consists of $mn+1$ numbers. Prove that it is possible to remove some of them so that remaining elements form either an increasing sequence of length $m+1$ or a decreasing sequence of length $n+1$. (*Hint*: Use the previous problem.)

Problem 91. Consider the set of all subsets of the set of nonnegative integers, ordered by inclusion. Does it contain a linearly ordered subset of continuum cardinality? Does it contain a subset of continuum cardinality such that no two of its elements are comparable?

An element of a poset X is called the *greatest* element of X if it is greater than any other element, and a *maximal* element if no greater element exist. If the set is not linearly ordered, these two notions are different: the greatest element is automatically maximal, but the converse is not necessarily true. (The box capable of containing any other box is not the same as the box that does not fit anywhere.)

The *least* element and a *minimal* elements are defined in a similar way.

It is easy to see that only one greatest element may exist in a given partially ordered set, while there may be several maximal elements.

Problem 92. Prove that no two maximal elements are comparable. Prove that in a finite poset X for any element x there exists a maximal element y greater than or equal to x.

2. Isomorphisms

Two partially ordered sets are called *isomorphic* if there exists an *isomorphism*, that is, a one-to-one correspondence between them respecting the order. (In particular, isomorphic sets have the same cardinality.) Let us say it again: a bijection $f : A \to B$ is an isomorphism of posets A and B if

$$a_1 \leq a_2 \Leftrightarrow f(a_1) \leq f(a_2)$$

for any elements $a_1, a_2 \in A$ (the sign "$\leq$" on the left means the ordering in the set A, and that on the right, in the set B).

It is clear that the isomorphism relation is reflexive (any poset is isomorphic to itself), symmetric (if X is isomorphic to Y, then Y is isomorphic to X) and transitive (two sets isomorphic to the third one are isomorphic). Thus, all partially ordered sets are split into classes of isomorphic sets, called *order types*. (Like with cardinalities, one has to be cautious here: there are too many isomorphic sets, and therefore it is not safe to speak about order types as sets.)

Theorem 12. *Finite linearly ordered sets containing equal number of elements are isomorphic.*

Proof. A finite linearly ordered set must contain the least element. Indeed, take any element; if it is not the least one, take a smaller one; if this one is not the least one either, take yet a smaller one, etc. We obtain a decreasing sequence $x > y > z > \cdots$, which must terminate somewhere.

Assign the number 1 to the least element, remove it from the set and take the least element among the rest; assign the number 2 to it, and so on. It is easy to see that the ordering of elements is the same as the ordering of numbers, that is, our set is isomorphic to the set $\{1, 2, \ldots, n\}$. □

Problem 93. Prove that the set of all positive divisors of the number 30 ordered by the relation "to be a divisor of" is isomorphic to the set of all subsets of the set $\{a, b, c\}$ ordered by inclusion.

Problem 94. Consider the set of all almost-zero sequences of non-negative integers, that is, the set of sequences such that all but a finite number of their elements are equal to zero. Introduce the component-wise ordering in this set: $(a_0, a_1, \ldots) \leq (b_0, b_1, \ldots)$ if $a_i \leq b_i$ for all i. Prove that this partially ordered set is isomorphic to the partially ordered set of all positive integers with the ordering "to be a divisor of".

If a bijection $f\colon A \to A$ (where A is a poset) is an isomorphism, f is frequently called an *automorphism* of A. The identity mapping is always an automorphism, but some partially ordered sets have more automorphisms. For example, the function $x \mapsto x+1$ is an automorphism of the ordered set $\mathbb{Z}$ of integers (with a usual ordering). The same formula does not define an automorphism of the set of positive integers because in this case the mapping is not one-to-one.

Problem 95. Show that the ordered set $\mathbb{N}$ of nonnegative integers has only one automorphism (the identity mapping).

Problem 96. Consider the set $P(A)$ of all the subsets of some k-element set A, partially ordered by inclusion. Find the number of automorphisms of $P(A)$.

Problem 97. Prove that the set of positive integers with partial order "to be a divisor of" has many automorphisms: the set of all automorphisms of this set has continuum cardinality.

Here are some examples of linearly ordered sets that have the same cardinality but are not isomorphic (by Theorem 12 they are all infinite).

- The interval $[0, 1]$ (with the usual ordering) is not isomorphic to the set $\mathbb{R}$ because the interval has the greatest element, but $\mathbb{R}$ does not. (Evidently, an isomorphism maps the greatest element to the greatest element.)

- The set $\mathbb{Z}$ (of integers with the usual ordering) is not isomorphic to the set $\mathbb{Q}$ (of rational numbers). Indeed, let $\alpha : \mathbb{Z} \to \mathbb{Q}$ be an isomorphism. Take two adjacent integers, say, 2 and 3. The isomorphism α would map them to two rational numbers $\alpha(2)$ and $\alpha(3)$ such that $\alpha(2) < \alpha(3)$ (because $2 < 3$). Then rational numbers between $\alpha(2)$ and $\alpha(3)$ should be images (under α) of some integers between 2 and 3—but there are no such integers.

- A more complicated example of nonisomorphic sets is $\mathbb{Z}$ and $\mathbb{Z}+\mathbb{Z}$. Take two copies of zero in $\mathbb{Z}+\mathbb{Z}$; as usual, denote them 0 and $\bar{0}$, with $0 < \bar{0}$. An isomorphism $\mathbb{Z}+\mathbb{Z} \to \mathbb{Z}$ would map them to some integers, a and b, such that $a < b$. Then all the elements between 0 and $\bar{0}$ (there are infinitely many of them, $1, 2, 3, \ldots, -\bar{3}, -\bar{2}, -\bar{1}$) should be mapped to integers between a and b, but there is only a finite number of them.

The nature of this example is different from the previous one because it is impossible to express the difference between the ordered sets $\mathbb{Z}$ and $\mathbb{Z}+\mathbb{Z}$ by a formula (which would be true in one set but false in another). One says that the ordered sets $\mathbb{Z}$ and $\mathbb{Z}+\mathbb{Z}$ are "elementary equivalent".

Problem 98. Prove that the linearly ordered sets $\mathbb{Z}\times\mathbb{N}$ and $\mathbb{Z}\times\mathbb{Z}$ (with the ordering described on page 46) are not isomorphic.

Problem 99. Are the linearly ordered sets $\mathbb{N}\times\mathbb{Z}$ and $\mathbb{Z}\times\mathbb{Z}$ isomorphic?

Problem 100. Are the linearly ordered sets $\mathbb{Q}\times\mathbb{Z}$ and $\mathbb{Q}\times\mathbb{N}$ isomorphic?

The mapping $x \mapsto \sqrt{2}x$ is an isomorphism between the intervals $(0,1)$ and $(0,\sqrt{2})$. However, it does not define an isomorphism between the sets of rational points of these intervals (that is, between $\mathbb{Q}\cap(0,1)$ and $\mathbb{Q}\cap(0,\sqrt{2})$) because multiplication by $\sqrt{2}$ maps rational numbers to irrational ones. Nevertheless, it is possible to construct an isomorphism between these sets. To do this, take increasing sequences of rational numbers $0 < x_1 < x_2 < \cdots$ and $0 < y_1 < y_2 < \cdots$ convergent to 1 and to $\sqrt{2}$, respectively, and consider a piecewise linear function f mapping x_i to y_i and linear on each interval $[x_i, x_{i+1}]$ (see Figure 5). It is easy to see that the function f is a desired isomorphism.

Problem 101. Show that the set of rational points of the interval $(0,1)$ is isomorphic to the set $\mathbb{Q}$. (*Hint*: Here it is also possible to construct a piecewise linear automorphism. This problem has, however, another solution: note that the mapping $x \mapsto 1/x$ sends rational numbers to rational numbers.)

The next problem requires a more sophisticated construction (perhaps the simplest solution is to use Theorem 13):

Problem 102. Prove that the set of binary-rational points of the interval $(0,1)$ is isomorphic to the set $\mathbb{Q}$. (The number is binary-rational if it has the form $m/2^n$ where m is an integer and n is a positive integer.)

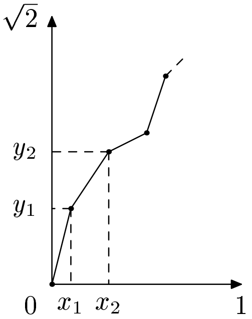


Figure 5. The piecewise linear isomorphism.

Two elements x, y of a linearly ordered set are called *adjacent* if $x < y$ and there are no elements between them, that is, there is no z such that $x < z < y$. A linearly ordered set is called *dense* if it has no adjacent elements (that is, there exists an element between any two of its elements).

Theorem 13. *Any two countable dense linearly ordered sets without the greatest and the least element are isomorphic.*

Proof. Let X and Y be two sets satisfying the hypothesis. Construct an isomorphism between them step by step. After n steps we would have two n-element subsets $X_n \subset X$ and $Y_n \subset Y$ (we call their elements "used") and an order-preserving one-to-one correspondence between used elements. A step consists in taking an unused element of one of the sets (say, X) and comparing it with all the used elements of X. It may be greater than all of them, smaller than all of them, or may get between two of them, say the ith and $(i+1)$th greatest elements. In all the cases we can find an unused element of Y being in the same position with respect to the elements of the set Y_n (greater than all, smaller than all, or between the ith and the $(i+1)$th greatest elements). Here we are using the hypothesis that Y has no greatest element, no least element, and no adjacent elements. After that, we add the chosen elements to X_n and Y_n and put them into correspondence.

To arrive finally to an isomorphism between the sets X and Y we must make sure that all the elements of both sets would be used at some step. This can be done as follows: since both sets are countable, enumerate their elements and take the unused element with the least number, from X in the odd-numbered steps and from Y in the even-numbered steps. This completes the proof. □

Problem 103. How many different (nonisomorphic) countable dense linearly ordered sets (if nothing is known about the greatest and the least element) are there? (*Answer*: Four.)

Problem 104. Give an example of two dense linearly ordered sets of continuum cardinality not isomorphic to each other. (*Hint*: Try $\mathbb{Q}+\mathbb{R}$ and $\mathbb{R}+\mathbb{Q}$.)

Theorem 14. *Any countable linearly ordered set is isomorphic to a subset of the set* $\mathbb{Q}$.

Note that we could replace $\mathbb{Q}$ by any countable dense set without the greatest and the least element because they are all isomorphic.

Proof. The proof of this theorem is similar to that of Theorem 13, but one must take new unused elements from one side only (from the given set), and look for their counterparts in the set of rational numbers.

Another proof: note that any countable linearly ordered set X is isomorphic to a subset of the $\mathbb{Q}\times X$, and the set $\mathbb{Q}\times X$ is a dense set without the least and the greatest element (and therefore is isomorphic to $\mathbb{Q}$ according to Theorem 13). □

3. Well-founded orderings

One of the possible forms of the Induction Principle reads as follows:

> *Let $A(n)$ be some property of a nonnegative integer n. Assume that we are able to prove $A(n)$ under the assumption that $A(m)$ is true for all $m<n$. Then $A(n)$ is true for all nonnegative integers.*

(Note that we suppose that $A(0)$ can be proved without any assumptions because no smaller nonnegative integers exist.)

What are partially ordered sets such that a similar principle is true for them? The following simple theorem gives an answer:

Theorem 15. *For a partially ordered set X the following three properties are equivalent:*

(a) *Any nonempty subset of the set X has a minimal element.*

(b) *There is no infinite strictly decreasing sequence $x_0 > x_1 > x_2 > \cdots$ of elements of the set X.*

(c) *The following induction principle is true for X: if (for every $x \in X$) a property $A(y)$ for all $y < x$ implies $A(x)$, then $A(x)$ holds for all $x \in X$. Formally it is written like this:*

$$\forall x(\forall y((y < x) \Rightarrow A(y)) \Rightarrow A(x)) \Rightarrow \forall x A(x).$$

Proof. Let us prove first that properties (a) and (b) are equivalent. If $x_0 > x_1 > x_2 > \cdots$ is an infinite decreasing sequence, then the set of its values has no minimal element. Thus, (a) implies (b). Conversely, if B is a nonempty set without a minimal element, then it is possible to construct an infinite decreasing sequence as follows. Take an arbitrary element $b_0 \in B$. By assumption, it is not minimal, so one can find $b_1 \in B$ such that $b_0 > b_1$. For the same reason, there exists $b_2 \in B$ such that $b_1 > b_2$, etc. The elements $b_0, b_1, \ldots$ form an infinite decreasing sequence.

Now derive the induction principle from the existence of a minimal element for any set. Let $A(x)$ be an arbitrary property which is not true for some element of the set X. Consider a nonempty set B of all the elements not possessing the property A. Let x be a minimal element of the set B. By assumption, B contains no smaller elements, and therefore all the elements $y < x$ possess property A. By the hypothesis, $A(x)$ must be true—a contradiction.

Now let us use the induction princliple to prove that any subset contains a minimal element. Let B be a subset without a minimal element. We prove by induction that B is empty. Take the property $A(x)$ to be $x \notin B$. If $A(y)$ is true for all $y < x$, then B contains no element less than x. Therefore, if x were an element of B (i.e., if $A(x)$ were false), then x would be a minimal element of B, but B has no such elements. □

The orderings with properties (a)–(c) are called *well-founded.* What examples of well-founded orderings do we know? First, it is our initial example, the set of nonnegative integers. Another example is the set $\mathbb{N} \times \mathbb{N}$ of pairs of nonnegative integers (the pair having a smaller second term is smaller; for breaking ties, compare the first terms). Indeed, let us check condition (b). It is convenient to reformulate it as follows: any sequence $u_0 \geq u_1 \geq u_2 \geq \cdots$ of elements of the set stabilizes (i.e., starting from some term, all its terms are equal); evidently, this is an equivalent formulation.

Let a sequence of pairs

$$(x_0, y_0) \geq (x_1, y_1) \geq (x_2, y_2) \geq \cdots$$

be given. By the definition of the ordering (second terms are compared first), we have $y_0 \geq y_1 \geq y_2 \geq \cdots$, and therefore the sequence y_i of nonnegative integers stabilizes. After this, the sequence x_i must be nonincreasing, and therefore it stabilizes, too.

The same argument is applicable to a more general situation.

Theorem 16. *Let A and B be two well-founded partially ordered sets. Then their product $A \times B$ with the following ordering:*

$$(a_1, b_1) \leq (a_2, b_2) \Leftrightarrow [(b_1 < b_2) \text{ or } (b_1 = b_2 \text{ and } a_1 \leq a_2)]$$

is well founded.

Proof. The second terms of the sequence $(a_0, b_0) \geq (a_1, b_1) \geq \cdots$ stabilize; then the first terms stabilize, too. □

This theorem implies a similar statement for $\mathbb{N} \times \mathbb{N} \times \mathbb{N}$, for $\mathbb{N}^k$ and, generally, for a product of finitely many well-founded posets.

It is even simpler to prove that the sum $A+B$ of two well-founded partially ordered sets A and B is well founded: the sequence $x_0 \geq x_1 \geq x_2 \geq \cdots$ is either contained in B (and then we use the well-foundedness of B) or contains an element of A. In the latter case, all the subsequent elements are elements of A, and we use the well-foundedness of A.

Writing a program (or solving a problem at a math competition), one often has to prove that some process cannot continue infinitely long. For example, having written a loop, we often want to be sure that the program will eventually leave it. To do this we can, for example, introduce some nonnegative integer parameter and ensure it decreases in every step. Then, if this parameter initially equals N, the number of iterations is at most N.

There are situations, though, when one cannot estimate the number of iterations but can nevertheless guarantee that the loop cannot be executed forever. This happens when there is a parameter taking its values in some well-founded set and decreasing at every iteration.

Here is an example of a problem using this idea:

Dr. Faust signed a contract with Mephistopheles: every day he gives Mephistopheles a coin, and in exchange gets any set of coins he wishes, but all the coins must be of a lesser value (coins come in a finite number of denominations). Dr. Faust is not allowed to change (or earn) money anywhere else, but can spend money as he wishes. Dr. Faust loses when he has no more coins left. Prove that Mephistopheles will eventually win, whatever set of coins Dr. Faust initially had.

Solution: Suppose there are k denominations of coins. The required parameter is defined as follows: count the coins of every type Dr. Faust has (n_1 being the number of the smallest coins, n_2, next smallest, and so on till n_k). Note that after each exchange the array $(n_1, \ldots, n_k)$ decreases with respect to the usual ordering: first compare the kth (last) terms, then the $(k-1)$th terms, and so on. The ordered set $\mathbb{N}^k$ is well founded, and therefore the process should stop.

Problem 105. A finite sequence of zeros and ones is given. Consider the following operation: a substring "01" can be replaced by the substring "100 … 00" (any number of zeros is allowed). Prove that such an operation can be carried out only finitely many times.

Problem 106. Consider the set of all finite strings of English letters with lexicographical order (see page 44). Is this ordered set well founded?

Problem 107. Consider the set A of all nonincreasing "almost-zero" sequences of nonnegative integers ("almost-zero" means that all but a finite number of terms are zero). Introduce the following ordering:

first compare the first terms; if they are equal, compare the second terms, etc. Prove that this set is well founded.

Problem 108. Consider the set of all polynomials in one variable x whose coefficients are nonnegative integers. Order it as follows: a polynomial P is greater than a polynomial Q if $P(x) > Q(x)$ for all sufficiently large x. Prove that this condition defines a linear ordering, and that the ordering obtained is well founded.

4. Well-ordered sets

A well-founded linearly ordered set is called a *well-ordered set.* For linear orderings the notions of minimal and least elements coincide, so in a well-ordered set every nonempty subset has the least element.

Note that if in a poset any nonempty subset has the least element, this poset is linearly ordered. Indeed, any two-element subset has the least element, and therefore any two elements are comparable.

Examples of well-ordered sets: $\mathbb{N}$, $\mathbb{N} + k$ (here k means a finite linearly ordered set of k elements), $\mathbb{N}+\mathbb{N}$, $\mathbb{N}\times\mathbb{N}$ (with linear ordering as defined above).

Our goal is to understand the structure of well-ordered sets. Let us start with several simple remarks:

- A well-ordered set has the least element. (Follows immediately from the definition.)
- For any element x of a well-ordered set (except the greatest one) there is a successor y (this means that $y > x$ but there is no z such that $y > z > x$). Indeed, if the set of all elements that are greater than x is nonempty, then this set has the least element y, and this is the one we need. It is natural to denote the successor of x by $x+1$, the next element by $x+2$, etc.
- It is possible that an element of a well-ordered set has no predecessor, i.e., it is not a successor of any element. For example, in the set $\mathbb{N} + \mathbb{N}$ there are two elements without predecessors—the least element and the least element of the

second copy of the set $\mathbb{N}$. Elements that have no predecessor are called *limit* elements.

- Any element of a well-ordered set has the form $x+n$, where x is a limit element, and n is a nonnegative integer (the meaning of the notation $x+n$ is explained above). Indeed, if x is not a limit element, then take its predecessor. If it is not a limit element either, we take its predecessor, etc., until we reach a limit element (the process cannot continue infinitely because the set is well ordered). It is clear that such a representation is unique (because an element cannot have more than one predecessor).

- Any subset of a well-ordered set bounded from above has the least upper bound. (As usual, the subset X of a partially ordered set A is called *bounded from above* if it has an *upper bound*, that is, an element $a \in A$ such that $x \leq a$ for all $x \in X$. The *least upper bound* is the least element, if it exists, in the set of all upper bounds.)

  Indeed, the set of all upper bounds is nonempty and therefore has the least element.

  Note that the symmetric question about the greatest lower bound (the greatest element in the set of all lower bounds) is trivial for a well-ordered set, because every subset has the least element.

Let A be a well-ordered set. Denote its least element by 0, the next element, 1, the next, 2, etc. If the set is finite, the process will terminate. If the set is infinite, then look whether we counted all the elements of the set A. If not, take the minimal element among the rest and denote it by ω. Denote the next element (if it exists) by $\omega+1$, then $\omega+2$, etc. If the process is not finished even now, then take the least element among the rest, call it $\omega \cdot 2$ and repeat the whole procedure. Then we will arrive at $\omega \cdot 3$, $\omega \cdot 4$, and so on. If there are still unnamed elements, the minimal unnamed element will be denoted ω^2. Then one obtains ω^2+1, ω^2+2, $\ldots$, $\omega^2+\omega$, $\ldots$, $\omega^2+\omega \cdot 2$, $\ldots$, $\omega^2 \cdot 2$, $\ldots$, $\omega^2 \cdot 3$, $\ldots$, ω^3, $\ldots$ (do not worry about the notation: it will be explained later).

What can one learn from these arguments? Let us try to formulate some statements. The following definition will be useful here: if a linearly ordered set A is split into two disjoint parts B and C such that any element of B is less than any element of C, then B is called an *initial segment* of the set A. In other words, a subset B of a linearly ordered set A is called an initial segment of A if any element of B is less than any element of $A \setminus B$. Yet another reformulation: $B \subset A$ is an initial segment if $a, b \in A, b \in B$ and $a \le b$ imply $a \in B$. Note that an initial segment may be empty or may coincide with the entire set A.

Here are some simple properties of initial segments:

- An initial segment of a well-ordered set (like any other subset, though) is a well-ordered set.
- An initial segment of an initial segment is an initial segment of the original set.
- A union of any family of initial segments (of some ordered set) is an initial segment (of the same set).
- If x is an arbitrary element of a well-ordered set A, then the sets $[0, x)$ (all the elements of A that are less than x) and $[0, x]$ (all the elements of A that are less than or equal to x) are both initial segments of A.
- Any initial segment I of a well-ordered set A is either equal to the entire set A or has the form $[0, x)$ for some $x \in A$. (Indeed, if $I \ne A$, then take the least element x of the set $A \setminus I$. Then all smaller elements belong to I, element x does not belong to I, and no element greater than x belongs to I, or else we would come to a contradiction with the definition of an initial segment.)
- Any two initial segments of a well-ordered set A are comparable with respect to inclusion, that is, one of them is a subset of the other. (Follows from the previous property.)
- Initial segments of a well-ordered set A, ordered by inclusion, form a well-ordered set. This set contains the greatest element (A); the remaining elements form an ordered set isomorphic to A. (Indeed, initial segments not equal to A all

have the form $[0, x)$, and the correspondence $[0, x) \leftrightarrow x$ is an isomorphism.)

Recall our discussion above. Its first part shows that if a well-ordered set A is infinite, then it has an initial segment isomorphic to ω. (We use notation ω instead of $\mathbb{N}$ to stress that we regard the set of natural numbers as an ordered set.)

The second part of the argument proves the following: A is either isomorphic to an initial segment of the set ω^2 or has an initial segment isomorphic to ω^2. (Here ω^2 means the well-ordered set of pairs of nonnegative integers: second terms of pairs are compared first; if they are equal, first terms are used.)

More generally, the following statement is true: for any two well-ordered sets one is isomorphic to an initial segment of the other. The proof more or less repeats the arguments above; to do it correctly, though, we need some preparation.

5. Transfinite induction

The terms "induction" and "recursion" are often used interchangeably. For example, the factorial $n! = 1 \cdot 2 \cdot 3 \cdots n$ can be defined as a function $f(n)$ such that $f(n) = n \cdot f(n-1)$ for $n > 0$, and $f(0) = 1$; this definition is often called both "recursive" and "inductive". We will try to distinguish between these words in the following way: if one *proves* something first for $n = 0$ and then for $n = 1, 2, \ldots$ so that every statement uses the previous one, then this is induction. If something is *defined* first for $n = 0$ and then for $n = 1, 2, \ldots$ so that definition for every n uses the values defined earlier, then this is recursion.

We want to consider inductive proofs and recursive definitions not only for natural numbers but also for other well-ordered sets.

We have already discussed inductive proofs when we spoke about well-founded sets (see Section 3 of Chapter 2); let us now give just one more example:

Theorem 17. *Let A be a well-ordered set, and $f : A \to A$ an increasing mapping (that is, $f(x) < f(y)$ if $x < y$). Then $f(x) \geq x$ for all $x \in A$.*

Proof. By the induction principle (Theorem 15, page 53) it is enough to prove inequality $f(x) \geq x$ assuming that $f(y) \geq y$ for all $y < x$. Suppose this is not true, and $f(x) < x$. Since f is increasing, $f(f(x)) < f(x)$. On the other hand, the element $y = f(x)$ is less than x, and therefore by the induction hypothesis $f(y) \geq y$, that is, $f(f(x)) \geq f(x)$.

One can use the existence of the least element directly and reformulate the above arguments as follows. Suppose the statement is not true. Take the least x such that $f(x) < x$. Then $f(f(x)) < f(x)$ because f is increasing, and therefore x is not the least, contrary to the assumption.

One more way to reformulate this proof is like this: if $x > f(x)$ then

$$x > f(x) > f(f(x)) > f(f(f(x))) > \cdots$$

because f is increasing, but there are no infinite strictly decreasing sequences in a well-founded set. □

Now we come to recursion. In the definition of the factorial the value $f(n)$ was expressed in terms of $f(n-1)$. In a more general situation the definition of $f(n)$ may refer to values of the function for several smaller values of the argument. One can, for example, define the function $f : \mathbb{N} \to \mathbb{N}$ by saying that $f(n)$ is one plus the sum of all previous values, that is

$$f(n) = f(0) + f(1) + \cdots + f(n-1) + 1.$$

This is a legal recursive definition (one should only clarify that an empty sum is taken to be zero, so that $f(0) = 1$).

Problem 109. What function f is given by this definition?

How is it possible to extend this scheme to arbitrary well-ordered sets? Let A be a well-ordered set. We want to give a recursive definition of some function $f : A \to B$ (here B is some set). Such a definition should relate the value $f(x)$ of the function f on an element $x \in A$ to its values $f(y)$ for all $y < x$. In other words, a recursive definition defines $f(x)$ under the assumption that the restriction of the function f to the initial segment $[0, x)$ is known. Here is an exact statement:

Theorem 18. *Let A be a well-ordered set, and B an arbitrary set. Let a recursive rule be given, that is, a mapping F whose arguments are an element $x \in A$ and a function $g : [0, x) \to B$, and whose value is an element of B. Then there exists exactly one function $f \colon A \to B$ such that*

$$f(x) = F(x, f|_{[0,x)})$$

for all $x \in A$. (*Here $f|_{[0,x)}$ means the restriction of the function f to the initial segment $[0, x)$ — we forget the values of the function f for all the arguments greater than or equal to x.*)

Proof. Informally one can argue as follows: the value of the function f on the least element is defined uniquely (restriction $f|_{[0,0)}$ is empty). Then the value of the function f on the next element is also defined uniquely because values of f on the preceding elements (more exactly, the only one preceding element) are already known, etc.

However, one has to express all this formally. It is done as follows: prove by induction the following statement about an arbitrary element $a \in A$:

> *There exists exactly one mapping f of the segment $[0, a]$ to the set B such that the recursive definition given above is true for every $x \in [0, a]$.*

We call a mapping $f : [0, a] \to B$ *sound* if it possesses the above property, i.e., if $f(x) = F(x, f|_{[0,x)})$ for each $x \le a$. Thus we are to prove that for any $a \in A$ there is a unique sound mapping of the segment $[0, a]$ to B.

We are reasoning by induction, and therefore we can assume that this statement is true for all $c < a$, that is, there exists a unique sound mapping $f_c : [0, c] \to B$ (soundness of f_c means that for all $d \le c$ the value $f_c(d)$ is the one prescribed by the recursive rule).

Consider the mappings f_{c_1} and f_{c_2} for two different c_1 and c_2. Suppose, for example, that $c_1 < c_2$, that is, the mapping f_{c_2} is defined on a larger segment $[0, c_2]$. The restriction of f_{c_2} to the smaller segment $[0, c_1]$ coincides with f_{c_1} because the restriction of a sound mapping to a smaller segment is evidently sound, and we have assumed uniqueness for the segment $[0, c_2]$.

Thus, all the mappings f_c are compatible, that is, any two of them take equal values on any element on which both are defined. Combining all the mappings f_c we obtain some mapping h defined in $[0,a)$. Applying the recursive rule to a and h, we get some value $b \in B$. Let $h(a) = b$. Then a mapping $h\colon [0,a] \to B$ is defined; it is easy to see that it is sound.

To finish the induction step, it is necessary to check that the sound mapping defined on $[0,a]$ is unique. Indeed, its restrictions to segments $[0,c]$ with $c < a$ must coincide with f_c, so that one needs only to check uniqueness at the point a. This is guaranteed by the recursive rule (expressing the value at the point a in terms of the preceding values). Thus an inductive proof is complete.

Note that the sound mappings of the segments $[0,a]$ for different a are compatible (restriction of a sound mapping to a smaller segment is sound, and we use the uniqueness), and therefore, they define a function $f\colon A \to B$ satisfying the recursive definition.

The existence is proved; uniqueness is clear, because restriction of the function to any segment $[0,a]$ is sound, and therefore defined uniquely. □

We will use this theorem to prove that for any two well-ordered sets one is always an initial segment of the other. Before doing that, we will have to improve the statement somehow. We must take into account the situation when the recursive rule is not everywhere defined. For example, define the sequence of real numbers by the relation $x_n = \tan x_{n-1}$ with the initial value $x_0 = a$. For some values of a the sequence may be finite because the tangent is not defined for the corresponding value of the argument.

Problem 110. Prove that the set of such "exceptional" values of a (where the sequence is finite) is countable.

A similar situation may occur in the general case:

Theorem 19. *Let the mapping F mentioned in Theorem 18 be partial* (*that is, F may be undefined for some x and some functions $g : [0,x) \to B$*). *Then there exists a function f such that*

- *either f is defined on the entire set A and satisfies the recursive definition,*
- *or f is defined on some initial segment $[0, a)$, satisfies the recursive definition on it, and the recursive rule is not applicable (i.e., the mapping F is not defined) for the point a and the function f.*

Proof. This theorem is a generalization and, at the same time, a corollary of Theorem 18. Indeed, add a special element $\perp$ ("undefined") to B and modify the recursive rule: the new rule gives $\perp$ every time the old rule was undefined. (If $\perp$ was among the values of the function for smaller values of the argument, then the new rule gives $\perp$ too.)

Use Theorem 18 for the modified rule to obtain some function f'. If this function never assumes the value $\perp$, then the first possibility takes place (with $f = f'$). If the function f' equals $\perp$ at some point, then it is equal to $\perp$ at all greater points. Construct a new function f equal to f' if f' is not equal to $\perp$, and undefined otherwise. The domain of f is some initial segment $[0, a)$, and the second possibility takes place. □

Problem 111. Prove the uniqueness for the function defined by a partial recursive rule. (You must formulate the exact statement first.)

Now we are ready to prove the theorem on comparison of well-ordered sets.

Theorem 20. *Let A and B be two well-ordered sets. Then either A is isomorphic to an initial segment of B, or B is isomorphic to an initial segment of A.*

Proof. Note first that an initial segment may coincide with the entire set, so the case when the sets A and B are isomorphic is also covered by the theorem.

Define the mapping f from A to B by the following recursive rule:

> *For $a \in A$, $f(a)$ is the least element of B not encountered among $f(a')$ for $a' < a$.*

This rule is not defined in case the values $f(a')$ for $a' < a$ constitute the entire set B. Using Theorem 19, one obtains the function f defined by this rule. Consider now two cases:

- The function f is defined on the entire set A. The recursive definition ensures monotonicity of f because $f(a)$ is defined as the least yet unused element; the larger a is, the fewer unused elements remain, and the least such element can only grow (the definition implies also that no equal values are possible). It remains to prove that the image of the function f is an initial segment of the set B. Indeed, let $b < f(a)$ for some $a \in A$; let us check that b is also a value of the function f. Indeed, by the recursive definition, $f(a)$ is the least yet unused value, so b is used, that is, equal to $f(a')$ for some $a' < a$.
- The function f is defined on some initial segment $[0, a)$ only. In this case this segment is isomorphic to B, and f is the required isomorphism. Indeed, since $f(a)$ is not defined, all the elements of B are values of the function f. On the other hand, f preserves the ordering (as we have seen).

Thus, the theorem is proved in both cases. □

Is it possible that A is isomorphic to an initial segment of B, and B is isomorphic to an initial segment of A? The answer is no, except for the trivial case when these initial segments are the sets A and B themselves. This is a consequence of the following statement:

Theorem 21. *No well-ordered set is isomorphic to its own initial segment* (*other than the entire set*).

Proof. Suppose a well-ordered set A is isomorphic to its own segment other than the entire set. As we have proved (see page 58), this segment has the form $[0, a)$ for some element $a \in A$. Let $f\colon A \to [0, a)$ be an isomorphism. The mapping f is strictly increasing, and by Theorem 17 the inequality $f(a) \ge a$ holds. This contradicts the fact that the set of values of the function f is $[0, a)$. □

If a set A is isomorphic to an initial segment of a set B, and the set B is isomorphic to an initial segment of the set A, then the composition of these isomorphisms gives an isomorphism between the set A and its initial segment (an initial segment of an initial segment is an initial segment). This initial segment must coincide with the entire set A, so that the sets A and B are isomorphic.

All this allows us to compare well-ordered sets. If A is isomorphic to an initial segment of B other than the entire B, then we say that the *order type of A is less than the order type of B*. If the sets A and B are isomorphic, we say that they have the same order type. If B is isomorphic to an initial segment of A (other than the entire A), then it we say that the *order type of A is greater than the order type of B*. Thus, we have just proved the following statement:

Theorem 22. *For any two well-ordered sets A and B exactly one of the three cases described above takes place.*

Forgetting for a moment about foundational problems of set theory, we can define the *order type* of a well-ordered set A as the class of well-ordered sets isomorphic to A. So, we have just defined a linear ordering on the class of order types of well-ordered sets. These order types are called *ordinal numbers*, or *ordinals*. This ordering is in fact a well-ordering. We reformulate this statement in a safe way, without using classes of sets:

Theorem 23. *Each nonempty family of well-ordered sets has the "least element", that is, a set isomorphic to initial segments of all other sets in the family.*

Proof. Take any set X in the family. If X is the least one, the theorem is proved. If not, consider all the sets of the family that are less than X, that is, are isomorphic to initial segments of the set X of the type $[0, x)$. Take the least element among all such x. The corresponding set will be the least one. □

It follows from the above theorems that the cardinalities of any two well-ordered sets are comparable (one set has the same cardinality as some subset of the other one). We will see now that any set can be well ordered (this important statement is called Zermelo's Theorem), and therefore cardinalities of any two sets are comparable.

6. Zermelo's Theorem

Theorem 24. *Any set can be well ordered.*

Proof. Proof of this theorem relies heavily on the axiom of choice; this proof attracted much criticism for the lack of constructivity. For countable sets it is easy to find a well-ordering (copy it from $\mathbb{N}$). But even for the set $\mathbb{R}$ of real numbers one cannot point out any specific well-ordering. We prove (using the axiom of choice) its existence but we cannot give a specific example of a well-ordering on $\mathbb{R}$.

Let us explain what form of the axiom of choice we are using. Let A be a given set. We admit that there exists a function φ defined on all subsets of the set A except A itself, such that for any subset it singles out an element of A outside this subset:

$$X \subsetneq A \Rightarrow \varphi(X) \in A \setminus X.$$

Having fixed this function, one can construct a well-ordering on A without any additional ambiguity. This is done as follows.

We define the element $a_0 = \varphi(\varnothing)$ to be the least element of the set A. The next element is $a_1 = \varphi(\{a_0\})$; it is different from a_0 by construction. The next element is $a_2 = \varphi(\{a_0, a_1\})$. If the set A is infinite, then this process can be continued to yield a sequence $\{a_0, a_1, \dots\}$ of elements of the set A. If some elements are still left, consider an element $a_\omega = \varphi(\{a_0, a_1, \dots\})$. Continue so until the set A is exhausted. When it is exhausted, the order in which its elements were chosen defines a well-ordering on A.

Surely, the last phrase should be made more precise—what does it mean "continue until exhausted"? We may wish to use transfinite recursion (the situation is quite similar: the next element is defined recursively if all the previous elements are known). This is indeed possible if some other well-ordered set B is given—then we would define a function $\varphi\colon B \to A$ by transfinite recursion and obtain a one-to-one correspondence either between A and a part of B or between B and a part of A. Everything is fine in the first case, but to make sure that this case occurs we need a well-ordered set B of cardinality not less than that of A—a dead end.

We can break it through like this: consider all the pieces of the future ordering and ensure they can be glued together.

Let $(S, \leq_S)$ be a subset of the set A with an ordering on it. We say that $(S, \leq_S)$ is a *sound fragment* if it is well ordered and

$$s = \varphi([0, s))$$

for each $s \in S$. Here $[0, s)$ is an initial fragment of the set S with respect to the ordering $\leq_S$.

Thus, the set $\{\varphi(\varnothing)\}$ is a sound fragment (we do not need to specify an ordering because there is only one element). The set $\{\varphi(\varnothing), \varphi(\{\varphi(\varnothing)\})\}$ (the first element is less than the second) is also a sound fragment. We may continue with this construction but we have to take an infinite (and having a very large cardinality) number of steps.

The plan is as follows: prove that any two sound fragments are compatible, and then consider the union of all sound fragments. It will be a sound fragment that coincides with the entire set A (or else one would expand it to obtain a sound fragment not taken into the union).

Lemma 1. *Let $(S, \leq_S)$ and $(T, \leq_T)$ be two sound fragments. Then one of them is an initial segment of the other, and the orderings are compatible ($\leq_S$ and $\leq_T$ coincide when both are defined).*

Note that by Theorem 20 one of the fragments is *isomorphic* to an initial segment of the other one. Let S be isomorphic to an initial segment of T, and let $h : S \to T$ be an isomorphism between S and some initial segment of T. Lemma 1 claims that the isomorphism h is in fact the identity mapping, that is, $h(x) = x$ for all $x \in S$. Let us prove this by induction on $x \in S$ (it is legal because S is well-ordered by the definition of the sound fragment). Induction hypothesis states that $h(y) = y$ for all $y < x$. We must prove that $h(x) = x$. Consider initial segments $[0, x)_S$ and $[0, h(x))_T$ (in $(S, \leq_S)$ and $(T, \leq_T)$, respectively). They are mapped to one another by the isomorphism h and therefore, by the induction hypothesis, coincide as sets. By the definition of the sound fragment $x = \varphi([0,x))$ and $h(x) = \varphi([0,h(x)))$, so that $x = h(x)$. Lemma 1 is proved. $\square$

Consider now the union of all sound fragments (i.e., the corresponding sets). On this union, a linear order is naturally defined: for any two elements there is a fragment containing them both (for each element there is its own fragment; take the bigger one), so it is possible to compare them. By Lemma 1 the ordering does not depend on the choice of a fragment.

Lemma 2. *This union is a sound fragment.*

To prove Lemma 2 we note that the union carries a linear order. In fact, this is a well-ordering. For a change, let us prove this using decreasing sequences. Let $x_0 \geq x_1 \geq \cdots$; take a sound fragment F containing x_0. Lemma 1 implies that all the x_i also belong to the fragment (because F is an initial segment in any larger fragment). F is well-ordered by definition, and therefore the sequence stabilizes. Lemma 2 is proved. $\square$

Lemma 2 can be reformulated as follows: there exists the largest sound fragment. One has to prove that this fragment (denote it by S) coincides with the entire set A. If $S \neq A$, then take the element $a = \varphi(S)$, which does not belong to S, and add it to S assuming it to be greater than all the elements of S. We obtain the ordered set S' (the sum of S and the one-element set); it is well ordered. The soundness condition is also satisfied (for a, by definition, and for all the other elements, because it is satisfied in S). Thus, we have obtained a larger sound fragment, which contradicts the maximality of S. This finishes the proof of Zermelo's Theorem. $\square$

As we have already mentioned, Zermelo's Theorem and Theorem 20 on the comparison of well-ordered sets imply the following statement:

Theorem 25. *For any two sets, one can be put into one-to-one correspondence with a subset of the other one.*

The notion of a well-ordered set was introduced by Cantor in his paper written in 1883. His other paper written in 1895–1897 contains a proof that any two well-ordered sets are comparable (one is isomorphic to an initial segment of the other one).

Several works by Cantor contain claims that it is possible to introduce a well-ordering on any set and compare cardinalities of any two sets (Theorems 24 and 25), but he never gave a clear proof of both claims. The first proof was given only in 1904 by the German mathematician E. Zermelo.

7. Transfinite induction and Hamel basis

The notion of a well-ordered set allows us to extend inductive arguments and recursive definitions applying them to sets of an arbitrary cardinality. The following linear algebra example shows how this can be done.

Any linearly independent set of vectors in a finite-dimensional linear space can be extended to a basis. Why? Let S be a finite set of linearly independent vectors. If S is not a basis, then there exists some x_0 that is not a linear combination of S-elements. Then we can add x_0 to S to get a bigger linearly independent set $S' = S \cup \{x_0\}$. If S' is not a basis, then there exists x_1 that is not a linear combination of S'-elements and can be added to S', etc. Finally we either get a basis or an infinite sequence of linearly independent vectors, and the latter case is impossible if the linear space has finite dimension.

Now we want to extend this argument to any linear space (not necessarily finite-dimensional) using transfinite induction.

Let V be an arbitrary linear space. A subset $S \subset V$ is called *linearly independent* if any nontrivial linear combination of its elements is nonzero. (Note that by linear combination we mean a *finite* linear combination: the sum of an infinite sequence is not defined unless some topology is fixed.) A linearly independent set S is called a *Hamel basis* of V if any element of V is a (finite) linear combination of elements of S.

If a linearly independent set S is maximal (i.e., it becomes dependent after we add any element), then S is a Hamel basis. (This can be proved in the same way as for finite-dimensional spaces.)

Theorem 26. *Any linearly independent set $S \subset V$ in any linear space V can be extended to a Hamel basis $S' \supset S$.*

Proof. Let S be a linearly independent subset of a linear space V. Consider a well-ordered set I whose cardinality is large enough (is greater than the cardinality of V). We define a (partial) function $f\colon I \to V$ by transfinite recursion:

> $f(i)$ = *an element of* V *that is not a linear combination of the elements of* S *and of the values* $f(j)$ *for* $j < i$.

This rule does not define $f(i)$ if such an element does not exist. (Note that we use the axiom of choice. Using it, we construct a function Φ that maps any subset $X \subset V$ to some element $v \in V$ that is not a linear combination of elements of X, assuming that such v exists. Then Φ is used in the recursive definition for f. However, the axiom of choice is needed anyway for Zermelo's Theorem.)

This definition guarantees that f is injective. Moreover, the union of S and the range of mapping f is a linearly independent set. Indeed, suppose a (finite) linear combination of elements of S and values of f equals zero. Without loss of generality we assume that all coefficients are nonzero. This combination involves $f(i)$ for some values of i. Consider the maximal value i_0 encountered. By definition, $f(i_0)$ is not a linear combination of S and earlier $f(i)$, so we come to a contradiction.

By assumption the set I has greater cardinality than V. Therefore the function f is defined on some proper initial segment $[0, i)$ of I, and the value $f(i)$ is undefined (see Theorem 19). According to the definition of f this means that all the elements of V are linear combinations of S-elements and values $f(x)$ for $x \in [0, i)$. As we have seen, these elements are linearly independent, hence we get a basis for E. □

We could also have avoided the use of an auxiliary set I of large cardinality by introducing a well-ordering on V. At each step we consider some element $v \in V$; if it is not a linear combination of current basis elements, then we add v to the basis; otherwise the basis remains unchanged.

Problem 112. Provide missing details in this proof.

Hamel bases can be used to construct some "pathological" examples.

Theorem 27. *There exists a (total) function $f\colon \mathbb{R} \to \mathbb{R}$ such that $f(x+y) = f(x) + f(y)$ for all x, y but f is not linear (i.e., f differs from the function $x \mapsto cx$ for each $c \in \mathbb{R}$).*

Proof. We regard $\mathbb{R}$ as a linear space over $\mathbb{Q}$ and consider its Hamel basis. (This basis has continuum cardinality, but this does not matter.) Let α be an element of the basis. Consider a function f that maps each real number x (i.e., each element of the linear space $\mathbb{R}$ over $\mathbb{Q}$) to its α-coordinate (so that $x = f(x)\alpha + \cdots$). This function is linear as a mapping of the $\mathbb{Q}$-linear space S to $\mathbb{Q}$; therefore $f(x+y) = f(x) + f(y)$ for any $x, y \in \mathbb{R}$. The function f is nonzero ($f(\alpha) = 1$) and its values are rational numbers; therefore f differs from cx for any c. □

Problem 113. Prove that any function f that satisfies the hypotheses of Theorem 27 is not bounded (even if restricted to an interval) and, moreover, its graph is dense in $\mathbb{R}^2$.

Theorem 28. *The additive groups $\mathbb{R}$ and $\mathbb{R} \oplus \mathbb{R}$ are isomorphic.*

Proof. Again we regard $\mathbb{R}$ as a vector space over $\mathbb{Q}$ and find a Hamel basis in it. Evidently, this basis is infinite (otherwise $\mathbb{R}$ would have been countable). Now consider a basis in $\mathbb{R} \oplus \mathbb{R}$ made (in a natural way) of two copies of the basis in $\mathbb{R}$. As we will see later (Section 9), for any infinite set B the set $B + B$ (made of two disjoint copies of B) has the same cardinality as B. It remains to note that two linear spaces over the same field that have the same dimension (i.e., Hamel bases of the same cardinality) are isomorphic. Therefore, the underlying additive groups are isomorphic, too. □

Problem 114. Prove that any basis in $\mathbb{R}$ (regarded as a linear space over $\mathbb{Q}$) has continuum cardinality. (*Hint*: Cf. section 9 of this chapter.)

As we have seen, transfinite induction allows us to prove the existence of a basis in any linear space (over any field). Later we will prove also that any two bases in a given linear space have the same cardinality, and therefore the notion of dimension is well defined for infinite-dimensional vector spaces (Theorem 36, page 82).

The existence of a Hamel basis can be used not only for constructing “pathological” examples. For instance, one can apply it to solve the well-known Hilbert’s Third Problem,, that is, to prove that there exist two polyhedra of the same volume that are not equidecomposable (cannot be decomposed into congruent polyhedra). (In dimension two things are different: any two polygons of the same area can be dissected into congruent polygons.)

Theorem 29. *A cube and a regular tetrahedron of the same volume are not equidecomposable.*

The same is (obviously) true for a cube and a regular tetrahedron having different volumes.

Proof. Suppose that for any polyhedron we define a quantity, named *pseudo-volume*, having the following properties:

- Like volume, the pseudo-volume is additive: if a polyhedron is composed of several others, its pseudo-volume is the sum of their pseudo-volumes.
- Congruent polyhedra have equal pseudo-volumes.
- The pseudo-volume of a cube is zero, while a regular tetrahedron has a nonzero pseudo-volume.

Existence of such a pseudo-volume will imply the statement we are to prove.

Assume that the pseudo-volume of a polyhedron is defined as the sum $\sum l_i\varphi(\alpha_i)$ taken over all edges, where l_i is the length of the ith edge, α_i is the dihedral angle at the ith edge, and φ is some function to be specified later. Congruent polyhedra have the same pseudo-volume by definition. What properties of φ are necessary to ensure that pseudo-volume is additive?

Imagine that a polyhedron with an edge e is cut into two polyhedra by a plane P containing e. Then the dihedral angle α is cut by P into two angles β and γ such that $\beta+\gamma=\alpha$. The formula for the pseudo-volume of the entire polyhedron includes the term $l\varphi(\alpha)$ (where l is the length of the edge e), while the sum of the pseudo-volumes of the two parts includes $l\varphi(\beta)+l\varphi(\gamma)$. Therefore, it is desirable that $\varphi(\alpha)=\varphi(\beta)+\varphi(\gamma)$.

On the other hand, the plane P can create new edges that are intersections of P and faces of a polyhedron. Let l' be the length of a new edge. Then the sum of the pseudo-volumes of two parts will include terms $l'\varphi(\alpha) + l'\varphi(\pi - \alpha)$ (corresponding to supplementary angles), and we want these terms to cancel out.

Now it is clear what properties of φ are desired. We want that $\varphi(\beta+\gamma) = \varphi(\beta)+\varphi(\gamma)$ and that $\varphi(\pi) = 0$. Then the pseudo-volume will be an additive function of a polyhedron. The rigorous proof of this additivity property requires a rigorous definition of a polyhedron, which is not so simple, so we omit this proof. But the statements look plausible, especially if we note that all dissections could be done by planes (this restriction may increase the number of parts, but this does not matter).

Therefore, the only thing necessary to finish the proof, is a function $\varphi\colon \mathbb{R} \to \mathbb{R}$ such that

- $\varphi(\beta+\gamma) = \varphi(\beta) + \varphi(\gamma)$ for any $\beta, \gamma \in \mathbb{R}$;
- $\varphi(\pi) = 0$;
- $\varphi(\pi/2) = 0$ (therefore the pseudo-volume of a cube is zero);
- $\varphi(\theta) \neq 0$, where θ is the dihedral angle of a regular tetrahedron.

(In fact the third property is an easy consequence of the first two.)

To construct such a function, let us note first that θ/π is an irrational number. Indeed, assume that the sides of our regular tetrahedron have length 2. Consider the heights of two faces of this tetrahedron that are perpendicular to the common edge of the faces. This two heights, together with the corresponding edge of the tetrahedron, form the isosceles triangle with sides $\sqrt{3}$, $\sqrt{3}$, 2. We need to prove that the angles of this triangle are irrational multiples of π. Let the angles be θ, β, β. The angle β is an acute angle in the right triangle with sides 1, $\sqrt{2}$ and $\sqrt{3}$; therefore $(\cos\beta + i\sin\beta) = (1+\sqrt{-2})/\sqrt{3}$. If $\beta/\pi \in \mathbb{Q}$, then the complex number $(1+\sqrt{-2})/\sqrt{3}$ is a root of unity. But this is not the case, since the ring $\mathbb{Z}[\sqrt{-2}]$ has unique factorization property. Therefore β (and θ) are not rational multiples of π.

Now we regard π and θ as elements of the linear space $\mathbb{R}$ over $\mathbb{Q}$. They are linearly independent, so we can construct a basis that contains θ and π and consider a $\mathbb{Q}$-linear mapping $\varphi\colon \mathbb{R} \to \mathbb{Q}$ that maps each real number x into its (rational) θ-coordinate. This function φ satisfies all our requirements. □

Problem 115. Prove that one can avoid using Hamel basis in these arguments by the following trick: note that it is enough to define φ only on the reals that are linear combinations of the angles of the polyhedra used in the decomposition, and that there are only finitely many of them.

8. Zorn's Lemma and its application

Modern textbooks usually replace transfinite induction by the so-called Zorn's Lemma. We show how this is done using the existence of a Hamel basis as an example.

Theorem 30 (Zorn's Lemma)**.** *Let Z be a partially ordered set with the following property: any chain has an upper bound. Then Z has a maximal element; moreover, for any $a \in Z$ there exists $b \geq a$ that is maximal in Z.* (Chain *is a linearly ordered subset: $a \leq b$ or $b \leq a$ for any two elements of a chain. An* upper bound *is an element that is greater than or equal to any element of the chain.*)

Proof. Note that the ordering on Z is a partial ordering, so we must distinguish between the greatest element (only one element can be the greatest element of Z) and maximal elements (many elements can be maximal).

The proof follows the scheme used to prove the existence of a basis, but in a more general situation where we consider elements of the ordered set Z instead of linearly independent families.

Let a be an element of Z. Assume that there is no maximal element b such that $b \geq a$. This means, by definition, that for any $b \geq a$ there exists an element $c > b$. Then $c > a$, and therefore there exists $d > c$, etc. If this process continues long enough, we come to a contradiction.

Let us see how this can be done. (In particular, we need to use the condition that each chain has an upper bound.) Let I be a well-ordered set of sufficiently large cardinality (larger than the cardinality of Z). Consider a strictly increasing function $f\colon I \to Z$ defined as follows. The value of f on the least element of a well-ordered set I equals a.

Now assume that the values $f(j)$ are defined for all $j < i$. We want to define $f(i)$. Monotonicity implies that the values $f(j)$ for all $j < i$ form a chain. Therefore, this chain has an upper bound s. Evidently, $a \le s$ (since a is an element of the chain). Consider an element $t > s$ (it exists since we have assumed that no element $s \ge a$ is maximal) and let $f(i) = t$. Note that f extended in this way is still monotone (and therefore injective). Therefore I has the same cardinality as some subset of Z, which contradicts our assumption.

This argument, however, is not formally correct: we simultaneously define some function and prove that it is monotone. The definition of $f(i)$ assumes that f is monotone on $[0, i)$. The formally correct argument goes as follows. We apply Theorem 19, assuming that $f(i)$ is undefined if f is not monotone on $[0, i)$. The function f constructed in this way is defined on some initial segment of I, either on I itself or on some proper initial segment $[0, i_0)$. But the latter case is impossible, because f is monotone and therefore $f(i_0)$ should be defined. □

As before (see Problem 112) we can avoid using an auxiliary set I of large cardinality. Applying Zermelo's Theorem, consider a well-ordering on Z. This well-ordering may have nothing in common with the given partial order on Z, so we denote it by $\prec$. Now we define a function $f\colon Z \to Z$ such that (1) $f(z) \ge a$ for each $z \in Z$; (2) f is monotone in the following sense: $x \prec y$ implies $f(x) \le f(y)$; and (3) $f(z)$ is never smaller than z (with respect to the original partial order $\le$).

Namely, for a $\prec$-minimal element $z_0 \in Z$, let $f(z_0)$ be either z_0 (if $z_0 > a$) or a. For any other z the value $f(z)$ is defined either as (some) upper bound α of values $\{f(z') \mid z' \prec z\}$ (note that this set is a chain since f is monotone) or as z itself (if $z > \alpha$).

The function f is defined on Z; its range is a chain (since f is monotone); therefore there is some $\beta \in Z$ that is an upper bound for the range of f. Evidently, $\beta \geq f(z_0) = a$. It remains to prove that β is a maximal element in Z. Indeed, if $\beta < z$ for some z, then $f(z) \leq \beta < z$, which is impossible by (3).

Problem 116. Provide missing details in these arguments.

Now we show how the existence of a Hamel basis can be proved using Zorn's Lemma. Let V be a linear space. Consider a partially ordered set Z whose elements are linearly independent subsets of V and $S \leq S'$ means $S \subset S'$.

Let us check that any chain in Z has an upper bound. A chain is a family of linearly independent sets; for any two chain elements, one is a subset of the other. Consider the union of all the elements of the chain. We need to show that this union is linearly independent (and therefore belongs to Z and is an upper bound of our chain).

Any nontrivial linear combination involves a finite number of vectors. Each vector belongs to some element of the chain. These elements form a linearly ordered finite subset of the chain; therefore there is the greatest element among them. It contains all the vectors from the linear combination, so this combination is not equal to zero (any element of the chain is a linearly independent set).

Applying Zorn's Lemma, we conclude that for any linearly independent set S there exists a maximal linearly independent set $S' \supset S$. It cannot be enlarged further and therefore S' is a basis in V.

Similar arguments can be used to prove the existence of an orthogonal basis in any Hilbert space. (Note that the definition of a basis in a Hilbert space is different: we consider countable linear combinations that are interpreted as sums of infinite series.)

One can also apply Zorn's Lemma to prove the existence of a transcendence basis in a field extension (i.e., a maximal algebraically independent set).

We mention one more application of Zorn's Lemma to partially ordered sets.

Theorem 31. *Any partial order can be extended to a linear (total) order.*

Proof. Let $(X, \leq)$ be a partially ordered set. We must prove that there exists a linear order relation $\leq'$ on X that is an extension of the given partial order (i.e., $x \leq y \Rightarrow x \leq' y$). Note that not every partial order can be extended to a well-ordering (for example, this cannot be done if we start with a linear order that is not a well-ordering).

To apply Zorn's Lemma we need to consider some partially ordered set Z whose elements are partial orders on X (i.e., subsets of $X \times X$ that are reflexive, transitive, and antisymmetric). The ordering on Z is an inclusion relation: $\leq_1$ is smaller than (or equal to) $\leq_2$, if $x \leq_1 y$ implies $x \leq_2 y$ for any $x, y \in X$.

It is easy to check that any chain in Z has an upper bound. Indeed, the union of a chain of partial orders is a partial order. (Let us check, for example, that this union is transitive. Assume that $x \leq_1 y$ according to some ordering $\leq_1$ that belongs to the chain, and $y \leq_2 z$ according to another ordering in the chain. Then one of the orderings $\leq_1$ and $\leq_2$ is an extension of the other. For example, let $\leq_1$ be an extension of $\leq_2$. Then $x \leq_1 y \leq_1 z$, and therefore $x \leq z$ if $\leq$ is the union of the chain.)

Applying Zorn's Lemma, for any given partial order we get a maximal partial order on X that extends the initial one. We denote this maximal order by $\leq$ (this does not cause a confusion since we do not need the initial partial order anymore). We have to show that $\leq$ is a total (linear) order. If it is not the case, there exist two elements $x, y \in X$ such that $x \not\leq y$ and $y \not\leq x$. Consider a new ordering relation $\leq'$ defined as follows: $a \leq' b$ if either (1) $a \leq b$ or (2) $a \leq x$ and $y \leq b$. It is easy to see that $x \leq' y$ and that $\leq'$ is a partial order. Evidently, $\leq'$ is reflexive. To prove that $\leq'$ is transitive, consider any a, b, c such that $a \leq' b$ and $b \leq' c$. According to the definition of $\leq'$ there are four possible cases (the two cases of $a \leq b$ are combined with the two cases for $b \leq c$):

$(1, 1)$ $a \leq b \leq c$ (trivial);

$(1, 2)$ $a \leq b \leq x$ and $y \leq c$ implies $a \leq' c$ according to (2);

$(2, 1)$ $a \leq x$ and $y \leq b \leq c$ implies $a \leq' c$ according to (2);

$(2, 2)$ $a \le x$, $y \le b$, $b \le x$, $y \le c$; this is impossible since in this case $y \le x$ contrary to our assumption.

A similar argument shows that $\le'$ is antisymmetric.

Therefore, $\le'$ is a partial order that extends $\le$ and is different from $\le$, so $\le$ is not a maximal element in Z—a contradiction. □

Problem 117. Prove that any binary relation R on any set X that has no cycles (cycle appears if xRx, or $xRyRx$, or $xRyRzRx$, etc.) can be extended to a linear order. (A computer scientist would say that "any acyclic directed graph can be topologically sorted".)

Problem 118. A set X on the plane is called *convex* if for any two points $u, v \in X$ the line segment uv belongs to X. Prove that for any two disjoint convex sets A and B there is a line l that separates them (i.e., A and B lie on the different sides of l; note that l is allowed to intersect A and B). (*Hint*: Using Zorn's Lemma, extend the given sets A and B to complementary disjoint convex sets A' and B'. Then prove that the boundary between A' and B' is a line.)

9. Operations on cardinals revisited

Now we can prove several theorems about operations on cardinals.

Theorem 32. *If A is infinite, then $A \times \mathbb{N}$ has the same cardinality as A.*

Proof. Zermelo's Theorem says that A can be well ordered. After that, as we know (see p. 57), each element of A has a unique representation as $z + n$, where z is a limit element (i.e., z has no predecessor) and n is a natural number. Therefore we get a one-to-one correspondence between A and $B \times \mathbb{N}$, where B is a set of limit elements. (Not exactly. The set of all elements $b + n$, where b is a maximal element of B, can be finite (instead of countable) if A has a maximal element. But we know already that adding a finite or a countable set does not change the cardinality of an infinite set.)

Therefore, $A \times \mathbb{N}$ has the same cardinality as $(B \times \mathbb{N}) \times \mathbb{N}$, or $B \times (\mathbb{N} \times \mathbb{N})$, or $B \times \mathbb{N}$ (a product of two countable sets is countable), i.e., the same cardinality as A. □

Recalling the Cantor–Bernstein Theorem, we conclude that all sets between A and $A \times \mathbb{N}$ have the same cardinality as A. (In particular, $|A| + |A| = |A|$ and $n|A| = A$ for any positive integer n.)

Here is another useful corollary:

Theorem 33. *The sum of two infinite cardinalities equals the greater of them.* (*If* $|A| \le |B|$ *and both* A *and* B *are infinite, then* $|A|+|B| = |B|$*.*)

Proof. We recall that for any A and B, either $|A| \le |B|$ or $|B| \le |A|$ (Theorem 25, page 68). Hence the notion of greater cardinality (for two given sets) is well defined.

If $|A| \le |B|$, then $|B| \le |A|+|B| \le |B|+|B| \le |B|\times\aleph_0 = |B|$ (the last inequality is guaranteed by Theorem 32). It remains to apply the Cantor–Bernstein Theorem to conclude that $|B| = |A + B|$. □

Now we are ready to prove an even stronger claim.

Theorem 34. *If* A *is infinite, then* $|A \times A| = |A|$*.*

Proof. Note that we have already proved this theorem for countable sets. (By the way, we know that it is also true for sets of continuum cardinality, but this does not matter now.) Therefore we can find some countable $B \subset A$ and a one-to-one correspondence between B and $B \times B$.

Consider the family Z of all infinite sets $B \subset A$ together with one-to-one correspondences between B and $B \times B$. (An element of Z is a pair $\langle B, f\rangle$, where B is an infinite subset of A and $f\colon B \to B \times B$ is a one-to-one correspondence (bijection)). The set Z is partially ordered by the following relation: $\langle B_1, f_1\rangle \le \langle B_2, f_2\rangle$ if $B_1 \subset B_2$ and the restriction of f_2 to B_1 coincides with f_1 (i.e., $f_1(x) = f_2(x)$ for $x \in B_1$); see Figure 6.

To apply Zorn's Lemma, we need to check first that any chain in Z has an upper bound. Assume that some sets (together with bijections) form a chain. Consider the union B of these sets. Since the bijections extend each other, we get a combined mapping $f\colon B \to B \times B$. The mapping f is an injection. Indeed, if b' and b'' are different elements of B and belong to different elements of the chain, then b' and b''

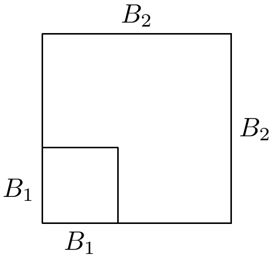


Figure 6. A mapping f_1 is a one-to-one correspondence between the smaller square and its side; f_2 adds to f_1 a one-to-one correspondence between $B_2 \setminus B_1$ and the remaining part of the larger square, i.e., $(B_2 \times B_2) \setminus (B_1 \times B_1)$.

both belong to the greater of the two chain elements. Therefore, $f(b') \neq f(b'')$.

Now let us prove that f is surjective. For any pair $\langle b', b'' \rangle \in B \times B$ consider the greater of the two chain elements that contain b' and b'', and recall that f induces a bijection between this set and its square.

The Zorn Lemma guarantees that Z has a maximal element $\langle B, f \rangle$. By definition, f is a one-to-one correspondence between B and $B \times B$ and $|B| = |B| \times |B|$.

Now there are two possibilities.

(1) A and B have the same cardinality. Then all four sets A, B, $A \times A$ and $B \times B$ have the same cardinality, and the theorem is proved.

(2) The cardinality of B is smaller than that of A. (Note that $|B| \leq |A|$ because $B \subset A$.) Let C be the remaining part of A, i.e., $C = A \setminus B$. Then $|A| = |B| + |C| = \max(|B|, |C|)$. Therefore, C has the same cardinality as A and greater cardinality than B. Let $C' \subset C$ be a part of C that has the same cardinality as B. By B' we denote the (disjoint) union $B' = B + C'$ (Figure 7).

Both parts of B' (i.e., B and C') have the same cardinality as B. Therefore, $B' \times B'$ consists of four parts; each has the same cardinality as $B \times B$ and, therefore, as B (by our assumption, f is a bijection between B and $B \times B$). The bijection f can be extended to a bijection $f' \colon B' \to B' \times B'$ by adding a bijection between C'

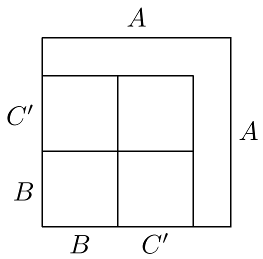


Figure 7. Extending a one-to-one correspondence from B to $B' = B + C'$.

and $(B' \times B') \setminus (B \times B)$ (this set consists of three parts that have cardinality $|B|$, so both sets have cardinality $|B|$).

Therefore the pair $\langle B', f' \rangle$ is larger than $\langle B, f \rangle$, but $\langle B, f \rangle$ was maximal according to Zorn's Lemma. Therefore, case (2) is impossible. □

Here are several corollaries of Theorem 34.

Theorem 35. **(a)** $A \times B = \max(|A|, |B|)$.

(b) *If A is infinite, then the set A^n whose elements are n-tuples made of A-elements, has the same cardinality as A.*

(c) *If A is infinite, then the set A^* whose elements are finite sequences of A-elements has the same cardinality as A.*

Proof. The first assertion: if $|A| \leq |B|$, then $|B| \leq |A| \times |B| \leq |B| \times |B| = |B|$, and $|A| \times |B| = |B|$ by the Cantor–Bernstein theorem.

The second assertion can be proved by induction on n. Indeed, if $|A^n| = |A|$, then $|A^{n+1}| = |A^n| \times |A| = |A| \times |A| = |A|$.

Finally, A^* equals $1 + A + A^2 + A^3 + \cdots$ (a finite sequence may have length $0, 1, 2, \ldots$); each part (except for the first, which is finite and can be ignored) has the same cardinality as A; therefore A^* has the same cardinality as $|A| \times \aleph_0 = |A|$. □

The statement (c) of Theorem 35 implies that the set of all finite subsets of an infinite set A has the same cardinality as A. (Indeed, a function that maps each finite sequence into the set of its elements is surjective; therefore the set of finite subsets of A is between A and A^*.)

Problem 119. Let A be an infinite set. Prove that $|A^A| = |2^A|$.

Problem 120. Consider the cardinality $\alpha = \aleph_0 + 2^{\aleph_0} + 2^{(2^{\aleph_0})} + \cdots$ (a countable sum of an increasing sequence of cardinalities can be defined in a natural way). Prove that α is a minimal cardinality that is greater than the cardinality of all sets $\mathbb{N}, P(\mathbb{N}), P(P(\mathbb{N})), \ldots$ (here $P(X)$ stands for the power set of X, i.e., the set of all subsets of X). Prove that $\alpha^{\aleph_0} = 2^\alpha > \alpha$.

Now we can prove that different Hamel bases have the same cardinality.

Theorem 36. *Any two bases in an infinite-dimensional vector space have the same cardinality.*

Proof. Consider any two bases, calling them "the first basis" and "the second basis", respectively. For each element of the first basis choose some representation of it as a linear combination of the vectors of the second basis. We get a function that maps each vector of the first basis into a finite subset of the second basis. As we have proved (Theorem 35), the range of this function has cardinality that does not exceed the cardinality of the second basis. On the other hand, the preimage of each value of this function is a linearly independent set of vectors that are linear combinations of a finite set; therefore this preimage is finite.

We see that the first basis is split into groups, each group is finite and the number of groups does not exceed the cardinality of the second basis. Therefore the cardinality of the first basis does not exceed the cardinality of the second multiplied by $\aleph_0$ (and this operation does not change the cardinailty).

Similar arguments apply in the other direction, and it remains to use the Cantor–Bernstein Theorem. □

10. Ordinals

As we have already mentioned, an *ordinal* is an order type of a well-ordered set, i.e., the class of all ordered sets that are order-isomorphic to it.

One can naturally define a linear order on ordinals. To compare two ordinals α and β, we consider their representatives A and B (so A belongs to the order type α and B belongs to the order type β). Then we apply Theorem 22 to see which of the three cases occurs. These three cases are: (1) A is order-isomorphic to a proper initial segment of B; (2) A and B are order-isomorphic; (3) B is order-isomorphic to a proper initial segment of A. In case (1) we say that $\alpha < \beta$; in case (2) we say that $\alpha = \beta$; and in case (3) we say that $\alpha > \beta$ (or $\beta < \alpha$).

Again we ignore difficulties related to the foundations of set theory (the class of isomorphic ordered sets is too big; see Section 6 of Chapter 1). Later we discuss possible workarounds. But first we mention basic properties of ordinals.

- The linear order defined on ordinals is a well-ordering: each nonempty family of ordinals has a minimal element (Theorem 23; we did not use the word "ordinal" there but instead spoke about ordered sets that represent ordinals).

- Let α be an ordinal. Consider the initial segment $[0, \alpha)$ that consists of all ordinals that are smaller than α according to our definition. This initial segment has order type α (i.e., it is order-isomorphic to elements of the class α). Indeed, let A be a well-ordered set having order type α. The ordinals smaller than α correspond to proper initial segments of A, i.e., segments $[0, a)$ for all $a \in A$. Thus, we get a one-to-one correspondence between the elements of A and the ordinals that are smaller than α. It is easy to see that this correspondence is an isomorphism of ordered sets.

  One may say that "each ordinal is order-isomorphic to the set of smaller ordinals". If we use von Neumann's approach, the ordinal *is* the set of smaller ordinals (see below). This approach allows us to avoid logical difficulties when defining ordinals.

- An ordinal α is called a *nonlimit* ordinal if it has a predecessor (the greatest ordinal smaller than α). If the predecessor does not exist, the ordinal is called a *limit* ordinal.
- Any bounded family of ordinals has the least upper bound. Indeed, let F be a bounded family of ordinals and let β be an upper bound of F (i.e., $\gamma \le \beta$ for any $\gamma \in F$). Let B be a well-ordered set of order type β. All ordinals in F are order-isomorphic to initial segments of B. If one of this initial segments equals B, then β is the greatest element of F (and, therefore, the least upper bound of F). If none of these initial segments equals B, then all the segments are $[0, b)$ for different elements $b \in B$. Consider the set S of all b's that correspond to ordinals in F. If S has no upper bound in B, then β is the least upper bound of F. If S has an upper bound in B, then it has the least upper bound s and the order type of the initial segment $[0, s)$ is the least upper bound of F.

One may say that ordinals form the "universal well-ordered class"; any well-ordered set is order-isomorphic to an initial segment of that class. But we must be careful. If we regarded "the set of all ordinals" as a set, this set would be the greatest well-ordered set. And there is no such thing: for any well-ordered set W there exists a greater one $(W + 1)$. This paradox (called the *Burali–Forti paradox*) shows that the class of all ordinals should not be treated as a set.

Problem 121. Prove that the least upper bound of a countable family of countable ordinals (i.e., order types of coundable well-ordered sets) is a countable ordinal.

There should be a way to speak about ordinals without the danger of falling into a contradiction. One possibility is to eliminate ordinals completely and speak about well-ordered sets that represent them. ("Equal ordinals" become "isomorphic well-ordered sets" after this translation.)

Another approach was suggested by von Neumann, and it is now a standard approach in axiomatic set theory. In this approach, each ordinal is the set of all smaller ordinals. For example, the minimal ordinal 0 has no smaller ordinals, i.e., 0 is the empty set $\varnothing$. For the successor of 0 (we denote it by 1) there exists exactly one smaller ordinal, 0, so

$$1 = \{0\} = \{\varnothing\}.$$

Similarly,

$$\begin{aligned}2 &= \{0, 1\} = \{\varnothing, \{\varnothing\}\},\\ 3 &= \{0, 1, 2\} = \{\varnothing, \{\varnothing\}, \{\varnothing, \{\varnothing\}\}\},\end{aligned}$$

etc. (Recall that we interpreted, say, 3 as the order type of well-ordered set with three elements, i.e., the class of all well-ordered sets with three elements. Now this class is not needed because it is replaced by its “canonical representative”.)

The first infinite ordinal is ω, the order type of the set of natural numbers. According to von Neumann, ω is the set of all smaller ordinals, i.e.,

$$\omega = \{0, 1, 2, 3, \dots\}.$$

The next ordinals are

$$\begin{aligned}\omega + 1 &= \{0, 1, 2, 3, \dots, \omega\},\\ \omega + 2 &= \{0, 1, 2, 3, \dots, \omega, \omega + 1\},\end{aligned}$$

etc. We will not go into details of this construction, since axiomatic set theory is beyond the scope of this book. However, let us mention that its most popular version is called the *Zermelo–Fraenkel* set theory (ZF). It assumes that all objects are sets and includes the so-called *axiom of extensionality* that says that two sets are the same if and only if they have the same elements. This approach looks strange to, say, an ecologist who may want to consider the population of birds in a given area as a set (without considering each bird as a set of something else). However, mathematicians are used to objects that are constructed in a rather complicated way (i.e., a real number is a set of equivalent fundamental sequences of rational numbers, while a rational number is a set of equivalent pairs of integers, while an integer is some yet another set, etc.).

Axiom of extensionality implies that there is only one set that has no elements. Another axiom, called the *axiom of foundation*, or the *axiom of regularity*, says that $\in$-relation is well-founded: each set X contains an element $x \in X$ that is $\in$-minimal in X, i.e., there is no $y \in X$ such that $y \in x$ ($x \cap X = \varnothing$). An immediate corollary: no set

x can be an element of itself (otherwise the foundation axiom would fail for $\{x\}$).

Problem 122. Prove that the axiom of foundation implies that there are no sets x, y, z such that $x \in y \in z \in x$.

A philosopher would explain the meaning of the axiom of foundation as follows: sets are constructed from sets that were constructed earlier, starting with the empty set. Therefore, while proving some property α for all sets, we may prove it by induction: first for the empty set, and then for any set whose elements satisfy α. This induction principle is equivalent to the axiom of foundation.

Ordinals are defined in ZF as follows. We say that a set x is *transitive* if each element of x is a subset of x, i.e., if $z \in y \in x$ implies $z \in x$. A set α is called an *ordinal* if α is transitive and all its elements are transitive. This requirement implies that the $\in$-relation defines a partial order on α. Then induction is used (as described above, applying the axiom of foundation) to prove that the $\in$-order is a linear (total) order. And using induction once more, we conclude that the $\in$-order is a well-ordering.

Problem 123. (**a**) Using the definition of an ordinal as a transitive set with transitive elements, prove that each element of an ordinal is an ordinal. (**b**) Let α be an ordinal (according to the definition above). Prove that the relation $\in$ defines a partial order on α. (**c**) Prove that for any two elements $a, b \in \alpha$ either $a \in b$, or $a = b$, or $b \in a$ (mutually excluding possibilities). (*Hint*: Use double induction over the well-founded relation $\in$ on α and the axiom of extensionality.) (**d**) Prove that an ordinal α is order-isomorphic to a proper initial segment of ordinal β if and only if $\alpha \in \beta$. Therefore the relation $<$ defined on ordinals (as well-ordered sets) coincides with the $\in$-relation. Prove that each ordinal is the set of all smaller ordinals.

Note also that the least upper bound of a set of ordinals (in the von Neumann sense) is the union of this set (since the $\leq$-relation on ordinals coincides with the $\subset$-relation.)

We will not develop this approach further and will contionue to use the naïve definition of an ordinal as a class of order-isomorphic well-ordered sets.

In the next section we define arithmetic operations on ordinals. But first we prove the following simple property of ordinals:

Theorem 37. *Let A be a subset of a well-ordered set B. Then the order type of A does not exceed the order type of B.*

Note that equality is possible even if A is a proper subset of B (i.e., $A \neq B$). For example, the set of all even natural numbers has the same order type (ω) as the set of all natural numbers.

Proof. Assume that A has a greater order type. Then B is isomorphic to some initial segment A' of A (and $A' \neq A$). Let a_0 be the upper bound of A' in A that does not belong to A', and let $f\colon B \to A'$ be an isomorphism. Then f (as a function $B \to B$) is a strictly increasing function, and therefore, $f(b) \geq b$ for all $b \in B$ (Theorem 17). In particular, $f(a_0) \geq a_0$, but $f(b)$ belongs to A' and therefore is less than a_0 according to our assumption. □

11. Ordinal arithmetic

We have defined the notions of sum and product of linearly ordered sets in Section 1 of this chapter. (Recall that in $A + B$ any element of A is smaller than any element of B, and in $A \times B$ we start with comparing second components and turn to the first components if the second ones are equal.) As we have seen, the sum and the product of two well-ordered sets are well-ordered sets and their orders are determined up to isomorphism by the order types of the operands.

We now prove some basic properties of addition:

- Addition is associative: $\alpha + (\beta + \gamma) = (\alpha + \beta) + \gamma$.
- Addition is not commutative, e.g., $1 + \omega = \omega$, but $\omega + 1 \neq \omega$.
- Evidently, $\alpha + 0 = 0 + \alpha = \alpha$.
- The sum increases when the second operand increases: if $\beta_1 < \beta_2$, then $\alpha + \beta_1 < \alpha + \beta_2$. Indeed, assume that β_1 is isomorphic to an initial segment of β_2 (other than β_2 itself). Combine this isomorphism with the identity mapping on α to get an isomorphism between $\alpha + \beta_1$ and some initial segment of $\alpha + \beta_2$ different from $\alpha + \beta_2$.

- The sum does not decrease when the first operand increases: if $\alpha_1 < \alpha_2$, then $\alpha_1 + \beta \leq \alpha_2 + \beta$. Indeed, $\alpha_1 + \beta$ is order-isomorphic to some subset of $\alpha_2 + \beta$. This subset is not an initial segment. However, we can apply Theorem 37 and get the desired inequality. (It is possible that $\alpha_1 + \beta = \alpha_2 + \beta$ even though $\alpha_1 < \alpha_2$.)
- The definition of addition can be treated as an extension of the previously used notation $\alpha + 1$ for the successor of α. Here 1 is the order type of a one-element set. The successor of $\alpha + 1$ is

$$(\alpha + 1) + 1 = \alpha + (1 + 1) = \alpha + 2,$$

  etc.
- If $\beta \leq \alpha$, there exists a unique ordinal γ such that $\beta + \gamma = \alpha$. Indeed, β is order-isomorphic to some initial segment of α; the remaining part of α is γ. (Such a γ is unique since addition is strictly increasing as a function of the second operand.) The ordinal γ is called the *difference* of ordinals α and β (one should rather say "the left difference" since addition is not commutative).
- What about a "right difference"? Unlike the left difference, sometimes it does not exist. Let α be an ordinal. Then the equation $\beta + 1 = \alpha$ (where β is regarded as an unknown variable) has solutions if and only if α is a nonlimit ordinal (i.e., when α has a maximal element).

The associativity law allows us to consider the sum of several operands: $\alpha + \beta + \gamma$ can be defined as $(\alpha + \beta) + \gamma$ or $\alpha + (\beta + \gamma)$. One can also define the sum of several ordinals directly (consider the disjoint union of the given sets with an appropriate order).

One can also define a sum $\alpha_1 + \alpha_2 + \cdots$ of a countable sequence of ordinals. In this sum any element of α_i is less than any element of α_j for $i < j$; the order inside α_i remains unchanged. To check that we get a well-ordering, consider any subset X of the union of α_i. To find a minimal element, find the least i such that $X \cap \alpha_i \neq \varnothing$ and then consider the minimal element in $X \cap \alpha_i$. (We apologize for mixing ordinals with disjoint well-ordered sets that represent them, and hope that this will not lead to a confusion.)

In this construction the set $\mathbb{N}$ of natural numbers (used as subscripts) can be replaced by any other well-ordered set I. If for any element i of a well-ordered set I a well-ordered set A_i is given, we can define the sum $\sum_{i\in I} A_i$ as the order type of the set of pairs $\langle a, i\rangle$ where $a \in A_i$. To compare two pairs we first compare their second components: if $i > i'$ then $\langle a, i\rangle > \langle a', i'\rangle$ (for any $a \in A_i$ and $a' \in A_{i'}$). If the second components are equal, we compare the first components: $\langle a, i\rangle > \langle a', i\rangle$ if $a > a'$ in A_i.

If all the A_i are order-isomorphic to some set A, the sum $\sum A_i$ becomes the product $A \times I$ (defined earlier).

We now prove some properties of multiplication.

- Multiplication is associative: $(\alpha\beta)\gamma = \alpha(\beta\gamma)$. (Indeed, both sides of the equality provide the same ordering on the set of triples; they are compared from right to left until we find a difference.)
- Multiplication is not commutative. For example, $2 \cdot \omega = \omega$ and $\omega \cdot 2 \neq \omega$.
- Evidently, $\alpha \cdot 0 = 0 \cdot \alpha = 0$ and $\alpha \cdot 1 = 1 \cdot \alpha = \alpha$.
- Distributivity: $\alpha(\beta + \gamma) = \alpha\beta + \alpha\gamma$ (just by definition), but $(\beta + \gamma)\alpha$ can differ from $\beta\alpha + \gamma\alpha$. For example, $(1 + 1) \cdot \omega = \omega \neq \omega + \omega = \omega \cdot (1 + 1)$.
- The product is an increasing function of the second operand if the first one is not 0. Let us (just for fun) derive this from the properties above: If $\beta_2 > \beta_1$, then $\beta_2 = \beta_1 + \delta$ for some $\delta \neq 0$. Then
$$\alpha\beta_2 = \alpha(\beta_1 + \delta) = \alpha\beta_1 + \alpha\delta > \alpha\beta_1.$$
- The product is a nondecreasing function of the first operand. Indeed, if $\alpha_1 < \alpha_2$, then $\alpha_1\beta$ is order-isomorphic to a subset of $\alpha_2\beta$. This subset is not an initial segment, but we may nevertheless apply Theorem 37.
- Any ordinal $\gamma < \alpha\beta$ has a unique representation of the form $\gamma = \alpha\beta' + \alpha'$, where $\beta' < \beta$ and $\alpha' < \alpha$.

(Indeed, consider sets A and B having order types α and β. Then $A \times B$ has order type $\alpha\beta$. Any ordinal $\gamma < \alpha \times \beta$ is order isomorphic to some proper initial segment $[0, \langle a, b\rangle)$ of the set $A \times B$. This initial segment consists of all pairs whose second element is smaller than b and all pairs where the second element is b and the first element is smaller than a. Therefore, γ is order-isomorphic to $A \times [0, b) + [0, a) = \alpha \times \beta' + \alpha'$ for $\beta' = [0, b)$ and $\alpha' = [0, a)$. We now prove that this representation is unique. Assume that $\alpha\beta' + \alpha' = \alpha\beta'' + \alpha''$. If $\beta' = \beta''$, then $\alpha' = \alpha''$ (see above about the left difference). Let us prove that $\beta' \neq \beta''$ is impossible. Indeed, if (say) $\beta' < \beta''$, then $\beta'' = \beta' + \delta$ and $\alpha' = \alpha\delta + \alpha''$, which is impossible because the left-hand side is smaller than α and the right-hand side is greater than or equal to α.)

- The ordinals β' and α' may be called the "quotient" and the "remainder" obtained when γ is "divided" by α. The similar "division" can be performed for any ordinals: Assume that $\alpha > 0$. Then any ordinal γ can be divided by α, i.e., γ can be represented as $\gamma = \alpha\tau + \rho$, where $\rho < \alpha$, and this representation is unique.

  (The existence follows from the previous property: one should take β large enough so that $\alpha\beta > \gamma$, for example, $\beta = \gamma + 1$. The uniqueness proof works without any changes.)

- Iterating division by some $\alpha > 0$, we may construct a *positional number system* with base α. Any ordinal $\gamma < \alpha^{k+1}$ (where k is a natural number) can be uniquely represented as

  $$\gamma = \alpha^k\beta_k + \alpha^{k-1}\beta_{k-1} + \cdots + \alpha\beta_1 + \beta_0,$$

  where $\beta_k, \ldots, \beta_1, \beta_0$ are ordinals less than α.

  (First we divide γ by α, then we divide the quotient by α again, etc. Or we may divide γ by α^k, then divide the remainder by α^{k-1}, etc.)

Problem 124. For which ordinals the equality $1 + \alpha = \alpha$ holds?

Problem 125. The same question for the equality $2 \cdot \alpha = \alpha$.

Problem 126. Which ordinals can be represented as $\omega \cdot \alpha$?

Problem 127. Prove that $\alpha + \beta = \beta$ if and only if $\alpha\omega \leq \beta$ (for any ordinals α and β).

Problem 128. Prove that if $\alpha + \beta = \beta + \alpha$ for some ordinals α and β, then there exists an ordinal γ and natural numbers m and n such that $\alpha = \gamma m$ and $\beta = \gamma n$.

Problem 129. Consider the following operation called "base change" from $k > 1$ to $l > k$. (The operation can be applied to natural numbers; k and l are natural numbers.) To apply base change to a natural number n we write n in the positional system with base k. Then we read this string of digits in the positional system with base l. For example, if we apply to number "five" the base change $2 \to 10$ (with $k = 2$ and $l = 10$), we get "one hundred and one".

Note that the base change makes any number bigger, except for the numbers less than or equal to k.

Now consider an arbitrary natural number n and apply the following sequence of operations: (base change $2 \to 3$)—(subtraction of 1)—(base change $3 \to 4$)—(subtraction of 1)—(base change $4 \to 5$)—(subtraction of 1)—etc.

Prove that this process, started from any n, will always terminate (i.e., we will come to zero and will not be able to subtract 1 from it).

(*Hint*: This problem does not involve ordinals in its statement. However, they can be useful: replace all the bases by ω and get a decreasing sequence of ordinals. This argument works for any sequence of bases.)

12. Recursive definitions and exponentiation

In the previous section we gave explicit definitions of the sum and product for any two ordinals. However, the same operations could be defined recursively.

Theorem 38. *Addition has the following properties* (α *and* β *are arbitrary ordinals*):

$$\begin{aligned} \alpha + 0 &= \alpha; \\ \alpha + (\beta + 1) &= (\alpha + \beta) + 1; \\ \alpha + \gamma &= \sup\{\alpha + \beta \mid \beta < \gamma\} \text{ \textit{for any limit ordinal} } \gamma \neq 0. \end{aligned}$$

These properties define the addition operation uniquely.

Proof. The first two properties are evidently true. Let us prove the third one. If $\beta < \gamma$, then $\alpha + \beta < \alpha + \gamma$; therefore $\alpha + \gamma$ is an upper bound of the set of all sums $\alpha + \beta$ (for all $\beta < \gamma$). We have to prove that $\alpha + \gamma$ is the *least* upper bound. Let $\tau < \alpha + \gamma$ be an ordinal. We check that $\tau < \alpha + \beta$ for some $\beta < \gamma$. If $\tau < \alpha$, this is evidently true. If $\tau \geq \alpha$, then $\tau = \alpha + \sigma$ for some σ, and $\alpha + \sigma < \alpha + \gamma$; therefore $\sigma < \gamma$. Since γ is a limit ordinal, $\sigma + 1$ is less than γ, so we let $\beta = \sigma + 1$.

It remains to prove that these three properties define addition uniquely (any operation that has all three of them coincides with addition). Indeed, these properties form a recursive definition over β; if two versions of addition are different for some β, we take the minimal β where they differ and come to a contradiction. □

Multiplication can be defined recursively as follows:

Theorem 39. *Multiplication has the following properties* (α *and* β *are arbitrary ordinals*):

$$\begin{aligned} \alpha 0 &= 0; \\ \alpha(\beta + 1) &= \alpha\beta + \alpha; \\ \alpha\gamma &= \sup\{\alpha\beta \mid \beta < \gamma\} \text{ \textit{for any limit ordinal} } \gamma \neq 0. \end{aligned}$$

These properties uniquely determine the multiplication operation.

Proof. The proof is similar to the proof of the previous theorem. We have to prove that if $\tau < \alpha\gamma$ for a limit ordinal γ, then $\tau < \alpha\beta$ for some $\beta < \gamma$. As we have seen (p. 89), $\tau = \alpha\gamma' + \alpha'$ for some $\gamma' < \gamma$; now we let $\beta = \gamma' + 1$. □

After addition and multiplication are defined, the next step is to define exponentiation. We have already defined α^n for positive integers n (as the product of n factors equal to α).

In more detail: if a well-ordered set A has order type α, then α^n is the order type of the set A^n whose elements are n-tuples of A-elements and whose order is the reverse lexicographical order (the comparison goes from right to left).

How should we define α^ω? One can consider the set $A^{\mathbb{N}}$ of infinite sequences of A-elements (where A has order type α) and define some well-ordering on $A^{\mathbb{N}}$. But it is not clear which well-ordering can be used. So let us drop this idea and try to define exponentiation recursively as follows (where α and β are arbitrary ordinals):

$$\begin{aligned}\alpha^0 &= 1;\\ \alpha^{\beta+1} &= \alpha^\beta \cdot \alpha;\\ \alpha^\gamma &= \sup\{\alpha^\beta \mid \beta < \gamma\} \text{ for any limit ordinal } \gamma \neq 0.\end{aligned}$$

Theorem 18 (on transfinite recursion) guarantees that these equations uniquely define some operation on ordinals that is called *exponentiation.*

Remark. Here we again approach the dangerous area where set-theoretic paradoxes may appear. Theorem 18 gives us a function defined on some well-ordered set. Here we try to define an operation that can be applied to any ordinal; but ordinals do not form a set (there are too many of them). Another problem is that Theorem 18 deals with functions taking values in a given set, but now we have no such set.

Axiomatic set theory resolves this problem by using a special axiom called *axiom of replacement*, but we will not go into details here. Instead we give an explicit definition of exponentiation which is free from these problems.

Let us look first at the ordinal α^ω (for some α). Let A be a well-ordered set of order type α. By definition, α^ω if the least upper bound of all α^n for all natural n. The ordinal α^n is the order type of the set A^n with reverse lexicographical order. To find the least upper bound, we regard A^k as an initial segment of A^l for $l > k$. For example, A^2

consists of pairs $\langle a_1, a_2 \rangle$ and is isomorphic to the initial segment of A^3 that consists of triples $\langle a_1, a_2, 0 \rangle$. (Here 0 stands for the least element in A.) Now we see that all A^n can be viewed as initial segments of the set A^∞ that consists of infinite sequences $a_0, a_1, \ldots$ of A-elements such that only finitely many of their terms differ from 0. (The last requirement makes the reverse lexicographical order possible: for any two given sequences we find the rightmost place where they differ and compare both terms at that place.) The union of all A^n (viewed as initial segments of A^∞) is A^∞. Therefore, A^∞ has order type α^ω according to our recursive definition.

This example motivates the following explicit definition of exponentiation.

Let A and B be well-ordered sets of order types α and β. Consider the set $[B \to A]$ whose elements are mappings from B to A having *finite support*. We say that a mapping $f\colon B \to A$ has finite support if it equals the least element of A everywhere except for a finite set. The set $[B \to A]$ can be ordered as follows: for any two different functions f_1 and f_2 we consider the greatest element $b \in B$ such that $f_1(b) \neq f_2(b)$ and compare $f_1(b)$ and $f_2(b)$.

Theorem 40. *This rule defines a well-ordering on the set* $[B \to A]$ *and its order type is* α^β.

Proof. It is easy to see that our rule defines a linear order on $[B \to A]$. Let us check that it is a well-ordering.

By the *support* of a mapping $f \in [B \to A]$ we mean the set of all $b \in B$ such that $f(b) > 0$ (where 0 stands for the least element of A). By the *rank* of f we mean the greatest element in the support of f (by the definition of $[B \to A]$ the support is finite). The rank is defined for all functions except the zero function (which is the least element of $[B \to A]$). If two elements of $[B \to A]$ have different ranks, the element with the greater rank is greater (according to the ordering in $[B \to A]$ defined above).

If $[B \to A]$ is not a well-ordered set, there exists a decreasing infinite sequence $f_0 > f_1 > f_2 > \cdots$ of nonzero elements. Consider the ranks of f_i. They form a nonincreasing sequence of elements of B. Since B is well-ordered, ranks stabilize (i.e., coincide for all

$f_n, f_{n+1}, f_{n+2}, \ldots$ for some n). We can ignore first n elements and assume that the ranks of all f_i are equal to some $b \in B$.

Consider the sequence $f_0(b), f_1(b), \ldots$. The definition of order in $[B \to A]$ implies that $f_0(b) \ge f_1(b) \ge f_2(b) \ge \cdots$; since A is well ordered, the sequence $f_0(b) \ge f_1(b) \ge f_2(b) \ge \cdots$ stabilizes. Again we can forget about finitely many terms and assume without loss of generality that all $f_i(b)$ are the same. Under this assumption the value $f_i(b)$ plays no role in comparisons, and can be replaced by 0. Then we get a decreasing sequence in $[B \to A]$ where the ranks of all elements are less than b. To complete the argument, we refer to the induction principle over B.

Let us say it again with more details. Consider all strictly decreasing infinite sequences in $[B \to A]$. (We assume that they exist and come to a contradiction.) For each sequence consider the rank of its first element. Let us choose a sequence with least possible rank. (Here we use that B is well ordered.) Let b be this minimal rank. Consider a strictly decreasing sequence that starts with an element of rank b. All the elements of the sequence have rank b (if some of them has smaller rank, we can remove an initial segment and get a decreasing sequence that starts with a smaller rank).

Consider all decreasing sequences $f_0 > f_1 > f_2 > \cdots$ whose elements have rank b and choose one of them that has the least value of $f_0(b)$. (Here we use that A is well ordered.) All terms of this sequence have the same value at b ($f_i(b) = f_0(b)$; otherwise we can drop some elements to get a sequence with a smaller value of $f_0(b)$). Replace this value by 0 (let $f_i(b) = 0$ for all i). We get an infinite strictly decreasing sequence of elements whose ranks are less than b, and this is impossible according to our assumptions.

It remains to prove that our explicit definition of exponentiation satisfies the recursive definition. For finite n it is evident.

Now let γ be a nonlimit ordinal: $\gamma = \beta + 1$. What is our (explicit) definition of α^γ in this case? Let B be a well-ordered set of type β. To get a well-ordered set of type γ we add a new element m to B so that it be greater than all the elements of B. Then we consider all mappings of $B \cup \{m\}$ to A having finite support. Any mapping g of this type can be regarded as a pair: its first element is a restriction of

g to B, and the second is the value $g(m)$ (which is an element of A). (Finite support remains finite if m is added or removed.)

To compare two mapping g and g', we first compare $g(m)$ and $g'(m)$; if they are equal, we compare the restrictions of g and g' to B. Therefore, the set $[B \cup \{m\} \to A]$ is isomorphic to $[B \to A] \times A$, so the second requirement of the recursive definition is satisfied.

Now let γ be a nonzero limit ordinal, and let C be a well-ordered set of order type γ. Let us look at the set $[C \to A]$. The elements of this set whose rank is less than some $c \in C$, form an initial segment in $[C \to A]$, and this initial segment is order-isomorphic to $[[0, c) \to A]$. The set $[C \to A]$ is the union of these initial segments (since any element of $[C \to A]$ has finite support). Therefore, the order type of $[C \to A]$ is the least upper bound of the order types of the sets $[[0, c) \to A]$, exactly as required by the recursive definition. □

Theorem 41. *For any countable ordinals α and β, the ordinals $\alpha+\beta$, $\alpha\beta$ and α^β are also countable.*

Proof. The sum and product are countable since the sum and product of any two countable sets is a countable set.

Exponentiation: If all elements of well-ordered sets A and B are numbered by integers, then any element f of $[B \to A]$ is determined by a finite list of integers (that includes all the elements of the support of f and all corresponding values), and the set of all finite lists of integers is countable. □

Problem 130. Give two proofs that $\alpha^{\beta+\gamma} = \alpha^\beta \cdot \alpha^\gamma$, first using transfinite induction and then using the explicit definition of exponentiation.

Problem 131. Prove that $(\alpha^\beta)^\gamma = \alpha^{\beta\gamma}$.

Problem 132. Prove that $\alpha^\beta \geq \alpha\beta$ for $\alpha \geq 2$.

Problem 133. Prove that if $\omega^\gamma = \alpha + \beta$ for some ordinals α, β and γ, then either $\beta = 0$ or $\beta = \omega^\gamma$.

Problem 134. Which ordinals cannot be represented as sums of two smaller ordinals?

Problem 135. Use the inductive definition of exponentiation to prove that α^β is countable for any countable α and β. (*Hint*: Recall that an upper bound of a countable family of countable ordinals is countable.)

Problem 136. Let $\alpha > 1$ be an arbitrary ordinal. Find the least ordinal $\beta > 0$ such that $\alpha\beta = \beta$. (*Hint*: What is the product of x and the power series $1 + x + x^2 + x^3 + \cdots$?)

Let us note the following important difference between exponentiation of ordinals and the previously defined operations (addition, multiplication). Defining a sum (or product) we introduced some well-ordering on the sum (or Cartesian product) of the corresponding sets. For exponentiation, we consider the set $[B \to A]$ whose definition depends not only on the underlying sets of A and B, but also on the orderings. This set differs from the set A^B of all functions of type $B \to A$ considered earlier. In particular, for countable A and B the set $[B \to A]$ is countable, whereas the set A^B has continuum cardinality.

The explicit definition of exponentiation as the order type of $[B \to A]$ allows us to understand the structure of ordinals that are smaller than α^β, i.e., the initial segments of $[B \to A]$.

Let f be an element of $[B \to A]$. Then f is a function with finite support taking nonzero values at the points $b_1 > b_2 > \cdots > b_k$. Let $a_1, a_2, \ldots, a_k$ be its values ($f(b_i) = a_i$).

Let g be any function smaller than f. Then $g(x) = 0$ for any $x > b_1$. The value $g(b_1)$ could be either less than a_1 or equal to a_1. These two possibilities split all functions $g < f$ into two classes. Any function of the first type (where $g(b_1) < a_1$) is less than any function of the second type.

Functions of both types have zero values outside $[0, b_1]$. First type functions take some value at b_1 that is less than a_1 and may have any values in $[0, b_1)$ (but only finitely many values are different from 0). Therefore, the set of all first type functions is isomorphic (as an ordered set) to $[[0, b_1) \to A] \times [0, a_1)$.

Second type functions g (such that $g(b_1) = a_1$) could be further divided into two subcategories: the first is formed by functions g such that $g(b_2) < a_2$; the second is formed by functions g such that $g(b_2) = a_2$. The first subcategory is isomorphic to $[[0, b_2) \to A] \times [0, a_2)$. The second subcategory can again be divided into two parts (where $g(b_3) < a_3$ and $g(b_3) = a_3$), etc. Therefore, $[0, f)$ is an ordered set isomorphic to

$$[[0, b_1) \to A] \times [0, a_1) + [[0, b_2) \to A] \times [0, a_2) + \cdots \\ + [[0, b_k) \to A] \times [0, a_k).$$

Reformulating this statement in terms of ordinals, we get the following

Theorem 42. *Any ordinal less than α^β can be represented as*

$$\alpha^{\beta_1}\alpha_1 + \alpha^{\beta_2}\alpha_2 + \cdots + \alpha^{\beta_k}\alpha_k,$$

where $\beta > \beta_1 > \beta_2 > \cdots > \beta_k$ and $\alpha_1, \alpha_2, \ldots, \alpha_k < \alpha$. This representation is unique, and any sum of this kind represents an ordinal less than α^β.

Proof. The existence of the representation is already proved. On the other hand, any sum of this kind is isomorphic to an initial segment in $[B \to A]$ (where A and B have order types α and β), and different sums correspond to different initial segments. □

This theorem is a generalization of the positional number system with base α for ordinals less than α^k (see page 90). It allows us to use any ordinal β in place of k.

In fact, we could define $[B \to A]$ as the set of (formal) sums of type

$$\alpha^{\beta_1}\alpha_1 + \alpha^{\beta_2}\alpha_2 + \cdots + \alpha^{\beta_k}\alpha_k$$

(where $\beta > \beta_1 > \cdots > \beta_k$ and $\alpha_1, \ldots, \alpha_k < \alpha$) ordered in a natural way.

Now we can give an explicit description of the ordinals

$$\omega^\omega, \omega^{(\omega^\omega)}, \ldots .$$

The first, ω^ω, is formed by two-level expressions

$$\omega^{b_1}a_1 + \omega^{b_2}a_2 + \cdots + \omega^{b_k}a_k,$$

where a_i and b_i are natural numbers and $b_1 > \cdots > b_k$. If we allow the use of any two-level expression (described above) as $b_1, \ldots, b_k$, then we obtain “three-level” expressions that are ordered as $\omega^{(\omega^\omega)}$. Using three-level expressions as $b_1, \ldots, b_k$, we get four-level expressions, etc. The union of all these sets consists of expressions of any finite height and is a well-ordered set of order type

$$\sup(\omega, \omega^\omega, \omega^{(\omega^\omega)}, \ldots).$$

This order type is denoted by ε_0.

Problem 137. Prove that

$$\varepsilon_0 = \omega + \omega^\omega + \omega^{(\omega^\omega)} + \cdots .$$

Problem 138. Consider the following operation called “total base change” from k to $l > k$ (here k and l are natural numbers). Operation can be applied to any natural number n and is performed as follows. We represent n in a positional number system with base k:

$$n = a_i k^i + a_{i-1} k^{i-1} + \cdots + a_0.$$

Then we represent all exponents $i, i-1, \ldots, 0$ also in positional system with base k, and so on. At the end all coefficients are less than k and all bases are k. Then we replace the base k by the base l and compute the value of this new expression.

(The difference between the “total base change” and the “base change” considered in Problem 129 is that now we use a positional number system with a changed base also for exponents.)

Now let us consider an arbitrary natural number n and apply the following sequence of operations: (total base change 2 → 3)—(subtraction of 1)—(total base change 3 → 4)—(subtraction of 1)—(total base change 4 → 5)—(subtraction of 1)—etc.

Prove that this process, started from any n, will always terminate (i.e., we will come to zero and will not be able to subtract 1).

(*Hint*: Replace all bases by ω and get a decreasing sequence of ordinals less than ε_0. Note that the statement of this problem does not mention ordinals at all; however, they are used in the solution.)

13. Application of ordinals

In most cases ordinals and transfinite induction can be replaced by Zorn’s Lemma. Usually this gives a less intuitive but formally simpler argument. However, in some cases Zorn's Lemma does not help much (or at least it is not easy to use it). In this section we give two examples of this type.

The first example is about Borel sets. For simplicity we consider only Borel sets of real numbers. Let us give the definition.

A family X of sets of real numbers is called a *σ-algebra* if it is closed under complement, countable unions and intersections. (This means that if $A \subset \mathbb{R}$ belongs to X, then $\mathbb{R} \setminus A$ belongs to X; if $A_0, A_1, \ldots$ belong to X, then $A_0 \cup A_1 \cup \cdots$ and $A_0 \cap A_1 \cap \cdots$ belong to X.)

Evidently, the family $P(\mathbb{R})$ of all subsets of $\mathbb{R}$ is a σ-algebra.

Theorem 43. *There exists the least* (*with respect to inclusion*) *σ-algebra that contains all closed intervals* $[a, b]$.

Proof. The formal proof is easy: consider all σ-algebras that contain all closed intervals. (As we have seen, there exists at least one such σ-algebra.) The intersection of all these σ-algebras is a σ-algebra that contains all closed intervals and (evidently) is the smallest algebra with this property. (Note that the intersection of any family of σ-algebras is a σ-algebra; this follows directly from the definition.) □

The elements of the least σ-algebra that contains all intervals are called *Borel sets.*

Problem 139. Prove that all open and all closed subsets of $\mathbb{R}$ are Borel sets. (*Hint*: An open set is a union of all its subsets $[p, q]$, where p, q are rational numbers.)

Problem 140. Prove that the preimage of a Borel set under a continuous function of type $\mathbb{R} \to \mathbb{R}$ is a Borel set.

Problem 141. Let $f_0, f_1, \ldots$ be a sequence of continuous functions of type $\mathbb{R} \to \mathbb{R}$. Prove that the set of all points x such that the sequence $f_0(x), f_1(x), \ldots$ has a limit, is a Borel set.

Borel sets play an important role in the *descriptive set theory.* But our current goal is very limited: we want to show the use of transfinite induction that cannot be easily replaced with Zorn's Lemma, by the example of the following theorem.

Theorem 44. *The family of all Borel sets has continuum cardinality.*

Proof. The class of all Borel sets can be constructed step by step. We start with closed intervals and their complements. In the next step we consider countable unions and intersection of sets already constructed.

Problem 142. Prove that all closed and all open subsets of $\mathbb{R}$ are among them.

Then we consider countable unions and intersections of sets already constructed, and so on.

More formally, let $\mathcal{B}_0 \subset P(\mathbb{R})$ be the family of all closed intervals and their complements. Then we define $\mathcal{B}_i$ by induction: $\mathcal{B}_{i+1}$ is the family of sets that are countable intersections or unions of sets from $\mathcal{B}_i$.

All elements of $\mathcal{B}_i$ are Borel sets (since a countable union or intersection of Borel sets is a Borel set). Is it true that any Borel set belongs to $\mathcal{B}_i$ for some natural i? Not necessarily: consider the sequence of sets $X_i \in \mathcal{B}_i$. The sets X_i are Borel sets; therefore the intersection $\bigcap_i X_i$ is a Borel set. But it may happen that it does not belong to any $\mathcal{B}_i$ for $i = 0, 1, 2, \ldots$.

Thus, we need to continue our construction and consider the class $\mathcal{B}_\omega$ defined as the union of all $\mathcal{B}_i$ for $i = 0, 1, \ldots$; then we consider $\mathcal{B}_{\omega+1}$, $\mathcal{B}_{\omega+2}$, etc. The union of these classes is called $\mathcal{B}_{\omega\cdot 2}$, and the construction goes on.

Here is the formal definition of $\mathcal{B}_\alpha$ for any ordinal α. It uses transfinite recursion. For $\alpha = \beta + 1$ the elements of $\mathcal{B}_\alpha$ are countable unions and intersections of the elements of $\mathcal{B}_\beta$. If α is a limit ordinal (and $\alpha \neq 0$), then $\mathcal{B}_\alpha$ is the union of $\mathcal{B}_\beta$ for all $\beta < \alpha$. (The class $\mathcal{B}_0$ has already been defined.)

The definition easily implies that $\mathcal{B}_\alpha \subset \mathcal{B}_\beta$ if $\alpha < \beta$; therefore the sequence $\mathcal{B}_\alpha$ is increasing. All classes $\mathcal{B}_\alpha$ are closed under complement (it is true for $\mathcal{B}_0$ due to our construction; then use transfinite induction). All elements of all $\mathcal{B}_\alpha$ are Borel sets, since we use only countable unions and intersections, and the class of Borel sets is closed under these operations. (Formally speaking, here we should use transfinite induction again.)

How long should we continue this construction? It turns out that the first uncountable ordinal is large enough.

Let $\aleph_1$ be the least uncountable ordinal. (This is a standard notation.) In other terms, $\aleph_1$ is the family of all countable ordinals ordered by the $<$-relation on ordinals.

Lemma. *The class $\mathcal{B}_{\aleph_1}$ is closed under countable intersections and unions* (*and therefore all Borel sets belong to $\mathcal{B}_{\aleph_1}$*).

Proof of the lemma. Let $B_0, B_1, \ldots$ be a sequence of sets that belong to $\mathcal{B}_{\aleph_1}$. The ordinal $\aleph_1$ is a limit one, hence $\mathcal{B}_{\aleph_1}$ is the union of smaller classes. Therefore each B_i belongs to some $\mathcal{B}_{\alpha_i}$, where α_i is an ordinal less than $\aleph_1$, i.e., a countable (or finite) ordinal. Let $\beta = \sup_i \alpha_i$. Then β is the least upper bound of a countable family of countable ordinals and is countable. (Indeed, view all α_i as initial segments of some greater ordinal, e.g., $\aleph_1$; then β is a countable union of countable sets.)

Now the claim is evident: all B_i are elements of $\mathcal{B}_\beta$; therefore their union (or intersection) belongs to $\mathcal{B}_{\beta+1}$ and to $\mathcal{B}_{\aleph_1}$ (since $\beta + 1$ is a countable ordinal less than $\aleph_1$).

Therefore the class $\mathcal{B}_{\aleph_1}$ is a σ-algebra that contains all closed intervals. The class of Borel sets is the smallest σ-algebra with that property; therefore, all Borel sets are elements of $\mathcal{B}_{\aleph_1}$. The lemma is proved. □

On the other hand, all classes $\mathcal{B}_\alpha$ contain Borel sets only, and therefore the class $\mathcal{B}_{\aleph_1}$ coincides with the class of all Borel sets (and with all subsequent classes $\mathcal{B}_\alpha$ for $\alpha > \aleph_1$).

What is the cardinality of $\mathcal{B}_\alpha$? The class $\mathcal{B}_0$ has continuum cardinality (closed segments are determined by their endpoints). If the class $\mathcal{B}_\alpha$ has continuum cardinality, then the next class $\mathcal{B}_{\alpha+1}$ also has continuum cardinality. Indeed, any element of $\mathcal{B}_{\alpha+1}$ is determined by a sequence of elements of $\mathcal{B}_\alpha$, and $\mathfrak{c}^{\aleph_0} = \mathfrak{c}$.

For a limit ordinal α the class $\mathcal{B}_\alpha$ is the union of preceding classes, and there are only countably many of them, since we consider only ordinals smaller than $\aleph_1$ (i.e., countable ordinals). Recalling that $\mathfrak{c}\aleph_0 = \mathfrak{c}$, we see that $\mathcal{B}_\alpha$ has continuum cardinality for any countable ordinal α. Finally, the class $\mathcal{B}_{\aleph_1}$ is the union of uncountably many preceding classes (namely, $\aleph_1$ classes), but $\aleph_1 \leq \mathfrak{c}$ and therefore $\mathfrak{c}\aleph_1 = \mathfrak{c}$.

Thus, we have proved that $\mathcal{B}_{\aleph_1}$, that is, the class of all Borel sets, has continuum cardinality. □

The traditional definition of Borel hierarchy is slightly different. Usually two classes are considered at the lowest level: open and closed sets. The next level contains two classes: F_σ is the class of countable unions of closed sets, while G_δ is the class of countable intersections of open sets. The next level contains the class of all countable intersections of F_σ-sets and all countable unions of G_δ-sets, etc. This approach is more natural from the topological viewpoint, since closed segments in our definition are chosen somehow arbitrarily. But the difference is not really important.

Problem 143. Prove that the intersection of two F_σ-sets is an F_σ-set, and the union of two G_δ-sets is a G_δ-set. (More generally, the classes F_σ and G_δ, as well as subsequent classes, are closed under finite unions and intersections.)

Problem 144. Prove that all F_σ- and G_δ-sets are elements of the class $\mathcal{B}_2$ (defined as above).

Problem 145. Prove that each set in $\mathcal{B}_2$ differs from some F_σ- or G_δ-set by at most a countable set.

Problem 146. Prove that each set in $\mathcal{B}_3$ is either a countable intersection of F_σ-sets or a countable union of G_δ-sets. Prove similar statements for higher levels of the hierarchy.

Problem 147. Prove that there exists an open set $U \subset \mathbb{R}^2$ such that every open set $V \subset \mathbb{R}$ is among the "vertical sections" $U_x = \{y \mid \langle x, y\rangle \in U\}$ of U. Prove that there exists a G_δ-set $U' \subset \mathbb{R}^2$ such that every G_δ-subset of $\mathbb{R}$ appears among its vertical sections. Prove similar statements for next levels of the hierarchy. (This problem uses the notion of G_δ-subset of $\mathbb{R}^2$ defined in a natural way.)

Problem 148. Prove that there exists a G_δ-set that is not an F_σ-set. Prove that there exists a countable union of G_δ-sets that is not a countable intersection of F_σ-sets etc. (*Hint*: Use the preceding problem.)

Ordinals often appear when we classify the elements of a set by their “ranks”. For example, we can define the rank of an element of a well-founded set as follows:

Theorem 45. *Let X be a well-founded set. There exists a unique function* rk, *defined on X and having ordinals as values, such that*

$$\operatorname{rk}(x) = \min\{\alpha \mid \alpha > \operatorname{rk}(y) \textit{ for all } y < x\}$$

(*for any $x \in X$*).

This theorem implies that rk is the least strictly increasing function on X whose values are ordinals.

Proof. We define a subset $X_\alpha \subset X$ using transfinite recursion (on α): X_α is the set of all $x \in X$ such that all smaller elements (in X) belong to X_β for some $\beta < \alpha$:

$$x \in X_\alpha \;\Leftrightarrow\; (\forall y < x)\,(\exists \beta < \alpha)\,(y \in X_\beta).$$

Note that “$<$” has two different meanings: the ordering of ordinals ($\beta < \alpha$) and the ordering in X ($y < x$).

The definition implies that $X_\beta \subset X_\gamma$ for $\beta < \gamma$. We prove that $X_\alpha = X$ for sufficiently large α. If this is not the case, then $\beta < \gamma$ implies $X_\beta \subsetneq X_\gamma$ (a minimal element of the nonempty set $X \setminus X_\beta$ belongs to X_γ). Therefore the mapping $\alpha \mapsto X_\alpha$ is injective, which is impossible (consider an ordinal that has cardinality greater than $P(X)$; there are too many smaller ordinals).

Now we define $\operatorname{rk}(x)$ as the least α such that $x \in X_\alpha$. If $\operatorname{rk}(x) = \alpha$ and $y < x$, then $\operatorname{rk}(y) < \alpha$. (Indeed, $x \in X_\alpha$ implies that any $y < x$ belongs to X_β for some $\beta < \alpha$ and $\operatorname{rk}(y) \le \beta < \alpha$.) On the other hand, if for some ordinal γ we have $\operatorname{rk}(y) < \gamma$ for all $y < x$, then $\operatorname{rk}(x) \le \gamma$. (Indeed, then any $y < x$ belongs to X_β for $\beta = \operatorname{rk}(y) < \gamma$ and therefore $x \in X_\gamma$ and $\operatorname{rk}(x) \le \gamma$.)

Therefore, the function rk has the required properties. It is obviously unique: if two rank functions differ, consider a minimal point where they differ and come to a contradiction. □

In particular, countable ordinals could be used to classify trees that have no infinite branches. We consider only *rooted trees with* *finite or countable branching* where each vertex has a finite or countable number of sons, and assume that such a tree has no infinite branches (a branch is a sequence of vertices where each vertex is a son of the preceding one).

Formally a tree of this type can be defined as a subset T of the set $\mathbb{N}^*$ (here $\mathbb{N}^*$ is the set of all finite sequences of natural numbers) such that any prefix of any element of T belongs to T. Elements of T are called *tree vertices*. The vertex y is a *son* of a vertex x if x is y without the last element. The vertex y is a *descendant* of x if x is a prefix of y.

An *infinite branch* in a tree T is an infinite sequence of natural numbers such that all its prefixes are vertices of T. If a tree has no infinite branches, the partial ordering

$$y < x \;\Leftrightarrow\; y \text{ is a descendant of } x$$

is well founded, and we can apply Theorem 45 to define ranks for all tree vertices. The rank of the *root vertex* (i.e., the empty sequence) is called the *tree rank* of T.

Theorem 46. (a) *Tree rank of any tree* (*as described above*) *is a countable ordinal.*

(b) *Any countable ordinal is the tree rank of some tree.*

Proof. (a) If the rank of some tree (i.e., the rank of its root) is uncountable, then one of the sons (say, x_1) of the root vertex has an uncountable rank, too. (Indeed, the least upper bound of a countable family of countable ordinals is a countable ordinal: a countable union of countable initial segments is countable.) Then one of the sons of x_1 (say, x_2) has an uncountable rank, etc. We get a sequence $x_1, x_2, \ldots$ of vertices; therefore our tree has an infinite branch.

(b) Let us prove this statement by induction. Let α be the least countable ordinal that is not the rank of any tree. For any smaller ordinal consider the corresponding tree and combine all these trees into one big tree (the roots of all trees become sons of the root of the new tree). This is possible since the set of ordinals that are less than α is countable. The root of the new tree has rank α. □

Problem 149. Consider a tree without infinite branches (as defined above). Assume that each *leaf* (vertex without sons) is labeled by a closed interval in $\mathbb{R}$ or by the complement of a closed interval in $\mathbb{R}$, and each nonleaf vertex is labeled either by $\cap$ or $\cup$ sign. Explain how this tree can be regarded as a representation of a Borel set. (*Hint*: Prove that there exists a unique labeling of vertices by sets that is consistent with the given labeling.) Prove that any Borel set corresponds to some tree.

Labeled trees (as defined in this problem) may be viewed as generalizations of formulas. Indeed, each tree is like an infinite formula that contains closed intervals, their complements and the union/intersection operations with a countable number of operands. (Finite trees correspond to usual finite formulas.) Borel sets are sets that can be represented by infinite formulas of this type.

Problem 150. Prove that the family of all Borel sets has continuum cardinality, using the representation of Borel sets by infinite formulas. (It is possible to eliminate ordinals and well-ordered sets in this proof and use only trees with no infinite branches.)

We end our book with a funny (though may be not so important) application of ordinals and transfinite recursion.

Theorem 47. *There exists a set of points on the plane that has exactly two common points with every line.*

Note that the set of points of any two parallel lines almost satisfies this requirement (for all lines that are not parallel to the given two lines), but it is not so easy to construct a set that has two intersection points with *each* line.

Proof. Let us reformulate the requirements for our set M as follows: (a) no three points of M are on the same line; (b) each line intersects M in at least two points.

We construct the set M using transfinite induction. Let α be the least ordinal of continuum cardinality. (If the Continuum Hypothesis is true, then $\alpha = \aleph_1$, but this is not important for us now.) Then the set of all ordinals that are less than α has continuum cardinality and can be put into a one-to-one correspondence with the set of all lines. Let l_β be the line that corresponds to an ordinal $\beta < \alpha$.

For each $\beta < \alpha$ we construct a set M_β that satisfies condition (a) (no three points are on the same line) and has a two-point intersection with the line l_β and with all lines l_γ for $\gamma < \beta$. This construction is monotone (bigger ordinals give bigger sets), and each point of M_β belongs to the line l_γ for some $\gamma \le \beta$.

How to construct M_β? First we consider the union of sets M_γ for all $\gamma < \beta$ and denote it by T. No three points of T are on the same line (indeed, if these points are from sets M_{γ_1}, M_{γ_2}, and M_{γ_3}, take the greatest ordinal γ among $\gamma_1, \gamma_2, \gamma_3$, and recall that the corresponding set M_γ satisfies (a)).

Now count the points in $l_\beta \cap T$. As we have proved, there are at most two of them. If there are two, everything is fine and we let $M_\beta = T$. If not, we have to add new points to T using points of l_β and trying not to violate condition (a). This means that we are not allowed to use points that are intersections of l_β and "dangerous" lines passing through two existing points of T.

How many dangerous lines are there? The induction hypothesis guarantees that every point of T belongs to the line l_γ for some $\gamma < \beta$. Therefore, the cardinality of T is at most $2\beta = \beta$ (each line provides two points in T), and $T \times T$ also has cardinality at most $\beta^2 = \beta < \mathfrak{c}$. Therefore, the set of dangerous lines has cardinality less than that of the continuum, and l_β has infinitely many points that could be added safely. In fact, we need at most two new points, and we add them to T getting M_β that has a two-element intersection with all the lines up to l_β and at most a two-element intersection with all other lines.

It remains to consider the union of all M_β for all $\beta < \alpha$ to get a desired set X. □

Problem 151. Find an error in the following "proof" that $\aleph_1 \neq \mathfrak{c}$ (the negation of the Continuum Hypothesis):

Assume that $\aleph_1 = \mathfrak{c}$. Then we can define an ordering on $[0, 1]$ that has order type $\aleph_1$ (and has no relation to the usual order). Consider then the characteristic function of this ordering, i.e., function f such that $f(x, y) = 1$ for $x < y$, and $f(x, y) = 0$ for $x \ge y$. (Here "<"

and "$\leq$" stand for the well-ordering, not the usual order.) Then for each x the function $y \mapsto f(x, y)$ equals 1 everywhere except for a countable number of points (since $[0, x]$ is countable); therefore $\int f(x, y)\, dy$ is defined and equals 1 for any x. A similar argument shows that $\int f(x, y)\, dx$ is defined and equals 0 for all y. Therefore,

$$\int \left(\int f(x, y)\, dy \right) dx \neq \int \left(\int f(x, y)\, dx \right) dy,$$

which contradicts the well-known Fubini Theorem.

Bibliography

[1] P. S. Aleksandrov, *Introduction to set theory and general topology*, "Nauka", Moscow, 1977. (Russian)

[2] N. Bourbaki, *Éléments de Mathématique* XXII, *Théorie des ensembles*, Hermann, Paris, 1957.

[3] G. Cantor, *Works in set theory*, Compiled by A. N. Kolmogorov, F. A. Medvedev, and A. P. Yushkevich, "Nauka", Moscow, 1985. (Russian)[1]

[4] P. J. Cohen, *Set theory and the continuum hypothesis*, Benjamin, New York, 1966.

[5] A. A. Fraenkel and Y. Bar-Hillel, *Foundations of set theory*, Studies in Logic and the Foundations of Mathematics, North-Holland, Amsterdam, 1958.

[6] *Handbook of mathematical logic*, Edited by Jon Barwise, with the cooperation of H. J. Keisler, K. Kunen, Y. N. Moschovakis, and A. S. Troelstra, Studies in Logic and the Foundations of Mathematics, Vol. 90, North-Holland, Amsterdam, 1977.

[7] F. Hausdorff, *Grundzüge der Mengenlehre*, Veit, Leipzig, 1914.

[8] T. J. Jech, *Lectures on set theory with particular emphasis on the method of forcing*, Springer-Verlag, Berlin, 1971.

[9] W. Just and M. Weese, *Discovering modern set theory, I. The basics*, Amer. Math. Soc., Providence, RI, 1996.

[1] *Editorial Note.* This collection consists mostly of selected works translated from the complete collection (Georg Cantor, *Gesammelte Abhandlungen mathematischen und philosophischen Inhalts*, Reprint of the 1932 original, Springer-Verlag, Berlin–New York, 1980). More on the content of the Russian book can be found in *Mathematical Reviews*, MR **87g:**01062.

[10] ———, *Discovering modern set theory, II. Set-theoretic tools for every mathematician*, Amer. Math. Soc., Providence, RI, 1997.

[11] A. Kechris, *Classical descriptive set theory*, Springer-Verlag, New York, 1995.

[12] K. Kuratowski and A. Mostowski, *Set theory*, North-Holland, Amsterdam, 1976.

[13] Yu. Manin, *A course in mathematical logic*, Springer-Verlag, New York, 1991.

[14] A. Mostowski, *Constructible sets with applications*, North-Holland, Amsterdam, 1969.

[15] J. R. Shoenfield, *Mathematical logic*, Addison-Wesley, Reading, MA, 1967.

Glossary

Felix BERNSTEIN, Feb. 24, 1878, Halle (Germany) – Dec. 3, 1956, Zurich (Switzerland), 16, 20

Félix Édouard Justin Émile BOREL, Jan. 7, 1871, Saint Affrique, Aveyron, Midi-Pyrénées (France) – Feb. 3, 1956, Paris (France), 100

Luitzen Egbertus Jan BROUWER, Feb. 27, 1881, Overschie (now in Rotterdam, Netherlands) – Dec. 2, 1966, Blaricum (Netherlands), 15

Cesare BURALI-FORTI, Aug. 13, 1861, Arezzo (Italy) – Jan. 21, 1931, Turin (Italy), 84

Georg Ferdinand Ludwig Philipp CANTOR, Mar. 3, 1845, St. Petersburg (Russia) – Jan. 6, 1918, Halle (Germany), 2, 15, 16, 20, 24–26, 29, 68

Paul Joseph COHEN, born Apr. 2, 1934, Long Branch, New Jersey (USA), 11

Julius Wilhelm Richard DEDEKIND, Oct. 6, 1831, Braunschweig (now Germany) – Feb. 12, 1916, Braunschweig (Germany), 13, 15

EUCLID of Alexandria, about 325 (?) BC – about 265 (?) BC, Alexandria (now Egypt), 29

Adolf Abraham Halevi FRAENKEL, Feb. 17, 1891, Munich (Germany) – Oct. 15, 1965, Jerusalem (Israel), 29, 85

Guido FUBINI, Jan. 19, 1879, Venice (Italy) – Jun. 6, 1943, New York (USA), 108

Galileo GALILEI, Feb. 15, 1564, Pisa (now Italy) – Jan. 8, 1642, Arcetri near Florence (now Italy), 17

Kurt GÖDEL, Apr. 28, 1906, Brünn, Austria-Hungary (now Brno, Czech Republic) – Jan. 14, 1978, Princeton (USA), 11

Georg Karl Wilhelm HAMEL, Sep. 12, 1877, Düren, Rheinland (Germany) – Oct. 4, 1954, Berlin (Germany), 69, 76

Charles HERMITE, Dec. 24, 1822, Dieuze, Lorraine (France) – Jan. 14, 1901, Paris (France), 25

David HILBERT, Jan. 23, 1862, Königsberg, Prussia (now Kaliningrad, Russia) – Feb. 14, 1943, Göttingen (Germany), 72

Julius KÖNIG, 1849, Györ (Hungary) – 1913, Budapest (Hungary), 40

Kazimierz KURATOWSKI, Feb. 2, 1896, Warsaw (Poland) – Jun. 18, 1980, Warsaw (Poland), 34

Carl Louis Ferdinand von LINDEMANN, Apr. 12, 1852, Hannover (now Germany) – Mar. 6, 1939, Munich (Germany), 25

Joseph LIOUVILLE, Mar. 24, 1809, Saint-Omer (France) – Sep. 8, 1882, Paris (France), 25

Nikolai Ivanovich LOBACHEVSKY , Dec. 1, 1792, Nizhnii Novgorod (Russia) – Feb. 24, 1856, Kazan (Russia), 29

Sir Isaac NEWTON, Jan. 4, 1643, Woolsthorpe, Lincolnshire (England) – Mar. 31, 1727, London (England), 7

Giuseppe PEANO, Aug. 27, 1858, Cuneo, Piemonte (Italy) – Apr. 20, 1932, Turin (Italy), 16

Frank Plumpton RAMSEY, Feb. 22, 1903, Cambridge, Cambridgeshire (England) – Jan. 19, 1930, London (England), 42

Bertrand Arthur William RUSSELL, May 18, 1872, Ravenscroft, Trelleck, Monmouthshire (Wales, UK) – Feb. 2, 1970, Penrhyndeudraeth, Merioneth (Wales, UK), 28

Friedrich Wilhelm Karl Ernst SCHRÖDER, Nov. 25, 1841, Mannhein (Germany) – Jun. 16, 1902, Karlsruhe (Germany), 20

John von NEUMANN, Dec. 28, 1903, Budapest (Hungary) – Feb. 8, 1957, Washington, D.C. (USA), 83, 84

Norbert WIENER, Nov. 26, 1894, Columbia, Missouri (USA) – Mar. 18, 1964, Stockholm (Sweden), 35

Ernst Friedrich Ferdinand ZERMELO, Jul. 27, 1871, Berlin (Germany) – May 21, 1953, Freiburg im Breisgau (Germany), 29, 66, 69, 85

Max ZORN, Jun. 6, 1906, Hamburg (Germany) – March 9, 1993, Bloomington, Indiana (USA), 74, 99

Index

Titles in This Series

17 **A. Shen and N. K. Vereshchagin,** Basic set theory, 2002

16 **Wolfgang Kühnel,** Differential geometry: curves—surfaces—manifolds, 2002

15 **Gerd Fischer,** Plane algebraic curves, 2001

14 **V. A. Vassiliev,** Introduction to topology, 2001

13 **Frederick J. Almgren, Jr.,** Plateau's problem: An invitation to varifold geometry, 2001

12 **W. J. Kaczor and M. T. Nowak,** Problems in mathematical analysis II: Continuity and differentiation, 2001

11 **Michael Mesterton-Gibbons,** An introduction to game-theoretic modelling, 2000

10 **John Oprea,** The mathematics of soap films: Explorations with Maple®, 2000

9 **David E. Blair,** Inversion theory and conformal mapping, 2000

8 **Edward B. Burger,** Exploring the number jungle: A journey into diophantine analysis, 2000

7 **Judy L. Walker,** Codes and curves, 2000

6 **Gérald Tenenbaum and Michel Mendès France,** The prime numbers and their distribution, 2000

5 **Alexander Mehlmann,** The game's afoot! Game theory in myth and paradox, 2000

4 **W. J. Kaczor and M. T. Nowak,** Problems in mathematical analysis I: Real numbers, sequences and series, 2000

3 **Roger Knobel,** An introduction to the mathematical theory of waves, 2000

2 **Gregory F. Lawler and Lester N. Coyle,** Lectures on contemporary probability, 1999

1 **Charles Radin,** Miles of tiles, 1999

The main notions of set theory (cardinals, ordinals, transfinite induction) are fundamental to all mathematicians, not only to those who specialize in mathematical logic or set-theoretic topology. Basic set theory is generally given a brief overview in courses on analysis, algebra, or topology, even though it is sufficiently important, interesting, and simple to merit its own leisurely treatment.

This book provides just that: a leisurely exposition for a diversified audience. It is suitable for a broad range of readers, from undergraduate students to professional mathematicians who want to finally find out what transfinite induction is and why it is always replaced by Zorn's Lemma.

The text introduces all main subjects of "naive" (nonaxiomatic) set theory: functions, cardinalities, ordered and well-ordered sets, transfinite induction and its applications, ordinals, and operations on ordinals. Included are discussions and proofs of the Cantor–Bernstein Theorem, Cantor's diagonal method, Zorn's Lemma, Zermelo's Theorem, and Hamel bases. With over 150 problems, the book is a complete and accessible introduction to the subject.

ISBN 0-8218-2731-6
9 780821 827314
STML/17

AMS on the Web
www.ams.org