1

Introduction and basic concepts

In this introductory chapter, we first give a sample of the problems and questions to be treated in the book. Then we explain some basic notions and techniques, mostly fundamental and simple ones common to most branches of mathematics. We assume that the reader is already familiar with many of them or has at least heard of them. Thus, we will mostly review the notions, give precise formal definitions, and point out various ways of capturing the meaning of these concepts by diagrams and pictures. A reader preferring a more detailed and thorough introduction to these concepts may refer to the book by Stewart and Tall [8], for instance.

Section 1.1 presents several problems to be studied later on in the book and some thoughts on the importance of mathematical problems and similar things.

Section 1.2 is a review of notation. It introduces some common symbols for operations with sets and numbers, such as $\cup$ for set union or $\sum$ for summation of a sequence of numbers. Most of the symbols are standard, and the reader should be able to go through this section fairly quickly, relying on the index to refresh memory later on.

In Section 1.3, we discuss mathematical induction, an important method for proving statements in discrete mathematics. Here it is sufficient to understand the basic principle; there will be many opportunities to see and practice various applications of induction in subsequent chapters. We will also say few words about mathematical proofs in general.

Section 1.4 recalls the notion of a function and defines special types of functions: injective functions, surjective functions, and bijections. These terms will be used quite frequently in the text.

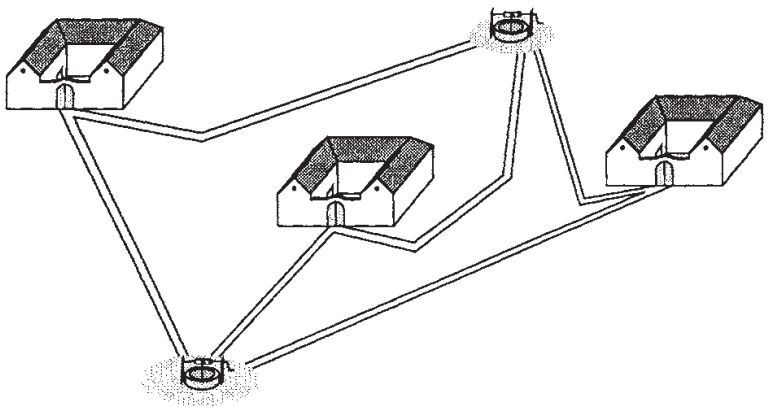
Sections 1.5 through 1.7 deal with relations and with special types of relations, namely equivalences and orderings. These again belong to the truly essential phrases in the vocabulary of mathematics. However, since they are simple general concepts which we have not yet fleshed out by many interesting particular examples, some readers may find them "too abstract"—a polite phrase for "boring"—on the first reading. Such readers may want to skim through these sections and return to them later. For instance, ordered sets (Section 1.7) are only needed for a full understanding of Section 6.2 and for some exercises in this book, but they certainly should be a part of any deeper mathematical education. (When learning a new language, say, it is not very thrilling to memorize the grammatical forms of the verb "to be", but after some time you may find it difficult to speak fluently knowing only "I am" and "he is". Well, this is what we have to do in this chapter: we must review some of the *language* of mathematics.)

1.1 An assortment of problems

Let us look at some of the problems we are going to consider in this book. Here we are going to present them in a popular form, so you may well know some of them as puzzles in recreational mathematics.

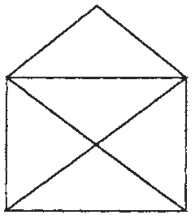
A well-known problem concerns three houses and three wells. Once upon a time, three fair white houses stood in a meadow in a distant kingdom, and there were three wells nearby, their water clean and fresh. All was well, until one day a seed of hatred was sown, fights started among the three households and would not cease, and no reconciliation was in sight. The people in each house insisted that they have three pathways leading from their gate to each well, three pathways which should not cross any of their neighbors' paths. Can they ever find paths that will satisfy everyone and let peace set in?

A solution would be possible if there were only two wells:

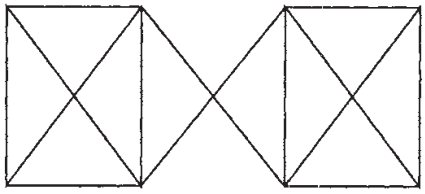


But with three wells, there is no hope (unless these proud men and women would be willing to use tunnels or bridges, which sounds quite unlikely). Can you state this as a mathematical problem and prove that it has no solution?

Essentially, this is a problem about drawing in the plane. Many other problems to be studied in this book can also be formulated in terms of drawing. Can one draw the following picture without lifting the pencil from the paper, drawing each line only once?

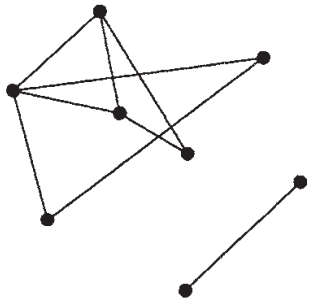


And what about this one?

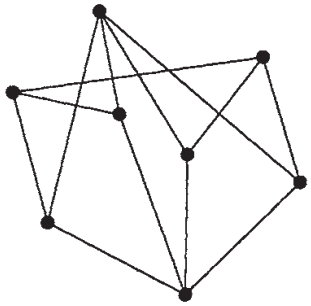


If not, why not? Is there a simple way to distinguish pictures that can be drawn in this way from those that cannot? (And, can you find nice accompanying stories to this problem and the ones below?)

For the subsequent set of problems, draw 8 dots in the plane in such a way that no 3 of them lie on a common line. (The number 8 is quite arbitrary; in general we could consider n such dots.) Connect some pairs of these points by segments, obtaining a picture like the following:

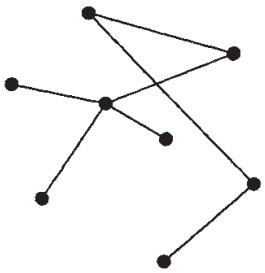


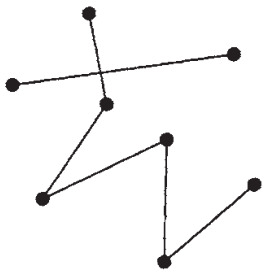
What is the maximum number of segments that can be drawn so that no triangle with vertices at the dots arises? The following picture has 13 segments:



Can you draw more segments for 8 dots with no triangle? Probably you can. But can you prove your result is already the best possible?

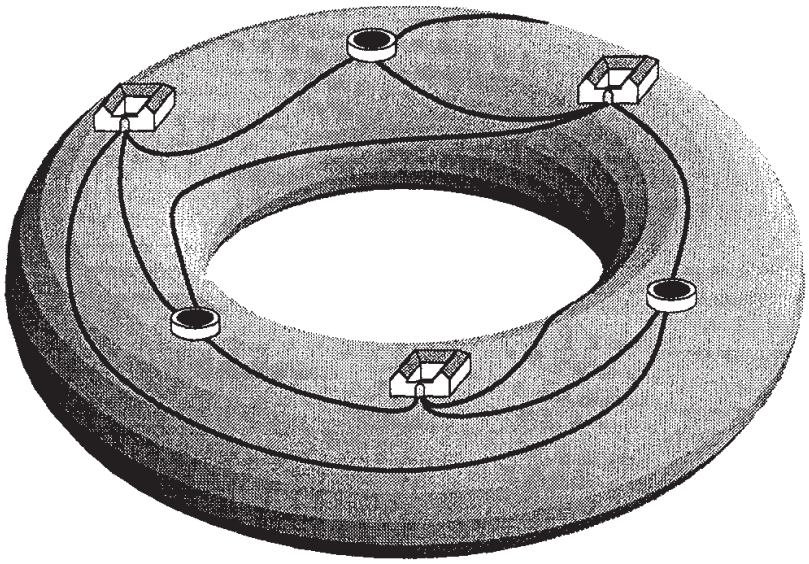
Next, suppose that we want to draw some segments so that any two dots can be connected by a path consisting of the drawn segments. The path is not allowed to make turns at the crossings of the segments, only at the dots, so the left picture below gives a valid solution while the right one doesn't:





What is the minimum number of segments we must draw? How many different solutions with this minimum number of segments are there? And how can we find a solution for which the total length of all the drawn segments is the smallest possible?

All these problems are popular versions of simple basic questions in *graph theory*, which is one of main subjects of this book (treated in Chapters 3, 4, and 5). For the above problems with 8 dots in the plane, it is easily seen that the way of drawing the dots is immaterial; all that matters is which pairs of dots are connected by a segment and which are not. Most branches of graph theory deal with problems which can be pictured geometrically but in which geometry doesn't really play a role. On the other hand, the problem about wells and houses belongs to a "truly" geometric part of graph theory. It is important that the paths should be built in the plane. If the houses and wells were on a tiny planet shaped like a tire-tube then the required paths would exist:



Another important theme of this book is *combinatorial counting*, treated in Chapters 2 and 10. The problems there usually begin with "How many ways are there..." or something similar. One question of this type was mentioned in our "8 dots" series (and it is a nice question—the whole of Chapter 7 is devoted to it). The reader has probably seen lots of such problems; let us add one more. How many ways are there to divide n identical coins into groups? For instance, 4 coins can be divided in 4 ways: $1+1+1+1$ (4 groups of 1 coin each), $1+1+2$, $1+3$, and 4 (all in one group, which is not really a "division" in the sense most people understand it, but what do you expect from mathematicians!). For this problem, we will not be able to give an exact formula; such a formula does exist but its derivation is far beyond the scope of this book. Nonetheless, we will at least derive estimates for the number in question. This number is a function of n, and the estimates will allow us to say "how fast" this function grows, compared to simple and well-known functions like n^2 or 2^n. Such a comparison of complicated functions to simple ones is the subject of the so-called *asymptotic analysis*, which will also be touched on below and which is important in many areas, for instance for comparing several algorithms which solve the same problem.

Although the problems presented may look like puzzles, each of them can be regarded as the starting point of a theory with numerous applications, both in mathematics and in practice.

In fact, distinguishing a good mathematical problem from a bad one is one of the most difficult things in mathematics, and the "quality" of a problem can often be judged only in hindsight, after the problem has been solved and the consequences of its solution mapped. What is a good

problem? It is one whose solution will lead to new insights, methods, or even a whole new fruitful theory. Many problems in recreational mathematics are not good in this sense, although their solution may require considerable skill or ingenuity.

A pragmatically minded reader might also object that the problems shown above are useless from a practical point of view. Why take a whole course about them, a skeptic might say, when I have to learn so many practically important things to prepare for my future career? Objections of this sort are quite frequent and cannot be simply dismissed, if only because the people controlling the funding are often pragmatically minded.

One possible answer is that for each of these puzzle-like problems, we can exhibit an eminently practical problem that is its cousin. For instance, the postal delivery service in a district must deliver mail to all houses, which means passing through each street at least once. What is the shortest route to take? Can it be found in a reasonable time using a supercomputer? Or with a personal computer? In order to understand this postal delivery problem, one should be familiar with simple results about drawing pictures without lifting a pencil from the paper.

Or, given some placement of components of a circuit on a board, is it possible to interconnect them in such a way that the connections go along the surface of the board and do not cross each other? What is the most economical placement of components and connections (using the smallest area of the board, say)? Such questions are typical of VLSI design (designing computer chips and similar things). Having learned about the three-wells problem and its relatives (or, scientifically speaking, about planar graphs) it is much easier to grasp ways of designing the layout of integrated circuits.

These "practical" problems also belong to graph theory, or to a mixture of graph theory and the design of efficient algorithms. This book doesn't provide a solution to them, but in order to comprehend a solution in some other book, or even to come up with a new good solution, one should master the basic concepts first.

We would also like to stress that the most valuable mathematical research was very seldom directly motivated by practical goals. Some great mathematical ideas of the past have only found applications quite recently. Mathematics does have impressive applications (it might be easier to list those human activities where it is not applied than those where it is), but anyone trying to restrict mathematical research to the directly applicable parts would be left with a lifeless fragment with most of the creative power gone.

Exercises are unnecessary in this section. Can you solve some of the problems sketched here, or perhaps all of them? Even if you try and get only partial results or fail completely, it will still be of great

help in reading further.

So what *is* this discrete mathematics they're talking about, the reader may (rightfully) ask? The adjective "discrete" here is an opposite of "continuous". Roughly speaking, objects in discrete mathematics, such as the natural numbers, are clearly separated and distinguishable from each other and we can perceive them individually (like trees in a forest which surrounds us). In contrast, for a typical "continuous" object, such as the set of all points on a line segment, the points are indiscernible (like the trees in a forest seen from a high-flying airplane). We can focus our attention on some individual points of the segment and see them clearly, but there are always many more points nearby that remain indistinguishable and form the totality of the segment.

According to this explanation, such parts of mathematics as algebra or set theory might also be considered "discrete". But in the common usage of the term, discrete mathematics is most often understood as mathematics dealing with finite sets. In many current university curricula, a course on discrete mathematics has quite a wide range, including some combinatorics, counting, graph theory, but also elements of mathematical logic, some set theory, basics from the theory of computing (finite automata, formal languages, elements of computer architecture), and other things. We prefer a more narrowly focussed scope, so perhaps a more descriptive title for this book would be "Invitation to combinatorics and graph theory", covering most of the contents. But the name of the course we have been teaching happened to be "Discrete mathematics" and we decided to stick to it.

1.2 Numbers and sets: notation

Number domains. For the set of all natural numbers, i.e. the set $\{1, 2, 3, \ldots\}$, we reserve the symbol $\mathbf{N}$. The letters n, m, k, i, j, p and possibly some others usually represent natural numbers.

Using the natural numbers, we may construct other well-known number domains: the integers, the rationals, and the reals (and also the complex numbers, but we will seldom hear about them here).

The *integer numbers* or simply *integers* arise from the natural numbers by adding the negative integer numbers and 0. The set of all integers is denoted by $\mathbf{Z}$.

The *rational numbers* are fractions with integer numerator and denominator. This set is usually denoted by $\mathbf{Q}$ but we need not introduce any symbol for it in this book. The construction of the set $\mathbf{R}$ of all *real numbers* is more complicated, and it is treated in introductory courses of mathematical analysis. Famous examples of real numbers which are not rational are numbers such as $\sqrt{2}$, some

important constants like π, and generally numbers whose decimal notation has an infinite and aperiodic sequence of digits following the decimal point, such as $0.12112111211112\ldots$.

The *closed interval* from a to b on the real axis is denoted by $[a, b]$, and the *open interval* with the same endpoints is written as (a, b).

Operations with numbers. Most symbols for operations with numbers, such as $+$ for addition, $\sqrt{\ }$ for square root, and so on, are generally well known. We write *division* either as a fraction, or sometimes with a slash, i.e. either in the form $\frac{a}{b}$ or as a/b.

We introduce two less common functions. For a real number x, the symbol $\lfloor x \rfloor$ is called[1] the *lower integer part of* x (or the *floor function* of x), and its value is the largest integer smaller than or equal to x. Similarly $\lceil x \rceil$, the *upper integer part of* x (or the *ceiling function*), denotes the smallest integer greater than or equal to x. For instance, $\lfloor 0.999 \rfloor = 0$, $\lfloor -0.1 \rfloor = -1$, $\lceil 0.01 \rceil = 1$, $\lceil \frac{17}{3} \rceil = 6$, $\lfloor \sqrt{2} \rfloor = 1$.

Later on, we will introduce some more operations and functions for numbers, which have an important combinatorial meaning and which we will investigate in more detail. Examples are $n!$ and $\binom{n}{k}$.

Sums and products. If $a_1, a_2, \ldots, a_n$ are real numbers, their sum $a_1 + a_2 + \cdots + a_n$ can also be written using the *summation sign* $\sum$, in the form

$$\sum_{i=1}^{n} a_i.$$

This notation somewhat resembles the `FOR` loop in many programming languages. Here are a few more examples:

$$\sum_{j=2}^{5} \frac{1}{2j} = \frac{1}{4} + \frac{1}{6} + \frac{1}{8} + \frac{1}{10}$$

$$\sum_{i=2}^{5} \frac{1}{2j} = \frac{1}{2j} + \frac{1}{2j} + \frac{1}{2j} + \frac{1}{2j} = \frac{2}{j}$$

$$\begin{aligned} \sum_{i=1}^{n} \sum_{j=1}^{n} (i+j) &= \sum_{i=1}^{n} ((i+1) + (i+2) + \cdots + (i+n)) \\ &= \sum_{i=1}^{n} (ni + (1 + 2 + \cdots + n)) \end{aligned}$$

[1] In the older literature, one often finds $[x]$ used for the same function.

$$= n\Big(\sum_{i=1}^{n} i\Big) + n(1+2+\cdots+n)$$
$$= 2n(1+2+\cdots+n).$$

Similarly as sums are written using $\sum$ (which is the capital Greek letter "sigma", from the word *s*um), products may be expressed using the sign $\prod$ (capital Greek "pi"). For example,

$$\prod_{i=1}^{n} \frac{i+1}{i} = \frac{2}{1}\cdot\frac{3}{2}\cdot\ldots\cdot\frac{n+1}{n} = n+1.$$

Sets. Another basic notion we will use is that of a set. Most likely you have already encountered sets in high school (and, thanks to the permanent modernization of the school system, maybe even in elementary school). Sets are usually denoted by capital letters:

$$A, B, \ldots, X, Y, \ldots, M, N, \ldots$$

and so on, and the elements of sets are mostly denoted by lowercase letters: $a, b, \ldots, x, y, \ldots, m, n, \ldots$.

The fact that a set X contains an element x is traditionally written using the symbol $\in$, which is a somewhat stylized Greek letter ε—"epsilon". The notation $x \in X$ is read "x is an element of X", "x belongs to X", "x is in X", and so on.

Let us remark that the concept of a set and the symbol $\in$ are so-called primitive notions. This means that we do not define them using other "simpler" notions (unlike the rational numbers, say, which are defined in terms of the integers). To understand the concept of a set, we rely on intuition (supported by numerous examples) in this book. It turned out at the beginning of the 20th century that if such an intuitive notion of a set is used completely freely, various strange situations, the so-called paradoxes, may arise.[2] In order to exclude such paradoxes, the theory of sets has been rebuilt on a formalized basis, where all properties of sets are derived formally from several precisely formulated basic assumptions (axioms). For the sets used in this text, which are mostly finite, we need not be afraid of any paradoxes, and so we can keep relying on the intuitive concept of a set.

[2]The most famous one is probably Russell's paradox. One possible formulation is about an army barber. An army barber is supposed to shave all soldiers who do not shave themselves—should he, as one of the soldiers, shave himself or not? This paradox can be translated into a rigorous mathematical language and it implies the inconsistency of notions like "the set of all sets".

The set with elements 1, 37, and 55 is written as $\{1, 37, 55\}$. This, and also the notations $\{37, 1, 55\}$ and $\{1, 37, 1, 55, 55, 1\}$, express the same thing. Thus, a multiple occurrence of the same element is ignored: the same element cannot be contained twice in the same set! Three dots (an ellipsis) in $\{2, 4, 6, 8, \ldots\}$ mean "and further similarly, using the same pattern", i.e. this notation means the set of all even natural numbers. The appropriate pattern should be apparent at first sight. For instance, $\{2^1, 2^2, 2^3, \ldots\}$ is easily understandable as the set of all powers of 2, while $\{2, 4, 8, \ldots\}$ may be less clear.

Ordered and unordered pairs. The symbol $\{x, y\}$ denotes the set containing exactly the elements x and y, as we already know. In this particular case, the set $\{x, y\}$ is sometimes called the *unordered pair* of x and y. Let us recall that $\{x, y\}$ is the same as $\{y, x\}$, and if $x = y$, then $\{x, y\}$ is a 1-element set.

We also introduce the notation (x, y) for the *ordered pair* of x and y. For this construct, the order of the elements x and y is important. We thus assume the following:

$$(x, y) = (z, t) \text{ if and only if } x = z \text{ and } y = t. \tag{1.1}$$

Interestingly, the ordered pair can be defined using the notion of unordered pair, as follows:

$$(x, y) = \{\{x\}, \{x, y\}\}.$$

Verify that ordered pairs defined in this way satisfy the condition (1.1). However, in this text it will be simpler for us to consider (x, y) as another primitive notion.

Similarly, we write $(x_1, x_2, \ldots, x_n)$ for the *ordered n-tuple* consisting of elements $x_1, x_2, \ldots, x_n$. A particular case of this convention is writing a point in the plane with coordinates x and y as (x, y), and similarly for points or vectors in higher-dimensional spaces.

Defining sets. More complicated and interesting sets are usually created from known sets using some rule. The sets of all squares of natural numbers can be written

$$\{i^2 \colon i \in \mathbf{N}\}$$

or also

$$\{n \in \mathbf{N} \colon \text{ there exists } k \in \mathbf{N} \text{ such that } k^2 = n\}$$

or using the symbol $\exists$ for "there exists":

$$\{n \in \mathbf{N}\colon \exists k \in \mathbf{N}\,(k^2 = n)\}.$$

Another example is a formal definition of the open interval (a, b) introduced earlier:

$$(a, b) = \{x \in \mathbf{R}\colon a < x < b\}.$$

Note that the symbol (a, b) may mean either the open interval, or also the ordered pair consisting of a and b. These two meanings must (and usually can) be distinguished by the context. This is not at all uncommon in mathematics: many symbols, like parentheses in this case, are used in several different ways. For instance, (a, b) also frequently denotes the greatest common divisor of natural numbers a and b (but we avoid this meaning in this book).

With modern typesetting systems, it is no problem to use any kind of alphabets and symbols including hieroglyphs, so one might think of changing the notation in such cases. But mathematics tends to be rather conservative and the existing literature is vast, and so such notational inventions are usually short-lived.

The empty set. An important set is the one containing no element at all. There is just one such set, and it is customarily denoted by $\emptyset$ and called the *empty set*. Let us remark that the empty set can be an element of another set. For example, $\{\emptyset\}$ is the set containing the empty set as an element, and so it is not the same set as $\emptyset$!

Set systems. In mathematics, we often deal with sets whose elements are other sets. For instance, we can define the set

$$M = \{\{1, 2\}, \{1, 2, 3\}, \{2, 3, 4\}, \{4\}\},$$

whose elements are 4 sets of natural numbers, more exactly 4 subsets of the set $\{1, 2, 3, 4\}$. One meets such sets in discrete mathematics quite frequently. To avoid saying a "set of sets", we use the notions *set system* or *family of sets*. We could thus say that M is a system of sets on the set $\{1, 2, 3, 4\}$. Such set systems are sometimes denoted by calligraphic capital letters, such as $\mathcal{M}$.

However, it is clear that such a distinction using various types of letters cannot always be quite consistent—what do we do if we encounter a set of sets of sets?

The system consisting of all possible subsets of some set X is denoted by the symbol[3] 2^X and called the *power set* of X. Another notation for the power set common in the literature is $\mathcal{P}(X)$.

Set size. A large part of this book is devoted to counting various kinds of objects. Hence a very important notation for us is that for the number of elements of a finite set X. We write it using the same symbol as for the absolute value of a number: $|X|$.

A more general notation for sums and products. Sometimes it is advantageous to use a more general way to write down a sum than using the pattern $\sum_{i=1}^{n} a_i$. For instance,

$$\sum_{i\in\{1,3,5,7\}} i^2$$

means the sum $1^2 + 3^2 + 5^2 + 7^2$. Under the summation sign, we first write the summation variable and then we write out the set of values over which the summation is to be performed. We have a lot of freedom in denoting this set of values. Sometimes it can in part be described by words, as in the following:

$$\sum_{\substack{i:\,1\le i\le 10\\ i \text{ a prime}}} i = 2+3+5+7.$$

Should the set of values for the summation be empty, we define the value of the sum as 0, no matter what appears after the summation sign. For example:

$$\sum_{i=1}^{0}(i+10) = 0, \qquad \sum_{\substack{i\in\{2,4,6,8\}\\ i \text{ odd}}} i^4 = 0.$$

A similar "set notation" can also be employed for products. An empty product, such as $\prod_{j:\,2\le j<1} 2^j$, is always defined as 1 (*not* 0 as for an empty sum).

Operations with sets. Using the primitive notion of set membership, $\in$, we can define further relations among sets and operations with sets. For example, two sets X and Y are considered identical (equal) if they have the same elements. In this case we write $X = Y$.

[3]This notation may look strange, but it is traditional and has its reasons. For instance, it helps to remember that an n-element set has 2^n subsets; see Proposition 2.1.2.

Other relations among sets can be defined similarly. If X, Y are sets, $X \subseteq Y$ (in words: "X is a subset of Y") means that each element of X also belongs to Y.

The notation $X \subset Y$ sometimes denotes that X is a subset of Y but X is not equal to Y. This distinction between $\subseteq$ and $\subset$ is not quite unified in the literature, and some authors may use $\subset$ synonymously with our $\subseteq$.

The notations $X \cup Y$ (the union of X and Y) and $X \cap Y$ (the intersection of X and Y) can be defined as follows:

$$X \cup Y = \{z\colon z \in X \text{ or } z \in Y\}, \qquad X \cap Y = \{z\colon z \in X \text{ and } z \in Y\}.$$

If we want to express that the sets X and Y in the considered union are disjoint, we write the union as $X \dot{\cup} Y$. The expression $X \setminus Y$ is the *difference* of the sets X and Y, i.e. the set of all elements belonging to X but not to Y.

Enlarged symbols $\cup$ and $\cap$ may be used in the same way as the symbols $\sum$ and $\prod$. So, if $X_1, X_2, \ldots, X_n$ are sets, their union can be written

$$\bigcup_{i=1}^{n} X_i \tag{1.2}$$

and similarly for intersection.

Note that this notation is possible (or correct) only because the operations of union and intersection are *associative*; that is, we have

$$X \cap (Y \cap Z) = (X \cap Y) \cap Z$$

and

$$X \cup (Y \cup Z) = (X \cup Y) \cup Z$$

for any triple X, Y, Z of sets. As a consequence, the way of "parenthesizing" the union of any 3, and generally of any n, sets is immaterial, and the common value can be denoted as in (1.2). The operations $\cup$ and $\cap$ are also *commutative*, in other words they satisfy the relations

$$X \cap Y = Y \cap X,$$

$$X \cup Y = Y \cup X.$$

The commutativity and the associativity of the operations $\cup$ and $\cap$ are complemented by their *distributivity*. For any sets X, Y, Z we have

$$X \cap (Y \cup Z) = (X \cap Y) \cup (X \cap Z),$$

$$X \cup (Y \cap Z) = (X \cup Y) \cap (X \cup Z).$$

The validity of these relations can be checked by proving that any element belongs to the left-hand side if and only if it belongs to the right-hand side. The relations can be generalized for an arbitrary number of sets as well. For instance,

$$A \cap \left(\bigcup_{i=1}^{n} X_i \right) = \bigcup_{i=1}^{n} (A \cap X_i);$$

$$A \cup \left(\bigcap_{i=1}^{n} X_i \right) = \bigcap_{i=1}^{n} (A \cup X_i).$$

Such relations can be proved by induction; see Section 1.3 below. Other popular relations for sets are

$$X \setminus (A \cup B) = (X \setminus A) \cap (X \setminus B) \quad \text{and} \quad X \setminus (A \cap B) = (X \setminus A) \cup (X \setminus B)$$

(the so-called *de Morgan laws*), and their generalizations

$$X \setminus \left(\bigcup_{i=1}^{n} A_i \right) = \bigcap_{i=1}^{n} (X \setminus A_i)$$

$$X \setminus \left(\bigcap_{i=1}^{n} A_i \right) = \bigcup_{i=1}^{n} (X \setminus A_i).$$

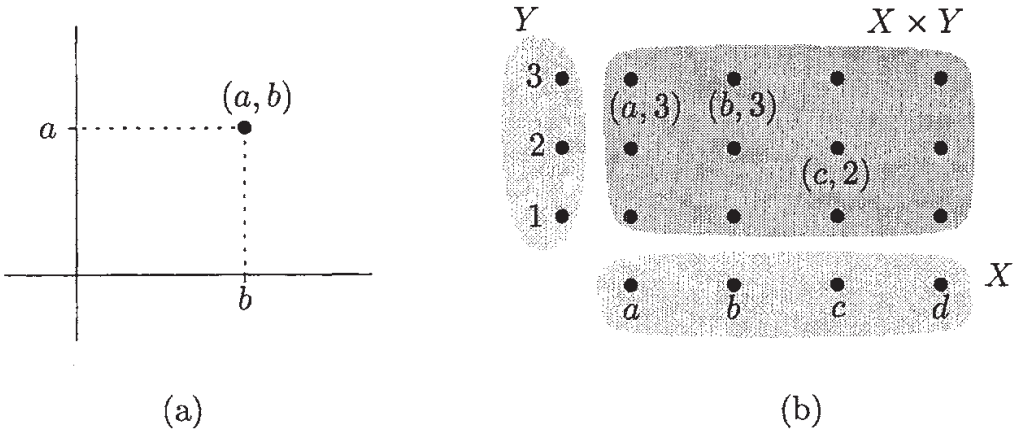
The last operation to be introduced here is the *Cartesian product*, denoted by $X \times Y$, of two sets X and Y. The Cartesian product of X and Y is the set of all ordered pairs of the form (x, y), where $x \in X$ and $y \in Y$. Written formally,

$$X \times Y = \{(x, y) \colon x \in X,\ y \in Y\}.$$

Note that generally $X \times Y$ is not the same as $Y \times X$, i.e. the operation is not commutative.

The name "Cartesian product" comes from a geometric interpretation. If, for instance, $X = Y = \mathbf{R}$, then $X \times Y$ can be interpreted as all points of the plane, since a point in the plane is uniquely described by an ordered pair of real numbers, namely its Cartesian coordinates[4]—the x-coordinate and the y-coordinate (Fig. 1.1(a)). This geometric view can also be useful for Cartesian products of sets whose elements are not numbers (Fig. 1.1(b)).

[4]These are named after their inventor, René Descartes.



(a) (b)

Fig. 1.1 Illustrating the Cartesian product: (a) $\mathbf{R} \times \mathbf{R}$; (b) $X \times Y$ for finite sets X, Y.

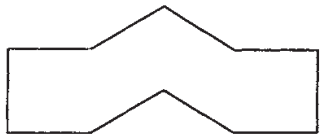
The Cartesian product of a set X with itself, i.e. $X \times X$, may also be denoted by X^2.

Exercises

1. Which of the following formulas are correct?
   (a) $\lfloor \frac{(n+1)^2}{2} \rfloor = \lfloor \frac{n^2}{2} \rfloor + n$,
   (b) $\lfloor \frac{n+k}{2} \rfloor = \lfloor \frac{n}{2} \rfloor + \lfloor \frac{k}{2} \rfloor$,
   (c) $\lceil (\lfloor x \rfloor) \rceil = \lceil x \rceil$ (for a real number x),
   (d) $\lceil (\lfloor x \rfloor + \lfloor y \rfloor) \rceil = \lfloor x \rfloor + \lfloor y \rfloor$.
2. *Prove that the equality $\lfloor \sqrt{x} \rfloor = \lfloor \sqrt{\lfloor x \rfloor} \rfloor$ holds for any positive real number x.
3. (a) Define a "parenthesizing" of a union of n sets $\bigcup_{i=1}^{n} X_i$. Similarly, define a "parenthesizing" of a sum of n numbers $\sum_{i=1}^{n} a_i$.
   (b) Prove that any two parenthesizings of the intersection $\bigcap_{i=1}^{n} X_i$ yield the same result.
   (c) How many ways are there to parenthesize the union of 4 sets $A \cup B \cup C \cup D$?
   (d) **Try to derive a formula or some other way to count the number of ways to parenthesize the union of n sets $\bigcup_{i=1}^{n} X_i$.
4. True or false? If $2^X = 2^Y$ holds for two sets X and Y, then $X = Y$.
5. Is a "cancellation" possible for the Cartesian product? That is, if $X \times Y = X \times Z$ holds for some sets X, Y, Z, does it necessarily follow that $Y = Z$?
6. Prove that for any two sets A, B we have
$$(A \setminus B) \cup (B \setminus A) = (A \cup B) \setminus (A \cap B).$$

7. *Consider the numbers $1, 2, \ldots, 1000$. Show that among any 501 of them, two numbers exist such that one divides the other one.

8. In this problem, you can test your ability to discover simple but "hidden" solutions. Divide the following figure into 7 parts, all of them congruent (they only differ by translation, rotation, and possibly by a mirror reflection). All the bounding segments in the figure have length 1, and the angles are 90, 120, and 150 degrees.



1.3 Mathematical induction and other proofs

Let us imagine that we want to calculate, say, the sum $1 + 2 + 2^2 + 2^3 + \cdots + 2^n = \sum_{i=0}^{n} 2^i$ (and that we can't remember a formula for the sum of a geometric progression). We suspect that one can express this sum by a nice general formula valid for all the n. By calculating numerical values for several small values of n, we can guess that the desired formula will most likely be $2^{n+1} - 1$. But even if we verify this for a million specific values of n with a computer, this is still no proof. The million-and-first number might, in principle, be a counterexample. The correctness of the guessed formula for all n can be proved by so-called *mathematical induction*. In our case, we can proceed as follows:

1. The formula $\sum_{i=0}^{n} 2^i = 2^{n+1} - 1$ holds for $n = 1$, as one can check directly.
2. Let us suppose that the formula holds for some value $n = n_0$. We prove that it also holds for $n = n_0 + 1$. Indeed, we have

$$\sum_{i=0}^{n_0+1} 2^i = \left(\sum_{i=0}^{n_0} 2^i\right) + 2^{n_0+1}.$$

The sum in parentheses equals $2^{n_0+1} - 1$ by our assumption (the validity for $n = n_0$). Hence

$$\sum_{i=0}^{n_0+1} 2^i = 2^{n_0+1} - 1 + 2^{n_0+1} = 2 \cdot 2^{n_0+1} - 1 = 2^{n_0+2} - 1.$$

This is the required formula for $n = n_0 + 1$.

This establishes the validity of the formula for an arbitrary n: by step 1, the formula is true for $n = 1$, by step 2 we may thus infer it is also true for $n = 2$ (using step 2 with $n_0 = 1$), then, again by step 2, the formula holds for $n = 3\ldots$, and in this way we can reach any natural number. Note that this argument only works because the value of n_0 in step 2 was quite arbitrary. We have made the step from n_0 to $n_0 + 1$, where any natural number could equally well appear as n_0.

Step 2 in this type of proof is called the *inductive step*. The assumption that the statement being proved is already valid for some value $n = n_0$ is called the *inductive hypothesis*.

One possible general formulation of the principle of mathematical induction is the following:

1.3.1 Proposition. *Let X be a set of natural numbers with the following properties:*

(i) The number 1 belongs to X.

(ii) If some natural number n is an element of X, then the number $n + 1$ belongs to X as well.

Then X is the set of all natural numbers ($X = \mathbf{N}$).

In applications of this scheme, X would be the set of all numbers n such that the statement being proved, $S(n)$, is valid for n.

The scheme of a proof by mathematical induction has many variations. For instance, if we need to prove some statement for all $n \geq 2$, the first step of the proof will be to check the validity of the statement for $n = 2$. As an inductive hypothesis, we can sometimes use the validity of the statement being proved not only for $n = n_0$, but for all $n \leq n_0$, and so on; these modifications are best mastered by examples.

Mathematical induction can either be regarded as a basic property of natural numbers (an axiom, i.e. something we take for granted without a proof), or be derived from the following other basic property (axiom): *Any nonempty subset of natural numbers possesses a smallest element.* This is expressed by saying that the usual ordering of natural numbers by magnitude is a *well-ordering*. In fact, the principle of mathematical induction and the well-ordering property are equivalent to each other,[5] and either one can be taken as a basic axiom for building the theory of natural numbers.

[5] Assuming that each natural number $n > 1$ has a unique predecessor $n - 1$.

Proof of Proposition 1.3.1 from the well-ordering property. For contradiction, let us assume that a set X satisfies both (i) and (ii), but it doesn't contain all natural numbers. Among all natural numbers n not lying in X, let us choose the smallest one and denote it by n_0. By condition (i) we know that $n_0 > 1$, and since n_0 was the smallest possible, the number $n_0 - 1$ is an element of X. However, using (ii) we get that n_0 is an element of X, which is a contradiction. □

Let us remark that this type of argument (saying "Let n_0 be the smallest number violating the statement we want to prove" and deriving a contradiction, namely that a yet smaller violating number must exist) sometimes replaces mathematical induction. Both ways, this one and induction, essentially do the same thing, and it depends on the circumstances or personal preferences which one is actually used.

We will use mathematical induction quite often. It is one of our basic proof methods, and the reader can thus find many examples and exercises on induction in subsequent chapters.

Mathematical proofs and not-quite proofs. Mathematical proof is an amazing invention. It allows one to establish the truth of a statement beyond any reasonable doubt, even when the statement deals with a situation so complicated that its truth is inaccessible to direct evidence. Hardly anyone can see directly that no two natural numbers m, n exist such that $\frac{m}{n} = \sqrt{2}$ and yet we can trust this fact completely, because it can be proved by a chain of simple logical steps.

Students often don't like proofs, even students of mathematics. One reason might be that they have never experienced satisfaction from understanding an elegant and clever proof or from making a nice proof by themselves. One of our main goals is to help the reader to acquire the skill of rigorously proving simple mathematical statements.

A possible objection is that most students will never need such proofs in their future jobs. We believe that learning how to prove mathematical theorems helps to develop useful habits in thinking, such as working with clear and precise notions, exactly formulating thoughts and statements, and not overlooking less obvious possibilities. For instance, such habits are invaluable for writing software that doesn't crash every time the circumstances become slightly non-standard.

The art of finding and writing proofs is mostly taught by examples,[6] by showing many (hopefully) correct and "good" proofs to the

[6] We will not even try to say what a proof is and how to do one!

student and by pointing out errors in the student's own proofs. The latter "negative" examples are very important, and since a book is a one-way communication device, we decided to include also a few negative examples in this book, i.e. students' attempts at proofs with mistakes which are, according to our experience, typical. These intentionally wrong proofs are presented in a special font like this. In the rest of this section, we discuss some common sources of errors. (We hasten to add that types of errors in proofs are as numerous as grains of sand, and by no means do we want to attempt any classification.)

One quite frequent situation is where the student doesn't understand the problem correctly. There may be subtleties in the problem's formulation which are easy to overlook, and sometimes a misunderstanding isn't the student's fault at all, since the author of the problem might very well have failed to see some double meaning. The only defense against this kind of misunderstanding is to pay the utmost attention to reading and understanding a problem before trying to solve it. Do a preliminary check: does the problem make sense in the way you understand it? Does it have a suspiciously trivial solution? Could there be another meaning?

With the current abundance of calculators and computers, errors are sometimes caused by the uncritical use of such equipment. Asked how many zeros does the decimal notation of the number $50! = 50 \cdot 49 \cdot 48 \cdot \ldots \cdot 1$ end with, a student answered 60, because a pocket calculator with an 8-digit display shows that $50! = 3.04140 \cdot 10^{64}$. Well, a more sophisticated calculator or computer programmed to calculate with integers with arbitrarily many digits would solve this problem correctly and calculate that

$$50!=30414093201713378043612608166064768844377641568960512000000000000$$

with 12 trailing zeros. Several software systems can even routinely solve such problems as finding a formula for the sum $1^2 \cdot 2^1 + 2^2 \cdot 2^2 + 3^2 \cdot 2^3 + \cdots + n^2 2^n$, or for the number of binary trees on n vertices (see Section 10.4). But even programmers of such systems can make mistakes and so it's better to double-check such results. Moreover, the capabilities of these systems are very limited; artificial intelligence researchers will have to make enormous progress before they can produce computers that can discover and prove a formula for the number of trailing zeros of $n!$, or solve a significant proportion of the exercises in this book, say.

Next, we consider the situation where a proof has been written down but it has a flaw, although its author believes it to be satisfactory.

In principle, proofs can be written down in such detail and in such a formal manner that they can be checked automatically by a computer. If such a completely detailed and formalized proof is wrong, some step has to be clearly false, but the catch is that formalizing proofs completely is very laborious and impractical. All textbook proofs and problem solutions are presented somewhat informally.

While some informality may be necessary for a reasonable presentation of a proof, it may also help to hide errors. Nevertheless, a good rule for writing and checking proofs is that *every statement in a correct proof should be literally true.* Errors can often be detected by isolating a specific false statement in the proof, a mistake in calculation, or a statement that makes no sense ("Let ℓ_1, ℓ_2 be two arbitrary lines in the 3-dimensional space, and let ρ be a plane containing both of them..." etc.). Once detected and brought out into the light, such errors become obvious to (almost) everyone. Still, they are frequent. If, while trying to come up with a proof, one discovers an idea seemingly leading to a solution and shouts "This must be IT!", caution is usually swept aside and one is willing to write down the most blatant untruths. (Unfortunately, the first idea that comes to mind is often nonsense, rather than "it", at least as far as the authors' own experience with problem solving goes.)

A particularly frequent mistake, common perhaps to all mathematicians of the world, is a *case omission.* The proof works for some objects it should deal with, but it fails in some cases the author overlooked. Such a case analysis is mostly problem specific, but one keeps encountering variations on favorite themes. Dividing an equation by $x - y$ is only allowed for $x \neq y$, and the $x = y$ case must be treated separately. An intersection of two lines in the plane can only be used in a proof if the lines are not parallel. Deducing $a^2 > b^2$ from $a > b$ may be invalid if we know nothing about the sign of a and b, and so on and so on.

Many proofs created by beginners are wrong because of a *confused application of theorems.* Something seems to follow from a theorem presented in class or in a textbook, say, but in reality the theorem says something slightly different, or some of its assumptions don't hold. Since we have covered no theorems worth mentioning so far, let

us give an artificial geometric example: "Since ABC is an isosceles triangle with the sides adjacent to A having equal length, we have $|AB|^2 + |AC|^2 = |BC|^2$ by the theorem of Pythagoras." Well, wasn't there something about a right angle in Pythagoras' theorem?

A rich source of errors and misunderstandings is *relying on unproved statements.*

Many proofs, including correct and even textbook ones, contain unproved statements intentionally, marked by clauses like "obviously...". In an honest proof, the meaning of such clauses should ideally be "I, the author of this proof, can see how to prove this rigorously, and since I consider this simple enough, I trust that you, my reader, can also fill in all the details without too much effort". Of course, in many mathematical papers, the reader's impression about the author's thinking is more in the spirit of "I can see it somehow since I've been working on this problem for years, and if you can't it's your problem". Hence omitting parts of proofs that are "clear" is a highly delicate social task, and one should always be very careful with it. Also, students shouldn't be surprised if their teacher insists that such an "obvious" part be proved in detail. After all, what would be a better hiding place for errors in a proof than in the parts that are missing?

A more serious problem concerns parts of a proof that are omitted unconsciously. Most often, the statement whose proof is missing is not even formulated explicitly.[7] For a teacher, it may be a very challenging task to convince the proof's author that something is wrong with the proof, especially when the unproved statement is actually true.

One particular type of incomplete proof, fairly typical of students' proofs in discrete mathematics, could be labeled as *mistaking the particular for the general.* To give an example, let us consider the following Mathematical Olympiad problem:

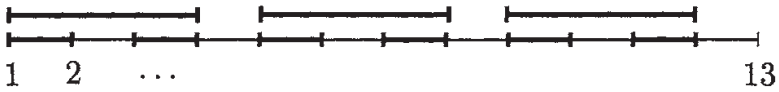
1.3.2 Problem. Let $n > 1$ be an integer. Let M be a set of closed intervals. Suppose that the endpoints u, v of each interval $[u, v] \in M$ are natural numbers satisfying $1 \leq u < v \leq n$, and, moreover, for any two distinct intervals $I, I' \in M$, one of the following possibilities occurs: $I \cap I' = \emptyset$, or $I \subset I'$, or $I' \subset I$ (i.e. two intervals must not partially overlap). Prove that $|M| \leq n - 1$.

An insufficient proof attempt. In order to construct an M as large as possible, we first insert as many unit-length intervals as possible, as in

[7] Even proofs by the greatest mathematicians of the past suffer from such incompleteness, partly because the notion of a proof has been developing over the ages (towards more rigor, that is).

the following figure:

These $\lfloor n/2 \rfloor$ intervals are all disjoint. Now any other interval in M must contain at least two of these unit intervals (or, for n odd, possibly the last unit interval plus the point that remains). Hence, to get the maximum number of intervals, we put in the next "layer" of shortest possible intervals, as illustrated below:



We continue in this manner, adding one layer after another, until we finally add the last layer consisting of the whole interval $[1, n]$:

It remains to show that the set M created in this way has at most $n-1$ intervals. We note that every interval I in the kth layer contains a point of the form $i+\frac{1}{2}$, $1 \leq i \leq n-1$, that was not contained in any interval of the previous layers, because the space between the two intervals in the previous layer was not covered before adding the kth layer. Therefore, $|M| \leq n-1$ as claimed. □

This "proof" looks quite clever (after all, the way of counting the intervals in the particular M constructed in the proof is quite elegant). So what's wrong with it? Well, we have shown that *one particular* M satisfies $|M| \leq n-1$. The argument tries to make the impression of showing that this particular M is the worst possible case, i.e. that no other M may have more intervals, but in reality it doesn't prove anything like that! For instance, the first step seems to argue that an M with the maximum possible number of intervals should contain $\lfloor n/2 \rfloor$ unit-length intervals. But this is not true, as is witnessed by $M = \{[1,2], [1,3], [1,4], \ldots, [1,n]\}$. Saving the "proof" above by justifying its various steps seems more difficult than finding another, correct, proof. Although the demonstrated "proof" contains some useful hints (the counting idea at the end of the proof can in fact be made to work for any M), it's still quite far from a valid solution.

The basic scheme of this "proof", apparently a very tempting one, says "this object X must be the worst one", and then proves that this particular X is OK. But the claim that nothing can be worse than X is not substantiated (although it usually looks plausible that by constructing this X, we "do the worst possible thing" concerning the statement being proved).

Another variation of "mistaking the particular for the general" often appears in proofs by induction, and is shown in several examples in Sections 4.1 and 5.3.

Exercises

1. Prove the following formulas by mathematical induction:

   (a) $1 + 2 + 3 + \cdots + n = n(n+1)/2$

   (b) $\sum_{i=1}^{n} i \cdot 2^i = (n-1)2^{n+1} + 2$.

2. The numbers $F_0, F_1, F_2, \ldots$ are defined as follows (this is a definition by mathematical induction, by the way): $F_0 = 0$, $F_1 = 1$, $F_{n+2} = F_{n+1} + F_n$ for $n = 0, 1, 2, \ldots$. Prove that for any $n \geq 0$ we have $F_n \leq ((1+\sqrt{5})/2)^{n-1}$ (see also Section 10.3).

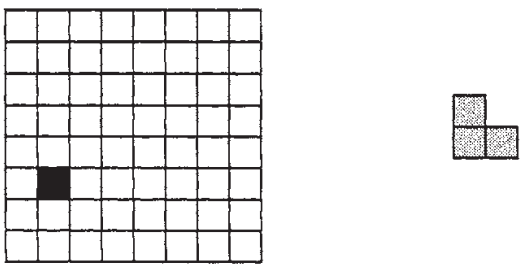
3. (a) Let us draw n lines in the plane in such a way that no two are parallel and no three intersect in a common point. Prove that the plane is divided into exactly $n(n+1)/2 + 1$ parts by the lines.

   (b) *Similarly, consider n planes in the 3-dimensional space in general position (no two are parallel, any three have exactly one point in common, and no four have a common point). What is the number of regions into which these planes partition the space?

4. Prove *de Moivre's theorem* by induction: $(\cos\alpha + \mathrm{i}\sin\alpha)^n = \cos(n\alpha) + \mathrm{i}\sin(n\alpha)$. Here i is the imaginary unit.

5. In ancient Egypt, fractions were written as sums of fractions with numerator 1. For instance, $\frac{3}{5} = \frac{1}{2} + \frac{1}{10}$. Consider the following algorithm for writing a fraction $\frac{m}{n}$ in this form ($1 \leq m < n$): write the fraction $\frac{1}{\lceil n/m \rceil}$, calculate the fraction $\frac{m}{n} - \frac{1}{\lceil n/m \rceil}$, and if it is nonzero repeat the same step. Prove that this algorithm always finishes in a finite number of steps.

6. *Consider a $2^n \times 2^n$ chessboard with one (arbitrarily chosen) square removed, as in the following picture (for $n = 3$):



   Prove that any such chessboard can be tiled without gaps or overlaps by L-shapes consisting of 3 squares each.

7. Let $n \geq 2$ be a natural number. We consider the following game. Two players write a sequence of 0s and 1s. They start with an empty line and alternate their moves. In each move, a player writes 0 or 1 to the end of the current sequence. A player loses if his digit completes a block of n consecutive digits that repeats itself in the sequence for the second time (the two occurrences of the block may overlap). For instance, for $n = 4$, a sequence produced by such a game may look as follows: 00100001101011110011 (the second player lost by the last move because 0011 is repeated).

   (a) Prove that the game always finishes after finitely many steps.

   (b) *Suppose that n is odd. Prove that the second player (the one who makes the second move) has a winning strategy.

   (c) *Show that for $n = 4$, the first player has a winning strategy. Unsolved question: Can you determine who has a winning strategy for some even $n > 4$?

8. *On an infinite sheet of white graph paper (a paper with a square grid), n squares are colored black. At moments $t = 1, 2, \ldots$, squares are recolored according to the following rule: each square gets the color occurring at least twice in the triple formed by this square, its top neighbor, and its right neighbor. Prove that after the moment $t = n$, all squares are white.

9. At time 0, a particle resides at the point 0 on the real line. Within 1 second, it divides into 2 particles that fly in opposite directions and stop at distance 1 from the original particle. Within the next second, each of these particles again divides into 2 particles flying in opposite directions and stopping at distance 1 from the point of division, and so on. Whenever particles meet they annihilate (leaving nothing behind). How many particles will there be at time $2^{11} + 1$?

10. *Let $M \subseteq \mathbf{R}$ be a set of real numbers, such that any nonempty subset of M has a smallest number and also a largest number. Prove that M is necessarily finite.

11. We will prove the following statement by mathematical induction: *Let $\ell_1, \ell_2, \ldots, \ell_n$ be $n \geq 2$ distinct lines in the plane, no two of which are parallel. Then all these lines have a point in common.*

    1. For $n = 2$ the statement is true, since any 2 nonparallel lines intersect.

    2. Let the statement hold for $n = n_0$, and let us have $n = n_0 + 1$ lines $\ell_1, \ldots, \ell_n$ as in the statement. By the inductive hypothesis, all these lines but the last one (i.e. the lines $\ell_1, \ell_2, \ldots, \ell_{n-1}$) have some point in common; let us denote this point by x. Similarly the $n - 1$ lines $\ell_1, \ell_2, \ldots, \ell_{n-2}, \ell_n$ have a point in common; let us denote it by y. The line ℓ_1 lies in both groups, so it contains both x and y. The same is true for the line ℓ_{n-2}. Now ℓ_1 and ℓ_{n-2} intersect at a single point only, and

so we must have $x = y$. Therefore all the lines $\ell_1, \ldots, \ell_n$ have a point in common, namely the point x.

Something must be wrong. What is it?

12. Let $n_1, n_2, \ldots, n_k$ be natural numbers, each of them at least 1, and let $n_1 + n_2 + \cdots + n_k = n$. Prove that $n_1^2 + n_2^2 + \cdots + n_k^2 \leq (n-k+1)^2 + k - 1$.

    "Solution": In order to make $\sum_{i=1}^{k} n_i^2$ as large as possible, we must set all the n_i but one to 1. The remaining one is therefore $n - k + 1$, and in this case the sum of squares is $(n - k + 1)^2 + k - 1$.

    Why isn't this a valid proof? *Give a correct proof.

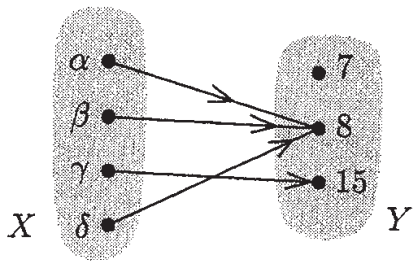
13. *Give a correct proof for Problem 1.3.2.

14. *Let $n > 1$ and k be given natural numbers. Let $I_1, I_2, \ldots, I_m$ be closed intervals (not necessarily all distinct), such that for each interval $I_j = [u_j, v_j]$, u_j and v_j are natural numbers with $1 \leq u_j < v_j \leq n$, and, moreover, no number is contained in more than k of the intervals $I_1, \ldots, I_m$. What is the largest possible value of m?

1.4 Functions

The notion of a function is a basic one in mathematics. It took a long time for today's view of functions to emerge. For instance, around the time when differential calculus was invented, only real or complex functions were considered, and an "honest" function had to be expressed by some formula, such as $f(x) = x^2 + 4$, $f(x) = \sqrt{\sin(x/\pi)}$, $f(x) = \int_0^x (\sin t/t)\mathrm{d}t$, $f(x) = \sum_{n=0}^{\infty}(x^n/n!)$, and so on. By today's standards, a real function may assign to each real number an arbitrary real number without any restrictions whatsoever, but this is a relatively recent invention.

Let X and Y be some quite arbitrary sets. Intuitively, a function f is "something" assigning a unique element of Y to each element of X. To depict a function, we can draw the sets X and Y, and draw an arrow from each element $x \in X$ to the element $y \in Y$ assigned to it:



Note that each element of X must have exactly one outgoing arrow, while the elements of Y may have none, one, or several ingoing arrows.

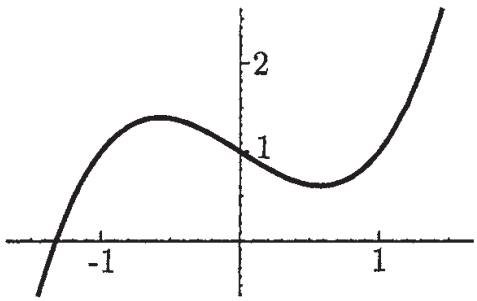
Instead of saying that a function is "something", it is better to define it using objects we already know, namely sets and ordered pairs.

1.4.1 Definition. *A function f from a set X into a set Y is a set of ordered pairs (x, y) with $x \in X$ and $y \in Y$ (in other words, a subset of the Cartesian product $X \times Y$), such that for any $x \in X$, f contains exactly one pair with first component x.*

Of course, an ordered pair (x, y) being in f means just that the element x is assigned the element y. We then write $y = f(x)$, and we also say that f *maps* x *to* y or that y is the *image* of x.

For instance, the function depicted in the above figure consists of the ordered pairs $(\alpha, 8)$, $(\beta, 8)$, $(\gamma, 15)$ and $(\delta, 8)$.

A function, as a subset of the Cartesian product $X \times Y$, is also drawn using a *graph*. We depict the Cartesian product as in Fig. 1.1, and then we mark the ordered pairs belonging to the function. This is perhaps the most usual way used in high school or in calculus. The following figure shows a graph of the function $f: \mathbf{R} \to \mathbf{R}$ given by $f(x) = x^3 - x + 1$:



The fact that f is a function from a set X into a set Y is written as follows:

$$f: X \to Y\,.$$

And the fact that the function f assigns some element y to an element x can also be written

$$f: x \mapsto y.$$

We could simply write $y = f(x)$ instead. So why this new notation? The symbol $\mapsto$ is advantageous when we want to speak about some function without naming it. (Those who have programmed in LISP, Mathematica, or a few other programming languages might recall the existence of unnamed functions in these languages.) For instance, it is not really correct to say "consider the function x^2", since we do not say what the variable is. In this particular case, one can be reasonably sure that we mean the function assigning x^2 to each real number x, but if we

say "consider the function $zy^2 + 5z^3y$", it is not clear whether we mean the dependence on y, on z, or on both. By writing $y \mapsto zy^2 + 5z^3y$, we indicate that we want to study the dependence on y, treating z as some parameter.

Instead of "function", the words "mapping" or "map" are used with the same meaning.[8]

Sometimes we also write $f(X)$ for the set $\{f(x)\colon x \in X\}$ (the set of those elements of Y that are images of something). Also other terms are usually introduced for functions. For example, X is called the *domain* and Y is the *range*, etc., but here we try to keep the terminology and formalism to a minimum.

We definitely need to mention that functions can be composed.

1.4.2 Definition (Composition of functions). *If $f\colon X \to Y$ and $g\colon Y \to Z$ are functions, we can define a new function $h\colon X \to Z$ by*

$$h(x) = g(f(x))$$

for each $x \in X$. In words, to find the value of $h(x)$, we first apply f to x and then we apply g to the result.

The function h (check that h is indeed a function) is called the composition of the functions g and f *and it is denoted by $g \circ f$. We thus have*

$$\big(g \circ f\big)(x) = g(f(x))$$

for each $x \in X$.

The composition of functions is associative but not commutative. For example, if $g \circ f$ is well defined, $f \circ g$ need not be. In order that two functions can be composed, the "middle set" must be the same.

Composing functions can get quite exciting. For example, consider the mapping $f\colon \mathbf{R}^2 \to \mathbf{R}^2$ (i.e. mapping the plane into itself) given by

$$f\colon (x, y) \mapsto \Big(\sin(ax) + b\sin(ay), \sin(cx) + d\sin(cy)\Big)$$

with $a = 2.879879$, $b = 0.765145$, $c = -0.966918$, $d = 0.744728$. Except for the rather hairy constants, this doesn't look like a very complicated function. But if one takes the initial point $p = (0.1, 0.1)$ and plots

[8] In some branches of mathematics, the word "function" is reserved for function into the set of real or complex numbers, and the word mapping is used for functions into arbitrary sets. For us, the words "function" and "mapping" will be synonymous.

Fig. 1.2 The "King's Dream" fractal (formula taken from the book by C. Pickover: *Chaos in Wonderland*, St Martin's Press, New York 1994).

the first several hundred thousand or million points of the sequence p, $f(p)$, $f(f(p))$, $f(f(f(p)))$,..., a picture like Fig. 1.2 emerges.[9] This is one of the innumerable species of the so-called *fractals*. There seems to be no universally accepted mathematical definition of a fractal, but fractals are generally understood as complicated point sets defined by iterations of relatively simple mappings. The reader can find colorful and more sophisticated pictures of various fractals in many books on the subject or download them from the Internet. Fractals can be not only pleasant to the eye (and suitable for killing an unlimited amount of time by playing with them on a personal computer) but also important for describing various phenomena in nature.

After this detour, let us return to the basic definitions concerning functions.

1.4.3 Definition (Important special types of functions). *A function $f: X \to Y$ is called*

- a one-to-one function *if $x \neq y$ implies $f(x) \neq f(y)$,*
- a function onto *if for every $y \in Y$ there exists $x \in X$ satisfying $f(x) = y$, and*

[9] To be quite honest, the way such pictures are generated by a computer is actually by iterating an *approximation* to the mapping given by the formula, because of limited numerical precision.

- a bijective function, *or* bijection, *if f is one-to-one and onto.*

A one-to-one function is also called an *injective function* or an *injection*, and a function onto is also called a *surjective function* or a *surjection.*

In a pictorial representation of functions by arrows, these types of functions can be recognized as follows:

- for a one-to-one function, each point $y \in Y$ has *at most one* ingoing arrow,
- for a function onto, each point $y \in Y$ has *at least one* ingoing arrow, and
- for a *bijection*, each point $y \in Y$ has *exactly one* ingoing arrow.

The fact that a function $f: X \to Y$ is one-to-one is sometimes expressed by the notation

$$f: X \hookrightarrow Y.$$

The $\hookrightarrow$ symbol is a combination of the inclusion sign $\subset$ with the mapping arrow $\to$. Why? If $f: X \hookrightarrow Y$ is an injective mapping, then the set $Z = f(X)$ can be regarded as a "copy" of the set X within Y (since f considered as a map $X \to Z$ is a bijection), and so an injective mapping $f: X \hookrightarrow Y$ can be thought of as a "generalized inclusion" of X in Y. This point can probably be best appreciated in more abstract and more advanced parts of mathematics like topology or algebra.

There are also symbols for functions onto and for bijections, but these are still much less standard in the literature than the symbol for an injective function, so we do not introduce them.

Since we will be interested in counting objects, bijections will be especially significant for us, for the following reason: if X and Y are sets and there exists a bijection $f: X \to Y$, then X and Y have the same number of elements. Let us give a simple example of using a bijection for counting (more sophisticated ones come later).

1.4.4 Example. How many 8-digit sequences consisting of digits 0 through 9 are there? How many of them contain an even number of odd digits?

Solution. The answer to the first question is 10^8. One easy way of seeing this is to note that each 8-digit sequence can be read as the decimal notation of an integer number between 0 and $10^8 - 1$, and conversely, each such integer can be written in decimal notation and, if necessary, padded with zeros on the left to produce an 8-digit sequence. This defines a bijection between the set $\{0, 1, \ldots, 10^8 - 1\}$ and the set of all 8-digit sequences.

Well, this bijection was perhaps too simple (or, rather, too customary) to impress anyone. What about the 8-digit sequences with an even

number of odd digits? Let E be the set of all these sequences (E for "even"), and let O be the remaining ones, i.e. those with an odd number of odd digits. Consider any sequence $s \in E$, and define another sequence, $f(s)$, by changing the first digit of s: 0 is changed to 1, 1 to 2,..., 8 to 9, and 9 to 0. It is easy to check that the modified sequence $f(s)$ has an odd number of odd digits and hence f is a mapping from E to O. From two different sequences $s, s' \in E$, we cannot get the same sequence by the described modification, so f is one-to-one. And any sequence $t \in O$ is obtained as $f(s)$ for some $s \in E$, i.e. s arises from t by changing the first digit "back", by replacing 1 by 0, 2 by 1,..., 9 by 8, and 0 by 9. Therefore, f is a bijection and $|E| = |O|$. Since $|E| + |O| = 10^8$, we finally have $|E| = 5 \cdot 10^7$. □

In the following proposition, we prove some simple properties of functions.

Proposition. *Let $f: X \to Y$ and $g: Y \to Z$ be functions. Then*

(i) If f, g are one-to-one, then $g \circ f$ is also a one-to-one function.

(ii) If f, g are functions onto, then $g \circ f$ is also a function onto.

(iii) If f, g are bijective functions, then $g \circ f$ is a bijection as well.

(iv) For any function $f: X \to Y$ there exist a set Z, a one-to-one function $h: Z \hookrightarrow Y$, and a function onto $g: X \to Z$, such that $f = h \circ g$. (So any function can be written as a composition of a one-to-one function and a function onto.)

Proof. Parts (i), (ii), (iii) are obtained by direct verification from the definition. As an example, let us prove (ii).

We choose $z \in Z$, and we are looking for an $x \in X$ satisfying $(g \circ f)(x) = z$. Since g is a function onto, there exists a $y \in Y$ such that $g(y) = z$. And since f is a function onto, there exists an $x \in X$ with $f(x) = y$. Such an x is the desired element satisfying $(g \circ f)(x) = z$.

The most interesting part is (iv). Let $Z = f(X)$ (so $Z \subseteq Y$). We define mappings $g: X \to Z$ and $h: Z \to Y$ as follows:

$$\begin{aligned} g(x) &= f(x) \quad && \text{for } x \in X \\ h(z) &= z \quad && \text{for } z \in Z. \end{aligned}$$

Clearly g is a function onto, h is one-to-one, and $f = h \circ g$. □

Finishing the remaining parts of the proof may be a good exercise for understanding the notions covered in this section.

Inverse function. If $f: X \to Y$ is a bijection, we can define a function $g: Y \to X$ by setting $g(y) = x$ if x is the unique element of X with $y = f(x)$. This g is called the *inverse function* of f, and it is commonly denoted by f^{-1}. Pictorially, the inverse function is

obtained by reversing all the arrows. Another equivalent definition of the inverse function is given in Exercise 4. It may look more complicated, but from a certain "higher" mathematical point of view it is better.

Exercises

1. Show that if X is a finite set, then a function $f: X \to X$ is one-to-one if and only if it is onto.

2. Find an example of:

   (a) A one-to-one function $f: \mathbf{N} \hookrightarrow \mathbf{N}$ which is not onto.

   (b) A function $f: \mathbf{N} \to \mathbf{N}$ which is onto but not one-to-one.

3. Decide which of the following functions $\mathbf{Z} \to \mathbf{Z}$ are injective and which are surjective: $x \mapsto 1 + x$, $x \mapsto 1 + x^2$, $x \mapsto 1 + x^3$, $x \mapsto 1 + x^2 + x^3$. Does anything in the answer change if we consider them as functions $\mathbf{R} \to \mathbf{R}$? (You may want to sketch their graphs and/or use some elementary calculus methods.)

4. For a set X, let $\mathrm{id}_X: X \to X$ denote the function defined by $\mathrm{id}_X(x) = x$ for all $x \in X$ (the *identity function*). Let $f: X \to Y$ be some function. Prove:

   (a) A function $g: Y \to X$ such that $g \circ f = \mathrm{id}_X$ exists if and only if f is one-to-one.

   (b) A function $g: Y \to X$ such that $f \circ g = \mathrm{id}_Y$ exists if and only if f is onto.

   (c) A function $g: Y \to X$ such that both $f \circ g = \mathrm{id}_Y$ and $g \circ f = \mathrm{id}_X$ exist if and only if f is a bijection.

   (d) If $f: X \to Y$ is a bijection, then the following three conditions are equivalent for a function $g: Y \to X$:

   (i) $g = f^{-1}$,
   (ii) $g \circ f = \mathrm{id}_X$, and
   (iii) $f \circ g = \mathrm{id}_Y$.

5. (a) If $g \circ f$ is an onto function, does g have to be onto? Does f have to be onto?

   (b) If $g \circ f$ is a one-to-one function, does g have to be one-to-one? Does f have to be one-to-one?

6. Prove that the following two statements about a function $f: X \to Y$ are equivalent (X and Y are some arbitrary sets):

   (i) f is one-to-one.

(ii) For any set Z and any two distinct functions $g_1: Z \to X$ and $g_2: Z \to X$, the composed functions $f \circ g_1$ and $f \circ g_2$ are also distinct.

(First, make sure you understand what it means that two functions are equal and what it means that they are distinct.)

7. In everyday mathematics, the number of elements of a set is understood in an intuitive sense and no definition is usually given. In the logical foundations of mathematics, however, the number of elements is defined via bijections: $|X| = n$ means that there exists a bijection from X to the set $\{1, 2, \ldots, n\}$. (Also other, alternative definitions of set size exist but we will consider only this one here.)

   (a) Prove that if X and Y have the same size according to this definition, then there exists a bijection from X to Y.

   (b) Prove that if X has size n according to this definition, and there exists a bijection from X to Y, then Y has size n too.

   (c) *Prove that a set cannot have two different sizes m and n, $m \neq n$, according to this definition. Be careful not to use the intuitive notion of "size" but only the definition via bijections. Proceed by induction.

1.5 Relations

It is remarkable how many mathematical notions can be expressed using sets and various set-theoretic constructions. It is not only remarkable but also surprising, since set theory, and even the notion of a set itself, are notions which appeared in mathematics relatively recently, and some 100 years ago, set theory was rejected even by some prominent mathematicians. Today, set theory has entered the mathematical vocabulary and it has become the language of all mathematics (and mathematicians), a language which helps to understand mathematics, with all its diversity, as a whole with common foundations.

We will show how more complicated mathematical notions can be built using the simplest set-theoretical tools. The key notion of a relation,[10] which we now introduce, is a common generalization of such diverse notions as equivalence, function, and ordering.

1.5.1 Definition. *A* relation *is a set of ordered pairs.*[11] *If X and Y are sets, any subset of the Cartesian product $X \times Y$ is called a* relation between X and Y. *The most important case is $X = Y$; then we speak of a* relation on X, *which is thus an arbitrary subset $R \subseteq X \times X$.*

[10] As a mathematical object; you know "relation" as a word in common language.

[11] In more detail, we could say a *binary relation* (since pairs of elements are being related). Sometimes also n-ary relations are considered for $n \neq 2$.

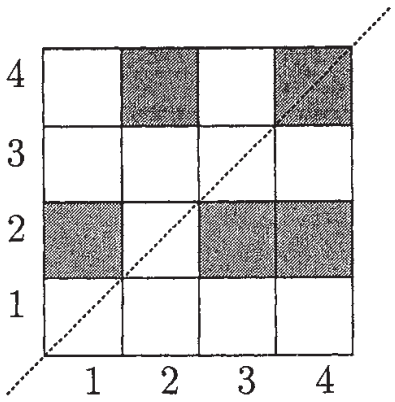


Fig. 1.3 A graphic presentation of the relation $R = \{(1,2), (2,4), (3,2), (4,2), (4,4)\}$ on the set $\{1,2,3,4\}$.

If an ordered pair (x,y) belongs to a relation R, i.e. $(x,y) \in R$, we say that x and y *are related by* R, and we also write xRy.

We have already seen an object which was a subset of a Cartesian product, namely a function. Indeed, a function is a special type of relation, where we require that any $x \in X$ is related to precisely one $y \in Y$. In a general relation, an $x \in X$ can be related to several elements of Y, or also to none.

Many symbols well known to the reader can be interpreted as relations in this sense. For instance, $=$ (equality) and $\geq$ (non-strict inequality) are both relations on the set $\mathbf{N}$ of all natural numbers. The first one consists of the pairs $(1,1), (2,2), (3,3), \ldots$, the second one of the pairs $(1,1)$, $(2,1)$, $(2,2)$, $(3,1)$, $(3,2)$, $(3,3)$, $(4,1), \ldots$. We could thus also write $(5,2) \in \geq$ instead of the usual $5 \geq 2$, which we usually don't do, however. Note that we had to specify the set on which the relation $\geq$, say, is considered: as a relation on $\mathbf{R}$ it would be a quite different set of ordered pairs.

Many interesting "real life" examples of relations come from various kinds of relationships among people, e.g. "to be the mother of", "to be the father of", "to be a cousin of" are relations on the set of all people, usually well defined although not always easy to determine.

A relation R on a set X can be captured pictorially in (at least) two quite different ways. The first way is illustrated in Fig. 1.3. The little squares correspond to ordered pairs in the Cartesian product, and for pairs belonging to the relation we have shaded the corresponding squares. This kind of picture emphasizes the definition of a relation on X and it captures its "overall shape".

This figure is also very close in spirit to an alternative way of describing a relation on a set X using the notion of a matrix.[12] If R is a relation on some n-element set $X = \{x_1, x_2, \ldots, x_n\}$ then R is completely described by an $n \times n$ matrix $A = (a_{ij})$, where

$$a_{ij} = 1 \quad \text{if } (x_i, x_j) \in R$$
$$a_{ij} = 0 \quad \text{if } (x_i, x_j) \notin R.$$

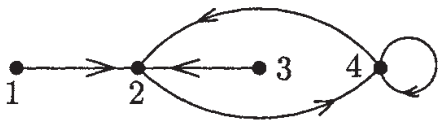
The matrix A is called the *adjacency matrix* of the relation R. For instance, for the relation in Fig. 1.3, the corresponding adjacency matrix would be

$$\begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 \end{pmatrix}.$$

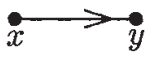
Note that this matrix is turned by 90 degrees compared to Fig. 1.3. This is because, for a matrix element, the first index is the number of a row and the second index is the number of a column, while for Cartesian coordinates it is customary for the first coordinate to determine the horizontal position and the second coordinate the vertical position.

The adjacency matrix is one possible computer representation of a relation on a finite set.

Another picture of the same relation as in Fig. 1.3 is shown below:



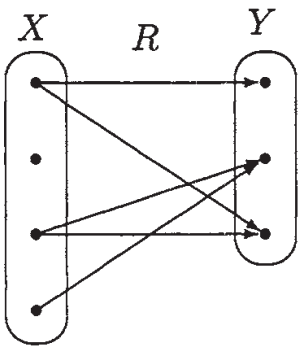
Here the dots correspond to elements of the set X. The fact that a given ordered pair (x, y) belongs to the relation R is marked by drawing an arrow from x to y:



and, in the case $x = y$, by a loop:

A relation between X and Y can be depicted in a similar way:

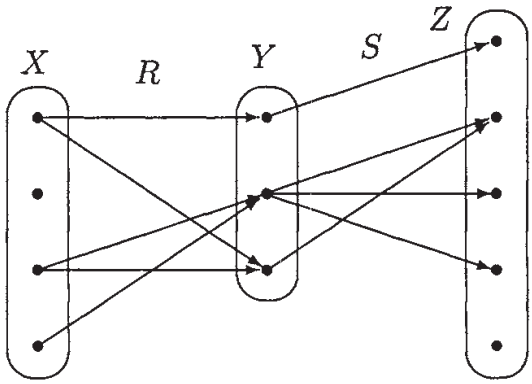
[12] An $n \times m$ matrix is a rectangular table of numbers with n rows and m columns. Any reader who hasn't met matrices yet can consult the Appendix for the definitions and basic facts, or, preferably, take a course of linear algebra or refer to a good textbook.



This way was suggested for drawing functions in Section 1.4.

Composition of relations. Let X, Y, Z be sets, let $R \subseteq X \times Y$ be a relation between X and Y, and let $S \subseteq Y \times Z$ be a relation between Y and Z. The *composition of the relations R and S* is the relation $T \subseteq X \times Z$ defined as follows: for given $x \in X$ and $z \in Z$, xTz holds if and only if there exists some $y \in Y$ such that xRy and ySz. The composition of relations R and S is usually denoted by $R \circ S$.

The composition of relations can be nicely illustrated using a drawing with arrows. In the following figure,



a pair (x, z) is in the relation $R \circ S$ whenever one can get from x to z along the arrows (i.e. via some $y \in Y$).

Have you noticed? Relations are composed in the same way as functions, but the notation is unfortunately different! For *relations* it is customary to write the composition "from left to right", and for functions it is usually written "from right to left". So if $f: X \to Y$ and $g: Y \to Z$ are functions, their composition is written $g \circ f$, but if we understood them as relations, we would write $f \circ g$ for the same thing! Both ways of notation have their reasons, such a notation has been established historically, and probably there is no point in trying to change it. In this text, we will talk almost exclusively about composing functions.

Similarly as for functions, the composition is not defined for arbitrary two relations. In order to compose relations, they must have the "middle" set in common (which was denoted by Y in the definition). In particular, it may happen that $R \circ S$ is defined while $S \circ R$ makes no sense! However, if both R and S are relations on the same set X,

their composition is always well defined. But also in this case the result of composing relations depends on the order, and $R \circ S$ is in general different from $S \circ R$—see Exercise 2.

Exercises

1. Describe the relation $R \circ R$, if R stands for
   (a) the equality relation "=" on the set $\mathbf{N}$ of all natural numbers,
   (b) the relation "less than or equal" ("$\leq$") on $\mathbf{N}$,
   (c) the relation "strictly less" ("$<$") on $\mathbf{N}$,
   (d) the relation "strictly less" ("$<$") on the set $\mathbf{R}$ of all real numbers.
2. Find relations R, S on some set X such that $R \circ S \neq S \circ R$.
3. For a relation R on a set X we define the symbol R^n by induction: $R^1 = R$, $R^{n+1} = R \circ R^n$.
   (a) Prove that if X is finite and R is a relation on it, then there exist $r, s \in \mathbf{N}$, $r < s$, such that $R^r = R^s$.
   (b) Find a relation R on a finite set such that $R^n \neq R^{n+1}$ for every $n \in \mathbf{N}$.
   (c) Show that if X is infinite, the claim (a) need not hold (i.e. a relation R may exist such that all the relations R^n, $n \in \mathbf{N}$, are distinct).
4. (a) Let $X = \{x_1, x_2, \ldots, x_n\}$ and $Y = \{y_1, y_2, \ldots, y_m\}$ be finite sets, and let $R \subseteq X \times Y$ be a relation. Generalize the definition of the adjacency matrix of a relation to this case.
   (b) *Let X, Y, Z be finite sets, let $R \subseteq X \times Y$ and $S \subseteq Y \times Z$ be relations, and let A_R and A_S be their adjacency matrices, respectively. If you have defined the adjacency matrix in (a) properly, the matrix product $A_R A_S$ should be well defined. Discover and describe the connection of the composed relation $R \circ S$ to the matrix product $A_R A_S$.
5. Prove the associativity of composing relations: if R, S, T are relations such that $(R \circ S) \circ T$ is well defined, then also $R \circ (S \circ T)$ is well defined and equals $(R \circ S) \circ T$.

1.6 Equivalences

Besides the functions, equivalences are another important special type of relations. Informally, an equivalence on a set X is a relation describing which pairs of elements of X are "of the same type" in some sense. For instance, let X be the set of all triangles in the plane. By saying that two triangles are related if and only if they are congruent (i.e. one can be transformed into the other by translation and rotation), we have defined one equivalence on X. Another

equivalence is defined by relating all pairs of similar triangles (two triangles are similar if one can be obtained from the other one by translation, rotation, and scaling; in other words, if their corresponding angles are the same). And a third equivalence arises by saying that each triangle is only related to itself.

These are three particular examples of equivalence relations, and the reader may look forward to many more examples later on. In general, in order to be called an equivalence, a relation must satisfy three conditions. These conditions are so useful that they each deserve a name.

1.6.1 Definition. *We say that a relation R on a set X is*

- reflexive *if xRx for every $x \in X$,*
- symmetric *if xRy implies yRx, for all $x, y \in X$,*
- transitive *if xRy and yRz imply xRz, for all $x, y, z \in X$.*

In a drawing like that in Fig. 1.3, a reflexive relation is one containing all squares on the diagonal (drawn by a dotted line). In drawing using arrows, a reflexive relation has loops at all points.

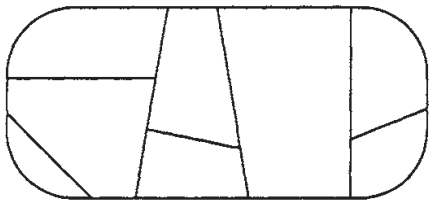
For a symmetric relation, a picture of the type in Fig. 1.3 has the diagonal as an axis of symmetry. In a picture using arrows, the arrows between two points always go in both directions:

The condition of transitivity can be well explained using arrows. If there are arrows $x \to y$ and $y \to z$, then the $x \to z$ arrow is present as well:

1.6.2 Definition. *We say that a relation R on X is an* equivalence *on X if it is reflexive, symmetric, and transitive.*

(You may want to contemplate for a while why these properties are natural for a relation that should express something like "being of the same type".) The notion of equivalence is a common generalization of notions expressing identity, isomorphism, similarity, etc. Relations of equivalence are often denoted by symbols like $=$, $\equiv$, $\simeq$, $\approx$, $\cong$, and so on.

Although an equivalence R on a set X is a special type of relation and we can thus depict it by either of the above methods, more often a picture similar to the one below is used:



The key to this type of drawing is the following notion of *equivalence class*. Let R be an equivalence on a set X and let x be an element of X. By the symbol $R[x]$, we denote the set of all elements $y \in X$ that are equivalent to x; in symbols, $R[x] = \{y \in X\colon xRy\}$. $R[x]$ is called the *equivalence class of R determined by x*.

1.6.3 Proposition. *For any equivalence R on X, we have*

(i) $R[x]$ *is nonempty for every* $x \in X$.

(ii) For any two elements $x, y \in X$, *either* $R[x] = R[y]$ *or* $R[x] \cap R[y] = \emptyset$.

(iii) The equivalence classes determine the relation R uniquely.

Before we start proving this, we should explain the meaning of (iii). It means the following: if R and S are two equivalences on X and if the equality $R[x] = S[x]$ holds for every element $x \in X$, then $R = S$.

Proof. The proof is simple using the three requirements in the definition of equivalence.

(i) The set $R[x]$ always contains x since R is a reflexive relation.

(ii) Let x, y be two elements. We distinguish two cases:

(a) If xRy, then we prove $R[x] \subseteq R[y]$ first. Indeed, if $z \in R[x]$, then we also know that zRx (by symmetry of R) and therefore zRy (by transitivity of R). Thus also $z \in R[y]$. By using symmetry again, we get that xRy implies $R[x] = R[y]$.

(b) Suppose that xRy doesn't hold. We show that $R[x] \cap R[y] = \emptyset$. We proceed by contradiction. Suppose that there exists $z \in R[x] \cap R[y]$. Then xRz and zRy (by symmetry of R), and so xRy (by transitivity of R), which is a contradiction.

(iii) This part is obvious, since the equivalence classes determine R as follows:

$$xRy \text{ if and only if } \{x, y\} \subseteq R[x].$$

□

The above proposition explains the preceding picture. It guarantees that the equivalence classes form a *partition* of the set X; that is, they are pairwise disjoint subsets of X whose union is the whole X. Conversely, any partition of X determines exactly one equivalence on X. That is, there exists a bijective mapping of the set of all equivalences on X onto the set of all partitions of X.

Exercises

1. Formulate the conditions for reflexivity of a relation, for symmetry of a relation, and for its transitivity using the adjacency matrix of the relation.

2. Prove that a relation R is transitive if and only if $R \circ R \subseteq R$.

3. (a) Prove that for any relation R, the relation $T = R \cup R \circ R \cup R \circ R \circ R \cup \ldots$ (the union of all multiple compositions of R) is transitive.

   (b) Prove that any transitive relation containing R as a subset also contains T.

   (c) Prove that if $|X| = n$, then $T = R \cup R \circ R \cup \cdots \cup \underbrace{R \circ R \circ \cdots \circ R}_{(n-1)\times}$.

   Remark. The relation T as in (a), (b) is the smallest transitive relation containing R, and it is called the *transitive closure* of R.

4. Let R and S be arbitrary equivalences on a set X. Decide which of the following relations are necessarily also equivalences (if yes, prove; if not, give a counterexample).

   (a) $R \cap S$

   (b) $R \cup S$

   (c) $R \setminus S$

   (d) $R \circ S$.

5. (a) Suppose that R is a transitive relation on the set $\mathbf{Z}$ of all integers, and we know that for any two integers $a, b \in \mathbf{Z}$, if $|a - b| = 2$ then aRb. Is *every* R satisfying these conditions necessarily an equivalence? (Note that a pair of elements can perhaps be in R even if it is not enforced by the given conditions!)

   (b) Suppose that R is a transitive relation on $\mathbf{Z}$, and we know that for any two integers $a, b \in \mathbf{Z}$, if $|a - b| \in \{3, 4\}$ then aRb. Is R necessarily an equivalence?

6. Call an equivalence $\sim$ on the set $\mathbf{Z}$ (the integers) a *congruence* if the following condition holds for all $a, x, y \in \mathbf{Z}$: if $x \sim y$ then also $a + x \sim a + y$.

   (a) Let q be a nonzero integer. Define a relation $\equiv_q$ on $\mathbf{Z}$ by letting $x \equiv_q y$ if and only if q divides $x - y$. Check that $\equiv_q$ is a congruence according to the above definition.

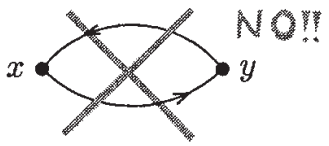
   (b) *Prove that any congruence on $\mathbf{Z}$ is either of the form $\equiv_q$ for some q or the *diagonal relation* $\{(x, x)\colon x \in \mathbf{Z}\}$.

   (c) Suppose we replaced the condition "$a+x \sim a+y$" in the definition of a congruence by "$ax \sim ay$". Would the claim in (a) remain true for this kind of "multiplicative congruence"? *And how about the claim in (b)?

1.7 Ordered sets

The reader will certainly be familiar with the ordering of natural numbers and of other number domains by magnitude (the "usual" ordering of numbers). In mathematics, such an ordering is considered as a special type of a relation, i.e. a set of pairs of numbers. In the case just mentioned, this relation is usually denoted by the symbol "$\leq$" ("less than or equal"). Various orderings can be defined on other sets too, such as the set of all words in some language, and one set can be ordered in many different and perhaps exotic ways.

Before introducing the general notion of an ordered set, we define an auxiliary notion. A relation R on a set X is called *antisymmetric*[13] if xRy and yRx imply that $x = y$, for all $x, y \in X$. When depicting a relation by arrows, the following situation is forbidden in an antisymmetric relation:



1.7.1 Definition. *An* ordering *on a set X is any relation on X that is reflexive, antisymmetric, and transitive. An* ordered set *is a pair (X, R), where X is a set and R is an ordering on X.*

Note the formal similarity of this definition to the definition of an equivalence. The definitions are almost the same, "only" the symmetry

[13] Sometimes this is called *weakly antisymmetric*, while for a *strongly antisymmetric* relation xRy and yRx never happen at the same time, i.e. xRx is also excluded.

has been replaced by antisymmetry. Yet equivalences and orderings are very different concepts.

For orderings, the symbols $\preceq$ or $\leq$ are commonly used. The first of them is useful, e.g. when we want to speak of some other ordering of the set of natural numbers than the usual ordering by magnitude, or if we consider some arbitrary ordering on a general set.

If we have some ordering $\preceq$, we define a derived relation of "strict inequality", $\prec$, as follows: $a \prec b$ if and only if $a \preceq b$ and $a \neq b$. Further we can introduce the "reverse inequality" $\succeq$, by letting $a \succeq b$ if and only if $b \preceq a$.

Examples. We have already mentioned several examples of ordered sets—these were $(\mathbf{N}, \leq)$, $(\mathbf{R}, \leq)$, and similar ones, where $\leq$ of course denotes the usual ordering, formally understood as a relation.

As is easy to check, if R is an ordering on a set X, and $Y \subseteq X$ is some subset of X, the relation $R \cap Y^2$ (the restriction of R on Y) is an ordering on Y. Intuitively, we order the elements of Y in the same way as before but we forget the others. This yields further examples of ordered sets, namely various subsets of real numbers with the usual ordering.

Linear orderings. The examples discussed so far have a significant feature in common: any two elements of the underlying set can be compared; in other words, for any two distinct elements x and y either $x \leq y$ or $y \leq x$ holds. This property is *not* a part of the definition of an ordering, and orderings having it are called *linear* orderings (sometimes the term *total ordering* is used with the same meaning).

Other examples of orderings. What do orderings which are not linear look like? For example, on any set X, we may define a relation Δ in which each element x is in relation with itself only, i.e. $\Delta = \{(x, x)\colon x \in X\}$. It is easily checked that this relation satisfies the definition of an ordering, but this is a rather dull example. Before giving more examples, let us insert a remark about terminology.

In order to emphasize that we speak of an ordering which is not necessarily linear, we sometimes use the longer term *partial ordering*. A partial ordering thus means exactly the same as ordering (without further adjectives), so a partial ordering may also happen to be linear. Similarly, instead of an ordered set, one sometimes speaks of a *partially ordered set*. To abbreviate this long term, the artificial word *poset* is frequently used.

Let us describe more interesting examples of partially ordered sets.

1.7.2 Example. Let us imagine we intend to buy a refrigerator, say. We simplify the complicated real situation by a mathematical abstraction, and we suppose that we only look at three numerical parameters of refrigerators: their cost, electricity consumption, and the volume of the inner space. If we consider two types of refrigerators, and if the first type is more expensive, consumes more power, and a smaller amount of food fits into it, then the second type can be considered a better one—a large majority of buyers of refrigerators would agree with that. On the other hand, someone may prefer a smaller and cheaper refrigerator, another may prefer a larger refrigerator even if it costs more, and an environmentally concerned customer may even buy an expensive refrigerator if it saves power.

The relation "to be clearly worse" (denote it by $\preceq$) in this sense is thus a partial ordering on refrigerators or, mathematically reformulated, on the set of triples (c, p, v) of real numbers (c stands for cost, p for power consumption, and v for volume), defined as follows:

$$\begin{aligned}&(c_1, p_1, v_1) \preceq (c_2, p_2, v_2) \text{ if and only if}\\ &c_1 \geq c_2,\ p_1 \geq p_2, \text{ and } v_1 \leq v_2.\end{aligned} \tag{1.3}$$

1.7.3 Example. For natural numbers a, b, the symbol $a|b$ means "a divides b". In other words, there exists a natural number c such that $b = ac$. The relation "$|$" is a partial ordering on $\mathbf{N}$. We leave the verification of this to the reader.

1.7.4 Example. Let X be a set. Recall that the symbol 2^X denotes the system of all subsets of the set X. The relation "$\subseteq$" (to be a subset) defines a partial ordering on 2^X.

Drawing partially ordered sets. Finite orderings can be drawn using arrows, as with any other relations. Typically, such drawings will contains lots of arrows. For instance, for a 10-element linearly ordered set we would have to draw $10 + 9 + \cdots + 1 = 55$ arrows and loops. A number of arrows can be reconstructed from transitivity, however: if we know that $x \preceq y$ and $y \preceq z$, then also $x \preceq z$, so we may leave out the arrow from x to z. Similarly, we need not draw the loops, since we know they are always there. For finite ordered sets, all the information is captured by the relation of "immediate predecessor", which we are now going to define.

Let $(X, \preceq)$ be an ordered set. We say that an element $x \in X$ is an *immediate predecessor of* an element $y \in X$ if

- $x \prec y$, and
- there is no $t \in X$ such that $x \prec t \prec y$.

Let us denote the just-defined relation of immediate predecessor by $\lhd$.

The claim that the ordering $\preceq$ can be reconstructed from the relation $\lhd$ may be formulated precisely as follows:

1.7.5 Proposition. *Let $(X, \preceq)$ be a finite ordered set, and let $\lhd$ be the corresponding immediate predecessor relation. Then for any two elements $x, y \in X$, $x \prec y$ holds if and only if there exist elements $x_1, x_2, \ldots, x_k \in X$ such that $x \lhd x_1 \lhd \cdots \lhd x_k \lhd y$ (possibly with $k = 0$, i.e. we may also have $x \lhd y$).*

Proof. One implication is easy to see: if we have $x \lhd x_1 \lhd \cdots \lhd x_k \lhd y$, then also $x \preceq x_1 \preceq \cdots \preceq x_k \preceq y$ (since the immediate predecessor relation is contained in the ordering relation), and by the transitivity of $\preceq$, we have $x \preceq y$.

The reverse implication is not difficult either, and we prove it by induction. We prove the following statement:

Lemma. *Let $x, y \in X$, $x \prec y$, be two elements such that there exist at most n elements $t \in X$ satisfying $x \prec t \prec y$ (i.e. "between" x and y). Then there exist $x_1, x_2, \ldots, x_k \in X$ such that $x \lhd x_1 \lhd \cdots \lhd x_k \lhd y$.*

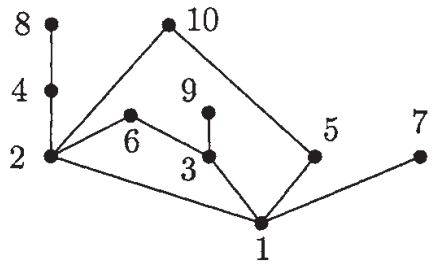
For $n = 0$, the assumption of this lemma asserts that there exists no t with $x \prec t \prec y$, and hence $x \lhd y$, which means that the statement holds (we choose $k = 0$).

Let the lemma hold for all n up to some n_0, and let us have $x \prec y$ such that the set $M_{xy} = \{t \in X\colon x \prec t \prec y\}$ has $n = n_0 + 1$ elements. Let us choose an element $u \in M_{xy}$, and consider the sets $M_{xu} = \{t \in X\colon x \prec t \prec u\}$ and M_{uy} defined similarly. By the transitivity of $\prec$ it follows that $M_{xu} \subset M_{xy}$ and $M_{uy} \subset M_{xy}$. Both M_{xu} and M_{uy} have at least one element less than M_{xy} (since $u \notin M_{xu}$, $u \notin M_{uy}$), and by the inductive hypothesis, we find elements $x_1, \ldots, x_k$ and $y_1, \ldots, y_\ell$ in such a way that $x \lhd x_1 \lhd \cdots \lhd x_k \lhd u$ and $u \lhd y_1 \lhd \cdots \lhd y_\ell \lhd y$. By combining these two "chains" we obtain the desired sequence connecting x and y. □

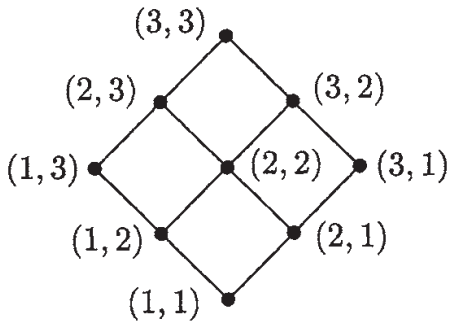
By the above proposition, it is enough to draw the relation of immediate predecessor by arrows. If we accept the convention that all arrows in the drawing will be directed upwards (this means that if $x \prec y$ then x is drawn higher than y), we need not even draw the direction of the arrows—it is enough to draw segments connecting

the points. Such a picture of a partially ordered set is called its *Hasse diagram*. The following figure shows a 7-element linearly ordered set, such as $(\{1, 2, \ldots, 7\}, \leq)$:

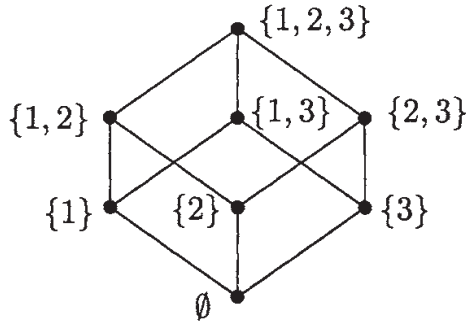
The next drawing depicts the set $\{1, 2, \ldots, 10\}$ ordered by the divisibility relation (see Example 1.7.3):



The following figure is a Hasse diagram of the set $\{1, 2, 3\} \times \{1, 2, 3\}$ with ordering $\preceq$ given by the rule $(a_1, b_1) \preceq (a_2, b_2)$ if and only if $a_1 \leq a_2$ and $b_1 \leq b_2$:



Finally, here is a Hasse diagram of the set of all subsets of $\{1, 2, 3\}$ ordered by inclusion:



Further examples and notions concerning posets are left to the exercises. The theory of finite posets is an important and flourishing branch of combinatorics. The reader can learn about it in Trotter [28].

Exercises

1. Describe all relations on a set X which are equivalences and (partial) orderings at the same time.

2. Let R and S be arbitrary partial orderings on a set X. Decide which of the following relations are necessarily partial orderings:

   (a) $R \cap S$

   (b) $R \cup S$

   (c) $R \setminus S$

   (d) $R \circ S$.

3. Verify that the relation (1.3) in Example 1.7.2 indeed defines a partial ordering.

4. *Let R be a relation on a set X such that there is no finite sequence of elements $x_1, x_2, \ldots, x_k$ of X satisfying $x_1 R x_2$, $x_2 R x_3, \ldots, x_{k-1} R x_k$, $x_k R x_1$ (we say that such an R is *acyclic*). Prove that there exists an ordering $\preceq$ on X such that $R \subseteq \prec$. You may assume that X is finite if this helps.

5. (a) Consider the set $\{1, 2, \ldots, n\}$ ordered by the divisibility relation $|$ (see Example 1.7.3). What is the maximum possible number of elements of a set $X \subseteq \{1, 2, \ldots, n\}$ which is ordered linearly by the relation $|$ (such a set X is called a *chain*)?

   (b) Solve the same question for the set $2^{\{1,2,\ldots,n\}}$ ordered by the relation $\subseteq$ (see Example 1.7.4).

6. Show that Proposition 1.7.5 does not hold for infinite sets.

7. Let $(X, \preceq)$ be a poset. An element $a \in X$ is called

   - a *largest element* of X if for every $x \in X$, $x \preceq a$ holds, and
   - a *maximal element* of X if there exists no $y \in X$ such that $a \prec y$.

   A *smallest element* and a *minimal element* are defined similarly.

   (a) Show that a largest element is always maximal, and find an example of a poset with a maximal element but no largest element.

   (b) Find a poset having no smallest element and no minimal element either, but possessing a largest element.

8. *Let $\preceq$ be any (partial) ordering on a set X. A *linear extension* of $\preceq$ is any linear ordering $\leq$ on X such that $x \preceq y$ implies $x \leq y$ for all $x, y \in X$. (If it didn't look so strange we could write this condition compactly as $\preceq \subseteq \leq$.) Prove that any partial ordering on a finite set X has at least one linear extension.

9. Let $(X, \leq)$, $(Y, \preceq)$ be ordered sets. We say that they are *isomorphic* (meaning that they "look the same" from the point of view of ordering) if there exists a bijection $f: X \to Y$ such that for every $x, y \in X$, we have $x \leq y$ if and only if $f(x) \preceq f(y)$.

   (a) Draw Hasse diagrams for all nonisomorphic 3-element posets.

   (b) Prove that any two n-element linearly ordered sets are isomorphic.

   (c) Find two nonisomorphic linear orderings of the set of all natural numbers.

   (d) Can you find infinitely many nonisomorphic linear orderings of $\mathbf{N}$? *Uncountably many (for readers knowing something about the cardinalities of infinite sets)?

10. *Show that for every finite poset $(X, \preceq)$ there exists a finite set A and a system $\mathcal{M}$ of subsets of A such that the ordered set $(\mathcal{M}, \subseteq)$ is isomorphic to $(X, \preceq)$. (Isomorphism of posets has been defined in Exercise 9.)

11. Let $(X, \preceq)$ be a poset and let $A \subseteq X$ be its subset. An element $s \in X$ is called a *supremum* of the set A if the following holds:

    - $a \preceq s$ for each $a \in A$,
    - if $a \preceq s'$ holds for all $a \in A$, where s' is some element of X, then $s \preceq s'$.

    The *infimum* of a subset $A \subseteq X$ is defined analogously, but with all inequalities going in the opposite direction.

    (a) Check that any subset $A \subseteq X$ has at most one supremum and at most one infimum. (The supremum of A, if it exists, is denoted by $\sup A$. Similarly $\inf A$ denotes the infimum.)

    (b) Which element is the supremum of the empty set (according to the definition just given)?

    (c) Find an example of a poset in which every nonempty subset has an infimum, but there are nonempty subsets having no supremum.

    (d) *Let $(X, \preceq)$ be a poset in which every subset (including the empty one) has a supremum. Show then that every subset has an infimum as well.

12. Consider the poset $(\mathbf{N}, |)$ (ordering by divisibility).

    (a) Decide whether each nonempty subset of $\mathbf{N}$ has a supremum.

    (b) Decide whether each nonempty finite subset of $\mathbf{N}$ has a supremum.

    (c) Decide whether each nonempty subset has an infimum.