A Course in Modern Analysis and Its Applications

Graeme L. Cohen
University of Technology, Sydney

CAMBRIDGE
UNIVERSITY PRESS

PUBLISHED BY THE PRESS SYNDICATE OF THE UNIVERSITY OF CAMBRIDGE
The Pitt Building, Trumpington Street, Cambridge, United Kingdom

CAMBRIDGE UNIVERSITY PRESS
The Edinburgh Building, Cambridge CB2 2RU, UK
40 West 20th Street, New York, NY 10011-4211, USA
477 Williamstown Road, Port Melbourne, VIC 3207, Australia
Ruiz de Alarcón 13, 28014 Madrid, Spain
Dock House, The Waterfront, Cape Town 8001, South Africa

http://www.cambridge.org

First published 2003

Printed in the United Kingdom at the University Press, Cambridge

Typeface Computer Modern 10/13. *System* LaTeX 2_ε [Typeset by the author]

A catalogue record for this book is available from the British Library

ISBN 0 521 81996 2 hardback
ISBN 0 521 52627 2 paperback

Contents

1

Prelude to Modern Analysis

1.1 Introduction

The primary purpose of this chapter is to review a number of topics from analysis, and some from algebra, that will be called upon in the following chapters. These are topics of a classical nature, such as appear in books on advanced calculus and linear algebra. For our treatment of modern analysis, we can distinguish four fundamental notions which will be particularly stressed in this chapter. These are

(a) set theory, of an elementary nature;
(b) the concept of a function;
(c) convergence of sequences; and
(d) some theory of vector spaces.

On a number of occasions in this chapter, we will also take the time to discuss the relationship of modern analysis to classical analysis. We begin this now, assuming some knowledge of the points (a) to (d) just mentioned.

Modern analysis is not a new brand of mathematics that replaces the old brand. It is totally dependent on the time-honoured concepts of classical analysis, although in parts it can be given without reference to the specifics of classical analysis. For example, whereas classical analysis is largely concerned with functions of a real or complex variable, modern analysis is concerned with functions whose domains and ranges are far more general than just sets of real or complex numbers. In fact, these functions can have domains and ranges which are themselves sets of functions. A function of this more general type will be called an operator or mapping. Importantly, very often any set will do as the domain of a mapping, with no specific reference to the nature of its elements.

This illustrates how modern analysis generalises the ideas of classical analysis. At the same time, in many ways modern analysis simplifies classical analysis because it uses a basic notation which is not cluttered with the symbolism that characterises many topics of a classical nature. Through this, the unifying aspect of modern analysis appears because when the symbolism of those classical topics is removed a surprising similarity becomes apparent in the treatments formerly thought to be peculiar to those topics.

Here is an example:

$$\int_a^b k(s,t)x(t)\,dt = f(s), \quad a \leqslant s \leqslant b,$$

is an *integral equation*; f and k are continuous functions and we want to solve this to find the continuous function x. The left-hand side shows that we have operated on the function x to give the function f, on the right. We can write the whole thing as

$$Kx = f,$$

where K is an operator of the type we just mentioned. Now the essence of the problem is clear. It has the same form as a matrix equation $A\mathbf{x} = \mathbf{b}$, for which the solution (sometimes) is $\mathbf{x} = A^{-1}\mathbf{b}$. In the same way, we would like the solution of the integral equation to be given simply as $x = K^{-1}f$. The two problems, stripped of their classical symbolism, appear to be two aspects of a more general study.

The process can be reversed, showing the strong applicability of modern analysis: when the symbolism of a particular branch of classical analysis is restored to results often obtained only because of the manipulative ease of the simplified notation, there arise results not formerly obtained in the earlier theory. In other cases, this procedure gives rise to results in one field which had not been recognised as essentially the same as well-known results in another field. The notations of the two branches had fully disguised the similarity of the results.

When this occurs, it can only be because there is some underlying structure which makes the two (or more) branches of classical analysis appear just as examples of some work in modern analysis. The basic entities in these branches, when extracted, are apparently combined together in a precisely corresponding manner in the several branches. This takes us back to our first point of the generalising nature of modern analysis and of the benefit of working with quite arbitrary sets. To combine the elements of these sets together requires some basic ground rules and this is why, very often and predominantly in this book, the sets are assumed to be vector spaces: simply because vector spaces are sets with certain rules attached allowing their elements to be combined together in a particular fashion.

We have indicated the relevance of set theory, functions and vector spaces in our work. The other point, of the four given above, is the springboard that takes us from algebra into analysis. In this book, we use in a very direct fashion the notion of a convergent sequence to generate virtually every result.

We might mention now, since we have been comparing classical and modern analysis, that another area of study, called functional analysis, may today be taken as identical with modern analysis. A functional is a mapping whose range is a set of real or complex numbers and functional analysis had a fairly specific meaning (the analysis of functionals) when the term was first introduced early in the 20th century. Other writers may make technical distinctions between the two terms but we will not.

In the review which follows, it is the aim at least to mention all topics required for an understanding of the subsequent chapters. Some topics, notably those connected with the points (a) to (d) above, are discussed in considerable detail, while others might receive little more than a definition and a few relevant properties.

1.2 Sets and numbers

A *set* is a concept so basic to modern mathematics that it is not possible to give it a precise definition without going deeply into the study of mathematical logic. Commonly, a set is described as any collection of objects but no attempt is made to say what a 'collection' is or what an 'object' is. We are forced in books of this type to accept sets as fundamental entities and to rely on an intuitive feeling for what a set is.

The objects that together make up a particular set are called *elements* or *members* of that set. The list of possible sets is as long as the imagination is vivid, or even longer (we are hardly being precise here) since, importantly, the elements of a set may themselves be sets.

Later in this chapter we will be looking with some detail into the properties of certain sets of numbers. We are going to rely on the reader's experience with numbers and not spend a great deal of time on the development of the real number system. In particular, we assume familiarity with

(a) the *integers*, or whole numbers, such as -79, -3, 0, 12, $4{,}063{,}180$;
(b) the *rational numbers*, such as $-\frac{5}{3}$, $\frac{11}{17}$, which are numbers expressible as a ratio of integers (the integers themselves also being examples);
(c) those numbers which are not rational, known as *irrational numbers*, such as $\sqrt{2}$, $\sqrt[3]{15}$, π;
(d) the *real numbers*, which are numbers that are either rational or irrational;
(e) the ordering of the real numbers, using the inequality signs $<$ and $>$ (and the use of the signs $\leqslant$ and $\geqslant$);
(f) the representation of the real numbers as points along a line; and
(g) the fact, in (f), that the real numbers fill the line, leaving no holes: to every point on the line there corresponds a real number.

The final point is a crucial one and may not appear to be so familiar. On reflection however, it will be seen to accord with experience, even when expressed in such a vague way. This is a crude formulation of what is known as the *completeness* of the real number system, and will be referred to again in some detail subsequently.

By way of review, we remark that we assume the ability to perform simple manipulations with inequalities. In particular, the following should be known. If a and b are real numbers and $a < b$, then

$$-a > -b;$$
$$\frac{1}{a} > \frac{1}{b}, \text{ if also } a > 0 \text{ or } b < 0;$$
$$\sqrt{a} < \sqrt{b}, \text{ if also } a \geqslant 0.$$

With regard to the third property, we stress that the use of the radical sign ($\sqrt{\ }$) always implies that the nonnegative root is to be taken. Bearing this comment in mind, we may define the *absolute value* $|a|$ of any real number a by

$$|a| = \sqrt{a^2}.$$

More commonly, and equivalently of course, we say that $|a|$ is a whenever $a > 0$ and $|a|$ is $-a$ whenever $a < 0$, while $|0| = 0$. For any real numbers a and b, we have

$$|a+b| \leqslant |a| + |b|, \qquad |ab| = |a|\,|b|.$$

These may be proved by considering the various combinations of positive and negative values for a and b.

We also assume a knowledge of *complex numbers*: numbers of the form $a + ib$ where a and b are real numbers and i is an imaginary unit, satisfying $i^2 = -1$.

This is a good place to review a number of definitions and properties connected with complex numbers. If $z = a+ib$ is a complex number, then we call the numbers a, b, $a - ib$ and $\sqrt{a^2 + b^2}$ the *real part*, *imaginary part*, *conjugate* and *modulus*, respectively, of z, and denote these by $\operatorname{Re} z$, $\operatorname{Im} z$, $\overline{z}$ and $|z|$, respectively. The following are some of the simple properties of complex numbers that we use. If z, z_1 and z_2 are complex numbers, then

$$\begin{aligned}
\overline{\overline{z}} &= z, \\
\overline{z_1 + z_2} &= \overline{z}_1 + \overline{z}_2, \\
\overline{z_1 z_2} &= \overline{z}_1\, \overline{z}_2, \\
|\operatorname{Re} z| \leqslant |z|,\ &|\operatorname{Im} z| \leqslant |z|, \\
z\overline{z} &= |z|^2, \\
|z_1 + z_2| &\leqslant |z_1| + |z_2|, \\
|z_1 z_2| &= |z_1|\,|z_2|.
\end{aligned}$$

It is essential to remember that, although z is a complex number, the numbers $\operatorname{Re} z$, $\operatorname{Im} z$ and $|z|$ are real. The final two properties in the above list are important generalisations of the corresponding properties just given for real numbers. They can be generalised further, in the natural way, to the sum or product of three or four or more complex numbers.

Real numbers, complex numbers, and other sets of numbers, all occur so frequently in our work that it is worth using special symbols to denote them.

Definition 1.2.1 The following symbols denote the stated sets:

$\mathbf{N}$, the set of all positive integers;
$\mathbf{Z}$, the set of all integers (positive, negative and zero);
$\mathbf{Q}$, the set of all rational numbers;
$\mathbf{R}$, the set of all real numbers;
$\mathbf{R}_+$, the set of all nonnegative real numbers;
$\mathbf{C}$, the set of all complex numbers.

Other sets will generally be denoted by ordinary capital letters and their elements by lower case letters; the same letter will not always refer to the same set or element. To indicate that an object x is an element of a set X, we will write $x \in X$; if x is not an element of X, we will write $x \notin X$. For example, $\sqrt{2} \in \mathbf{R}$ but $\sqrt{2} \notin \mathbf{Z}$. A statement such as $x, y \in X$ will be used as an abbreviation for the two statements $x \in X$ and $y \in X$. To show the elements of a set we always enclose them in braces and give either a complete listing (for example, $\{1, 2, 3\}$ is the set consisting of the integers 1, 2 and 3), or an indication of a pattern (for example, $\{1, 2, 3, \dots\}$ is the set $\mathbf{N}$), or a description of a rule of formation following a colon (for example, $\{x : x \in \mathbf{R},\ x \geqslant 0\}$ is the set $\mathbf{R}_+$). Sometimes we use an abbreviated notation (for example, $\{n : n = 2m,\ m \in \mathbf{N}\}$ and $\{2n : n \in \mathbf{N}\}$ both denote the set of all even positive integers).

An important aspect in the understanding of sets is that the order in which their elements are listed is irrelevant. For example, $\{1, 2, 3\}$, $\{3, 1, 2\}$, $\{2, 1, 3\}$ are different ways of writing the same set. However, on many occasions we need to be able to specify the first position, the second position, and so on, and for this we need a new notion. We speak of *ordered pairs* of two elements, *ordered triples* of three elements, and, generally, *ordered n-tuples* of n elements with this property that each requires for its full determination a list of its elements and the order in which they are to be listed. The elements, in their right order, are enclosed in parentheses (rather than braces, as for sets). For example, $(1, 2, 3)$, $(3, 1, 2)$, $(2, 1, 3)$ are different ordered triples. This is not an unfamiliar notion. In ordinary three-dimensional coordinate geometry, the coordinates of a point provide an example of an ordered triple: the three ordered triples just given would refer to three different points in space.

We give now a number of definitions which help us describe various manipulations to be performed with sets.

Definition 1.2.2

(a) A set S is called a *subset* of a set X, and this is denoted by $S \subseteq X$ or $X \supseteq S$, if every element of S is also an element of X.

(b) Two sets X and Y are called *equal*, and this is denoted by $X = Y$, if each is a subset of the other; that is, if both $X \subseteq Y$ and $Y \subseteq X$. Otherwise, we write $X \neq Y$.

(c) A set which is a subset of any other set is called a *null set* or *empty set*, and is denoted by $\varnothing$.

(d) A set S is called a *proper* subset of a set X if $S \subseteq X$, but $S \neq X$.

(e) The *union* of two sets X and Y, denoted by $X \cup Y$, is the set of elements belonging to at least one of X and Y; that is,

$$X \cup Y = \{x : x \in X \text{ or } x \in Y \text{ (or both)}\}.$$

(f) The *intersection* of two sets X and Y, denoted by $X \cap Y$, is the set of elements belonging to both X and Y; that is,

$$X \cap Y = \{x : x \in X \text{ and } x \in Y\}.$$

(g) The *cartesian product* of two sets X and Y, denoted by $X \times Y$, is the set of all ordered pairs, the first elements of which belong to X and the second elements to Y; that is,

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

(h) The *complement* of a set X, denoted by $\sim X$, is the set of elements that do not belong to X; that is, $\sim X = \{x : x \notin X\}$. The complement of X relative to a set Y is the set $Y \cap \sim X$; this is denoted by $Y \backslash X$.

For some simple examples illustrating parts of this definition, we let $X = \{1, 3, 5\}$ and $Y = \{1, 4\}$. Then

$$\begin{aligned} X \cup Y &= \{1, 3, 4, 5\}, \qquad X \cap Y = \{1\}, \\ X \times Y &= \{(1,1), (1,4), (3,1), (3,4), (5,1), (5,4)\}, \\ Y \times X &= \{(1,1), (1,3), (1,5), (4,1), (4,3), (4,5)\}. \end{aligned}$$

We see that in general $X \times Y \neq Y \times X$. The set $Y \backslash X$ is the set of elements of Y that do not belong to X, so here $Y \backslash X = \{4\}$.

The definitions of union, intersection and cartesian product of sets can be extended to more than two sets. Suppose we have n sets X_1, X_2, $\ldots$, X_n. Their union, intersection and cartesian product are defined as

$$\begin{aligned} X_1 \cup X_2 \cup \cdots \cup X_n &= \bigcup_{k=1}^{n} X_k \\ &= \{x : x \in X_k \text{ for at least one } k = 1,\ 2,\ \ldots,\ n\}, \\ X_1 \cap X_2 \cap \cdots \cap X_n &= \bigcap_{k=1}^{n} X_k \\ &= \{x : x \in X_k \text{ for all } k = 1,\ 2,\ \ldots,\ n\}, \\ X_1 \times X_2 \times \cdots \times X_n &= \prod_{k=1}^{n} X_k \\ &= \{(x_1, x_2, \ldots, x_n) : x_k \in X_k \text{ for } k = 1,\ 2,\ \ldots,\ n\}, \end{aligned}$$

respectively (the cartesian product being a set of ordered n-tuples). The notations in the middle are similar to the familiar sigma notation for addition, where we write

$$x_1 + x_2 + \cdots + x_n = \sum_{k=1}^{n} x_k,$$

when x_1, x_2, ..., x_n are numbers.

For cartesian products only, there is a further simplification of notation when all the sets are equal. If $X_1 = X_2 = \cdots = X_n = X$, then in place of $\prod_{k=1}^{n} X_k$ or $\prod_{k=1}^{n} X$ we write simply X^n, as suggested by the $\times$ notation, but note that there is no suggestion of multiplication: X^n is a set of n-tuples. In particular, it is common to write $\mathbf{R}^n$ for the set of all n-tuples of real numbers and $\mathbf{C}^n$ for the set of all n-tuples of complex numbers.

It is necessary to make some comments regarding the definition of an empty set in Definition 1.2.2(c). These are gathered together as a theorem.

Theorem 1.2.3

(a) *All empty sets are equal.*
(b) *The empty set has no elements.*
(c) *The only set with no elements is the empty set.*

To prove (a), we suppose that $\varnothing_1$ and $\varnothing_2$ are any two empty sets. Since an empty set is a subset of any other set, we must have both $\varnothing_1 \subseteq \varnothing_2$ and $\varnothing_2 \subseteq \varnothing_1$. By the definition of equality of sets, it follows that $\varnothing_1 = \varnothing_2$. This proves (a) and justifies our speaking of 'the' empty set in the remainder of the theorem. We prove (b) by contradiction. Suppose $x \in \varnothing$. Since for any set X we have $\varnothing \subseteq X$ and $\varnothing \subseteq {\sim}X$, we must have both $x \in X$ and $x \in {\sim}X$. This surely contradicts the existence of x, proving (b). Finally, we prove (c), again by contradiction. Suppose X is a set with no elements and suppose $X \neq \varnothing$. Since $\varnothing \subseteq X$, this means that X is not a subset of $\varnothing$. Then there must be an element of X which is not in $\varnothing$. But X has no elements so this is the contradiction we need. □

All this must seem a bit peculiar if it has not been met before. In defence, it may be pointed out that sets were only introduced intuitively in the first place and that the inclusion in the concept of 'a set with no elements' is a necessary addition (possibly beyond intuition) to provide consistency elsewhere. For example, if two sets X and Y have no elements in common and we wish to speak of their intersection, we can now happily say $X \cap Y = \varnothing$. (Two such sets are called *disjoint.*)

Manipulations with sets often make use of the following basic results.

Theorem 1.2.4 *Let X, Y and Z be sets. Then*

(a) $\sim(\sim X) = X$,
(b) $X \cup Y = Y \cup X$ *and* $X \cap Y = Y \cap X$ (*commutative rules*),
(c) $X \cup (Y \cup Z) = (X \cup Y) \cup Z$ *and* $X \cap (Y \cap Z) = (X \cap Y) \cap Z$ (*associative rules*),
(d) $X \cup (Y \cap Z) = (X \cup Y) \cap (X \cup Z)$ *and* $X \cap (Y \cup Z) = (X \cap Y) \cup (X \cap Z)$ (*distributive rules*).

We will prove only the second distributive rule. To show that two sets are equal we must make use of the definition of equality in Definition 1.2.2(b).

First, suppose $x \in X \cap (Y \cup Z)$. Then $x \in X$ and $x \in Y \cup Z$. That is, x is a member of X and of either Y or Z (or both). If $x \in Y$ then $x \in X \cap Y$; if $x \in Z$ then $x \in X \cap Z$. At least one of these must be true, so $x \in (X \cap Y) \cup (X \cap Z)$. This proves that $X \cap (Y \cup Z) \subseteq (X \cap Y) \cup (X \cap Z)$. Next, suppose $x \in (X \cap Y) \cup (X \cap Z)$. Then $x \in X \cap Y$ or $x \in X \cap Z$ (or both). In both cases, $x \in X \cap (Y \cup Z)$ since in both cases $x \in X$, and $Y \subseteq Y \cup Z$ and $Z \subseteq Y \cup Z$. Thus $X \cap (Y \cup Z) \supseteq (X \cap Y) \cup (X \cap Z)$.

Then it follows that $X \cap (Y \cup Z) = (X \cap Y) \cup (X \cap Z)$, completing this part of the proof. □

The following theorem gives two of the more important relationships between sets.

Theorem 1.2.5 (De Morgan's Laws) *Let X, Y and Z be sets. Then*

$$Z \backslash (X \cap Y) = Z \backslash X \cup Z \backslash Y \quad \textit{and} \quad Z \backslash (X \cup Y) = Z \backslash X \cap Z \backslash Y.$$

There is a simpler form of de Morgan's laws for ordinary complements:

$$\sim(X \cap Y) = \sim X \cup \sim Y \quad \text{and} \quad \sim(X \cup Y) = \sim X \cap \sim Y.$$

To prove the first of these, suppose $x \in \sim(X \cap Y)$. Then $x \notin X \cap Y$ so either $x \notin X$ or $x \notin Y$. That is, $x \in \sim X$ or $x \in \sim Y$, so $x \in \sim X \cup \sim Y$. This proves that $\sim(X \cap Y) \subseteq \sim X \cup \sim Y$. Suppose next that $x \in \sim X \cup \sim Y$. If $x \in \sim X$ then $x \notin X$ so $x \notin X \cap Y$, since $X \cap Y \subseteq X$. That is, $x \in \sim(X \cap Y)$. The same is true if $x \in \sim Y$. Thus $\sim X \cup \sim Y \subseteq \sim(X \cap Y)$, so we have proved that $\sim(X \cap Y) = \sim X \cup \sim Y$.

We can use this, the definition of relative complement, and a distributive rule from Theorem 1.2.4 to prove the first result of Theorem 1.2.5:

$$\begin{aligned} Z\backslash(X \cap Y) &= Z \cap {\sim}(X \cap Y) = Z \cap ({\sim}X \cup {\sim}Y) \\ &= (Z \cap {\sim}X) \cup (Z \cap {\sim}Y) = Z\backslash X \cup Z\backslash Y. \end{aligned}$$

The second of de Morgan's laws is proved similarly. □

Review exercises 1.2

(1) Let a and b be real numbers. Show that

(a) $||a| - |b|| \leqslant |a - b|$,
(b) $|a - b| < \epsilon$ if and only if $b - \epsilon < a < b + \epsilon$,
(c) if $a < b + \epsilon$ for every $\epsilon > 0$ then $a \leqslant b$.

(2) Suppose $A \cup B = X$. Show that $X \times Y = (A \times Y) \cup (B \times Y)$, for any set Y.

(3) For any sets A and B, show that

(a) $A\backslash B = A$ if and only if $A \cap B = \varnothing$,
(b) $A\backslash B = \varnothing$ if and only if $A \subseteq B$.

1.3 Functions or mappings

We indicated in Section 1.1 how fundamental the concept of a function is in modern analysis. (It is equally important in classical analysis but may be given a restricted meaning there, as we remark below.) A function is often described as a rule which associates with an element in one set a unique element in another set; we will give a definition which avoids the undefined term 'rule'. In this definition we will include all associated terms and notations that will be required. Examples and general discussion will follow.

Definition 1.3.1 Let X and Y be any two nonempty sets (which may be equal).

(a) A *function* f from X into Y is a subset of $X \times Y$ with the property that for each $x \in X$ there is precisely one element (x, y) in the subset f. We write $f\colon X \to Y$ to indicate that f is a function from X into Y.
(b) The set X is called the *domain* of the function $f\colon X \to Y$.

(c) If $(x, y) \in f$ for some function $f\colon X \to Y$ and some $x \in X$, then we call y the *image* of x under f, and we write $y = f(x)$.

(d) Let S be a subset of X. The set

$$\{y : y \in Y,\ y = f(x) \text{ for some } x \in S\},$$

which is a subset of Y, is called the *image* of the set S under $f\colon X \to Y$, and is denoted by $f(S)$. The subset $f(X)$ of Y is called the *range* of f.

(e) When $f(X) = Y$, we say that the function f is from X *onto* Y (rather than into Y) and we call f an *onto* function.

(f) If, for $x_1, x_2 \in X$, we have $f(x_1) = f(x_2)$ only when $x_1 = x_2$, then we call the function $f\colon X \to Y$ *one-to-one*.

(g) An onto function is also said to be *surjective*, or a *surjection*. A one-to-one function is also said to be *injective*, or an *injection*. A function that is both injective and surjective is called *bijective*, or a *bijection*.

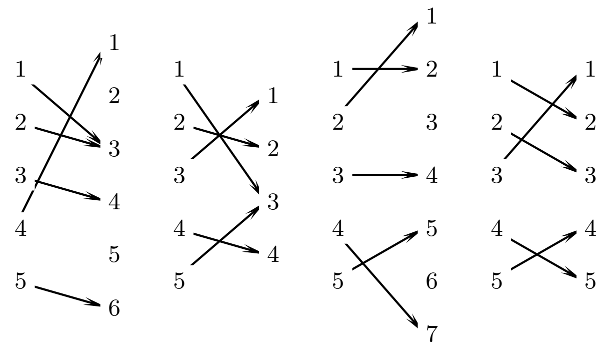
Enlarging on the definition in (a), we see that a function f from a set X into a set Y is itself a set, namely a set of ordered pairs chosen from $X \times Y$ in such a way that distinct elements of f cannot have distinct second elements with the same first element. In (c), we see that the common method of denoting a function as $y = f(x)$ is no more than an alternative, and more convenient, way of writing $(x, y) \in f$. Notice the different roles played by the sets X and Y. The set X is fully used up in that every $x \in X$ has an image $f(x) \in Y$, but the set Y need not be used up in that there may be a $y \in Y$, or many such, which is not the image of any $x \in X$. Paraphrasing (e), when in fact each $y \in Y$ is the image of some $x \in X$, then the function is called ‘onto’. Notice that the same term ‘image’ is used slightly differently in (c) and (d), but this will not cause any confusion.

It follows from Definition 1.2.2(b) that two functions f and g from X into Y are equal if and only if $f(x) = g(x)$ for all $x \in X$.

In Figure 1, four functions

$$f_k\colon X \to Y_k, \qquad k = 1,\ 2,\ 3,\ 4,$$

are illustrated. Each has domain $X = \{1, 2, 3, 4, 5\}$. The function $f_1\colon X \to Y_1$ has $Y_1 = \{1, 2, 3, 4, 5, 6\}$ and the function is the subset $\{(1, 3), (2, 3), (3, 4), (4, 1), (5, 6)\}$ of $X \times Y_1$, as indicated by arrows giving the images of the elements of X. The range of f_1 is the set $f_1(X) = \{1, 3, 4, 6\}$. The other functions may be similarly described.



$f_1 \colon X \to Y_1$ $f_2 \colon X \to Y_2$ $f_3 \colon X \to Y_3$ $f_4 \colon X \to Y_4$

Figure 1

For all four functions, each element of X is the tail of an arrow and of only one arrow, while the elements of the Y's may be at the head of more than one arrow or perhaps not at the head of any arrow. This situation is typical of any function. The elements of Y_2 and Y_4 are all at heads of arrows, so the functions f_2 and f_4 are both onto. Observe that $f_1(1) = 3$ and $f_1(2) = 3$. Also, $f_2(1) = 3$ and $f_2(5) = 3$. This situation does not apply to the functions f_3 and f_4: each element of Y_3 and Y_4 is at the head of at most one arrow, so the functions f_3 and f_4 are both one-to-one.

Only the function f_4 is both one-to-one and onto: it is a bijection. This is a highly desirable situation which we pursue further in Chapters 5 and 7, though we briefly mention the reason now. Only for the function f_4 of the four functions can we simply reverse the directions of the arrows to give another function from a Y into X. We will denote this function temporarily by $g \colon Y_4 \to X$. In full:

$$\begin{aligned} f_4 &= \{(1,2),(2,3),(3,1),(4,5),(5,4)\}, \\ g &= \{(1,3),(2,1),(3,2),(4,5),(5,4)\}. \end{aligned}$$

The function g is also a bijection, and has the characteristic properties

$$\begin{aligned} g(f_4(x)) &= x \quad \text{for each } x \in X, \\ f_4(g(y)) &= y \quad \text{for each } y \in Y_4. \end{aligned}$$

We call g the inverse of the function f_4, and denote it by f_4^{-1}. The precise definition of this term follows.

Definition 1.3.2 For any bijection $f\colon X \to Y$, the *inverse function* $f^{-1}\colon Y \to X$ is the function for which

$$f^{-1}(y) = x \text{ whenever } f(x) = y,$$

where $x \in X$ and $y \in Y$.

It follows readily that if f is a function possessing an inverse function, then f^{-1} also has an inverse function and in fact $(f^{-1})^{-1} = f$.

It is sometimes useful in other contexts to speak of the inverse of a function when it is one-to-one but not necessarily onto. This could be applied to the function $f_3\colon X \to Y_3$, above. We can reverse the arrows there to give a function h, but the domain of h would only be $f_3(X)$ and not the whole of Y_3.

The following definition gives us an important method of combining two functions together to give a third function.

Definition 1.3.3 Let $f\colon X \to Y$ and $g\colon Y \to Z$ be two functions. The *composition* of f with g is the function $g \circ f\colon X \to Z$ given by

$$(g \circ f)(x) = g(f(x)), \quad x \in X.$$

Note carefully that the composition $g \circ f$ is only defined when the range of f is a subset of the domain of g. It should be clear that in general the composition of g with f, that is, the function $f \circ g$, does not exist when $g \circ f$ does, and even if it does exist it need not equal $g \circ f$.

For example, consider the functions f_1 and f_4 above. Since $Y_4 = X$, we may form the composition $f_1 \circ f_4$ (but not $f_4 \circ f_1$). We have

$$(f_1 \circ f_4)(1) = f_1(f_4(1)) = f_1(2) = 3,$$

and so on; in full, $f_1 \circ f_4 = \{(1,3), (2,4), (3,3), (4,6), (5,1)\}$.

There are some other terms which require mention. For a function itself, of the general nature given here, we will prefer the terms *map* and *mapping.* The use of the word 'function' will be restricted to the classical sense in which the domain and range are essentially sets of numbers. These are the traditional real-valued or complex-valued functions of one or more real variables. (We do not make use in this book of functions of a complex variable.) The terms *functional* and *operator* will be used later for special types of mappings.

We will generally reserve the usual letters f, g, etc., for the traditional types of functions, and also later for functionals, and we will use letters such as A and B for mappings.

Review exercises 1.3

(1) Let $f = \{(2,2),(3,1),(4,3)\}$, $g = \{(1,6),(2,8),(3,6)\}$. Does f^{-1} exist? Does g^{-1} exist? If so, write out the function in full. Does $f \circ g$ exist? Does $g \circ f$ exist? If so, write out the function in full.

(2) Define a function $f\colon \mathbf{R} \to \mathbf{R}$ by $f(x) = 5x - 2$, for $x \in \mathbf{R}$. Show that f is one-to-one and onto. Find f^{-1}.

(3) For functions $f\colon X \to Y$ and $g\colon Y \to Z$, show that

   (a) $g \circ f\colon X \to Z$ is one-to-one if f and g are both one-to-one,
   (b) $g \circ f\colon X \to Z$ is onto if f and g are both onto.

1.4 Countability

Our aim is to make a basic distinction between finite and infinite sets and then to show how infinite sets can be distinguished into two types, called countable and uncountable. These are very descriptive names: countable sets are those whose elements can be listed and then counted. This has to be made precise of course, but essentially it means that although in an infinite set the counting process would never end, any particular element of the set would eventually be included in the count. The fact that there are uncountable sets will soon be illustrated by an important example.

Two special terms are useful here. Two sets X and Y are called *equivalent* if there exists a one-to-one mapping from X onto Y. Such a mapping is a bijection, but in this context is usually called a *one-to-one correspondence* between X and Y. Notice that these are two-way terms, treating the two sets interchangeably. This is because a bijection has an inverse, so that if $f\colon X \to Y$ is a one-to-one correspondence between X and Y, then so is $f^{-1}\colon Y \to X$, and either serves to show that X and Y are equivalent. Any set is equivalent to itself: the *identity mapping* $I\colon X \to X$, where $I(x) = x$ for each $x \in X$, gives a one-to-one correspondence between X and itself. It is also not difficult to prove, using the notion of composition of mappings, that if X and Y are equivalent sets and Y and Z are equivalent sets, then also X and Z are equivalent sets. See Review Exercise 1.3(3).

We now define a *finite* set as one that is empty or is equivalent to the set $\{1, 2, 3, \dots, n\}$ for some positive integer n. A set that is not finite is called an *infinite* set. Furthermore:

Definition 1.4.1 *Countable* sets are sets that are finite or that are equivalent to the set $\mathbf{N}$ of positive integers. Sets that are not countable are called *uncountable*.

It follows that the set $\mathbf{N}$ itself is countable.

For the remainder of this section, we will be referring only to infinite sets. It will be easy to see that some of the results apply equally to finite sets.

According to the definition, if X is a countable set then there is a one-to-one correspondence between $\mathbf{N}$ and X, that is, a mapping $f\colon \mathbf{N} \to X$ which is one-to-one and onto. Thus X is the set of images, under f, of elements of $\mathbf{N}$:

$$X = \{f(1), f(2), f(3), \dots\},$$

and no two of these images are equal. This displays the sense in which the elements of X may be counted: each is the image of precisely one positive integer. It is therefore permissible, when speaking of a countable set X, to write $X = \{x_1, x_2, x_3, \dots\}$, implying that any element of X will eventually be included in the list x_1, x_2, x_3, $\dots$.

In proving below that a given set is countable, we will generally be satisfied to indicate how the set may be counted or listed, and will not give an actual mapping which confirms the equivalence of the set with $\mathbf{N}$. For example, the set $\mathbf{Z}$ of all integers is countable, since we may write

$$\mathbf{Z} = \{0, -1, 1, -2, 2, -3, 3, \dots\}$$

and it is clear with this arrangement how the integers may be counted. It now follows that any other set is countable if it can be shown to be equivalent to $\mathbf{Z}$. In fact, any countable set may be used in this way to prove that other sets are countable.

The next theorem gives two important results which will cover most of our applications. The second uses a further extension of the notion of a union of sets, this time to a countable number of sets: if X_1, X_2, $\dots$, are sets, then

$$\bigcup_{k=1}^{\infty} X_k = \{x : x \in X_k \text{ for at least one } k = 1,\ 2,\ 3,\ \dots\}.$$

Theorem 1.4.2 *If X_1, X_2, ... are countable sets, then*

(a) $\prod_{k=1}^{n} X_k$ *is countable for any integer $n \geqslant 2$,*
(b) $\bigcup_{k=1}^{\infty} X_k$ *is countable.*

Our proof of (a) will require mathematical induction. We show first that $X_1 \times X_2$ is countable. Recall that $X_1 \times X_2$ is the set of all ordered pairs (x_1, x_2), where $x_1 \in X_1$ and $x_2 \in X_2$. Since X_1 and X_2 are countable, we may list their elements and write, using a double subscript notation for convenience,

$$X_1 = \{x_{11}, x_{12}, x_{13}, \dots\}, \qquad X_2 = \{x_{21}, x_{22}, x_{23}, \dots\}.$$

(The first subscript is the set number of any element, the second subscript is the element number in that set.) Writing the elements of $X_1 \times X_2$ down in the following array

$$\begin{array}{ccccccccc}
(x_{11}, x_{21}) & & (x_{11}, x_{22}) & \rightarrow & (x_{11}, x_{23}) & & (x_{11}, x_{24}) & \rightarrow & \dots \\
\downarrow & \nearrow & & \swarrow & & \nearrow & & & \\
(x_{12}, x_{21}) & & (x_{12}, x_{22}) & & (x_{12}, x_{23}) & & (x_{12}, x_{24}) & & \dots \\
& \swarrow & & \nearrow & & & & & \\
(x_{13}, x_{21}) & & (x_{13}, x_{22}) & & (x_{13}, x_{23}) & & (x_{13}, x_{24}) & & \dots \\
\downarrow & \nearrow & & & & & & & \\
(x_{14}, x_{21}) & & (x_{14}, x_{22}) & & (x_{14}, x_{23}) & & (x_{14}, x_{24}) & & \dots \\
\vdots & & \vdots & & \vdots & & \vdots & &
\end{array}$$

and then counting them in the order indicated (those whose subscripts total 5, then those whose subscripts total 6, then those whose subscripts total 7, ...) proves that $X_1 \times X_2$ is countable.

Now assume that $X_1 \times X_2 \times \cdots \times X_{n-1}$ is countable for $n > 2$ and let this set be Y. Then $Y \times X_n$ can be shown to be countable exactly as we showed $X_1 \times X_2$ to be countable. Now, $Y \times X_n$ is the set of ordered pairs $\{((x_1, x_2, \dots, x_{n-1}), x_n) : x_k \in X_k,\ k = 1,\ 2,\ \dots,\ n\}$. The mapping $f \colon Y \times X_n \to X_1 \times X_2 \times \cdots \times X_n$ given by

$$f(((x_1, x_2, \dots, x_{n-1}), x_n)) = (x_1, x_2, \dots, x_{n-1}, x_n)$$

is clearly a one-to-one correspondence, and this establishes that $X_1 \times X_2 \times \cdots \times X_n$, or $\prod_{k=1}^{n} X_k$, is countable. The induction is complete, and (a) is proved.

The proof of (b) uses a similar method of counting. As before, we write $X_k = \{x_{k1}, x_{k2}, x_{k3}, \dots\}$, for $k \in \mathbf{N}$. We write down the elements of $\bigcup_{k=1}^{\infty} X_k$ in the array

$$\begin{array}{ccccccccc}
x_{11} & & x_{12} & \to & x_{13} & & x_{14} & \to & \dots \\
\downarrow & \nearrow & & \swarrow & & \nearrow & & & \\
x_{21} & & x_{22} & & x_{23} & & x_{24} & & \dots \\
 & \swarrow & & \nearrow & & & & & \\
x_{31} & & x_{32} & & x_{33} & & x_{34} & & \dots \\
\downarrow & \nearrow & & & & & & & \\
x_{41} & & x_{42} & & x_{43} & & x_{44} & & \dots \\
\vdots & & \vdots & & \vdots & & \vdots & &
\end{array}$$

and count them in the order indicated (those whose subscripts total 2, then 3, then 4, ...), this time taking care that any x's belonging to more than one X_k are counted only once. This proves (b), a result which is often expressed by saying: the union of countably many countable sets is itself a countable set. □

It should be clear that the proof of (b) covers the cases where there are only finitely many sets X_k, and where some of these are finite sets. In particular, it implies that the union of two countable sets is countable.

We now prove two fundamental results.

Theorem 1.4.3

(a) *The set* $\mathbf{Q}$ *of rational numbers is countable.*
(b) *The set* $\mathbf{R}$ *of real numbers is uncountable.*

To prove (a), for each $k \in \mathbf{N}$ let X_k be the set of all rational numbers that can be expressed as p/k where $p \in \mathbf{Z}$. That is,

$$X_k = \left\{ \frac{0}{k}, \frac{-1}{k}, \frac{1}{k}, \frac{-2}{k}, \frac{2}{k}, \dots \right\}.$$

Writing X_k in this way shows that X_k is countable for each k. Any rational number belongs to X_k for some k, so $\bigcup_{k=1}^{\infty} X_k = \mathbf{Q}$. Hence, $\mathbf{Q}$ is countable, by Theorem 1.4.2(b).

We now prove (b), that $\mathbf{R}$ is uncountable, giving our first example of an uncountable set. The proof relies on the statement that every real number has a decimal expansion. (The following observations are relevant to this. Any real number x has a decimal expansion which, for nonnegative numbers, has the form

$$x = m.n_1n_2n_3\dots = m + \frac{n_1}{10} + \frac{n_2}{10^2} + \frac{n_3}{10^3} + \cdots,$$

where m, n_1, n_2, n_3, ... are integers with $0 \leqslant n_k \leqslant 9$ for each k. The number is rational if and only if its decimal expansion either terminates or becomes periodic: for example, $\frac{1}{8} = 0.125000\ldots$ terminates and $\frac{1887}{4950} = 0.38121212\ldots$ is periodic, whereas $\sqrt{2} = 1.4142135\ldots$ is neither terminating nor periodic, being irrational. One problem with decimal expansions is that they are not unique for all real numbers. For example, we also have $\frac{1}{8} = 0.124999\ \ldots$.)

The proof that $\mathbf{R}$ is uncountable is a proof by contradiction. We suppose that $\mathbf{R}$ is countable. Then the elements of $\mathbf{R}$ can be counted, and all will be included in the count. In particular, all real numbers between 0 and 1 will be counted. Let the set $\{x_1, x_2, x_3, \ldots\}$ serve to list all these numbers between 0 and 1 and give these numbers their decimal expansions, say

$$\begin{aligned} x_1 &= 0.n_{11}n_{12}n_{13}\ldots, \\ x_2 &= 0.n_{21}n_{22}n_{23}\ldots, \\ x_3 &= 0.n_{31}n_{32}n_{33}\ldots, \\ &\vdots \end{aligned}$$

the double subscript notation again being convenient. Consider the number

$$y = 0.r_1r_2r_3\ldots,$$

where

$$r_k = \begin{cases} 2, & n_{kk} = 1, \\ 1, & n_{kk} \neq 1, \end{cases}$$

for $k \in \mathbf{N}$. This choice of values (which may be replaced by many other choices) ensures that $r_k \neq n_{kk}$ for any k. Hence, $y \neq x_1$ (since these numbers differ in their first decimal place), $y \neq x_2$ (since these numbers differ in their second decimal place), and so on. That is, $y \neq x_j$ for any j. The choice of 1's and 2's in the decimal expansion of y ensures that there is no ambiguity with 0's and 9's. But y is a number between 0 and 1 and the set $\{x_1, x_2, x_3, \ldots\}$ was supposed to include all such numbers. This is the contradiction which proves that $\mathbf{R}$ is uncountable. □

We will not prove here the very reasonable statement that a subset of a countable set is itself a countable set, possibly finite. This result was used already in the preceding paragraph and may now be used to prove further that the set $\mathbf{C}$ of all complex numbers is uncountable: if this were not true then the subset of $\mathbf{C}$ consisting of all complex numbers with zero imaginary part would be countable, but this subset is $\mathbf{R}$.

On the other hand, the set

$$X = \{z : z = x + iy,\ x, y \in \mathbf{Q}\}$$

of all complex numbers with rational real and imaginary parts is countable. This follows using the two theorems above. For $\mathbf{Q}$ is countable, so $\mathbf{Q} \times \mathbf{Q}$ is countable, and there is a natural one-to-one correspondence between X and $\mathbf{Q} \times \mathbf{Q}$, namely the mapping $f\colon \mathbf{Q} \times \mathbf{Q} \to X$ given by $f((x, y)) = x + iy$, $x, y \in \mathbf{Q}$.

Presumably, uncountable sets are bigger than countable sets, but is $\mathbf{N} \times \mathbf{N}$ bigger than $\mathbf{N}$? To make this notion precise, and thus to be able to compare the sizes of different sets, we introduce cardinality.

Definition 1.4.4 Any set X has an associated symbol called its *cardinal number*, denoted by $|X|$. If X and Y are sets then we write $|X| = |Y|$ if X is equivalent to Y; we write $|X| \leqslant |Y|$ if X is equivalent to a subset of Y; and we write $|X| < |Y|$ if $|X| \leqslant |Y|$ but X is not equivalent to Y. We specify that the cardinal number of a finite set is the number of its elements (so, in particular, $|\varnothing| = 0$), and we write $|\mathbf{N}| = \aleph_0$ and $|\mathbf{R}| = \mathbf{c}$.

There is a lot in this definition. First, it defines how to use the symbols $=$, $<$ and $\leqslant$ in connection with this object called the cardinal number of a set. For finite sets, these turn out to be our usual uses of these symbols. Then, for two specific infinite sets, special symbols are given as their cardinal numbers.

Any infinite countable set is equivalent to $\mathbf{N}$, by definition, so any infinite countable set has cardinal number $\aleph_0$ (pronounced 'aleph null'). So, for example, $|\mathbf{N} \times \mathbf{N}| = |\mathbf{Q}| = \aleph_0$. It is not difficult to see that $n < \aleph_0$ for any $n \in \mathbf{N}$ and that $\aleph_0 < \mathbf{c}$. This is the sense in which uncountable sets are bigger than countable sets.

The arithmetic of cardinal numbers is quite unlike ordinary arithmetic. We will not pursue the details here but will, for interest, list some of the main results. We define addition and multiplication of cardinal numbers by: $|X| + |Y| = |X \cup Y|$ and $|X| \cdot |Y| = |X \times Y|$, where X and Y are any sets, and we define $|Y|^{|X|}$ to be the cardinal number of the *power set* Y^X, which is the set of all functions from X into Y. Then:

$$1 + \aleph_0 = \aleph_0, \quad \aleph_0 + \aleph_0 = \aleph_0, \quad \aleph_0 \cdot \aleph_0 = \aleph_0,$$
$$\mathbf{c} + \mathbf{c} = \mathbf{c}, \quad \mathbf{c} \cdot \mathbf{c} = \mathbf{c}, \quad 2^{\aleph_0} = \mathbf{c}.$$

The famous *continuum hypothesis* is that there is no cardinal number α satisfying $\aleph_0 < \alpha < \mathbf{c}$. All efforts to prove this, or to disprove it by finding a set with cardinal number strictly between those of $\mathbf{N}$ and $\mathbf{R}$, had been unsuccessful. In 1963, it was shown that the existence of such a set could neither be proved nor disproved within the usual axioms of set theory. (Those 'usual' axioms have not been discussed here).

Review exercises 1.4

(1) Define a function $f\colon \mathbf{Z} \to \mathbf{N}$ by

$$f(n) = \begin{cases} 2n+1, & n \geqslant 0, \\ -2n, & n < 0. \end{cases}$$

Show that f determines a one-to-one correspondence between $\mathbf{Z}$ and $\mathbf{N}$.

(2) Suppose X is an uncountable set and Y is a countable set. Show that $X\backslash Y$ is uncountable.

(3) Show that the set of all polynomial functions with rational coefficients is countable.

1.5 Point sets

In this section, we are concerned only with sets of real numbers. Because real numbers can conveniently be considered as points on a line, such sets are known as *point sets* and their elements as *points*.

The simplest point sets are *intervals*, for which we have special notations. Let a and b be real numbers, with $a < b$. The point set $\{x : a \leqslant x \leqslant b\}$ is a *closed* interval, denoted by $[a, b]$, and the point set $\{x : a < x < b\}$ is an *open* interval, denoted by (a, b). There are also the *half-open* intervals $\{x : a \leqslant x < b\}$ and $\{x : a < x \leqslant b\}$, denoted by $[a, b)$ and $(a, b]$, respectively. In all cases, the numbers a and b are called *endpoints* of the intervals. Closed intervals contain their endpoints as members, but open intervals do not. The following point sets are *infinite* intervals: $\{x : a < x\}$, denoted by (a, ∞); $\{x : a \leqslant x\}$, denoted by $[a, \infty)$; $\{x : x < b\}$, denoted by $(-\infty, b)$; and $\{x : x \leqslant b\}$, denoted by $(-\infty, b]$. These have only one endpoint, which may or may not be a member of the set. The use of the signs ∞ and $-\infty$ is purely conventional and does not imply that these things are numbers. The set $\mathbf{R}$ itself is sometimes referred to as the infinite interval $(-\infty, \infty)$.

A special name is given to an open interval of the form $(a - \delta, a + \delta)$, where δ is a positive number. This is called the *δ-neighbourhood* of a.