From: Munkres, Topology (2nd Ed)

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 $\bar{\mathbb{R}}_+$ denote the nonnegative reals. Consider the functions

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

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 $\bar{\mathbb{R}}_+$. 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.

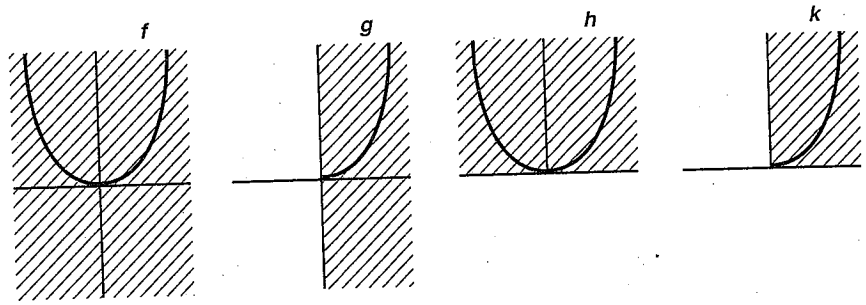

Figure 2.1

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.

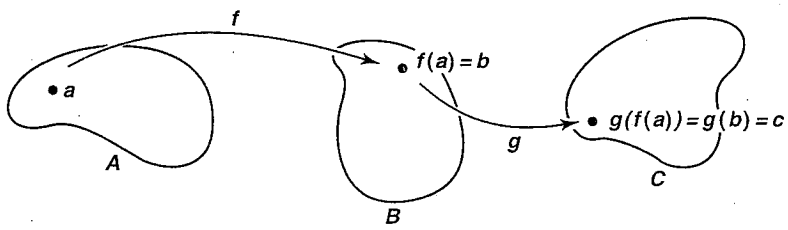

Figure 2.2

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

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

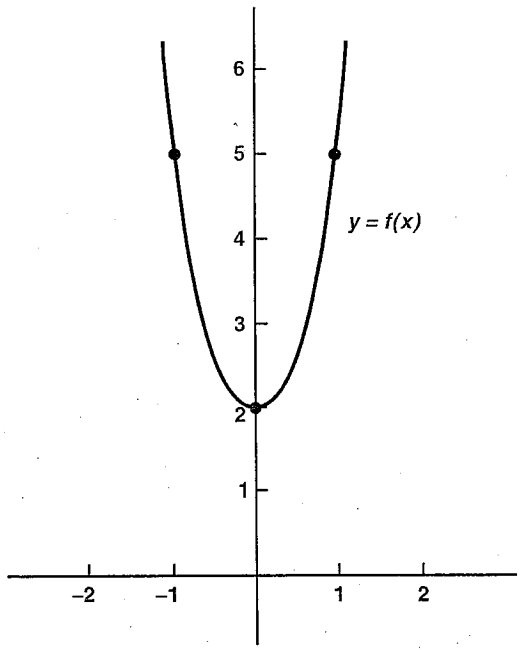

Figure 2.3

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^{-1} 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 $E = E'$.

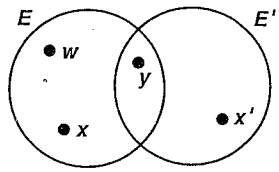

Figure 3.1

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:

(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 $\mathbb{R}^m$ denotes the set of all m-tuples of real numbers; it is often called ***euclidean m-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.

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.

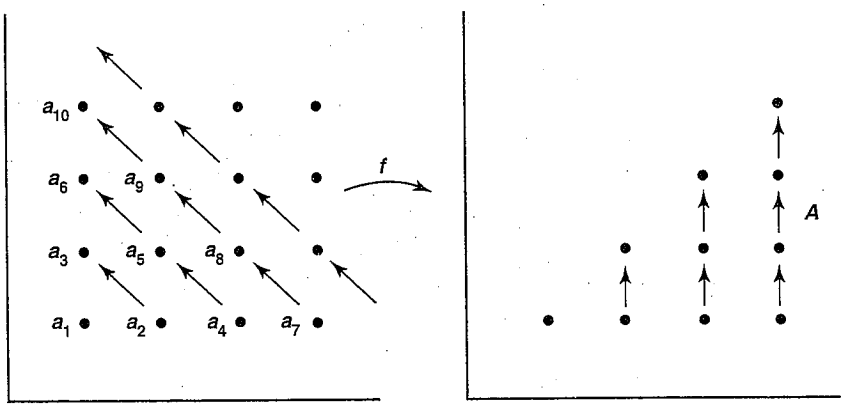

Figure 7.1

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. ∎

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

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

Axiom of choice. *Given a collection $\mathcal{A}$ of disjoint nonempty sets, there exists a set C consisting of exactly one element from each element of $\mathcal{A}$; that is, a set C such that C is contained in the union of the elements of $\mathcal{A}$, and for each $A \in \mathcal{A}$, the set $C \cap A$ contains a single element.*

The set C can be thought of as having been obtained by choosing one element from each of the sets in $\mathcal{A}$.

The axiom of choice certainly seems an innocent-enough assertion. And, in fact, most mathematicians today accept it as part of the set theory on which they base their mathematics. But in years past a good deal of controversy raged around this particular assertion concerning set theory, for there are theorems one can prove with its aid that some mathematicians were reluctant to accept. One such is the well-ordering theorem, which we shall discuss shortly. For the present we shall simply use the choice axiom to clear up the difficulty we mentioned in the preceding proof. First, we prove an easy consequence of the axiom of choice:

Lemma 9.2 (Existence of a choice function). *Given a collection $\mathcal{B}$ of nonempty sets (not necessarily disjoint), there exists a function*

$$c : \mathcal{B} \longrightarrow \bigcup_{B \in \mathcal{B}} B$$

such that $c(B)$ is an element of B, for each $B \in \mathcal{B}$.

The function c is called a ***choice function*** for the collection $\mathcal{B}$.

The difference between this lemma and the axiom of choice is that in this lemma the sets of the collection $\mathcal{B}$ are not required to be disjoint. For example, one can allow $\mathcal{B}$ to be the collection of *all* nonempty subsets of a given set.

Proof of the lemma. Given an element B of $\mathcal{B}$, we define a set B' as follows:

$$B' = \{(B, x) \mid x \in B\}.$$

That is, B' is the collection of all ordered pairs, where the first coordinate of the ordered pair is the set B, and the second coordinate is an element of B. The set B' is a subset of the cartesian product

$$\mathcal{B} \times \bigcup_{B \in \mathcal{B}} B.$$

Because B contains at least one element x, the set B' contains at least the element (B, x), so it is nonempty.

Now we claim that if B_1 and B_2 are two different sets in $\mathcal{B}$, then the corresponding sets B_1' and B_2' are disjoint. For the typical element of B_1' is a pair of the form (B_1, x_1)

for each $x \in J$, where S_x is the section of J by x.

Proof.

(a) If h and k map sections of J, or all of J, into C and satisfy $(*)$ for all x in their respective domains, show that $h(x) = k(x)$ for all x in both domains.

(b) If there exists a function $h : S_\alpha \to C$ satisfying $(*)$, show that there exists a function $k : S_\alpha \cup \{\alpha\} \to C$ satisfying $(*)$.

(c) If $K \subset J$ and for all $\alpha \in K$ there exists a function $h_\alpha : S_\alpha \to C$ satisfying $(*)$, show that there exists a function

$$k : \bigcup_{\alpha \in K} S_\alpha \longrightarrow C$$

satisfying $(*)$.

(d) Show by transfinite induction that for every $\beta \in J$, there exists a function $h_\beta : S_\beta \to C$ satisfying $(*)$. [*Hint:* If β has an immediate predecessor α, then $S_\beta = S_\alpha \cup \{\alpha\}$. If not, S_β is the union of all S_α with $\alpha < \beta$.]

(e) Prove the theorem.

11. Let A and B be two sets. Using the well-ordering theorem, prove that either they have the same cardinality, or one has cardinality greater than the other. [*Hint:* If there is no surjection $f : A \to B$, apply the preceding exercise.]

*§11 The Maximum Principle†

We have already indicated that the axiom of choice leads to the deep theorem that every set can be well-ordered. The axiom of choice has other consequences that are even more important in mathematics. Collectively referred to as "maximum principles," they come in many versions. Formulated independently by a number of mathematicians, including F. Hausdorff, K. Kuratowski, S. Bochner, and M. Zorn, during the years 1914–1935, they were typically proved as consequences of the well-ordering theorem. Later, it was realized that they were in fact *equivalent* to the well-ordering theorem. We consider several of them here.

First, we make a definition. Given a set A, a relation $\prec$ on A is called a ***strict partial order*** on A if it has the following two properties:

(1) (Nonreflexivity) The relation $a \prec a$ never holds.

(2) (Transitivity) If $a \prec b$ and $b \prec c$, then $a \prec c$.

These are just the second and third of the properties of a simple order (see §3); the comparability property is the one that is omitted. In other words, a strict partial order behaves just like a simple order except that it need not be true that for every pair of distinct points x and y in the set, either $x \prec y$ or $y \prec x$.

If $\prec$ is a strict partial order on a set A, it can easily happen that some subset B of A is simply ordered by the relation; all that is needed is for every pair of elements of B to be comparable under $\prec$.

†This section will be assumed in Chapters 5 and 14.

Now we can state the following principle, which was first formulated by Hausdorff in 1914.

Theorem (The maximum principle). *Let A be a set; let $\prec$ be a strict partial order on A. Then there exists a maximal simply ordered subset B of A.*

Said differently, there exists a subset B of A such that B is simply ordered by $\prec$ and such that no subset of A that properly contains B is simply ordered by $\prec$.

EXAMPLE 1. If $\mathcal{A}$ is any collection of sets, the relation "is a proper subset of" is a strict partial order on $\mathcal{A}$. Suppose that $\mathcal{A}$ is the collection of all circular regions (interiors of circles) in the plane. One maximal simply ordered subcollection of $\mathcal{A}$ consists of all circular regions with centers at the origin. Another maximal simply ordered subcollection consists of all circular regions bounded by circles tangent from the right to the y-axis at the origin. See Figure 11.1.

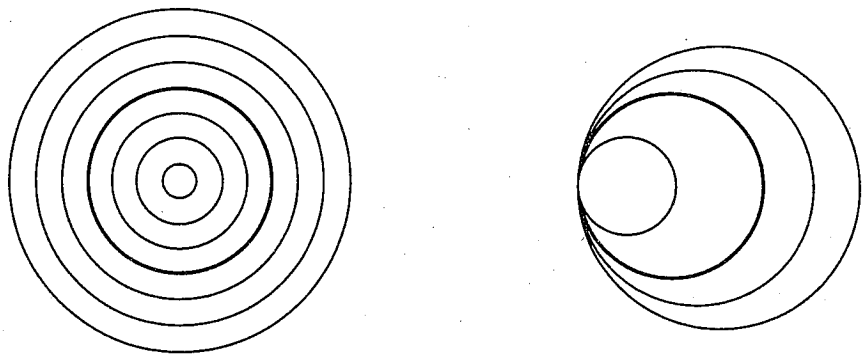

Figure 11.1

EXAMPLE 2. If (x_0, y_0) and (x_1, y_1) are two points of the plane $\mathbb{R}^2$, define

$$(x_0, y_0) \prec (x_1, y_1)$$

if $y_0 = y_1$ and $x_0 < x_1$. This is a partial ordering of $\mathbb{R}^2$ under which two points are comparable only if they lie on the same horizontal line. The maximal simply ordered sets are the horizontal lines in $\mathbb{R}^2$.

One can give an intuitive "proof" of the maximum principle that is rather appealing. It involves a step-by-step procedure, which one can describe in physical terms as follows. Suppose we take a box, and put into it some of the elements of A according to the following plan: First we pick an arbitrary element of A and put it in the box. Then we pick another element of A. If it is comparable with the element in the box, we put it in the box too; otherwise, we throw it away. At the general step, we will have a collection of elements in the box and a collection of elements that have been tossed away. Take one of the remaining elements of A. If it is comparable with everything in the box, toss it in the box, too; otherwise, throw it away. Similarly continue. After

you have checked all the elements of A, the elements you have in the box will be comparable with one another, and thus they will form a simply ordered set. Every element not in the box will be noncomparable with at least one element in the box, for that was why it was tossed away. Hence, the simply ordered set in the box is maximal, for no larger subset of A can satisfy the comparability condition.

Now of course the weak point in the preceding "proof" comes when we said, "After you have checked all the elements of A." How do you know you ever "get through" checking all the elements of A? If A should happen to be countable, it is not hard to make this intuitive proof into a real proof. Let us take the countably infinite case; the finite case is even easier. Index the elements of A bijectively with the positive integers, so that $A = \{a_1, a_2 \dots\}$. This indexing gives a way of deciding what order to test the elements of A in, and how to know when one has tested them all.

Now we define a function $h : \mathbb{Z}_+ \to \{0, 1\}$, by letting it assign the value 0 to i if we "put a_i in the box," and the value 1 if we "throw a_i away." This means that $h(1) = 0$, and for $i > 1$, we have $h(i) = 0$ if and only if a_i is comparable with every element of the set

$$\{a_j \mid j < i \text{ and } h(j) = 0\}.$$

By the principle of recursive definition, this formula determines a unique function $h : \mathbb{Z}_+ \to \{0, 1\}$. It is easy to check that the set of those a_j for which $h(j) = 0$ is a maximal simply ordered subset of A.

If A is not countable, a variant of this procedure will work, *if we allow ourselves to use the well-ordering theorem.* Instead of indexing the elements of A with the set $\mathbb{Z}_+$, we index them (in a bijective fashion) with the elements of some well-ordered set J, so that $A = \{a_\alpha \mid \alpha \in J\}$. For this we need the well-ordering theorem, so that we know there is a bijection between A and some well-ordered set J. Then we can proceed as in the previous paragraph, letting α replace i in the argument. Strictly speaking, you need to generalize the principle of recursive definition to well-ordered sets as well, but that is not particularly difficult. (See the Supplementary Exercises.)

Thus, the well-ordering theorem implies the maximum principle.

Although the maximum principle of Hausdorff was the first to be formulated and is probably the simplest to understand, there is another such principle that is nowadays the one most frequently quoted. It is popularly called "Zorn's Lemma," although Kuratowski (1922) and Bochner (1922) preceded Zorn (1935) in enunciating and proving versions of it. For a history and discussion of the tangled history of these ideas, see [C] or [Mo]. To state this principle, we need some terminology.

Definition. Let A be a set and let $\prec$ be a strict partial order on A. If B is a subset of A, an ***upper bound*** on B is an element c of A such that for every b in B, either $b = c$ or $b \prec c$. A ***maximal element*** of A is an element m of A such that for no element a of A does the relation $m \prec a$ hold.

Zorn's Lemma. *Let A be a set that is strictly partially ordered. If every simply ordered subset of A has an upper bound in A, then A has a maximal element.*

Zorn's lemma is an easy consequence of the maximum principle: Given A, the maximum principle implies that A has a maximal simply ordered subset B. The hypothesis of Zorn's lemma tells us that B has an upper bound c in A. The element c is then automatically a maximal element of A. For if $c \prec d$ for some element d of A, then the set $B \cup \{d\}$, which properly contains B, is simply ordered because $b \prec d$ for every $b \in B$. This fact contradicts maximality of B.

It is also true that the maximum principle is an easy consequence of Zorn's lemma. See Exercises 5–7.

One final remark. We have defined what we mean by a strict partial order on a set, but we have not said what a partial order itself is. Let $\prec$ be a strict partial order on a set A. Suppose that we define $a \preceq b$ if either $a \prec b$ or $a = b$. Then the relation $\preceq$ is called a ***partial order*** on A. For example, the inclusion relation $\subset$ on a collection of sets is a partial order, whereas proper inclusion is a strict partial order.

Many authors prefer to deal with partial orderings rather than strict partial orderings; the maximum principle and Zorn's lemma are often expressed in these terms. Which formulation is used is simply a matter of taste and convenience.

Exercises

1. If a and b are real numbers, define $a \prec b$ if $b - a$ is positive and rational. Show this is a strict partial order on $\mathbb{R}$. What are the maximal simply ordered subsets?

2. (a) Let $\prec$ be a strict partial order on the set A. Define a relation on A by letting $a \preceq b$ if either $a \prec b$ or $a = b$. Show that this relation has the following properties, which are called the ***partial order axioms***:

(i) $a \preceq a$ for all $a \in A$.

(ii) $a \preceq b$ and $b \preceq a \Longrightarrow a = b$.

(iii) $a \preceq b$ and $b \preceq c \Longrightarrow a \preceq c$.

(b) Let P be a relation on A that satisfies properties (i)–(iii). Define a relation S on A by letting aSb if aPb and $a \neq b$. Show that S is a strict partial order on A.

3. Let A be a set with a strict partial order $\prec$; let $x \in A$. Suppose that we wish to find a maximal simply ordered subset B of A that contains x. One plausible way of attempting to define B is to let B equal the set of all those elements of A that are *comparable* with x;

$$B = \{y \mid y \in A \text{ and either } x \prec y \text{ or } y \prec x\}.$$

But this will not always work. In which of Examples 1 and 2 will this procedure succeed and in which will it not?

4. Given two points (x_0, y_0) and (x_1, y_1) of $\mathbb{R}^2$, define

$$(x_0, y_0) \prec (x_1, y_1)$$