Chapter 1

Set Theory and Logic

We adopt, as most mathematicians do, the naive point of view regarding set theory. We shall assume that what is meant by a *set* of objects is intuitively clear, and we shall proceed on that basis without analyzing the concept further. Such an analysis properly belongs to the foundations of mathematics and to mathematical logic, and it is not our purpose to initiate the study of those fields.

Logicians have analyzed set theory in great detail, and they have formulated axioms for the subject. Each of their axioms expresses a property of sets that mathematicians commonly accept, and collectively the axioms provide a foundation broad enough and strong enough that the rest of mathematics can be built on them.

It is unfortunately true that careless use of set theory, relying on intuition alone, can lead to contradictions. Indeed, one of the reasons for the axiomatization of set theory was to formulate rules for dealing with sets that would avoid these contradictions. Although we shall not deal with the axioms explicitly, the rules we follow in dealing with sets derive from them. In this book, you will learn how to deal with sets in an "apprentice" fashion, by observing how we handle them and by working with them yourself. At some point of your studies, you may wish to study set theory more carefully and in greater detail; then a course in logic or foundations will be in order.

§1 Fundamental Concepts

Here we introduce the ideas of set theory, and establish the basic terminology and notation. We also discuss some points of elementary logic that, in our experience, are apt to cause confusion.

Basic Notation

Commonly we shall use capital letters $A, B, \ldots$ to denote sets, and lowercase letters $a, b, \ldots$ to denote the ***objects*** or ***elements*** belonging to these sets. If an object a belongs to a set A, we express this fact by the notation

$$a \in A.$$

If a does not belong to A, we express this fact by writing

$$a \notin A.$$

The equality symbol $=$ is used throughout this book to mean *logical identity*. Thus, when we write $a = b$, we mean that "a" and "b" are symbols for the same object. This is what one means in arithmetic, for example, when one writes $\frac{2}{4} = \frac{1}{2}$. Similarly, the equation $A = B$ states that "A" and "B" are symbols for the same set; that is, A and B consist of precisely the same objects.

If a and b are different objects, we write $a \neq b$; and if A and B are different sets, we write $A \neq B$. For example, if A is the set of all nonnegative real numbers, and B is the set of all positive real numbers, then $A \neq B$, because the number 0 belongs to A and not to B.

We say that A is a ***subset*** of B if every element of A is also an element of B; and we express this fact by writing

$$A \subset B.$$

Nothing in this definition requires A to be different from B; in fact, if $A = B$, it is true that both $A \subset B$ and $B \subset A$. If $A \subset B$ and A is different from B, we say that A is a ***proper subset*** of B, and we write

$$A \subsetneq B.$$

The relations $\subset$ and $\subsetneq$ are called ***inclusion*** and ***proper inclusion***, respectively. If $A \subset B$, we also write $B \supset A$, which is read "B ***contains*** A."

How does one go about specifying a set? If the set has only a few elements, one can simply list the objects in the set, writing "A is the set consisting of the elements a, b, and c." In symbols, this statement becomes

$$A = \{a, b, c\},$$

where braces are used to enclose the list of elements.

The usual way to specify a set, however, is to take some set A of objects and some *property* that elements of A may or may not possess, and to form the set consisting of all elements of A having that property. For instance, one might take the set of real numbers and form the subset B consisting of all even integers. In symbols, this statement becomes

$$B = \{x \mid x \text{ is an even integer}\}.$$

Here the braces stand for the words "the set of," and the vertical bar stands for the words "such that." The equation is read "B is the set of all x such that x is an even integer."

The Union of Sets and the Meaning of "or"

Given two sets A and B, one can form a set from them that consists of all the elements of A together with all the elements of B. This set is called the ***union*** of A and B and is denoted by $A \cup B$. Formally, we define

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

But we must pause at this point and make sure exactly what we mean by the statement "$x \in A$ or $x \in B$."

In ordinary everyday English, the word "or" is ambiguous. Sometimes the statement "P or Q" means "P or Q, or both" and sometimes it means "P or Q, but not both." Usually one decides from the context which meaning is intended. For example, suppose I spoke to two students as follows:

> "Miss Smith, every student registered for this course has taken either a course in linear algebra or a course in analysis."
>
> "Mr. Jones, either you get a grade of at least 70 on the final exam or you will flunk this course"

In the context, Miss Smith knows perfectly well that I mean "everyone has had linear algebra or analysis, or both," and Mr. Jones knows I mean "either he gets at least 70 or he flunks, but not both." Indeed, Mr. Jones would be exceedingly unhappy if both statements turned out to be true!

In mathematics, one cannot tolerate such ambiguity. One has to pick just one meaning and stick with it, or confusion will reign. Accordingly, mathematicians have agreed that they will use the word "or" in the first sense, so that the statement "P or Q" always means "P or Q, or both." If one means "P or Q, but not both," then one has to include the phrase "but not both" explicitly.

With this understanding, the equation defining $A \cup B$ is unambiguous; it states that $A \cup B$ is the set consisting of all elements x that belong to A or to B or to both.

The Intersection of Sets, the Empty Set, and the Meaning of "If . . . Then"

Given sets A and B, another way one can form a set is to take the common part of A and B. This set is called the ***intersection*** of A and B and is denoted by $A \cap B$. Formally, we define

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

But just as with the definition of $A \cup B$, there is a difficulty. The difficulty is not in the meaning of the word "and"; it is of a different sort. It arises when the sets A and B happen to have no elements in common. What meaning does the symbol $A \cap B$ have in such a case?

To take care of this eventuality, we make a special convention. We introduce a special set that we call the ***empty set***, denoted by $\varnothing$, which we think of as "the set having no elements."

Using this convention, we express the statement that A and B have no elements in common by the equation

$$A \cap B = \varnothing.$$

We also express this fact by saying that A and B are ***disjoint***.

Now some students are bothered by the notion of an "empty set." "How," they say, "can you have a set with nothing in it?" The problem is similar to that which arose many years ago when the number 0 was first introduced.

The empty set is only a convention, and mathematics could very well get along without it. But it is a very convenient convention, for it saves us a good deal of awkwardness in stating theorems and in proving them. Without this convention, for instance, one would have to prove that the two sets A and B do have elements in common before one could use the notation $A \cap B$. Similarly, the notation

$$C = \{x \mid x \in A \text{ and } x \text{ has a certain property}\}$$

could not be used if it happened that no element x of A had the given property. It is much more convenient to agree that $A \cap B$ and C equal the empty set in such cases.

Since the empty set $\varnothing$ is merely a convention, we must make conventions relating it to the concepts already introduced. Because $\varnothing$ is thought of as "the set with no elements," it is clear we should make the convention that for each object x, the relation $x \in \varnothing$ does not hold. Similarly, the definitions of union and intersection show that for every set A we should have the equations

$$A \cup \varnothing = A \quad \text{and} \quad A \cap \varnothing = \varnothing.$$

The inclusion relation is a bit more tricky. Given a set A, should we agree that $\varnothing \subset A$? Once more, we must be careful about the way mathematicians use the English language. The expression $\varnothing \subset A$ is a shorthand way of writing the sentence, "Every element that belongs to the empty set also belongs to the set A." Or to put it more

formally, "For every object x, if x belongs to the empty set, then x also belongs to the set A."

Is this statement true or not? Some might say "yes" and others say "no." You will never settle the question by argument, only by agreement. This is a statement of the form "If P, then Q," and in everyday English the meaning of the "if . . . then" construction is ambiguous. It always means that if P is true, then Q is true also. Sometimes that is all it means; other times it means something more: that if P is false, Q must be false. Usually one decides from the context which interpretation is correct.

The situation is similar to the ambiguity in the use of the word "or." One can reformulate the examples involving Miss Smith and Mr. Jones to illustrate the ambiguity. Suppose I said the following:

> "Miss Smith, if any student registered for this course has not taken a course in linear algebra, then he has taken a course in analysis."
>
> "Mr. Jones, if you get a grade below 70 on the final, you are going to flunk this course."

In the context, Miss Smith understands that if a student in the course has not had linear algebra, then he has taken analysis, but if he has had linear algebra, he may or may not have taken analysis as well. And Mr. Jones knows that if he gets a grade below 70, he will flunk the course, but if he gets a grade of at least 70, he will pass.

Again, mathematics cannot tolerate ambiguity, so a choice of meanings must be made. Mathematicians have agreed always to use "if . . . then" in the first sense, so that a statement of the form "If P, then Q" means that if P is true, Q is true also, but if P is false, Q may be either true or false.

As an example, consider the following statement about real numbers:

$$\textit{If } x > 0, \textit{ then } x^3 \neq 0.$$

It is a statement of the form, "If P, then Q," where P is the phrase "$x > 0$" (called the ***hypothesis*** of the statement) and Q is the phrase "$x^3 \neq 0$" (called the ***conclusion*** of the statement). This is a true statement, for in every case for which the hypothesis $x > 0$ holds, the conclusion $x^3 \neq 0$ holds as well.

Another true statement about real numbers is the following:

$$\textit{If } x^2 < 0, \textit{ then } x = 23;$$

in every case for which the hypothesis holds, the conclusion holds as well. Of course, it happens in this example that there are no cases for which the hypothesis holds. A statement of this sort is sometimes said to be ***vacuously true***.

To return now to the empty set and inclusion, we see that the inclusion $\varnothing \subset A$ does hold for every set A. Writing $\varnothing \subset A$ is the same as saying, "If $x \in \varnothing$, then $x \in A$," and this statement is vacuously true.

Contrapositive and Converse

Our discussion of the "if . . . then" construction leads us to consider another point of elementary logic that sometimes causes difficulty. It concerns the relation between a *statement*, its *contrapositive*, and its *converse*.

Given a statement of the form "If P, then Q," its ***contrapositive*** is defined to be the statement "If Q is not true, then P is not true." For example, the contrapositive of the statement

$$\textit{If } x > 0, \textit{ then } x^3 \neq 0,$$

is the statement

$$\textit{If } x^3 = 0, \textit{ then it is not true that } x > 0.$$

Note that both the statement and its contrapositive are true. Similarly, the statement

$$\textit{If } x^2 < 0, \textit{ then } x = 23,$$

has as its contrapositive the statement

$$\textit{If } x \neq 23, \textit{ then it is not true that } x^2 < 0.$$

Again, both are true statements about real numbers.

These examples may make you suspect that there is some relation between a statement and its contrapositive. And indeed there is; they are two ways of saying precisely the same thing. Each is true if and only if the other is true; they are *logically equivalent.*

This fact is not hard to demonstrate. Let us introduce some notation first. As a shorthand for the statement "If P, then Q," we write

$$P \Longrightarrow Q,$$

which is read "P implies Q." The contrapositive can then be expressed in the form

$$(\text{not } Q) \Longrightarrow (\text{not } P),$$

where "not Q" stands for the phrase "Q is not true."

Now the only way in which the statement "$P \Rightarrow Q$" can fail to be correct is if the hypothesis P is true and the conclusion Q is false. Otherwise it is correct. Similarly, the only way in which the statement (not Q) $\Rightarrow$ (not P) can fail to be correct is if the hypothesis "not Q" is true and the conclusion "not P" is false. This is the same as saying that Q is false and P is true. And this, in turn, is precisely the situation in which $P \Rightarrow Q$ fails to be correct. Thus, we see that the two statements are either both correct or both incorrect; they are logically equivalent. Therefore, we shall accept a proof of the statement "not $Q \Rightarrow$ not P" as a proof of the statement "$P \Rightarrow Q$."

There is another statement that can be formed from the statement $P \Rightarrow Q$. It is the statement

$$Q \Longrightarrow P,$$

which is called the ***converse*** of $P \Rightarrow Q$. One must be careful to distinguish between a statement's converse and its contrapositive. Whereas a statement and its contrapositive are logically equivalent, the truth of a statement says nothing at all about the truth or falsity of its converse. For example, the true statement

$$\textit{If } x > 0, \textit{ then } x^3 \neq 0,$$

has as its converse the statement

$$\textit{If } x^3 \neq 0, \textit{ then } x > 0,$$

which is false. Similarly, the true statement

$$\textit{If } x^2 < 0, \textit{ then } x = 23,$$

has as its converse the statement

$$\textit{If } x = 23, \textit{ then } x^2 < 0,$$

which is false.

If it should happen that both the statement $P \Rightarrow Q$ and its converse $Q \Rightarrow P$ are true, we express this fact by the notation

$$P \Longleftrightarrow Q,$$

which is read "P holds if and only if Q holds."

Negation

If one wishes to form the contrapositive of the statement $P \Rightarrow Q$, one has to know how to form the statement "not P," which is called the ***negation*** of P. In many cases, this causes no difficulty; but sometimes confusion occurs with statements involving the phrases "for every" and "for at least one." These phrases are called *logical quantifiers*.

To illustrate, suppose that X is a set, A is a subset of X, and P is a statement about the general element of X. Consider the following statement:

(*) *For every $x \in A$, statement P holds.*

How does one form the negation of this statement? Let us translate the problem into the language of sets. Suppose that we let B denote the set of all those elements x of X for which P holds. Then statement (*) is just the statement that A is a subset of B. What is its negation? Obviously, the statement that A is *not* a subset of B; that is, the statement that there exists at least one element of A that does not belong to B. Translating back into ordinary language, this becomes

For at least one $x \in A$, statement P does not hold.

Therefore, to form the negation of statement (*), one replaces the quantifier "for every" by the quantifier "for at least one," and one replaces statement P by *its* negation.

The process works in reverse just as well; the negation of the statement

For at least one $x \in A$, statement Q holds,

is the statement

For every $x \in A$, statement Q does not hold.

The Difference of Two Sets

We return now to our discussion of sets. There is one other operation on sets that is occasionally useful. It is the ***difference*** of two sets, denoted by $A - B$, and defined as the set consisting of those elements of A that are not in B. Formally,

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

It is sometimes called the ***complement*** of B relative to A, or the complement of B *in* A.

Our three set operations are represented schematically in Figure 1.1.

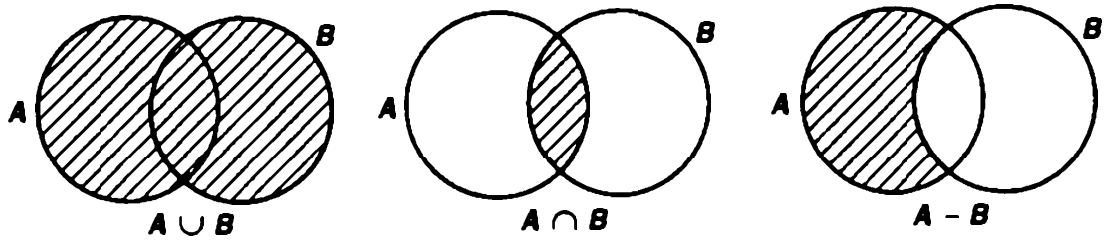

Figure 1.1

Rules of Set Theory

Given several sets, one may form new sets by applying the set-theoretic operations to them. As in algebra, one uses parentheses to indicate in what order the operations are to be performed. For example, $A \cup (B \cap C)$ denotes the union of the two sets A and $B \cap C$, while $(A \cup B) \cap C$ denotes the intersection of the two sets $A \cup B$ and C. The sets thus formed are quite different, as Figure 1.2 shows.

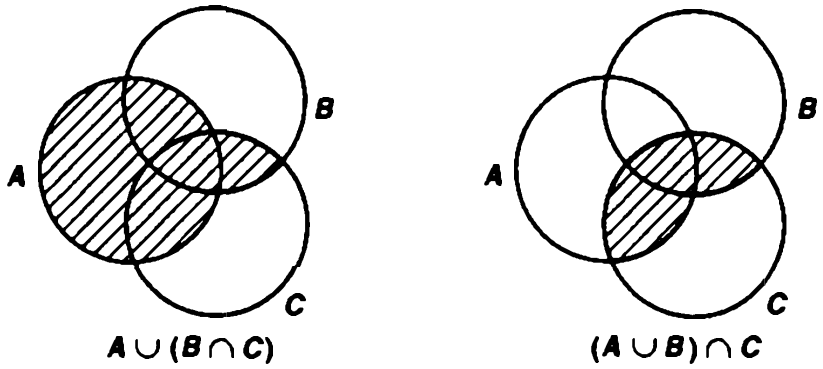

Figure 1.2

Sometimes different combinations of operations lead to the same set; when that happens, one has a rule of set theory. For instance, it is true that for any sets A, B, and C the equation

$$A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$$

holds. The equation is illustrated in Figure 1.3; the shaded region represents the set in question, as you can check mentally. This equation can be thought of as a "distributive law" for the operations $\cap$ and $\cup$.

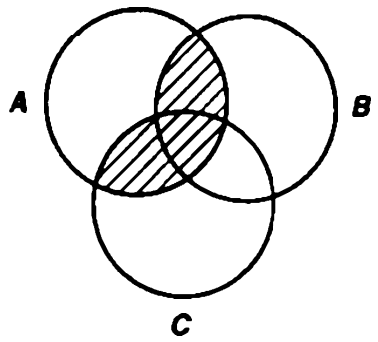

Figure 1.3

Other examples of set-theoretic rules include the second "distributive law,"

$$A \cup (B \cap C) = (A \cup B) \cap (A \cup C),$$

and *DeMorgan's laws*,

$$A - (B \cup C) = (A - B) \cap (A - C),$$
$$A - (B \cap C) = (A - B) \cup (A - C).$$

We leave it to you to check these rules. One can state other rules of set theory, but these are the most important ones. DeMorgan's laws are easier to remember if you verbalize them as follows:

> *The complement of the union equals the intersection of the complements.*
> *The complement of the intersection equals the union of the complements.*

Collections of Sets

The objects belonging to a set may be of any sort. One can consider the set of all even integers, and the set of all blue-eyed people in Nebraska, and the set of all decks of playing cards in the world. Some of these are of limited mathematical interest, we admit! But the third example illustrates a point we have not yet mentioned: namely, that the objects belonging to a set may *themselves* be sets. For a deck of cards is itself a set, one consisting of pieces of pasteboard with certain standard designs printed on them. The set of all decks of cards in the world is thus a set whose elements are themselves sets (of pieces of pasteboard).

We now have another way to form new sets from old ones. Given a set A, we can consider sets whose elements are subsets of A. In particular, we can consider the set of all subsets of A. This set is sometimes denoted by the symbol $\mathcal{P}(A)$ and is called the ***power set*** of A (for reasons to be explained later).

When we have a set whose elements are sets, we shall often refer to it as a ***collection*** of sets and denote it by a script letter such as $\mathcal{A}$ or $\mathcal{B}$. This device will help us in keeping things straight in arguments where we have to consider objects, and sets of objects, and collections of sets of objects, all at the same time. For example, we might use $\mathcal{A}$ to denote the collection of all decks of cards in the world, letting an ordinary capital letter A denote a deck of cards and a lowercase letter a denote a single playing card.

A certain amount of care with notation is needed at this point. We make a distinction between the object a, which is an *element* of a set A, and the one-element set $\{a\}$, which is a *subset* of A. To illustrate, if A is the set $\{a, b, c\}$, then the statements

$$a \in A, \qquad \{a\} \subset A, \qquad \text{and} \qquad \{a\} \in \mathcal{P}(A)$$

are all correct, but the statements $\{a\} \in A$ and $a \subset A$ are not.

Arbitrary Unions and Intersections

We have already defined what we mean by the union and the intersection of two sets. There is no reason to limit ourselves to just two sets, for we can just as well form the union and intersection of arbitrarily many sets.

Given a collection $\mathcal{A}$ of sets, the ***union*** of the elements of $\mathcal{A}$ is defined by the equation

$$\bigcup_{A \in \mathcal{A}} A = \{x \mid x \in A \text{ for at least one } A \in \mathcal{A}\}.$$

The ***intersection*** of the elements of $\mathcal{A}$ is defined by the equation

$$\bigcap_{A \in \mathcal{A}} A = \{x \mid x \in A \text{ for every } A \in \mathcal{A}\}.$$

There is no problem with these definitions if one of the elements of $\mathcal{A}$ happens to be the empty set. But it is a bit tricky to decide what (if anything) these definitions mean if we allow $\mathcal{A}$ to be the empty collection. Applying the definitions literally, we see that no element x satisfies the defining property for the union of the elements of $\mathcal{A}$. So it is reasonable to say that

$$\bigcup_{A \in \mathcal{A}} A = \varnothing$$

if $\mathcal{A}$ is empty. On the other hand, every x satisfies (vacuously) the defining property for the intersection of the elements of $\mathcal{A}$. The question is, every x in what set? If one has a given large set X that is specified at the outset of the discussion to be one's "universe of discourse," and one considers only subsets of X throughout, it is reasonable to let

$$\bigcap_{A \in \mathcal{A}} A = X$$

when $\mathcal{A}$ is empty. Not all mathematicians follow this convention, however. To avoid difficulty, *we shall not define the intersection when $\mathcal{A}$ is empty.*

Cartesian Products

There is yet another way of forming new sets from old ones; it involves the notion of an "ordered pair" of objects. When you studied analytic geometry, the first thing you did was to convince yourself that after one has chosen an x-axis and a y-axis in the plane, every point in the plane can be made to correspond to a unique ordered pair (x, y) of real numbers. (In a more sophisticated treatment of geometry, the plane is more likely to be *defined* as the set of all ordered pairs of real numbers!)

The notion of ordered pair carries over to general sets. Given sets A and B, we define their cartesian product $A \times B$ to be the set of all ordered pairs (a, b) for which a is an element of A and b is an element of B. Formally,

$$A \times B = \{(a, b) \mid a \in A \text{ and } b \in B\}.$$

This definition assumes that the concept of "ordered pair" is already given. It can be taken as a primitive concept, as was the notion of "set"; or it can be given a definition in terms of the set operations already introduced. One definition in terms of set operations is expressed by the equation

$$(a, b) = \{\{a\}, \{a, b\}\};$$

it defines the ordered pair (a, b) as a collection of sets. If $a \neq b$, this definition says that (a, b) is a collection containing two sets, one of which is a one-element set and the other a two-element set. The *first coordinate* of the ordered pair is defined to be the element belonging to both sets, and the *second coordinate* is the element belonging to only one of the sets. If $a = b$, then (a, b) is a collection containing only one set $\{a\}$, since $\{a, b\} = \{a, a\} = \{a\}$ in this case. Its first coordinate and second coordinate both equal the element in this single set.

I think it is fair to say that most mathematicians think of an ordered pair as a primitive concept rather than thinking of it as a collection of sets!

Let us make a comment on notation. It is an unfortunate fact that the notation (a, b) is firmly established in mathematics with two entirely different meanings. One meaning, as an ordered pair of objects, we have just discussed. The other meaning is the one you are familiar with from analysis; if a and b are real numbers, the symbol (a, b) is used to denote the interval consisting of all numbers x such that $a < x < b$. Most of the time, this conflict in notation will cause no difficulty because the meaning will be clear from the context. Whenever a situation occurs where confusion is possible, we shall adopt a different notation for the ordered pair (a, b), denoting it by the symbol

$$a \times b$$

instead.

Exercises

1. Check the distributive laws for $\cup$ and $\cap$ and DeMorgan's laws.
2. Determine which of the following statements are true for all sets A, B, C, and D. If a double implication fails, determine whether one or the other of the possible implications holds. If an equality fails, determine whether the statement becomes true if the "equals" symbol is replaced by one or the other of the inclusion symbols $\subset$ or $\supset$.
   (a) $A \subset B$ and $A \subset C \Leftrightarrow A \subset (B \cup C)$.
   (b) $A \subset B$ or $A \subset C \Leftrightarrow A \subset (B \cup C)$.
   (c) $A \subset B$ and $A \subset C \Leftrightarrow A \subset (B \cap C)$.
   (d) $A \subset B$ or $A \subset C \Leftrightarrow A \subset (B \cap C)$.
   (e) $A - (A - B) = B$.
   (f) $A - (B - A) = A - B$.
   (g) $A \cap (B - C) = (A \cap B) - (A \cap C)$.
   (h) $A \cup (B - C) = (A \cup B) - (A \cup C)$.
   (i) $(A \cap B) \cup (A - B) = A$.
   (j) $A \subset C$ and $B \subset D \Rightarrow (A \times B) \subset (C \times D)$.
   (k) The converse of (j).
   (l) The converse of (j), assuming that A and B are nonempty.
   (m) $(A \times B) \cup (C \times D) = (A \cup C) \times (B \cup D)$.
   (n) $(A \times B) \cap (C \times D) = (A \cap C) \times (B \cap D)$.
   (o) $A \times (B - C) = (A \times B) - (A \times C)$.
   (p) $(A - B) \times (C - D) = (A \times C - B \times C) - A \times D$.
   (q) $(A \times B) - (C \times D) = (A - C) \times (B - D)$.
3. (a) Write the contrapositive and converse of the following statement: "If $x < 0$, then $x^2 - x > 0$," and determine which (if any) of the three statements are true.
   (b) Do the same for the statement "If $x > 0$, then $x^2 - x > 0$."
4. Let A and B be sets of real numbers. Write the negation of each of the following statements:
   (a) For every $a \in A$, it is true that $a^2 \in B$.
   (b) For at least one $a \in A$, it is true that $a^2 \in B$.
   (c) For every $a \in A$, it is true that $a^2 \notin B$.
   (d) For at least one $a \notin A$, it is true that $a^2 \in B$.
5. Let $\mathcal{A}$ be a nonempty collection of sets. Determine the truth of each of the following statements and of their converses:
   (a) $x \in \bigcup_{A \in \mathcal{A}} A \Rightarrow x \in A$ for at least one $A \in \mathcal{A}$.
   (b) $x \in \bigcup_{A \in \mathcal{A}} A \Rightarrow x \in A$ for every $A \in \mathcal{A}$.
   (c) $x \in \bigcap_{A \in \mathcal{A}} A \Rightarrow x \in A$ for at least one $A \in \mathcal{A}$.
   (d) $x \in \bigcap_{A \in \mathcal{A}} A \Rightarrow x \in A$ for every $A \in \mathcal{A}$.
6. Write the contrapositive of each of the statements of Exercise 5.

7. Given sets A, B, and C, express each of the following sets in terms of A, B, and C, using the symbols $\cup$, $\cap$, and $-$.

$$D = \{x \mid x \in A \text{ and } (x \in B \text{ or } x \in C)\},$$
$$E = \{x \mid (x \in A \text{ and } x \in B) \text{ or } x \in C\},$$
$$F = \{x \mid x \in A \text{ and } (x \in B \Rightarrow x \in C)\}.$$

8. If a set A has two elements, show that $\mathcal{P}(A)$ has four elements. How many elements does $\mathcal{P}(A)$ have if A has one element? Three elements? No elements? Why is $\mathcal{P}(A)$ called the power set of A?
9. Formulate and prove DeMorgan's laws for arbitrary unions and intersections.
10. Let $\mathbb{R}$ denote the set of real numbers. For each of the following subsets of $\mathbb{R} \times \mathbb{R}$, determine whether it is equal to the cartesian product of two subsets of $\mathbb{R}$.
    (a) $\{(x, y) \mid x \text{ is an integer}\}$.
    (b) $\{(x, y) \mid 0 < y \leq 1\}$.
    (c) $\{(x, y) \mid y > x\}$.
    (d) $\{(x, y) \mid x \text{ is not an integer and } y \text{ is an integer}\}$.
    (e) $\{(x, y) \mid x^2 + y^2 < 1\}$.

§2 Functions

The concept of *function* is one you have seen many times already, so it is hardly necessary to remind you how central it is to all mathematics. In this section, we give the precise mathematical definition, and we explore some of the associated concepts.

A function is usually thought of as a *rule* that assigns to each element of a set A, an element of a set B. In calculus, a function is often given by a simple formula such as $f(x) = 3x^2 + 2$ or perhaps by a more complicated formula such as

$$f(x) = \sum_{k=1}^{\infty} x^k.$$

One often does not even mention the sets A and B explicitly, agreeing to take A to be the set of all real numbers for which the rule makes sense and B to be the set of all real numbers.

As one goes further in mathematics, however, one needs to be more precise about what a function is. Mathematicians *think* of functions in the way we just described, but the definition they use is more exact. First, we define the following:

Definition. A *rule of assignment* is a subset r of the cartesian product $C \times D$ of two sets, having the property that each element of C appears as the first coordinate of *at most one* ordered pair belonging to r.

Thus, a subset r of $C \times D$ is a rule of assignment if

$$[(c, d) \in r \text{ and } (c, d') \in r] \Longrightarrow [d = d'].$$

We think of r as a way of assigning, to the element c of C, the element d of D for which $(c, d) \in r$.

Given a rule of assignment r, the ***domain*** of r is defined to be the subset of C consisting of all first coordinates of elements of r, and the ***image set*** of r is defined as the subset of D consisting of all second coordinates of elements of r. Formally,

$$\text{domain } r = \{c \mid \text{there exists } d \in D \text{ such that } (c, d) \in r\},$$
$$\text{image } r = \{d \mid \text{there exists } c \in C \text{ such that } (c, d) \in r\}.$$

Note that given a rule of assignment r, its domain and image are entirely determined.

Now we can say what a function is.

Definition. A ***function*** f is a rule of assignment r, together with a set B that contains the image set of r. The domain A of the rule r is also called the ***domain*** of the function f; the image set of r is also called the ***image set*** of f; and the set B is called the ***range*** of f.[†]

If f is a function having domain A and range B, we express this fact by writing

$$f : A \longrightarrow B,$$

which is read "f is a function from A to B," or "f is a mapping from A into B," or simply "f maps A into B." One sometimes visualizes f as a geometric transformation physically carrying the points of A to points of B.

If $f : A \to B$ and if a is an element of A, we denote by $f(a)$ the unique element of B that the rule determining f assigns to a; it is called the ***value*** of f at a, or sometimes the ***image*** of a under f. Formally, if r is the rule of the function f, then $f(a)$ denotes the unique element of B such that $(a, f(a)) \in r$.

Using this notation, one can go back to defining functions almost as one did before, with no lack of rigor. For instance, one can write (letting $\mathbb{R}$ denote the real numbers)

> "Let f be the function whose rule is $\{(x, x^3 + 1) \mid x \in \mathbb{R}\}$ and whose range is $\mathbb{R}$,"

or one can equally well write

> "Let $f : \mathbb{R} \to \mathbb{R}$ be the function such that $f(x) = x^3 + 1$."

Both sentences specify precisely the same function. But the sentence "Let f be the function $f(x) = x^3 + 1$" is no longer adequate for specifying a function because it specifies neither the domain nor the range of f.

[†]Analysts are apt to use the word "range" to denote what we have called the "image set" of f They avoid giving the set B a name.

Definition. If $f : A \to B$ and if A_0 is a subset of A, we define the ***restriction*** of f to A_0 to be the function mapping A_0 into B whose rule is

$$\{(a, f(a)) \mid a \in A_0\}.$$

It is denoted by $f|A_0$, which is read "f restricted to A_0."

EXAMPLE 1. Let $\mathbb{R}$ denote the real numbers and let $\tilde{\mathbb{R}}_+$ denote the nonnegative reals. Consider the functions

$$\begin{array}{llll} f : \mathbb{R} \longrightarrow \mathbb{R} & \text{defined by} & f(x) = x^2, \\ g : \tilde{\mathbb{R}}_+ \longrightarrow \mathbb{R} & \text{defined by} & g(x) = x^2, \\ h : \mathbb{R} \longrightarrow \tilde{\mathbb{R}}_+ & \text{defined by} & h(x) = x^2, \\ k : \tilde{\mathbb{R}}_+ \longrightarrow \tilde{\mathbb{R}}_+ & \text{defined by} & k(x) = x^2. \end{array}$$

The function g is different from the function f because their rules are different subsets of $\mathbb{R} \times \mathbb{R}$; it is the restriction of f to the set $\tilde{\mathbb{R}}_+$. The function h is also different from f, even though their rules are the same set, because the range specified for h is different from the range specified for f. The function k is different from all of these. These functions are pictured in Figure 2.1

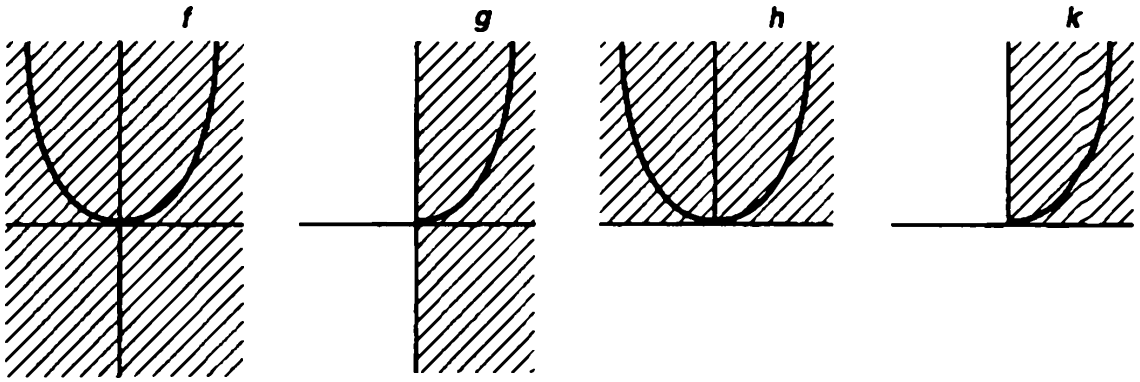

Figure 2.1

Restricting the domain of a function and changing its range are two ways of forming a new function from an old one. Another way is to form the composite of two functions.

Definition. Given functions $f : A \to B$ and $g : B \to C$, we define the ***composite*** $g \circ f$ of f and g as the function $g \circ f : A \to C$ defined by the equation $(g \circ f)(a) = g(f(a))$.

Formally, $g \circ f : A \to C$ is the function whose rule is

$$\{(a, c) \mid \text{For some } b \in B,\ f(a) = b \text{ and } g(b) = c\}.$$

We often picture the composite $g \circ f$ as involving a physical movement of the point a to the point $f(a)$, and then to the point $g(f(a))$, as illustrated in Figure 2.2.

Note that $g \circ f$ is defined only when the range of f *equals* the domain of g.

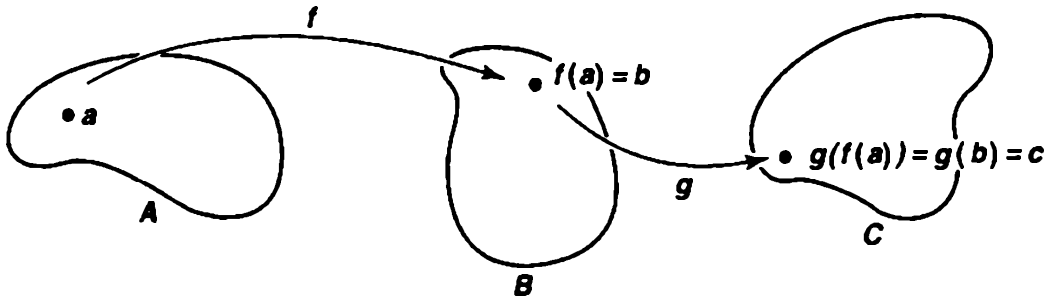

Figure 2.2

EXAMPLE 2. The composite of the function $f : \mathbb{R} \to \mathbb{R}$ given by $f(x) = 3x^2 + 2$ and the function $g : \mathbb{R} \to \mathbb{R}$ given by $g(x) = 5x$ is the function $g \circ f : \mathbb{R} \to \mathbb{R}$ given by

$$(g \circ f)(x) = g(f(x)) = g(3x^2 + 2) = 5(3x^2 + 2).$$

The composite $f \circ g$ can also be formed in this case; it is the quite different function $f \circ g : \mathbb{R} \to \mathbb{R}$ given by

$$(f \circ g)(x) = f(g(x)) = f(5x) = 3(5x)^2 + 2.$$

Definition. A function $f : A \to B$ is said to be ***injective*** (or ***one-to-one***) if for each pair of distinct points of A, their images under f are distinct. It is said to be ***surjective*** (or f is said to map A ***onto*** B) if every element of B is the image of some element of A under the function f. If f is both injective and surjective, it is said to be ***bijective*** (or is called a ***one-to-one correspondence***).

More formally, f is injective if

$$[f(a) = f(a')] \Longrightarrow [a = a'],$$

and f is surjective if

$$[b \in B] \Longrightarrow [b = f(a) \text{ for at least one } a \in A].$$

Injectivity of f depends only on the rule of f; surjectivity depends on the range of f as well. You can check that the composite of two injective functions is injective, and the composite of two surjective functions is surjective; it follows that the composite of two bijective functions is bijective.

If f is bijective, there exists a function from B to A called the ***inverse*** of f. It is denoted by f^{-1} and is defined by letting $f^{-1}(b)$ be that unique element a of A for which $f(a) = b$. Given $b \in B$, the fact that f is surjective implies that there *exists* such an element $a \in A$; the fact that f is injective implies that there is *only one* such element a. It is easy to see that if f is bijective, f^{-1} is also bijective.

EXAMPLE 3. Consider again the functions f, g, h, and k of Figure 2.1. The function $f : \mathbb{R} \to \mathbb{R}$ given by $f(x) = x^2$ is neither injective nor surjective. Its restriction g to the nonnegative reals is injective but not surjective. The function $h : \mathbb{R} \to \bar{\mathbb{R}}_+$ obtained from f

by changing the range is surjective but not injective. The function $k : \bar{\mathbb{R}}_+ \to \bar{\mathbb{R}}_+$ obtained from f by restricting the domain *and* changing the range is both injective and surjective, so it has an inverse. Its inverse is, of course, what we usually call the *square-root function*.

A useful criterion for showing that a given function f is bijective is the following, whose proof is left to the exercises:

Lemma 2.1. *Let $f : A \to B$. If there are functions $g : B \to A$ and $h : B \to A$ such that $g(f(a)) = a$ for every a in A and $f(h(b)) = b$ for every b in B, then f is bijective and $g = h = f^{-1}$.*

Definition. Let $f : A \to B$. If A_0 is a subset of A, we denote by $f(A_0)$ the set of all images of points of A_0 under the function f; this set is called the ***image*** of A_0 under f. Formally,

$$f(A_0) = \{b \mid b = f(a) \text{ for at least one } a \in A_0\}.$$

On the other hand, if B_0 is a subset of B, we denote by $f^{-1}(B_0)$ the set of all elements of A whose images under f lie in B_0; it is called the ***preimage*** of B_0 under f (or the "counterimage," or the "inverse image," of B_0). Formally,

$$f^{-1}(B_0) = \{a \mid f(a) \in B_0\}.$$

Of course, there may be no points a of A whose images lie in B_0; in that case, $f^{-1}(B_0)$ is empty.

Note that if $f : A \to B$ is bijective and $B_0 \subset B$, we have two meanings for the notation $f^{-1}(B_0)$. It can be taken to denote the *preimage* of B_0 under the function f or to denote the *image* of B_0 under the function $f^{-1} : B \to A$. These two meanings give precisely the same subset of A, however, so there is, in fact, no ambiguity.

Some care is needed if one is to use the f and f^{-1} notation correctly. The operation f^{-1}, for instance, when applied to subsets of B, behaves very nicely; it preserves inclusions, unions, intersections, and differences of sets. We shall use this fact frequently. But the operation f, when applied to subsets of A, preserves only inclusions and unions. See Exercises 2 and 3.

As another situation where care is needed, we note that it is not in general true that $f^{-1}(f(A_0)) = A_0$ and $f(f^{-1}(B_0)) = B_0$. (See the following example.) The relevant rules, which we leave to you to check, are the following: If $f : A \to B$ and if $A_0 \subset A$ and $B_0 \subset B$, then

$$A_0 \subset f^{-1}(f(A_0)) \quad \text{and} \quad f(f^{-1}(B_0)) \subset B_0.$$

The first inclusion is an equality if f is injective, and the second inclusion is an equality if f is surjective.

EXAMPLE 4. Consider the function $f : \mathbb{R} \to \mathbb{R}$ given by $f(x) = 3x^2 + 2$ (Figure 2.3). Let $[a, b]$ denote the closed interval $a \leq x \leq b$. Then

$$f^{-1}(f([0, 1])) = f^{-1}([2, 5]) = [-1, 1], \quad \text{and}$$
$$f(f^{-1}([0, 5])) = f([-1, 1]) = [2, 5].$$

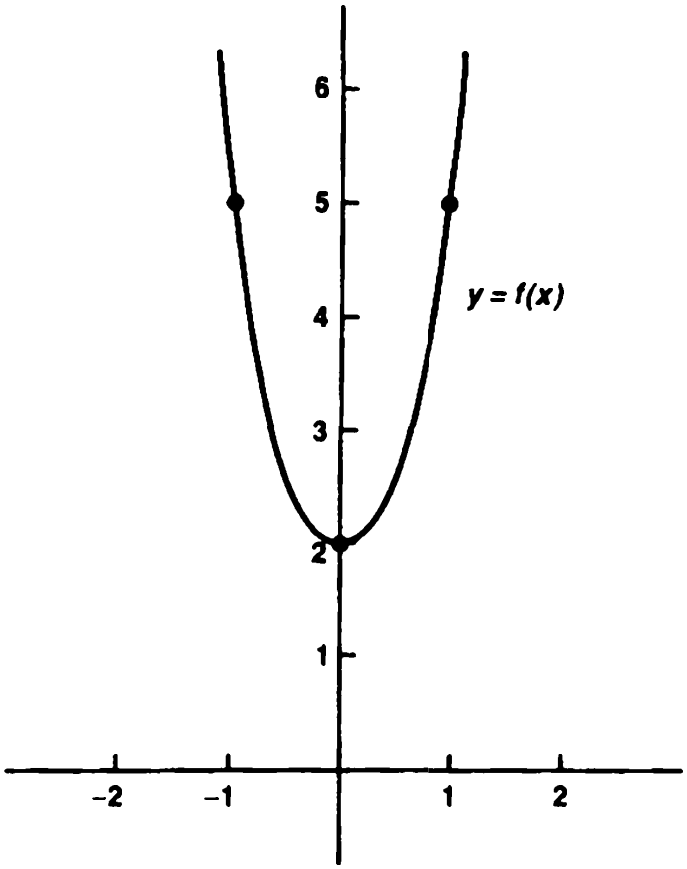

Figure 2.3

Exercises

1. Let $f : A \to B$. Let $A_0 \subset A$ and $B_0 \subset B$.
   (a) Show that $A_0 \subset f^{-1}(f(A_0))$ and that equality holds if f is injective.
   (b) Show that $f(f^{-1}(B_0)) \subset B_0$ and that equality holds if f is surjective.
2. Let $f : A \to B$ and let $A_i \subset A$ and $B_i \subset B$ for $i = 0$ and $i = 1$. Show that f^- preserves inclusions, unions, intersections, and differences of sets:
   (a) $B_0 \subset B_1 \Rightarrow f^{-1}(B_0) \subset f^{-1}(B_1)$.
   (b) $f^{-1}(B_0 \cup B_1) = f^{-1}(B_0) \cup f^{-1}(B_1)$.
   (c) $f^{-1}(B_0 \cap B_1) = f^{-1}(B_0) \cap f^{-1}(B_1)$.
   (d) $f^{-1}(B_0 - B_1) = f^{-1}(B_0) - f^{-1}(B_1)$.
   Show that f preserves inclusions and unions only:
   (e) $A_0 \subset A_1 \Rightarrow f(A_0) \subset f(A_1)$.

(f) $f(A_0 \cup A_1) = f(A_0) \cup f(A_1)$.
(g) $f(A_0 \cap A_1) \subset f(A_0) \cap f(A_1)$; show that equality holds if f is injective.
(h) $f(A_0 - A_1) \supset f(A_0) - f(A_1)$; show that equality holds if f is injective.

3. Show that (b), (c), (f), and (g) of Exercise 2 hold for arbitrary unions and intersections.
4. Let $f : A \to B$ and $g : B \to C$.
   (a) If $C_0 \subset C$, show that $(g \circ f)^{-1}(C_0) = f^{-1}(g^{-1}(C_0))$.
   (b) If f and g are injective, show that $g \circ f$ is injective.
   (c) If $g \circ f$ is injective, what can you say about injectivity of f and g?
   (d) If f and g are surjective, show that $g \circ f$ is surjective.
   (e) If $g \circ f$ is surjective, what can you say about surjectivity of f and g?
   (f) Summarize your answers to (b)–(e) in the form of a theorem.
5. In general, let us denote the ***identity function*** for a set C by i_C. That is, define $i_C : C \to C$ to be the function given by the rule $i_C(x) = x$ for all $x \in C$. Given $f : A \to B$, we say that a function $g : B \to A$ is a ***left inverse*** for f if $g \circ f = i_A$; and we say that $h : B \to A$ is a ***right inverse*** for f if $f \circ h = i_B$.
   (a) Show that if f has a left inverse, f is injective; and if f has a right inverse, f is surjective.
   (b) Give an example of a function that has a left inverse but no right inverse.
   (c) Give an example of a function that has a right inverse but no left inverse.
   (d) Can a function have more than one left inverse? More than one right inverse?
   (e) Show that if f has both a left inverse g and a right inverse h, then f is bijective and $g = h = f^{-1}$.
6. Let $f : \mathbb{R} \to \mathbb{R}$ be the function $f(x) = x^3 - x$. By restricting the domain and range of f appropriately, obtain from f a bijective function g. Draw the graphs of g and g^{-1}. (There are several possible choices for g.)

§3 Relations

A concept that is, in some ways, more general than that of function is the concept of a *relation*. In this section, we define what mathematicians mean by a relation, and we consider two types of relations that occur with great frequency in mathematics: *equivalence relations* and *order relations*. Order relations will be used throughout the book; equivalence relations will not be used until §22.

Definition. A ***relation*** on a set A is a subset C of the cartesian product $A \times A$.

If C is a relation on A, we use the notation xCy to mean the same thing as $(x, y) \in C$. We read it "x is in the relation C to y."

A rule of assignment r for a function $f : A \to A$ is also a subset of $A \times A$. But it is a subset of a very special kind: namely, one such that each element of A appears as the first coordinate of an element of r exactly once. *Any* subset of $A \times A$ is a relation on A.

EXAMPLE 1. Let P denote the set of all people in the world, and define $D \subset P \times P$ by the equation

$$D = \{(x, y) \mid x \text{ is a descendant of } y\}.$$

Then D is a relation on the set P. The statements "x is in the relation D to y" and "x is a descendant of y" mean precisely the same thing, namely, that $(x, y) \in D$. Two other relations on P are the following:

$$B = \{(x, y) \mid x \text{ has an ancestor who is also an ancestor of } y\},$$
$$S = \{(x, y) \mid \text{the parents of } x \text{ are the parents of } y\}.$$

We can call B the "blood relation" (pun intended), and we can call S the "sibling relation." These three relations have quite different properties. The blood relationship is symmetric, for instance (if x is a blood relative of y, then y is a blood relative of x), whereas the descendant relation is not. We shall consider these relations again shortly.

Equivalence Relations and Partitions

An ***equivalence relation*** on a set A is a relation C on A having the following three properties:

(1) (Reflexivity) xCx for every x in A.

(2) (Symmetry) If xCy, then yCx.

(3) (Transitivity) If xCy and yCz, then xCz.

EXAMPLE 2. Among the relations defined in Example 1, the descendant relation D is neither reflexive nor symmetric, while the blood relation B is not transitive (I am not a blood relation to my wife, although my children are!) The sibling relation S is, however, an equivalence relation, as you may check.

There is no reason one must use a capital letter—or indeed a letter of any sort—to denote a relation, even though it *is* a set. Another symbol will do just as well. One symbol that is frequently used to denote an equivalence relation is the "tilde" symbol $\sim$. Stated in this notation, the properties of an equivalence relation become

(1) $x \sim x$ for every x in A.

(2) If $x \sim y$, then $y \sim x$.

(3) If $x \sim y$ and $y \sim z$, then $x \sim z$.

There are many other symbols that have been devised to stand for particular equivalence relations; we shall meet some of them in the pages of this book.

Given an equivalence relation $\sim$ on a set A and an element x of A, we define a certain subset E of A, called the ***equivalence class*** determined by x, by the equation

$$E = \{y \mid y \sim x\}.$$

Note that the equivalence class E determined by x contains x, since $x \sim x$. Equivalence classes have the following property:

Lemma 3.1. *Two equivalence classes E and E' are either disjoint or equal.*

Proof. Let E be the equivalence class determined by x, and let E' be the equivalence class determined by x'. Suppose that $E \cap E'$ is not empty; let y be a point of $E \cap E'$. See Figure 3.1. We show that $E = E'$.

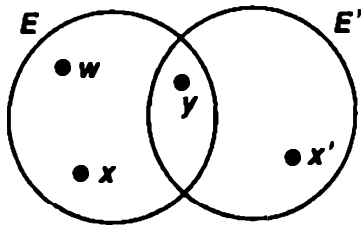

Figure 3.1

By definition, we have $y \sim x$ and $y \sim x'$. Symmetry allows us to conclude that $x \sim y$ and $y \sim x'$; from transitivity it follows that $x \sim x'$. If now w is any point of E, we have $w \sim x$ by definition; it follows from another application of transitivity that $w \sim x'$. We conclude that $E \subset E'$.

The symmetry of the situation allows us to conclude that $E' \subset E$ as well, so that $E = E'$. ■

Given an equivalence relation on a set A, let us denote by $\mathcal{E}$ the collection of all the equivalence classes determined by this relation. The preceding lemma shows that distinct elements of $\mathcal{E}$ are disjoint. Furthermore, the union of the elements of $\mathcal{E}$ equals all of A because every element of A belongs to an equivalence class. The collection $\mathcal{E}$ is a particular example of what is called a partition of A:

Definition. A ***partition*** of a set A is a collection of disjoint nonempty subsets of A whose union is all of A.

Studying equivalence relations on a set A and studying partitions of A are really the same thing. Given any partition $\mathcal{D}$ of A, there is exactly one equivalence relation on A from which it is derived.

The proof is not difficult. To show that the partition $\mathcal{D}$ comes from some equivalence relation, let us define a relation C on A by setting xCy if x and y belong to the same element of $\mathcal{D}$. Symmetry of C is obvious; reflexivity follows from the fact that the union of the elements of $\mathcal{D}$ equals all of A; transitivity follows from the fact that distinct elements of $\mathcal{D}$ are disjoint. It is simple to check that the collection of equivalence classes determined by C is precisely the collection $\mathcal{D}$.

To show there is only one such equivalence relation, suppose that C_1 and C_2 are two equivalence relations on A that give rise to the same collection of equivalence classes $\mathcal{D}$. Given $x \in A$, we show that yC_1x if and only if yC_2x, from which we conclude that $C_1 = C_2$. Let E_1 be the equivalence class determined by x relative to the relation C_1; let E_2 be the equivalence class determined by x relative to the relation C_2. Then E_1 is an element of $\mathcal{D}$, so that it must equal the unique element D of $\mathcal{D}$ that

contains x. Similarly, E_2 must equal D. Now by definition, E_1 consists of all y such that yC_1x; and E_2 consists of all y such that yC_2x. Since $E_1 = D = E_2$, our result is proved.

EXAMPLE 3 Define two points in the plane to be equivalent if they lie at the same distance from the origin. Reflexivity, symmetry, and transitivity hold trivially. The collection $\mathcal{E}$ of equivalence classes consists of all circles centered at the origin, along with the set consisting of the origin alone.

EXAMPLE 4 Define two points of the plane to be equivalent if they have the same y-coordinate. The collection of equivalence classes is the collection of all straight lines in the plane parallel to the x-axis.

EXAMPLE 5. Let $\mathcal{L}$ be the collection of all straight lines in the plane parallel to the line $y = -x$. Then $\mathcal{L}$ is a partition of the plane, since each point lies on exactly one such line. The partition $\mathcal{L}$ comes from the equivalence relation on the plane that declares the points (x_0, y_0) and (x_1, y_1) to be equivalent if $x_0 + y_0 = x_1 + y_1$.

EXAMPLE 6. Let $\mathcal{L}'$ be the collection of *all* straight lines in the plane. Then $\mathcal{L}'$ is not a partition of the plane, for distinct elements of $\mathcal{L}'$ are not necessarily disjoint; two lines may intersect without being equal.

Order Relations

A relation C on a set A is called an ***order relation*** (or a ***simple order***, or a ***linear order***) if it has the following properties:

(1) (Comparability) For every x and y in A for which $x \neq y$, either xCy or yCx.

(2) (Nonreflexivity) For no x in A does the relation xCx hold.

(3) (Transitivity) If xCy and yCz, then xCz.

Note that property (1) does not by itself exclude the possibility that for some pair of elements x and y of A, both the relations xCy and yCx hold (since "or" means "one or the other, or both"). But properties (2) and (3) combined do exclude this possibility; for if both xCy and yCx held, transitivity would imply that xCx, contradicting nonreflexivity.

EXAMPLE 7. Consider the relation on the real line consisting of all pairs (x, y) of real numbers such that $x < y$. It is an order relation, called the "usual order relation," on the real line. A less familiar order relation on the real line is the following: Define xCy if $x^2 < y^2$, or if $x^2 = y^2$ and $x < y$. You can check that this is an order relation.

EXAMPLE 8. Consider again the relationships among people given in Example 1. The blood relation B satisfies none of the properties of an order relation, and the sibling relation S satisfies only (3). The descendant relation D does somewhat better, for it satisfies both (2) and (3); however, comparability still fails. Relations that satisfy (2) and (3) occur often enough in mathematics to be given a special name. They are called *strict partial order* relations; we shall consider them later (see §11).

As the tilde, $\sim$, is the generic symbol for an equivalence relation, the "less than" symbol, $<$, is commonly used to denote an order relation. Stated in this notation, the properties of an order relation become

(1) If $x \neq y$, then either $x < y$ or $y < x$.

(2) If $x < y$, then $x \neq y$.

(3) If $x < y$ and $y < z$, then $x < z$.

We shall use the notation $x \leq y$ to stand for the statement "either $x < y$ or $x = y$"; and we shall use the notation $y > x$ to stand for the statement "$x < y$." We write $x < y < z$ to mean "$x < y$ and $y < z$."

Definition. If X is a set and $<$ is an order relation on X, and if $a < b$, we use the notation (a, b) to denote the set

$$\{x \mid a < x < b\};$$

it is called an ***open interval*** in X. If this set is empty, we call a the ***immediate predecessor*** of b, and we call b the ***immediate successor*** of a.

Definition. Suppose that A and B are two sets with order relations $<_A$ and $<_B$ respectively. We say that A and B have the same ***order type*** if there is a bijective correspondence between them that preserves order; that is, if there exists a bijective function $f : A \to B$ such that

$$a_1 <_A a_2 \Longrightarrow f(a_1) <_B f(a_2).$$

EXAMPLE 9. The interval $(-1, 1)$ of real numbers has the same order type as the set $\mathbb{R}$ of real numbers itself, for the function $f : (-1, 1) \to \mathbb{R}$ given by

$$f(x) = \frac{x}{1 - x^2}$$

is an order-preserving bijective correspondence, as you can check. It is pictured in Figure 3.2.

EXAMPLE 10. The subset $A = \{0\} \cup (1, 2)$ of $\mathbb{R}$ has the same order type as the subset

$$[0, 1) = \{x \mid 0 \leq x < 1\}$$

of $\mathbb{R}$. The function $f : A \to [0, 1)$ defined by

$$f(0) = 0,$$
$$f(x) = x - 1 \quad \text{for } x \in (1, 2)$$

is the required order-preserving correspondence.

One interesting way of defining an order relation, which will be useful to us later in dealing with some examples, is the following:

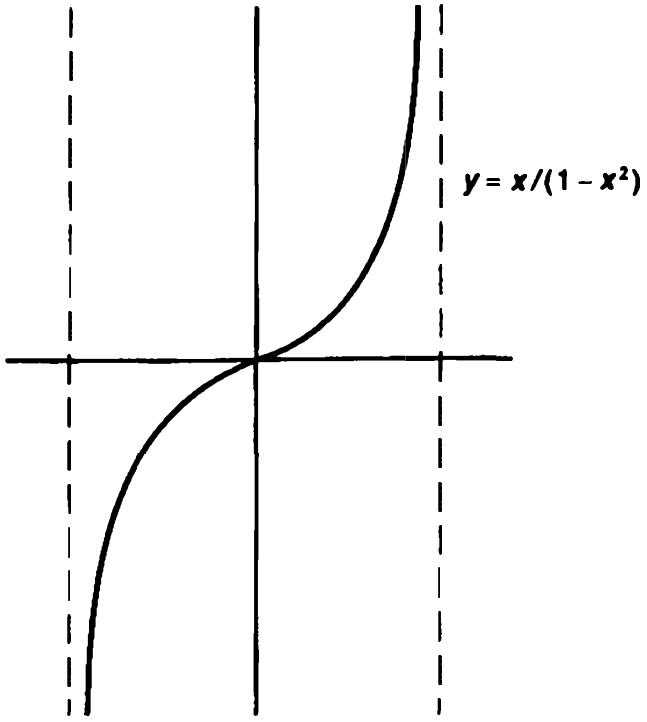

Figure 3.2

Definition. Suppose that A and B are two sets with order relations $<_A$ and $<_B$ respectively. Define an order relation $<$ on $A \times B$ by defining

$$a_1 \times b_1 < a_2 \times b_2$$

if $a_1 <_A a_2$, or if $a_1 = a_2$ and $b_1 <_B b_2$. It is called the ***dictionary order relation*** on $A \times B$.

Checking that this is an order relation involves looking at several separate cases; we leave it to you.

The reason for the choice of terminology is fairly evident. The rule defining $<$ is the same as the rule used to order the words in the dictionary. Given two words, one compares their first letters and orders the words according to the order in which their first letters appear in the alphabet. If the first letters are the same, one compares their second letters and orders accordingly. And so on.

EXAMPLE 11. Consider the dictionary order on the plane $\mathbb{R} \times \mathbb{R}$. In this order, the point p is less than every point lying above it on the vertical line through p, and p is less than every point to the right of this vertical line.

EXAMPLE 12 Consider the set $[0, 1)$ of real numbers and the set $\mathbb{Z}_+$ of positive integers, both in their usual orders; give $\mathbb{Z}_+ \times [0, 1)$ the dictionary order. This set has the same order type as the set of nonnegative reals; the function

$$f(n \times t) = n + t - 1$$

is the required bijective order-preserving correspondence. On the other hand, the set $[0, 1) \times \mathbb{Z}_+$ in the dictionary order has quite a different order type; for example, every element of this ordered set has an immediate successor. These sets are pictured in Figure 3.3.

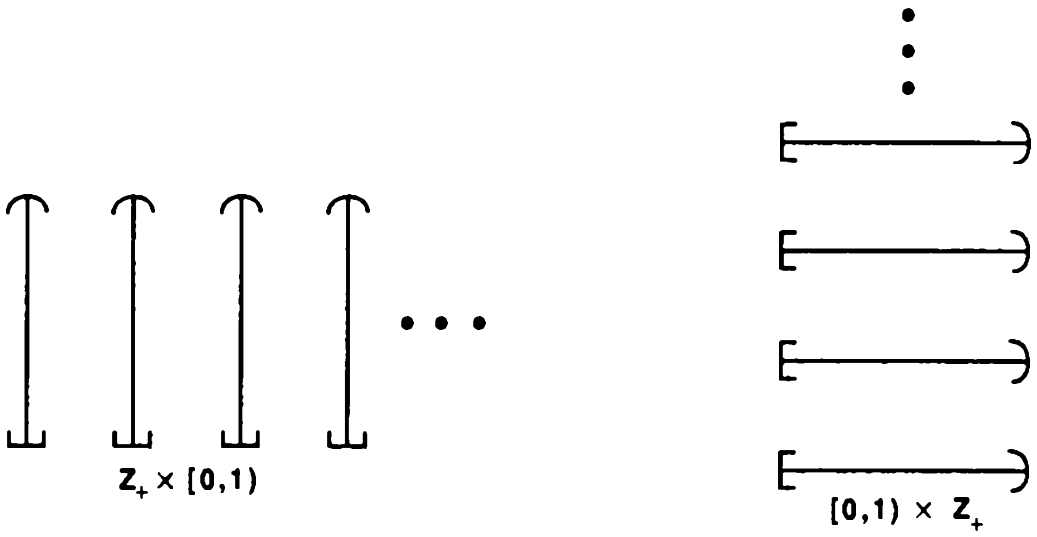

Figure 3.3

One of the properties of the real numbers that you may have seen before is the "least upper bound property." One can define this property for an arbitrary ordered set. First, we need some preliminary definitions.

Suppose that A is a set ordered by the relation $<$. Let A_0 be a subset of A. We say that the element b is the ***largest element*** of A_0 if $b \in A_0$ and if $x \leq b$ for every $x \in A_0$. Similarly, we say that a is the ***smallest element*** of A_0 if $a \in A_0$ and if $a \leq x$ for every $x \in A_0$. It is easy to see that a set has at most one largest element and at most one smallest element.

We say that the subset A_0 of A is ***bounded above*** if there is an element b of A such that $x \leq b$ for every $x \in A_0$; the element b is called an ***upper bound*** for A_0. If the set of all upper bounds for A_0 has a smallest element, that element is called the ***least upper bound***, or the ***supremum***, of A_0. It is denoted by $\sup A_0$; it may or may not belong to A_0. If it does, it is the largest element of A_0.

Similarly, A_0 is ***bounded below*** if there is an element a of A such that $a \leq x$ for every $x \in A_0$; the element a is called a ***lower bound*** for A_0. If the set of all lower bounds for A_0 has a largest element, that element is called the ***greatest lower bound***, or the ***infimum***, of A_0. It is denoted by $\inf A_0$; it may or may not belong to A_0. If it does, it is the smallest element of A_0.

Now we can define the least upper bound property.

Definition. An ordered set A is said to have the ***least upper bound property*** if every nonempty subset A_0 of A that is bounded above has a least upper bound. Analogously, the set A is said to have the ***greatest lower bound property*** if every nonempty subset A_0 of A that is bounded below has a greatest lower bound.

We leave it to the exercises to show that A has the least upper bound property if and only if it has the greatest lower bound property.

EXAMPLE 13. Consider the set $A = (-1, 1)$ of real numbers in the usual order. Assuming the fact that the real numbers have the least upper bound property, it follows that

the set A has the least upper bound property. For, given any subset of A having an upper bound *in* A, it follows that its least upper bound (in the real numbers) must be in A. For example, the subset $\{-1/2n \mid n \in \mathbb{Z}_+\}$ of A, though it has no largest element, does have a least upper bound in A, the number 0.

On the other hand, the set $B = (-1, 0) \cup (0, 1)$ does not have the least upper bound property. The subset $\{-1/2n \mid n \in \mathbb{Z}_+\}$ of B is bounded above by any element of $(0, 1)$, but it has no least upper bound in B.

Exercises

Equivalence Relations

1. Define two points (x_0, y_0) and (x_1, y_1) of the plane to be equivalent if $y_0 - x_0^2 = y_1 - x_1^2$. Check that this is an equivalence relation and describe the equivalence classes.
2. Let C be a relation on a set A. If $A_0 \subset A$, define the ***restriction*** of C to A_0 to be the relation $C \cap (A_0 \times A_0)$. Show that the restriction of an equivalence relation is an equivalence relation.
3. Here is a "proof" that every relation C that is both symmetric and transitive is also reflexive: "Since C is symmetric, aCb implies bCa. Since C is transitive, aCb and bCa together imply aCa, as desired." Find the flaw in this argument.
4. Let $f : A \to B$ be a surjective function. Let us define a relation on A by setting $a_0 \sim a_1$ if

$$f(a_0) = f(a_1).$$

   (a) Show that this is an equivalence relation.
   (b) Let A^* be the set of equivalence classes. Show there is a bijective correspondence of A^* with B.

5. Let S and S' be the following subsets of the plane:

$$S = \{(x, y) \mid y = x + 1 \text{ and } 0 < x < 2\},$$
$$S' = \{(x, y) \mid y - x \text{ is an integer}\}.$$

   (a) Show that S' is an equivalence relation on the real line and $S' \supset S$. Describe the equivalence classes of S'.
   (b) Show that given any collection of equivalence relations on a set A, their intersection is an equivalence relation on A.
   (c) Describe the equivalence relation T on the real line that is the intersection of all equivalence relations on the real line that contain S. Describe the equivalence classes of T.

Order Relations

6. Define a relation on the plane by setting

$$(x_0, y_0) < (x_1, y_1)$$

if either $y_0 - x_0^2 < y_1 - x_1^2$, or $y_0 - x_0^2 = y_1 - x_1^2$ and $x_0 < x_1$. Show that this is an order relation on the plane, and describe it geometrically.

7. Show that the restriction of an order relation is an order relation.

8. Check that the relation defined in Example 7 is an order relation.

9. Check that the dictionary order is an order relation.

10. (a) Show that the map $f : (-1, 1) \to \mathbb{R}$ of Example 9 is order preserving.
(b) Show that the equation $g(y) = 2y/[1 + (1 + 4y^2)^{1/2}]$ defines a function $g : \mathbb{R} \to (-1, 1)$ that is both a left and a right inverse for f.

11. Show that an element in an ordered set has at most one immediate successor and at most one immediate predecessor. Show that a subset of an ordered set has at most one smallest element and at most one largest element.

12. Let $\mathbb{Z}_+$ denote the set of positive integers. Consider the following order relations on $\mathbb{Z}_+ \times \mathbb{Z}_+$:

(i) The dictionary order.

(ii) $(x_0, y_0) < (x_1, y_1)$ if either $x_0 - y_0 < x_1 - y_1$, or $x_0 - y_0 = x_1 - y_1$ and $y_0 < y_1$.

(iii) $(x_0, y_0) < (x_1, y_1)$ if either $x_0 + y_0 < x_1 + y_1$, or $x_0 + y_0 = x_1 + y_1$ and $y_0 < y_1$.

In these order relations, which elements have immediate predecessors? Does the set have a smallest element? Show that all three order types are different.

13. Prove the following:
Theorem. If an ordered set A has the least upper bound property, then it has the greatest lower bound property.

14. If C is a relation on a set A, define a new relation D on A by letting $(b, a) \in D$ if $(a, b) \in C$.
(a) Show that C is symmetric if and only if $C = D$.
(b) Show that if C is an order relation, D is also an order relation.
(c) Prove the converse of the theorem in Exercise 13.

15. Assume that the real line has the least upper bound property.
(a) Show that the sets

$$[0, 1] = \{x \mid 0 \le x \le 1\},$$
$$[0, 1) = \{x \mid 0 \le x < 1\}$$

have the least upper bound property.
(b) Does $[0, 1] \times [0, 1]$ in the dictionary order have the least upper bound property? What about $[0, 1] \times [0, 1)$? What about $[0, 1) \times [0, 1]$?

§4 The Integers and the Real Numbers

Up to now we have been discussing what might be called the *logical foundations* for our study of topology—the elementary concepts of set theory. Now we turn to what we might call the *mathematical foundations* for our study—the integers and the real number system. We have already used them in an informal way in the examples and exercises of the preceding sections. Now we wish to deal with them more formally.

One way of establishing these foundations is to *construct* the real number system, using only the axioms of set theory—to build them with one's bare hands, so to speak. This way of approaching the subject takes a good deal of time and effort and is of greater logical than mathematical interest.

A second way is simply to assume a set of axioms for the real numbers and work from these axioms. In the present section, we shall sketch this approach to the real numbers. Specifically, we shall give a set of axioms for the real numbers and shall indicate how the familiar properties of real numbers and the integers are derived from them. But we shall leave most of the proofs to the exercises. If you have seen all this before, our description should refresh your memory. If not, you may want to work through the exercises in detail in order to make sure of your knowledge of the mathematical foundations.

First we need a definition from set theory.

Definition. A ***binary operation*** on a set A is a function f mapping $A \times A$ into A.

When dealing with a binary operation f on a set A, we usually use a notation different from the standard functional notation introduced in §2. Instead of denoting the value of the function f at the point (a, a') by $f(a, a')$, we usually write the symbol for the function *between* the two coordinates of the point in question, writing the value of the function at (a, a') as afa'. Furthermore (just as was the case with relations), it is more common to use some symbol other than a letter to denote an operation. Symbols often used are the plus symbol $+$, the multiplication symbols $\cdot$ and $\circ$, and the asterisk $*$; however, there are many others.

Assumption

We assume there exists a set $\mathbb{R}$, called the set of ***real numbers***, two binary operations $+$ and $\cdot$ on $\mathbb{R}$, called the addition and multiplication operations, respectively, and an order relation $<$ on $\mathbb{R}$, such that the following properties hold:

Algebraic Properties

(1) $(x + y) + z = x + (y + z)$,
$(x \cdot y) \cdot z = x \cdot (y \cdot z)$ for all x, y, z in $\mathbb{R}$.

(2) $x + y = y + x$,
$x \cdot y = y \cdot x$ for all x, y in $\mathbb{R}$.

(3) There exists a unique element of $\mathbb{R}$ called ***zero***, denoted by 0, such that $x+0=x$ for all $x \in \mathbb{R}$.
There exists a unique element of $\mathbb{R}$ called ***one***, different from 0 and denoted by 1, such that $x \cdot 1 = x$ for all $x \in \mathbb{R}$.

(4) For each x in $\mathbb{R}$, there exists a unique y in $\mathbb{R}$ such that $x + y = 0$.
For each x in $\mathbb{R}$ different from 0, there exists a unique y in $\mathbb{R}$ such that $x \cdot y = 1$.

(5) $x \cdot (y+z) = (x \cdot y) + (x \cdot z)$ for all $x, y, z \in \mathbb{R}$.

A Mixed Algebraic and Order Property

(6) If $x > y$, then $x + z > y + z$.
If $x > y$ and $z > 0$, then $x \cdot z > y \cdot z$.

Order Properties

(7) The order relation $<$ has the least upper bound property.

(8) If $x < y$, there exists an element z such that $x < z$ and $z < y$.

From properties (1)–(5) follow the familiar "laws of algebra." Given x, one denotes by $-x$ that number y such that $x + y = 0$; it is called the ***negative*** of x. One defines the ***subtraction operation*** by the formula $z - x = z + (-x)$. Similarly, given $x \neq 0$, one denotes by $1/x$ that number y such that $x \cdot y = 1$; it is called the ***reciprocal*** of x. One defines the ***quotient*** z/x by the formula $z/x = z \cdot (1/x)$. The usual laws of signs, and the rules for adding and multiplying fractions, follow as theorems. These laws of algebra are listed in Exercise 1 at the end of the section. We often denote $x \cdot y$ simply by xy.

When one adjoins property (6) to properties (1)–(5), one can prove the usual "laws of inequalities," such as the following:

$$\text{If } x > y \text{ and } z < 0, \text{ then } x \cdot z < y \cdot z.$$
$$-1 < 0 \text{ and } 0 < 1.$$

The laws of inequalities are listed in Exercise 2.

We define a number x to be ***positive*** if $x > 0$, and to be ***negative*** if $x < 0$. We denote the positive reals by $\mathbb{R}_+$ and the nonnegative reals (for reasons to be explained later) by $\bar{\mathbb{R}}_+$. Properties (1)–(6) are familiar properties in modern algebra. Any set with two binary operations satisfying (1)–(5) is called by algebraists a ***field***; if the field has an order relation satisfying (6), it is called an ***ordered field***.

Properties (7) and (8), on the other hand, are familiar properties in topology. They involve only the order relation; any set with an order relation satisfying (7) and (8) is called by topologists a ***linear continuum***.

Now it happens that when one adjoins to the axioms for an ordered field [properties (1)–(6)] the axioms for a linear continuum [properties (7) and (8)], the resulting list contains some redundancies. Property (8), in particular, can be proved as a consequence of the others; given $x < y$ one can show that $z = (x + y)/(1 + 1)$ satisfies the requirements of (8). Therefore, in the standard treatment of the real numbers, properties (1)–(7) are taken as axioms, and property (8) becomes a theorem. We have

included (8) in our list merely to emphasize the fact that it and the least upper bound property are the two crucial properties of the order relation for $\mathbb{R}$. From these two properties many of the topological properties of $\mathbb{R}$ may be derived, as we shall see in Chapter 3.

Now there is nothing in this list as it stands to tell us what an integer is. We now *define* the integers, using only properties (1)–(6).

Definition. A subset A of the real numbers is said to be ***inductive*** if it contains the number 1, and if for every x in A, the number $x+1$ is also in A. Let $\mathcal{A}$ be the collection of all inductive subsets of $\mathbb{R}$. Then the set $\mathbb{Z}_+$ of ***positive integers*** is defined by the equation

$$\mathbb{Z}_+ = \bigcap_{A \in \mathcal{A}} A.$$

Note that the set $\mathbb{R}_+$ of positive real numbers is inductive, for it contains 1 and the statement $x > 0$ implies the statement $x + 1 > 0$. Therefore, $\mathbb{Z}_+ \subset \mathbb{R}_+$, so the elements of $\mathbb{Z}_+$ are indeed positive, as the choice of terminology suggests. Indeed, one sees readily that 1 is the smallest element of $\mathbb{Z}_+$, because the set of all real numbers x for which $x \geq 1$ is inductive.

The basic properties of $\mathbb{Z}_+$, which follow readily from the definition, are the following:

(1) $\mathbb{Z}_+$ is inductive.

(2) (Principle of induction). If A is an inductive set of positive integers, then $A = \mathbb{Z}_+$.

We define the set $\mathbb{Z}$ of ***integers*** to be the set consisting of the positive integers $\mathbb{Z}_+$, the number 0, and the negatives of the elements of $\mathbb{Z}_+$. One proves that the sum, difference, and product of two integers are integers, but the quotient is not necessarily an integer. The set $\mathbb{Q}$ of quotients of integers is called the set of ***rational numbers***.

One proves also that, given the integer n, there is no integer a such that $n < a < n + 1$.

If n is a positive integer, we use the symbol S_n to denote the set of all positive integers less than n; we call it a ***section*** of the positive integers. The set S_1 is empty, and S_{n+1} denotes the set of positive integers between 1 and n, inclusive. We also use the notation

$$\{1, \ldots, n\} = S_{n+1}$$

for the latter set.

Now we prove two properties of the positive integers that may not be quite so familiar, but are quite useful. They may be thought of as alternative versions of the induction principle.

Theorem 4.1 (Well-ordering property). *Every nonempty subset of $\mathbb{Z}_+$ has a smallest element.*

Proof. We first prove that, for each $n \in \mathbb{Z}_+$, the following statement holds: *Every nonempty subset of* $\{1, \ldots, n\}$ *has a smallest element.*

Let A be the set of all positive integers n for which this statement holds. Then A contains 1, since if $n = 1$, the only nonempty subset of $\{1, \ldots, n\}$ is the set $\{1\}$ itself. Then, supposing A contains n, we show that it contains $n + 1$. So let C be a nonempty subset of the set $\{1, \ldots, n + 1\}$. If C consists of the single element $n + 1$, then that element is the smallest element of C. Otherwise, consider the set $C \cap \{1, \ldots, n\}$, which is nonempty. Because $n \in A$, this set has a smallest element, which will automatically be the smallest element of C also. Thus A is inductive, so we conclude that $A = \mathbb{Z}_+$; hence the statement is true for all $n \in \mathbb{Z}_+$.

Now we prove the theorem. Suppose that D is a nonempty subset of $\mathbb{Z}_+$. Choose an element n of D. Then the set $A = D \cap \{1, \ldots, n\}$ is nonempty, so that A has a smallest element k. The element k is automatically the smallest element of D as well. ■

Theorem 4.2 (Strong induction principle). *Let A be a set of positive integers. Suppose that for each positive integer n, the statement $S_n \subset A$ implies the statement $n \in A$. Then $A = \mathbb{Z}_+$.*

Proof. If A does not equal all of $\mathbb{Z}_+$, let n be the smallest positive integer that is not in A. Then every positive integer less than n is in A, so that $S_n \subset A$. Our hypothesis implies that $n \in A$, contrary to assumption. ■

Everything we have done up to now has used only the axioms for an ordered field, properties (1)–(6) of the real numbers. At what point do you need (7), the least upper bound axiom?

For one thing, you need the least upper bound axiom to prove that the set $\mathbb{Z}_+$ of positive integers has no upper bound in $\mathbb{R}$. This is the ***Archimedean ordering property*** of the real line. To prove it, we assume that $\mathbb{Z}_+$ has an upper bound and derive a contradiction. If $\mathbb{Z}_+$ has an upper bound, it has a least upper bound b. There exists $n \in \mathbb{Z}_+$ such that $n > b - 1$; for otherwise, $b - 1$ would be an upper bound for $\mathbb{Z}_+$ smaller than b. Then $n + 1 > b$, contrary to the fact that b is an upper bound for $\mathbb{Z}_+$.

The least upper bound axiom is also used to prove a number of other things about $\mathbb{R}$. It is used for instance to show that $\mathbb{R}$ has the greatest lower bound property. It is also used to prove the existence of a unique positive square root $\sqrt{x}$ for every positive real number. This fact, in turn, can be used to demonstrate the existence of real numbers that are not rational numbers; the number $\sqrt{2}$ is an easy example.

We use the symbol 2 to denote $1 + 1$, the symbol 3 to denote $2 + 1$, and so on through the standard symbols for the positive integers. It is a fact that this procedure assigns to each positive integer a unique symbol, but we never need this fact and shall not prove it.

Proofs of these properties of the integers and real numbers, along with a few other properties we shall need later, are outlined in the exercises that follow.

Exercises

1. Prove the following "laws of algebra" for $\mathbb{R}$, using only axioms (1)–(5):
   (a) If $x + y = x$, then $y = 0$.
   (b) $0 \cdot x = 0$. [*Hint:* Compute $(x + 0) \cdot x$.]
   (c) $-0 = 0$.
   (d) $-(-x) = x$.
   (e) $x(-y) = -(xy) = (-x)y$.
   (f) $(-1)x = -x$.
   (g) $x(y - z) = xy - xz$.
   (h) $-(x + y) = -x - y$; $-(x - y) = -x + y$.
   (i) If $x \neq 0$ and $x \cdot y = x$, then $y = 1$.
   (j) $x/x = 1$ if $x \neq 0$.
   (k) $x/1 = x$.
   (l) $x \neq 0$ and $y \neq 0 \Rightarrow xy \neq 0$.
   (m) $(1/y)(1/z) = 1/(yz)$ if $y, z \neq 0$.
   (n) $(x/y)(w/z) = (xw)/(yz)$ if $y, z \neq 0$.
   (o) $(x/y) + (w/z) = (xz + wy)/(yz)$ if $y, z \neq 0$.
   (p) $x \neq 0 \Rightarrow 1/x \neq 0$.
   (q) $1/(w/z) = z/w$ if $w, z \neq 0$.
   (r) $(x/y)/(w/z) = (xz)/(yw)$ if $y, w, z \neq 0$.
   (s) $(ax)/y = a(x/y)$ if $y \neq 0$.
   (t) $(-x)/y = x/(-y) = -(x/y)$ if $y \neq 0$.
2. Prove the following "laws of inequalities" for $\mathbb{R}$, using axioms (1)–(6) along with the results of Exercise 1:
   (a) $x > y$ and $w > z \Rightarrow x + w > y + z$.
   (b) $x > 0$ and $y > 0 \Rightarrow x + y > 0$ and $x \cdot y > 0$.
   (c) $x > 0 \Leftrightarrow -x < 0$.
   (d) $x > y \Leftrightarrow -x < -y$.
   (e) $x > y$ and $z < 0 \Rightarrow xz < yz$.
   (f) $x \neq 0 \Rightarrow x^2 > 0$, where $x^2 = x \cdot x$.
   (g) $-1 < 0 < 1$
   (h) $xy > 0 \Leftrightarrow x$ and y are both positive or both negative.
   (i) $x > 0 \Rightarrow 1/x > 0$.
   (j) $x > y > 0 \Rightarrow 1/x < 1/y$.
   (k) $x < y \Rightarrow x < (x + y)/2 < y$.
3. (a) Show that if $\mathcal{A}$ is a collection of inductive sets, then the intersection of the elements of $\mathcal{A}$ is an inductive set.
   (b) Prove the basic properties (1) and (2) of $\mathbb{Z}_+$.
4. (a) Prove by induction that given $n \in \mathbb{Z}_+$, every nonempty subset of $\{1, \ldots, n\}$ has a largest element.
   (b) Explain why you cannot conclude from (a) that every nonempty subset of $\mathbb{Z}_+$ has a largest element.

5. Prove the following properties of $\mathbb{Z}$ and $\mathbb{Z}_+$:
(a) $a, b \in \mathbb{Z}_+ \Rightarrow a + b \in \mathbb{Z}_+$. [*Hint:* Show that given $a \in \mathbb{Z}_+$, the set $X = \{x \mid x \in \mathbb{R} \text{ and } a + x \in \mathbb{Z}_+\}$ is inductive.]
(b) $a, b \in \mathbb{Z}_+ \Rightarrow a \cdot b \in \mathbb{Z}_+$.
(c) Show that $a \in \mathbb{Z}_+ \Rightarrow a - 1 \in \mathbb{Z}_+ \cup \{0\}$. [*Hint:* Let $X = \{x \mid x \in \mathbb{R} \text{ and } x - 1 \in \mathbb{Z}_+ \cup \{0\}\}$; show that X is inductive.]
(d) $c, d \in \mathbb{Z} \Rightarrow c + d \in \mathbb{Z}$ and $c - d \in \mathbb{Z}$. [*Hint.* Prove it first for $d = 1$.]
(e) $c, d \in \mathbb{Z} \Rightarrow c \cdot d \in \mathbb{Z}$.

6. Let $a \in \mathbb{R}$. Define inductively

$$a^1 = a,$$
$$a^{n+1} = a^n \cdot a$$

for $n \in \mathbb{Z}_+$. (See §7 for a discussion of the process of inductive definition.) Show that for $n, m \in \mathbb{Z}_+$ and $a, b \in \mathbb{R}$,

$$a^n a^m = a^{n+m},$$
$$(a^n)^m = a^{nm},$$
$$a^m b^m = (ab)^m.$$

These are called the ***laws of exponents***. [*Hint:* For fixed n, prove the formulas by induction on m.]

7. Let $a \in \mathbb{R}$ and $a \neq 0$. Define $a^0 = 1$, and for $n \in \mathbb{Z}_+$, $a^{-n} = 1/a^n$. Show that the laws of exponents hold for $a, b \neq 0$ and $n, m \in \mathbb{Z}$.

8. (a) Show that $\mathbb{R}$ has the greatest lower bound property.
(b) Show that $\inf\{1/n \mid n \in \mathbb{Z}_+\} = 0$.
(c) Show that given a with $0 < a < 1$, $\inf\{a^n \mid n \in \mathbb{Z}_+\} = 0$. [*Hint:* Let $h = (1 - a)/a$, and show that $(1 + h)^n \geq 1 + nh$.]

9. (a) Show that every nonempty subset of $\mathbb{Z}$ that is bounded above has a largest element.
(b) If $x \notin \mathbb{Z}$, show there is exactly one $n \in \mathbb{Z}$ such that $n < x < n + 1$.
(c) If $x - y > 1$, show there is at least one $n \in \mathbb{Z}$ such that $y < n < x$.
(d) If $y < x$, show there is a rational number z such that $y < z < x$.

10. Show that every positive number a has exactly one positive square root, as follows:
(a) Show that if $x > 0$ and $0 \leq h < 1$, then

$$(x + h)^2 \leq x^2 + h(2x + 1),$$
$$(x - h)^2 \geq x^2 - h(2x).$$

(b) Let $x > 0$. Show that if $x^2 < a$, then $(x + h)^2 < a$ for some $h > 0$; and if $x^2 > a$, then $(x - h)^2 > a$ for some $h > 0$.

(c) Given $a > 0$, let B be the set of all real numbers x such that $x^2 < a$. Show that B is bounded above and contains at least one positive number. Let $b = \sup B$; show that $b^2 = a$.
(d) Show that if b and c are positive and $b^2 = c^2$, then $b = c$.

11. Given $m \in \mathbb{Z}$, we say that m is ***even*** if $m/2 \in \mathbb{Z}$, and m is ***odd*** otherwise.
(a) Show that if m is odd, $m = 2n + 1$ for some $n \in \mathbb{Z}$. [*Hint:* Choose n so that $n < m/2 < n + 1$.]
(b) Show that if p and q are odd, so are $p \cdot q$ and p^n, for any $n \in \mathbb{Z}_+$.
(c) Show that if $a > 0$ is rational, then $a = m/n$ for some $m, n \in \mathbb{Z}_+$ where not both m and n are even. [*Hint:* Let n be the smallest element of the set $\{x \mid x \in \mathbb{Z}_+ \text{ and } x \cdot a \in \mathbb{Z}_+\}$.]
(d) *Theorem.* $\sqrt{2}$ *is irrational.*

§5 Cartesian Products

We have already defined what we mean by the cartesian product $A \times B$ of two sets. Now we introduce more general cartesian products.

Definition. Let $\mathcal{A}$ be a nonempty collection of sets. An ***indexing function*** for $\mathcal{A}$ is a surjective function f from some set J, called the ***index set***, to $\mathcal{A}$. The collection $\mathcal{A}$, together with the indexing function f, is called an ***indexed family of sets***. Given $\alpha \in J$, we shall denote the set $f(\alpha)$ by the symbol A_α. And we shall denote the indexed family itself by the symbol

$$\{A_\alpha\}_{\alpha \in J},$$

which is read "the family of all A_α, as α ranges over J." Sometimes we write merely $\{A_\alpha\}$, if it is clear what the index set is.

Note that although an indexing function is required to be surjective, it is not required to be *injective*. It is entirely possible for A_α and A_β to be the same set of $\mathcal{A}$, even though $\alpha \neq \beta$.

One way in which indexing functions are used is to give a new notation for arbitrary unions and intersections of sets. Suppose that $f : J \to \mathcal{A}$ is an indexing function for $\mathcal{A}$; let A_α denote $f(\alpha)$. Then we define

$$\bigcup_{\alpha \in J} A_\alpha = \{x \mid \text{for at least one } \alpha \in J, x \in A_\alpha\},$$

and

$$\bigcap_{\alpha \in J} A_\alpha = \{x \mid \text{for every } \alpha \in J, x \in A_\alpha\}.$$

These are simply new notations for previously defined concepts; one sees at once (using the surjectivity of the index function) that the first equals the union of all the elements of $\mathcal{A}$ and the second equals the intersection of all the elements of $\mathcal{A}$.

Two especially useful index sets are the set $\{1, \ldots, n\}$ of positive integers from 1 to n, and the set $\mathbb{Z}_+$ of all positive integers. For these index sets, we introduce some special notation. If a collection of sets is indexed by the set $\{1, \ldots, n\}$, we denote the indexed family by the symbol $\{A_1, \ldots, A_n\}$, and we denote the union and intersection, respectively, of the members of this family by the symbols

$$A_1 \cup \cdots \cup A_n \quad \text{and} \quad A_1 \cap \cdots \cap A_n.$$

In the case where the index set is the set $\mathbb{Z}_+$, we denote the indexed family by the symbol $\{A_1, A_2, \ldots\}$, and the union and intersection by the respective symbols

$$A_1 \cup A_2 \cup \cdots \quad \text{and} \quad A_1 \cap A_2 \cap \cdots.$$

Definition. Let m be a positive integer. Given a set X, we define an ***m-tuple*** of elements of X to be a function

$$\mathbf{x} : \{1, \ldots, m\} \to X.$$

If $\mathbf{x}$ is an m-tuple, we often denote the value of $\mathbf{x}$ at i by the symbol x_i rather than $\mathbf{x}(i)$ and call it the ith ***coordinate*** of $\mathbf{x}$. And we often denote the function $\mathbf{x}$ itself by the symbol

$$(x_1, \ldots, x_m).$$

Now let $\{A_1, \ldots, A_m\}$ be a family of sets indexed with the set $\{1, \ldots, m\}$. Let $X = A_1 \cup \cdots \cup A_m$. We define the ***cartesian product*** of this indexed family, denoted by

$$\prod_{i=1}^{m} A_i \quad \text{or} \quad A_1 \times \cdots \times A_m,$$

to be the set of all m-tuples $(x_1, \ldots, x_m)$ of elements of X such that $x_i \in A_i$ for each i.

EXAMPLE 1. We now have two definitions for the symbol $A \times B$. One definition is, of course, the one given earlier, under which $A \times B$ denotes the set of all ordered pairs (a, b) such that $a \in A$ and $b \in B$. The second definition, just given, defines $A \times B$ as the set of all functions $\mathbf{x} : \{1, 2\} \to A \cup B$ such that $\mathbf{x}(1) \in A$ and $\mathbf{x}(2) \in B$. There is an obvious bijective correspondence between these two sets, under which the ordered pair (a, b) corresponds to the function $\mathbf{x}$ defined by $\mathbf{x}(1) = a$ and $\mathbf{x}(2) = b$. Since we commonly denote this function $\mathbf{x}$ in "tuple notation" by the symbol (a, b), the notation itself suggests the correspondence. Thus for the cartesian product of two sets, the general definition of cartesian product reduces essentially to the earlier one.

EXAMPLE 2. How does the cartesian product $A \times B \times C$ differ from the cartesian products $A \times (B \times C)$ and $(A \times B) \times C$? Very little. There are obvious bijective correspondences between these sets, indicated as follows

$$(a, b, c) \longleftrightarrow (a, (b, c)) \longleftrightarrow ((a, b), c).$$

Definition. Given a set X, we define an ***ω-tuple*** of elements of X to be a function

$$\mathbf{x} : \mathbb{Z}_+ \longrightarrow X;$$

we also call such a function a ***sequence***, or an ***infinite sequence***, of elements of X. If $\mathbf{x}$ is an ω-tuple, we often denote the value of $\mathbf{x}$ at i by x_i rather than $\mathbf{x}(i)$, and call it the ith ***coordinate*** of $\mathbf{x}$. We denote $\mathbf{x}$ itself by the symbol

$$(x_1, x_2, \dots) \quad \text{or} \quad (x_n)_{n \in \mathbb{Z}_+}.$$

Now let $\{A_1, A_2, \dots\}$ be a family of sets, indexed with the positive integers; let X be the union of the sets in this family. The ***cartesian product*** of this indexed family of sets, denoted by

$$\prod_{i \in \mathbb{Z}_+} A_i \quad \text{or} \quad A_1 \times A_2 \times \cdots,$$

is defined to be the set of all ω-tuples $(x_1, x_2, \dots)$ of elements of X such that $x_i \in A_i$ for each i.

Nothing in these definitions requires the sets A_i to be different from one another. Indeed, they may all equal the same set X. In that case, the cartesian product $A_1 \times \cdots \times A_m$ is just the set of *all* m-tuples of elements of X, which we denote by X^m. Similarly, the product $A_1 \times A_2 \times \cdots$ is just the set of all ω-tuples of elements of X, which we denote by X^ω.

Later we will define the cartesian product of an *arbitrary* indexed family of sets.

EXAMPLE 3. If $\mathbb{R}$ is the set of real numbers, then $\mathbb{R}^n$ denotes the set of all n-tuples of real numbers; it is often called ***euclidean n-space*** (although Euclid would never recognize it). Analogously, $\mathbb{R}^\omega$ is sometimes called "infinite-dimensional euclidean space"; it is the set of all ω-tuples $(x_1, x_2, \dots)$ of real numbers, that is, the set of all functions $\mathbf{x} : \mathbb{Z}_+ \to \mathbb{R}$.

Exercises

1. Show there is a bijective correspondence of $A \times B$ with $B \times A$.
2. (a) Show that if $n > 1$ there is bijective correspondence of

$$A_1 \times \cdots \times A_n \quad \text{with} \quad (A_1 \times \cdots \times A_{n-1}) \times A_n.$$

   (b) Given the indexed family $\{A_1, A_2, \dots\}$, let $B_i = A_{2i-1} \times A_{2i}$ for each positive integer i. Show there is bijective correspondence of $A_1 \times A_2 \times \cdots$ with $B_1 \times B_2 \times \cdots$.
3. Let $A = A_1 \times A_2 \times \cdots$ and $B = B_1 \times B_2 \times \cdots$.
   (a) Show that if $B_i \subset A_i$ for all i, then $B \subset A$. (Strictly speaking, if we are given a function mapping the index set $\mathbb{Z}_+$ into the union of the sets B_i, we must change its range before it can be considered as a function mapping $\mathbb{Z}_+$ into the union of the sets A_i. We shall ignore this technicality when dealing with cartesian products).

(b) Show the converse of (a) holds if B is nonempty.
(c) Show that if A is nonempty, each A_i is nonempty. Does the converse hold? (We will return to this question in the exercises of §19.)
(d) What is the relation between the set $A \cup B$ and the cartesian product of the sets $A_i \cup B_i$? What is the relation between the set $A \cap B$ and the cartesian product of the sets $A_i \cap B_i$?

4. Let $m, n \in \mathbb{Z}_+$. Let $X \neq \varnothing$.
(a) If $m \leq n$, find an injective map $f : X^m \to X^n$.
(b) Find a bijective map $g : X^m \times X^n \to X^{m+n}$.
(c) Find an injective map $h : X^n \to X^\omega$.
(d) Find a bijective map $k : X^n \times X^\omega \to X^\omega$.
(e) Find a bijective map $l : X^\omega \times X^\omega \to X^\omega$.
(f) If $A \subset B$, find an injective map $m : (A^\omega)^n \to B^\omega$.

5. Which of the following subsets of $\mathbb{R}^\omega$ can be expressed as the cartesian product of subsets of $\mathbb{R}$?
(a) $\{\mathbf{x} \mid x_i \text{ is an integer for all } i\}$.
(b) $\{\mathbf{x} \mid x_i \geq i \text{ for all } i\}$.
(c) $\{\mathbf{x} \mid x_i \text{ is an integer for all } i \geq 100\}$.
(d) $\{\mathbf{x} \mid x_2 = x_3\}$.

§6 Finite Sets

Finite sets and infinite sets, countable sets and uncountable sets, these are types of sets that you may have encountered before. Nevertheless, we shall discuss them in this section and the next, not only to make sure you understand them thoroughly, but also to elucidate some particular points of logic that will arise later on. First we consider finite sets.

Recall that if n is a positive integer, we use S_n to denote the set of positive integers less than n; it is called a *section* of the positive integers. The sets S_n are the prototypes for what we call the finite sets.

Definition. A set is said to be ***finite*** if there is a bijective correspondence of A with some section of the positive integers. That is, A is finite if it is empty or if there is a bijection

$$f : A \longrightarrow \{1, \ldots, n\}$$

for some positive integer n. In the former case, we say that A has ***cardinality 0***; in the latter case, we say that A has ***cardinality n***.

For instance, the set $\{1, \ldots, n\}$ itself has cardinality n, for it is in bijective correspondence with itself under the identity function.

Now note carefully: *We have not yet shown that the cardinality of a finite set is uniquely determined by the set.* It is of course clear that the empty set must have cardinality zero. But as far as we know, there might exist bijective correspondences of a given nonempty set A with two different sets $\{1, \ldots, n\}$ and $\{1, \ldots, m\}$. The possibility may seem ridiculous, for it is like saying that it is possible for two people to count the marbles in a box and come out with two different answers, *both correct.* Our experience with counting in everyday life suggests that such is impossible, and in fact this is easy to prove when n is a small number such as 1, 2, or 3. But a direct proof when n is 5 million would be impossibly demanding.

Even empirical demonstration would be difficult for such a large value of n. One might, for instance, construct an experiment by taking a freight car full of marbles and hiring 10 different people to count them independently. If one thinks of the physical problems involved, it seems likely that the counters would not all arrive at the same answer. Of course, the conclusion one could draw is that at least one person made a mistake. But that would mean assuming the correctness of the result one was trying to demonstrate empirically An alternative explanation could be that there do exist bijective correspondences between the given set of marbles and two different sections of the positive integers.

In real life, we accept the first explanation. We simply take it on faith that our experience in counting comparatively small sets of objects demonstrates a truth that holds for arbitrarily large sets as well.

However, in mathematics (as opposed to real life), one does not have to take this statement on faith. If it is formulated in terms of the existence of bijective correspondences rather than in terms of the physical act of counting, it is capable of mathematical proof. We shall prove shortly that if $n \neq m$, there do not exist bijective functions mapping a given set A onto both the sets $\{1, \ldots, n\}$ and $\{1, \ldots, m\}$.

There are a number of other "intuitively obvious" facts about finite sets that are capable of mathematical proof; we shall prove some of them in this section and leave the rest to the exercises. Here is an easy fact to start with:

Lemma 6.1. *Let n be a positive integer. Let A be a set; let a_0 be an element of A. Then there exists a bijective correspondence f of the set A with the set $\{1, \ldots, n+1\}$ if and only if there exists a bijective correspondence g of the set $A - \{a_0\}$ with the set $\{1, \ldots, n\}$.*

Proof. There are two implications to be proved. Let us first assume that there is a bijective correspondence

$$g : A - \{a_0\} \longrightarrow \{1, \ldots, n\}.$$

We then define a function $f : A \longrightarrow \{1, \ldots, n+1\}$ by setting

$$f(x) = g(x) \quad \text{for } x \in A - \{a_0\},$$
$$f(a_0) = n + 1.$$

One checks at once that f is bijective.

To prove the converse, assume there is a bijective correspondence

$$f : A \longrightarrow \{1, \ldots, n+1\}.$$

If f maps a_0 to the number $n+1$, things are especially easy; in that case, the restriction $f|A - \{a_0\}$ is the desired bijective correspondence of $A - \{a_0\}$ with $\{1, \ldots, n\}$. Otherwise, let $f(a_0) = m$, and let a_1 be the point of A such that $f(a_1) = n+1$. Then $a_1 \neq a_0$. Define a new function

$$h : A \longrightarrow \{1, \ldots, n+1\}$$

by setting

$$\begin{aligned} h(a_0) &= n+1, \\ h(a_1) &= m, \\ h(x) &= f(x) \qquad \text{for } x \in A - \{a_0\} - \{a_1\}. \end{aligned}$$

See Figure 6.1. It is easy to check that h is a bijection.

Now we are back in the easy case; the restriction $h|A - \{a_0\}$ is the desired bijection of $A - \{a_0\}$ with $\{1, \ldots, n\}$. ■

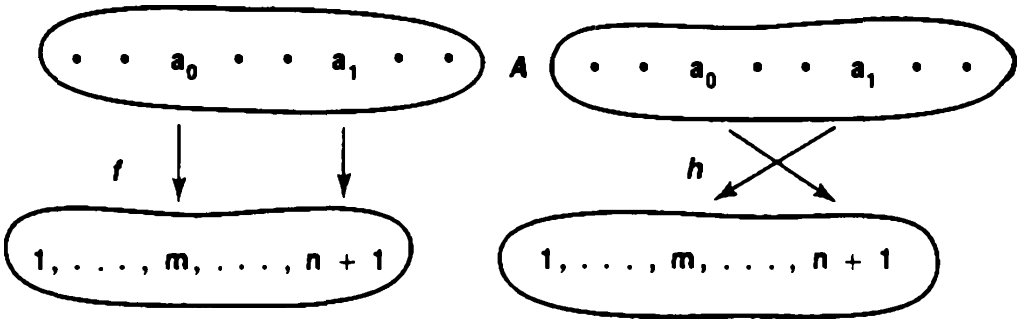

Figure 6.1

From this lemma a number of useful consequences follow:

Theorem 6.2. *Let A be a set; suppose that there exists a bijection $f : A \to \{1, \ldots, n\}$ for some $n \in \mathbb{Z}_+$. Let B be a proper subset of A. Then there exists no bijection $g : B \to \{1, \ldots, n\}$; but (provided $B \neq \emptyset$) there does exist a bijection $h : B \to \{1, \ldots, m\}$ for some $m < n$.*

Proof. The case in which $B = \emptyset$ is trivial, for there cannot exist a bijection of the empty set B with the nonempty set $\{1, \ldots, n\}$.

We prove the theorem "by induction." Let C be the subset of $\mathbb{Z}_+$ consisting of those integers n for which the theorem holds. We shall show that C is inductive. From this we conclude that $C = \mathbb{Z}_+$, so the theorem is true for all positive integers n.

First we show the theorem is true for $n = 1$. In this case A consists of a single element $\{a\}$, and its only proper subset B is the empty set.

Now assume that the theorem is true for n; we prove it true for $n + 1$. Suppose that $f : A \to \{1, \ldots, n + 1\}$ is a bijection, and B is a nonempty proper subset of A. Choose an element a_0 of B and an element a_1 of $A - B$. We apply the preceding lemma to conclude there is a bijection

$$g : A - \{a_0\} \longrightarrow \{1, \ldots, n\}.$$

Now $B - \{a_0\}$ is a proper subset of $A - \{a_0\}$, for a_1 belongs to $A - \{a_0\}$ and not to $B - \{a_0\}$. Because the theorem has been assumed to hold for the integer n, we conclude the following:

(1) There exists no bijection $h : B - \{a_0\} \to \{1, \ldots, n\}$.

(2) Either $B - \{a_0\} = \varnothing$, or there exists a bijection

$$k : B - \{a_0\} \longrightarrow \{1, \ldots, p\} \qquad \text{for some } p < n.$$

The preceding lemma, combined with (1), implies that there is no bijection of B with $\{1, \ldots, n + 1\}$. This is the first half of what we wanted to proved. To prove the second half, note that if $B - \{a_0\} = \varnothing$, there is a bijection of B with the set $\{1\}$; while if $B - \{a_0\} \neq \varnothing$, we can apply the preceding lemma, along with (2), to conclude that there is a bijection of B with $\{1, \ldots, p + 1\}$. In either case, there is a bijection of B with $\{1, \ldots, m\}$ for some $m < n + 1$, as desired. The induction principle now shows that the theorem is true for all $n \in \mathbb{Z}_+$. ■

Corollary 6.3. *If A is finite, there is no bijection of A with a proper subset of itself.*

Proof. Assume that B is a proper subset of A and that $f : A \to B$ is a bijection. By assumption, there is a bijection $g : A \to \{1, \ldots, n\}$ for some n. The composite $g \circ f^{-1}$ is then a bijection of B with $\{1, \ldots, n\}$. This contradicts the preceding theorem. ■

Corollary 6.4. *$\mathbb{Z}_+$ is not finite.*

Proof. The function $f : \mathbb{Z}_+ \to \mathbb{Z}_+ - \{1\}$ defined by $f(n) = n + 1$ is a bijection of $\mathbb{Z}_+$ with a proper subset of itself. ■

Corollary 6.5. *The cardinality of a finite set A is uniquely determined by A.*

Proof. Let $m < n$. Suppose there are bijections

$$f : A \longrightarrow \{1, \ldots, n\},$$
$$g : A \longrightarrow \{1, \ldots, m\}.$$

Then the composite

$$g \circ f^{-1} : \{1, \ldots, n\} \longrightarrow \{1, \ldots, m\}$$

is a bijection of the finite set $\{1, \ldots, n\}$ with a proper subset of itself, contradicting the corollary just proved. ■

Corollary 6.6. *If B is a subset of the finite set A, then B is finite. If B is a proper subset of A, then the cardinality of B is less than the cardinality of A.*

Corollary 6.7. *Let B be a nonempty set. Then the following are equivalent:*

(1) *B is finite.*

(2) *There is a surjective function from a section of the positive integers onto B.*

(3) *There is an injective function from B into a section of the positive integers.*

Proof. (1) $\Longrightarrow$ (2). Since B is nonempty, there is, for some n, a bijective function $f : \{1, \ldots, n\} \to B$.

(2) $\Longrightarrow$ (3). If $f : \{1, \ldots, n\} \to B$ is surjective, define $g : B \to \{1, \ldots, n\}$ by the equation

$$g(b) = \text{smallest element of } f^{-1}(\{b\}).$$

Because f is surjective, the set $f^{-1}\{(b)\}$ is nonempty; then the well-ordering property of $\mathbb{Z}_+$ tells us that $g(b)$ is uniquely defined. The map g is injective, for if $b \neq b'$, then the sets $f^{-1}(\{b\})$ and $f^{-1}(\{b'\})$ are disjoint, so their smallest elements must be different.

(3) $\Longrightarrow$ (1). If $g : B \to \{1, \ldots, n\}$ is injective, then changing the range of g gives a bijection of B with a subset of $\{1, \ldots, n\}$. It follows from the preceding corollary that B is finite. ■

Corollary 6.8. *Finite unions and finite cartesian products of finite sets are finite.*

Proof. We first show that if A and B are finite, so is $A \cup B$. The result is trivial if A or B is empty. Otherwise, there are bijections $f : \{1, \ldots, m\} \to A$ and $g : \{1, \ldots, n\} \to B$ for some choice of m and n. Define a function $h : \{1, \ldots, m + n\} \to A \cup B$ by setting $h(i) = f(i)$ for $i = 1, 2, \ldots, m$ and $h(i) = g(i - m)$ for $i = m + 1, \ldots, m + n$. It is easy to check that h is surjective, from which it follows that $A \cup B$ is finite.

Now we show by induction that finiteness of the sets $A_1, \ldots, A_n$ implies finiteness of their union. This result is trivial for $n = 1$. Assuming it true for $n - 1$, we note that $A_1 \cup \cdots \cup A_n$ is the union of the two finite sets $A_1 \cup \cdots \cup A_{n-1}$ and A_n, so the result of the preceding paragraph applies.

Now we show that the cartesian product of two finite sets A and B is finite. Given $a \in A$, the set $\{a\} \times B$ is finite, being in bijective correspondence with B. The set $A \times B$ is the union of these sets; since there are only finitely many of them, $A \times B$ is a finite union of finite sets and thus finite.

To prove that the product $A_1 \times \cdots \times A_n$ is finite if each A_i is finite, one proceeds by induction. ■

Exercises

1. (a) Make a list of all the injective maps
$$f : \{1, 2, 3\} \longrightarrow \{1, 2, 3, 4\}.$$
Show that none is bijective. (This constitutes a *direct* proof that a set A of cardinality three does not have cardinality four.)
(b) How many injective maps
$$f : \{1, \ldots, 8\} \longrightarrow \{1, \ldots, 10\}$$
are there? (You can see why one would not wish to try to prove *directly* that there is no bijective correspondence between these sets.)
2. Show that if B is not finite and $B \subset A$, then A is not finite.
3. Let X be the two-element set $\{0, 1\}$. Find a bijective correspondence between X^ω and a proper subset of itself.
4. Let A be a nonempty finite simply ordered set.
(a) Show that A has a largest element. [*Hint:* Proceed by induction on the cardinality of A.]
(b) Show that A has the order type of a section of the positive integers.
5. If $A \times B$ is finite, does it follow that A and B are finite?
6. (a) Let $A = \{1, \ldots, n\}$. Show there is a bijection of $\mathcal{P}(A)$ with the cartesian product X^n, where X is the two-element set $X = \{0, 1\}$.
(b) Show that if A is finite, then $\mathcal{P}(A)$ is finite.
7. If A and B are finite, show that the set of all functions $f : A \to B$ is finite.

§7 Countable and Uncountable Sets

Just as sections of the positive integers are the prototypes for the finite sets, the set of all the positive integers is the prototype for what we call the *countably infinite* sets. In this section, we shall study such sets; we shall also construct some sets that are neither finite nor countably infinite. This study will lead us into a discussion of what we mean by the process of "inductive definition."

Definition. A set A is said to be ***infinite*** if it is not finite. It is said to be ***countably infinite*** if there is a bijective correspondence
$$f : A \longrightarrow \mathbb{Z}_+.$$

EXAMPLE 1. The set $\mathbb{Z}$ of all integers is countably infinite. One checks easily that the function $f : \mathbb{Z} \to \mathbb{Z}_+$ defined by
$$f(n) = \begin{cases} 2n & \text{if } n > 0, \\ -2n + 1 & \text{if } n \leq 0 \end{cases}$$
is a bijection.

EXAMPLE 2. The product $\mathbb{Z}_+ \times \mathbb{Z}_+$ is countably infinite. If we represent the elements of the product $\mathbb{Z}_+ \times \mathbb{Z}_+$ by the integer points in the first quadrant, then the left-hand portion of Figure 7.1 suggests how to "count" the points, that is, how to put them in bijective correspondence with the positive integers. A picture is not a proof, of course, but this picture suggests a proof. First, we define a bijection $f : \mathbb{Z}_+ \times \mathbb{Z}_+ \to A$, where A is the subset of $\mathbb{Z}_+ \times \mathbb{Z}_+$ consisting of pairs (x, y) for which $y \leq x$, by the equation

$$f(x, y) = (x + y - 1, y).$$

Then we construct a bijection of A with the positive integers, defining $g : A \to \mathbb{Z}_+$ by the formula

$$g(x, y) = \frac{1}{2}(x - 1)x + y.$$

We leave it to you to show that f and g are bijections.

Another proof that $\mathbb{Z}_+ \times \mathbb{Z}_+$ is countably infinite will be given later.

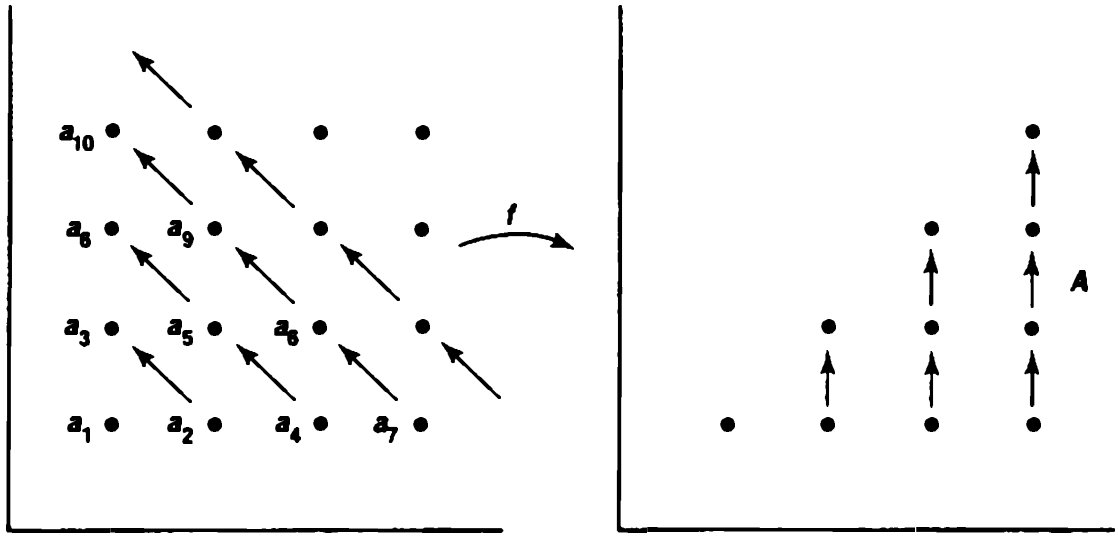

Figure 7.1

Definition. A set is said to be ***countable*** if it is either finite or countably infinite. A set that is not countable is said to be ***uncountable***.

There is a very useful criterion for showing that a set is countable. It is the following:

Theorem 7.1. *Let B be a nonempty set. Then the following are equivalent:*

(1) *B is countable.*

(2) *There is a surjective function $f : \mathbb{Z}_+ \to B$.*

(3) *There is an injective function $g : B \to \mathbb{Z}_+$.*

Proof. (1) $\Longrightarrow$ (2). Suppose that B is countable. If B is countably infinite, there is a bijection $f : \mathbb{Z}_+ \to B$ by definition, and we are through. If B is finite, there is a

bijection $h : \{1, \ldots, n\} \to B$ for some $n \geq 1$. (Recall that $B \neq \varnothing$.) We can extend h to a surjection $f : \mathbb{Z}_+ \to B$ by defining

$$f(i) = \begin{cases} h(i) & \text{for } 1 \leq i \leq n, \\ h(1) & \text{for } i > n. \end{cases}$$

(2) $\Longrightarrow$ (3). Let $f : \mathbb{Z}_+ \to B$ be a surjection. Define $g : B \to \mathbb{Z}_+$ by the equation

$$g(b) = \text{smallest element of } f^{-1}(\{b\}).$$

Because f is surjective, $f^{-1}(\{b\})$ is nonempty; thus g is well defined. The map g is injective, for if $b \neq b'$, the sets $f^{-1}(\{b\})$ and $f^{-1}(\{b'\})$ are disjoint, so their smallest elements are different.

(3) $\Longrightarrow$ (1). Let $g : B \to \mathbb{Z}_+$ be an injection; we wish to prove B is countable. By changing the range of g, we can obtain a bijection of B with a subset of $\mathbb{Z}_+$. Thus to prove our result, it suffices to show that every subset of $\mathbb{Z}_+$ is countable. So let C be a subset of $\mathbb{Z}_+$.

If C is finite, it is countable by definition. So what we need to prove is that every infinite subset C of $\mathbb{Z}_+$ is countably infinite. This statement is certainly plausible. For the elements of C can easily be arranged in an infinite sequence; one simply takes the set $\mathbb{Z}_+$ in its usual order and "erases" all the elements of $\mathbb{Z}_+$ that are not in C!

The plausibility of this argument may make one overlook its informality. Providing a formal proof requires a certain amount of care. We state this result as a separate lemma, which follows. ■

Lemma 7.2. *If C is an infinite subset of $\mathbb{Z}_+$, then C is countably infinite.*

Proof. We define a bijection $h : \mathbb{Z}_+ \to C$. We proceed by induction. Define $h(1)$ to be the smallest element of C; it exists because every nonempty subset C of $\mathbb{Z}_+$ has a smallest element. Then assuming that $h(1), \ldots, h(n-1)$ are defined, define

$$h(n) = \text{ smallest element of } [C - h(\{1, \ldots, n-1\})].$$

The set $C - h(\{1, \ldots, n-1\})$ is not empty; for if it were empty, then $h : \{1, \ldots, n-1\} \to C$ would be surjective, so that C would be finite (by Corollary 6.7). Thus $h(n)$ is well defined. By induction, we have defined $h(n)$ for all $n \in \mathbb{Z}_+$.

To show that h is injective is easy. Given $m < n$, note that $h(m)$ belongs to the set $h(\{1, \ldots, n-1\})$, whereas $h(n)$, by definition, does not. Hence $h(n) \neq h(m)$.

To show that h is surjective, let c be any element of C; we show that c lies in the image set of h. First note that $h(\mathbb{Z}_+)$ cannot be contained in the finite set $\{1, \ldots, c\}$, because $h(\mathbb{Z}_+)$ is infinite (since h is injective). Therefore, there is an n in $\mathbb{Z}_+$, such that $h(n) > c$. Let m be the *smallest* element of $\mathbb{Z}_+$, such that $h(m) \geq c$. Then for all $i < m$, we must have $h(i) < c$. Thus, c does not belong to the set $h(\{1, \ldots, m-1\})$. Since $h(m)$ is defined as the smallest element of the set $C - h(\{1, \ldots, m-1\})$, we must have $h(m) \leq c$. Putting the two inequalities together, we have $h(m) = c$, as desired. ■

There is a point in the preceding proof where we stretched the principles of logic a bit. It occurred at the point where we said that "using the induction principle" we had defined the function h for all positive integers n. You may have seen arguments like this used before, with no questions raised concerning their legitimacy. We have already used such an argument ourselves, in the exercises of §4, when we defined a^n.

But there is a problem here. After all, the induction principle states only that if A is an inductive set of positive integers, then $A = \mathbb{Z}_+$. To use the principle to prove a theorem "by induction," one begins the proof with the statement "Let A be the set of all positive integers n for which the theorem is true," and then one goes ahead to prove that A is inductive, so that A must be all of $\mathbb{Z}_+$.

In the preceding theorem, however, we were not really proving a theorem by induction, but defining something by induction. How then should we start the proof? Can we start by saying, "Let A be the set of all integers n for which the function h is defined"? But that's silly; the symbol h has no *meaning* at the outset of the proof. It only takes on meaning in the course of the proof. So something more is needed.

What is needed is another principle, which we call the ***principle of recursive definition***. In the proof of the preceding theorem, we wished to assert the following:

Given the infinite subset C of $\mathbb{Z}_+$, there is a unique function $h : \mathbb{Z}_+ \to C$ satisfying the formula:

$$(*) \qquad \begin{aligned} h(1) &= \text{smallest element of } C, \\ h(i) &= \text{smallest element of } [C - h(\{1, \ldots, i-1\})] \quad \text{for all } i > 1. \end{aligned}$$

The formula $(*)$ is called a ***recursion formula*** for h; it defines the function h *in terms of itself*. A definition given by such a formula is called a ***recursive definition***.

Now one can get into logical difficulties when one tries to define something recursively. Not all recursive formulas make sense. The recursive formula

$$h(i) = \text{smallest element of } [C - h(\{1, \ldots, i+1\})],$$

for example, is self-contradictory; although $h(i)$ necessarily is an element of the set $h(\{1, \ldots, i+1\})$, this formula says that it does not belong to the set. Another example is the classic paradox:

> Let the barber of Seville shave every man of Seville who does not shave himself.
> Who shall shave the barber?

In this statement, the barber appears twice, once in the phrase "the barber of Seville" and once as an element of the set "men of Seville"; this definition of whom the barber shall shave is a recursive one. It also happens to be self-contradictory.

Some recursive formulas do make sense, however. Specifically, one has the following principle:

Principle of recursive definition. *Let A be a set. Given a formula that defines $h(1)$ as a unique element of A, and for $i > 1$ defines $h(i)$ uniquely as an element of A in terms of the values of h for positive integers less than i, this formula determines a unique function $h : \mathbb{Z}_+ \to A$.*

This principle is the one we actually used in the proof of Lemma 7.2. You can simply accept it on faith if you like. It may however be proved rigorously, using the principle of induction. We shall formulate it more precisely in the next section and indicate how it is proved. Mathematicians seldom refer to this principle specifically. They are much more likely to write a proof like our proof of Lemma 7.2 above, a proof in which they invoke the "induction principle" to define a function when what they are really using is the principle of recursive definition. We shall avoid undue pedantry in this book by following their example.

Corollary 7.3. *A subset of a countable set is countable.*

Proof. Suppose $A \subset B$, where B is countable. There is an injection f of B into $\mathbb{Z}_+$; the restriction of f to A is an injection of A into $\mathbb{Z}_+$. ■

Corollary 7.4. *The set $\mathbb{Z}_+ \times \mathbb{Z}_+$ is countably infinite.*

Proof. In view of Theorem 7.1, it suffices to construct an injective map $f : \mathbb{Z}_+ \times \mathbb{Z}_+ \to \mathbb{Z}_+$. We define f by the equation

$$f(n, m) = 2^n 3^m.$$

It is easy to check that f is injective. For suppose that $2^n 3^m = 2^p 3^q$. If $n < p$, then $3^m = 2^{p-n} 3^q$, contradicting the fact that 3^m is odd for all m. Therefore, $n = p$. As a result, $3^m = 3^q$, Then if $m < q$, it follows that $1 = 3^{q-m}$, another contradiction. Hence $m = q$. ■

EXAMPLE 3. The set $\mathbb{Q}_+$ of positive rational numbers is countably infinite. For we can define a surjection $g : \mathbb{Z}_+ \times \mathbb{Z}_+ \to \mathbb{Q}_+$ by the equation

$$g(n, m) = m/n.$$

Because $\mathbb{Z}_+ \times \mathbb{Z}_+$ is countable, there is a surjection $f : \mathbb{Z}_+ \to \mathbb{Z}_+ \times \mathbb{Z}_+$. Then the composite $g \circ f : \mathbb{Z}_+ \to \mathbb{Q}_+$ is a surjection, so that $\mathbb{Q}_+$ is countable. And, of course, $\mathbb{Q}_+$ is infinite because it contains $\mathbb{Z}_+$.

We leave it as an exercise to show the set $\mathbb{Q}$ of *all* rational numbers is countably infinite.

Theorem 7.5. *A countable union of countable sets is countable.*

Proof. Let $\{A_n\}_{n \in J}$ be an indexed family of countable sets, where the index set J is either $\{1, \ldots, N\}$ or $\mathbb{Z}_+$. Assume that each set A_n is nonempty, for convenience; this assumption does not change anything.

Because each A_n is countable, we can choose, for each n, a surjective function $f_n : \mathbb{Z}_+ \to A_n$. Similarly, we can choose a surjective function $g : \mathbb{Z}_+ \to J$. Now define

$$h : \mathbb{Z}_+ \times \mathbb{Z}_+ \to \bigcup_{n \in J} A_n$$

by the equation

$$h(k, m) = f_{g(k)}(m).$$

It is easy to check that h is surjective. Since $\mathbb{Z}_+ \times \mathbb{Z}_+$ is in bijective correspondence with $\mathbb{Z}_+$, the countability of the union follows from Theorem 7.1. ■

Theorem 7.6. *A finite product of countable sets is countable.*

Proof. First let us show that the product of two countable sets A and B is countable. The result is trivial if A or B is empty. Otherwise, choose surjective functions $f : \mathbb{Z}_+ \to A$ and $g : \mathbb{Z}_+ \to B$. Then the function $h : \mathbb{Z}_+ \times \mathbb{Z}_+ \to A \times B$ defined by the equation $h(n, m) = (f(n), g(m))$ is surjective, so that $A \times B$ is countable.

In general, we proceed by induction. Assuming that $A_1 \times \cdots \times A_{n-1}$ is countable if each A_i is countable, we prove the same thing for the product $A_1 \times \cdots \times A_n$. First, note that there is a bijective correspondence

$$g : A_1 \times \cdots \times A_n \longrightarrow (A_1 \times \cdots \times A_{n-1}) \times A_n$$

defined by the equation

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

Because the set $A_1 \times \cdots \times A_{n-1}$ is countable by the induction assumption and A_n is countable by hypothesis, the product of these two sets is countable, as proved in the preceding paragraph. We conclude that $A_1 \times \cdots \times A_n$ is countable as well. ■

It is very tempting to assert that countable products of countable sets should be countable; but this assertion is in fact not true:

Theorem 7.7. *Let X denote the two element set $\{0, 1\}$. Then the set X^ω is uncountable.*

Proof. We show that, given any function

$$g : \mathbb{Z}_+ \longrightarrow X^\omega,$$

g is not surjective. For this purpose, let us denote $g(n)$ as follows :

$$g(n) = (x_{n1}, x_{n2}, x_{n3}, \ldots x_{nm}, \ldots),$$

where each x_{ij} is either 0 or 1. Then we define an element $\mathbf{y} = (y_1, y_2, \ldots, y_n, \ldots)$ of X^ω by letting

$$y_n = \begin{cases} 0 & \text{if } x_{nn} = 1, \\ 1 & \text{if } x_{nn} = 0. \end{cases}$$

(If we write the numbers x_{ni} in a rectangular array, the particular elements x_{nn} appear as the diagonal entries in this array; we choose **y** so that its nth coordinate *differs* from the diagonal entry x_{nn}.)

Now **y** is an element of X^ω, and **y** does not lie in the image of g; given n, the point $g(n)$ and the point **y** differ in at least one coordinate, namely, the nth. Thus, g is not surjective. ■

The cartesian product $\{0, 1\}^\omega$ is one example of an uncountable set. Another is the set $\mathcal{P}(\mathbb{Z}_+)$, as the following theorem implies:

Theorem 7.8. *Let A be a set. There is no injective map $f : \mathcal{P}(A) \to A$, and there is no surjective map $g : A \to \mathcal{P}(A)$.*

Proof. In general, if B is a nonempty set, the existence of an injective map $f : B \to C$ implies the existence of a surjective map $g : C \to B$; one defines $g(c) = f^{-1}(c)$ for each c in the image set of f, and defines g arbitrarily on the rest of C.

Therefore, it suffices to prove that given a map $g : A \to \mathcal{P}(A)$, the map g is not surjective. For each $a \in A$, the image $g(a)$ of a is a subset of A, which may or may not contain the point a itself. Let B be the subset of A consisting of all those points a such that $g(a)$ does not contain a;

$$B = \{a \mid a \in A - g(a)\}.$$

Now, B may be empty, or it may be all of A, but that does not matter. We assert that B is a subset of A that does not lie in the image of g. For suppose that $B = g(a_0)$ for some $a_0 \in A$. We ask the question: Does a_0 belong to B or not? By definition of B,

$$a_0 \in B \iff a_0 \in A - g(a_0) \iff a_0 \in A - B.$$

In either case, we have a contradiction. ■

Now we have proved the existence of uncountable sets. But we have not yet mentioned the most familiar uncountable set of all—the set of real numbers. You have probably seen the uncountability of $\mathbb{R}$ demonstrated already. If one assumes that every real number can be represented uniquely by an infinite decimal (with the proviso that a representation ending in an infinite string of 9's is forbidden), then the uncountability of the reals can be proved by a variant of the diagonal procedure used in the proof of Theorem 7.7. But this proof is in some ways not very satisfying. One reason is that the infinite decimal representation of a real number is not at all an elementary consequence of the axioms but requires a good deal of labor to prove. Another reason is that the uncountability of $\mathbb{R}$ does not, in fact, depend on the infinite decimal expansion of $\mathbb{R}$ or indeed on any of the algebraic properties of $\mathbb{R}$; it depends on only the order properties of $\mathbb{R}$. We shall demonstrate the uncountability of $\mathbb{R}$, using only its order properties, in a later chapter.

Exercises

1. Show that $\mathbb{Q}$ is countably infinite.
2. Show that the maps f and g of Examples 1 and 2 are bijections.
3. Let X be the two-element set $\{0, 1\}$. Show there is a bijective correspondence between the set $\mathcal{P}(\mathbb{Z}_+)$ and the cartesian product X^ω.
4. (a) A real number x is said to be ***algebraic*** (over the rationals) if it satisfies some polynomial equation of positive degree

$$x^n + a_{n-1}x^{n-1} + \cdots + a_1 x + a_0 = 0$$

with rational coefficients a_i. Assuming that each polynomial equation has only finitely many roots, show that the set of algebraic numbers is countable.

(b) A real number is said to be ***transcendental*** if it is not algebraic. Assuming the reals are uncountable, show that the transcendental numbers are uncountable. (It is a somewhat surprising fact that only two transcendental numbers are familiar to us: e and π. Even proving these two numbers transcendental is highly nontrivial.)

5. Determine, for each of the following sets, whether or not it is countable. Justify your answers.
   (a) The set A of all functions $f : \{0, 1\} \to \mathbb{Z}_+$.
   (b) The set B_n of all functions $f : \{1, \ldots, n\} \to \mathbb{Z}_+$.
   (c) The set $C = \bigcup_{n \in \mathbb{Z}_+} B_n$.
   (d) The set D of all functions $f : \mathbb{Z}_+ \to \mathbb{Z}_+$.
   (e) The set E of all functions $f : \mathbb{Z}_+ \to \{0, 1\}$.
   (f) The set F of all functions $f : \mathbb{Z}_+ \to \{0, 1\}$ that are "eventually zero." [We say that f is ***eventually zero*** if there is a positive integer N such that $f(n) = 0$ for all $n \geq N$.]
   (g) The set G of all functions $f : \mathbb{Z}_+ \to \mathbb{Z}_+$ that are eventually 1.
   (h) The set H of all functions $f : \mathbb{Z}_+ \to \mathbb{Z}_+$ that are eventually constant.
   (i) The set I of all two-element subsets of $\mathbb{Z}_+$.
   (j) The set J of all finite subsets of $\mathbb{Z}_+$.
6. We say that two sets A and B ***have the same cardinality*** if there is a bijection of A with B.
   (a) Show that if $B \subset A$ and if there is an injection

$$f : A \longrightarrow B,$$

then A and B have the same cardinality. [*Hint:* Define $A_1 = A$, $B_1 = B$, and for $n > 1$, $A_n = f(A_{n-1})$ and $B_n = f(B_{n-1})$. (Recursive definition again!) Note that $A_1 \supset B_1 \supset A_2 \supset B_2 \supset A_3 \supset \cdots$. Define a bijection $h : A \to B$ by the rule

$$h(x) = \begin{cases} f(x) & \text{if } x \in A_n - B_n \text{ for some } n, \\ x & \text{otherwise.} \end{cases}$$

(b) *Theorem (Schroeder-Bernstein theorem). If there are injections $f : A \to C$ and $g : C \to A$, then A and C have the same cardinality.*

7. Show that the sets D and E of Exercise 5 have the same cardinality.

8. Let X denote the two-element set $\{0, 1\}$; let $\mathcal{B}$ be the set of *countable* subsets of X^ω. Show that X^ω and $\mathcal{B}$ have the same cardinality.

9. (a) The formula

$$(*) \qquad \begin{aligned} h(1) &= 1, \\ h(2) &= 2, \\ h(n) &= [h(n+1)]^2 - [h(n-1)]^2 \quad \text{for } n \geq 2 \end{aligned}$$

is not one to which the principle of recursive definition applies. Show that nevertheless there does exist a function $h : \mathbb{Z}_+ \to \mathbb{R}$ satisfying this formula. [*Hint:* Reformulate $(*)$ so that the principle will apply and require h to be positive.]

(b) Show that the formula $(*)$ of part (a) does not determine h uniquely. [*Hint:* If h is a positive function satisfying $(*)$, let $f(i) = h(i)$ for $i \neq 3$, and let $f(3) = -h(3)$.]

(c) Show that there is no function $h : \mathbb{Z}_+ \to \mathbb{R}$ satisfying the formula

$$\begin{aligned} h(1) &= 1, \\ h(2) &= 2, \\ h(n) &= [h(n+1)]^2 + [h(n-1)]^2 \quad \text{for } n \geq 2. \end{aligned}$$

*§8 The Principle of Recursive Definition

Before considering the general form of the principle of recursive definition, let us first prove it in a specific case, that of Lemma 7.2. That should make the underlying idea of the proof much clearer when we consider the general case.

So, given the infinite subset C of $\mathbb{Z}_+$, let us consider the following recursion formula for a function $h : \mathbb{Z}_+ \to C$:

$$(*) \qquad \begin{aligned} h(1) &= \text{smallest element of } C, \\ h(i) &= \text{smallest element of } [C - h(\{1, \ldots, i-1\})] \quad \text{for } i > 1. \end{aligned}$$

We shall prove that there exists a unique function $h : \mathbb{Z}_+ \to C$ satisfying this recursion formula.

The first step is to prove that there exist functions defined on *sections* $\{1, \ldots, n\}$ of $\mathbb{Z}_+$ that satisfy $(*)$:

Lemma 8.1. *Given $n \in \mathbb{Z}_+$, there exists a function*

$$f : \{1, \ldots, n\} \to C$$

that satisfies $(*)$ for all i in its domain.