Springer Undergraduate Mathematics Series

Springer-Verlag London Ltd.

Advisory Board

Other books in this series

Geoff Smith

Introductory Mathematics: Algebra and Analysis

With 19 Figures

Springer

Geoff Smith, MA, MSc, PhD
Department of Mathematical Sciences, University of Bath, Claverton Down,
Bath BA2 7AY, UK

Cover illustration elements reproduced by kind permission of:
Aptech Systems, Inc., Publishers of the GAUSS Mathematical and Statistical System, 23804 S.E. Kent-Kangley Road, Maple Valley, WA 98038, USA. Tel: (206) 432 - 7855 Fax (206) 432 -7832 email: info@aptech.com URL: www.aptech.com
American Statistical Association: Chance Vol 8 No 1,1995 article by KS and KW Heiner 'Tree Rings of the Northern Shawangunks' page 32 fig 2
Springer-Verlag: Mathematica in Education and Research Vol 4 Issue 3 1995 article by Roman E Maeder, Beatrice Amrhein and Oliver Gloor 'Illustrated Mathematica: Visualization of Mathematical Objects' page 9 fig 11, originally published as a CD ROM 'Illustrated Mathematics' by TELOS: ISBN 0-387-14222-3, German edition by Birkhauser: ISBN 3-7643-5100-4.
Mathematica in Education and Research Vol 4 Issue 3 1995 article by Richard J Gaylord and Kazume Nishidate 'Traffic Engineering with Cellular Automata' page 35 fig 2. Mathematica in Education and Research Vol 5 Issue 2 1996 article by Michael Trott 'The Implicitization of a Trefoil Knot' page 14.
Mathematica in Education and Research Vol 5 Issue 2 1996 article by Lee de Cola 'Coins, Trees, Bars and Bells: Simulation of the Binomial Process' page 19 fig 3. Mathematica in Education and Research Vol 5 Issue 2 1996 article by Richard Gaylord and Kazume Nishidate 'Contagious Spreading' page 33 fig 1. Mathematica in Education and Research Vol 5 Issue 2 1996 article by Joe Buhler and Stan Wagon 'Secrets of the Madelung Constant' page 50 fig 1.

British Library Cataloguing in Publication Data
Smith, Geoffrey Charles
Introductory mathematics: algebra and analysis. - (Springer undergraduate mathematics series)
1. Mathematics 2. Algebra 3. Mathematical analysis
I. Title
510
ISBN 978-3-540-76178-5 ISBN 978-1-4471-0619-7 (eBook)
DOI 10.1007/978-1-4471-0619-7
Library of Congress Cataloging-in-Publication Data
Smith, Geoff, 1953-
Introductory mathematics: algebra and analysis / Geoff Smith
p. cm. -- (Springer undergraduate mathematics series)
Includes index.

1. Algebra 2. Mathematical analysis I. Title. II. Series.
QA154.2.S546 1998 97-41196
510-dc21 CIP

Springer Undergraduate Mathematics Series ISSN 1615-2085

Typesetting: Camera ready by author

12/3830-54 Printed on acid-free paper SPIN 11012368

To my current wife

Preface

This book is a gentle and relaxed introduction to the two branches of pure mathematics which dominate the early stages of the subject as it is taught to undergraduates in many countries. It is not a substitute for more advanced texts, and has no pretensions to comprehensiveness. There are several places where I would have liked to press on further, but you have to stop somewhere. It should, I hope, be easy to read, and to this end the style is decidedly more colloquial than is traditional in text books. I only hope that the language will not date too quickly. Thus this is not a book in the relentless theorem–proof style; it contains discursive commentary.

The ways in which pure mathematicians express themselves, and the cumulative nature of the subject, may make pure mathematics seem daunting to a beginner. The mathematical mode of expression and the deductive method are vital to pure mathematics. We wish to explore strange geometries, new algebraic systems, and infinite dimensional spaces. There is no point in embarking on this enterprise unless we are prepared to be ruthlessly precise, since otherwise, no-one will have any idea what we are talking about (if anything).

These exotic spaces and objects are not part of everyday experience, unlike, for example a dog. If we mention that "there is a dog in the garden", we do not expect the response "what is a *dog*, what is a *garden*, what does *is* mean in this sentence, why have you used the indefinite article, and what is the contribution of the word *there*?" We know a lot about dogs and gardens, and do not need to put the sentence under scrutiny in order to understand the meaning. However, if instead someone says "every linear group is either virtually solvable, or contains a free subgroup of rank 2", then either you have to live in a world where these terms are as familiar as dogs and gardens, or you have to take the remark apart, and analyze every part of it until you understand what it asserts. Of course there is little point in doing this unless you happen to know that linear groups are very interesting – which, incidentally, they are.

There is a web site which supports this book.

`http://www.maths.bath.ac.uk/~masgcs/book1/`

If that ever changes, a link to the new site will be put in place. At the web site you will find additional exercises and solutions, and corrections to any errors that are discovered after publication.

The material in this book is not always arranged in a logically perfect sequence. This is deliberate, and is a consequence of trying to make the book accessible. The ideal way to read the book is from cover to cover. Chapter 1 establishes notation and sets the scene, and Chapter 2 concerns mathematical proof – many readers will want to read that before proceeding with the main topics. I have tried to make subsequent chapters (fairly) independent, though Chapter 6 should definitely be read before either Chapter 7 or Chapter 8. In consequence of the partial independence, some material is repeated in different parts of the book, though the treatments vary.

I also felt that this book should contain early chapters on complex numbers and matrices. These topics are basic to university mathematics, engineering and science, and are rarely or barely taught at secondary level.

It is standard practice to thank everyone who had anything to do with the creation of a book. I therefore wish to thank and congratulate my parents Eileen and Roy. This volume is being published in the year of their golden wedding anniversary, and Springer-Verlag have kindly agreed to celebrate this event by supplying a cover of appropriate colour.

Inductively, if I thank anyone, I thank their parents too, thereby acknowledging the assistance of a vast number of people, creatures, single-celled organisms and amino acids. Please note that this (thinly spread) gratitude was expressed with considerable economy (see Section 2.1).

Despite the millions of generations who have already been thanked, there are some deserving cases who have not yet qualified. I also acknowledge the help and support of various colleagues at the University of Bath. My TeX and LaTeX guru is Fran Burstall, and Teck Sin How provided figures at amazing speed. James Davenport and Aaron Wilson helped to weed out errors, and suggested improvements. I must also thank John Toland who persuaded me to write this book with his usual combination of charm, flattery and threats – and supplied the beautiful question that constitutes Exercise 8.4. Any remaining errors are mine, and copyright.

I would also like to thank my bright, enthusiastic and frustrated students, without whom this book would not have been necessary, and my wife Olga Markovna Tabachnikova, without whom my salary would have been sufficient.

GCS, Bath, 11-xi-1997.

Added at second printing I thank the following people who have reported typographical errors in the first printing of this book: Verity Jeffery (Meridian School), Prof Charles F. Miller III (Melbourne University), Martyn Partridge (Intertype), Carrie Rutherford (Q.M.W., London) and Aaron Wilson (University of Bath). These errors have been eliminated. I also wish to thank Prof Edward Fraenkel FRS (University of Bath) for his tactful attempts to improve my grammar.

In addition to solutions of problems and amplifications on material in the book, the web site now contains supplementary material on many topics, some of which were suggested by Gregory Sankaran and Wafaa Shabana. This material includes Cardinality and Countability, Functions, Preimages, Unions and Intersections, the Inclusion-Exclusion Principle, Injections and Surjections, Fermat's Two Squares Theorem, Group Actions (and exercises) and the Integers modulo N.

Added at third printing A few more errors have been dealt with thanks to Dr Victoria Gould (University of York) and Prof Dave Johnson (University of the West Indies). The proof of Proposition 6.9 replaces the garbled mush which disgraced the first two printings.

In this latest iteration we give Dedekind and von Neumann proper credit for creating the natural numbers. However, the provenance of the whole numbers is slightly controversial since according to Kronecker the integers were made by God.

Contents

Preface vii

1. **Sets, Functions and Relations** 1
   1.1 Sets 1
   1.2 Subsets 2
   1.3 Well-known Sets 3
   1.4 Rationals, Reals and Pictures 6
   1.5 Set Operations 8
   1.6 Sets of Sets 11
   1.7 Paradox 14
   1.8 Set-theoretic Constructions 15
   1.9 Notation 16
   1.10 Venn Diagrams 17
   1.11 Quantifiers and Negation 19
   1.12 Informal Description of Maps 21
   1.13 Injective, Surjective and Bijective Maps 22
   1.14 Composition of Maps 23
   1.15 Graphs and Respectability Reclaimed 29
   1.16 Characterizing Bijections 30
   1.17 Sets of Maps 31
   1.18 Relations 31
   1.19 Intervals 37

2. **Proof** 39
   2.1 Induction 39
   2.2 Complete Induction 43

2.3 Counter-examples and Contradictions 47
2.4 Method of Descent 50
2.5 Style 53
2.6 Implication 54
2.7 Double Implication 54
2.8 The Master Plan 56

3. Complex Numbers and Related Functions 57
3.1 Motivation 57
3.2 Creating the Complex Numbers 62
3.3 A Geometric Interpretation 70
3.4 Sine, Cosine and Polar Form 76
3.5 *e* 80
3.6 Hyperbolic Sine and Hyperbolic Cosine 85
3.7 Integration Tricks 88
3.8 Extracting Roots and Raising to Powers 89
3.9 Logarithm 90
3.10 Power Series 92

4. Vectors and Matrices 95
4.1 Row Vectors 95
4.2 Higher Dimensions 97
4.3 Vector Laws 98
4.4 Lengths and Angles 98
4.5 Position Vectors 103
4.6 Matrix Operations 104
4.7 Laws of Matrix Algebra 106
4.8 Identity Matrices and Inverses 108
4.9 Determinants 110
4.10 Geometry of Determinants 119
4.11 Linear Independence 120
4.12 Vector Spaces 121
4.13 Transposition 123

5. Group Theory 125
5.1 Permutations 125
5.2 Inverse Permutations 130
5.3 The Algebra of Permutations 131
5.4 The Order of a Permutation 133
5.5 Permutation Groups 135
5.6 Abstract Groups 136
5.7 Subgroups 142

5.8 Cosets 143
5.9 Cyclic Groups 146
5.10 Isomorphism 148
5.11 Homomorphism 151

6. Sequences and Series 153
6.1 Denary and Decimal Sequences 153
6.2 The Real Numbers 156
6.3 Notation for Sequences 158
6.4 Limits of Sequences 160
6.5 The Completeness Axiom 169
6.6 Limits of Sequences Revisited 170
6.7 Series 174

7. Mathematical Analysis 181
7.1 Continuity 181
7.2 Limits 194

8. Creating the Real Numbers 197
8.1 Dedekind's Construction 197
8.2 Construction via Cauchy Sequences 199
8.3 A Sting in the Tail: p-adic numbers 202

Further Reading 204

Solutions 205

Index 213

List of Figures

1.1 Venn diagram for three sets 17

1.2 $A = B = \mathbb{R}$ gives the familiar picture 29

3.1 Nose-down parabolic graph of a quadratic polynomial with two roots 60

3.2 The Argand diagram 64

3.3 Angle measure defined by arc length 74

3.4 Geometry of complex multiplication. Note that $|\mathbf{u}||\mathbf{v}| = |\mathbf{uv}|$ 78

3.5 $\log a = \int_1^a 1/x\,dx$ 81

4.1 Geometric vector addition: $\mathbf{u} + \mathbf{v} = \mathbf{w}$ 96

4.2 The angle between the vectors is θ 100

4.3 A dot product in the Argand diagram 101

4.4 Line through points with position vectors $\mathbf{a}$ and $\mathbf{b}$ 103

4.5 Line (or plane) perpendicular to $\mathbf{c}$ 104

5.1 x then y 126

5.2 y then x 127

5.3 Six objects being permuted 127

6.1 Picture of a limit 161

7.1 A snapshot of δ doing the job for a particular ε 183

7.2 A picture of the Intermediate Value Theorem 194

7.3 Genesis of differentiation 196

The Greek Alphabet

Lower Case	Upper Case	Name	Roman Equivalent
α	A	alpha	a
β	B	beta	b
γ	$\varGamma$	gamma	g (hard)
δ	$\varDelta$	delta	d
ϵ or ε	E	epsilon	e
ζ	Z	zeta	z
η	H	eta	e
θ	$\varTheta$	theta	th
ι	I	iota	i
κ	K	kappa	k
λ	$\varLambda$	lambda	l
μ	M	mu	m
ν	N	nu	n
ξ	$\varXi$	xi	xy (xigh in Britain, xee in the USA)
o	O	omicron	o (short)
π or ϖ	$\varPi$	pi	p
ρ	P	rho	r
σ or ς	$\varSigma$	sigma	s
τ	T	tau	t
υ	$\varUpsilon$	upsilon	u
ϕ or φ	$\varPhi$	phi	f
χ	X	chi	ch (as in a Scottish loch)
ψ	$\varPsi$	psi	ps
ω	$\varOmega$	omega	o (long)

Pronunciation, particularly syllable stress, varies widely.

I κομμϵνδ πρακτισινγ κρυδϵ τρανσλιτϵρατιον ιντο Γρϵϵκ το λϵαρν θϵ "αβ" φαιρλι κυικλϵϵ. Χϵατ ιφ νϵσϵσαρϵϵ.

Γϵοφφ Σμιθ

1
Sets, Functions and Relations

1.1 Sets

We use the language of sets when doing mathematics. Roughly speaking, a *set* is a collection of objects. Suppose that the objects we wish to think about are a, b and c. The collection of all these objects is a set and we write it as

$$\{a, b, c\}.$$

We may care to give this set a name, say S. We write

$$S = \{a, b, c\}.$$

The objects which comprise a set are called its *elements* or *members*. A special symbol is used to describe the fact that an object is an element of a set – it is a straightened version of the small Greek letter epsilon – and it is written $\in$. Thus we may write $a \in S$. In speech, we vocalize this statement as "a is a member of S", or "a is an element of S", or when pushed for time, even "a in S" or "S contains a". Elements of a set are often numbers, but they could be points in space, or lines, or even other sets. By not being specific about what the elements are, we win in various ways. Our theory will be applicable in many contexts and, by stripping the ideas to the bare essentials, we can see exactly what is going on. The process of throwing away unnecessary clutter is called *abstraction* – and it underlies the whole of mathematics. Children learn the abstract idea of number quite early. There is no need to develop one theory of arithmetic for counting apples and another for oranges – you just construct one

G. Smith, *Introductory Mathematics: Algebra and Analysis*

general theory of counting and apply it where appropriate. Fractions, negative numbers and 0 can then be added to the theory without too much difficulty.

1.2 Subsets

Definition 1.1

Suppose that A and B are sets. We say A is a *subset* of B if whenever $x \in A$, then $x \in B$.

We use a rounded version of the inequality symbol to describe this situation:

$$A \subseteq B.$$

Notice that if

$$A \subseteq B, \text{ and } B \subseteq C,$$

then $A \subseteq C$. We capture this property by saying that $\subseteq$ is *transitive*. We will develop this notion further in Section 1.18. Notice also that for any set A we have $A \subseteq A$.

We often want to be able to state that two sets are the same, and we accomplish this by inserting an equality symbol between them. We need a rule to decide when two sets are the same. It turns out that it is convenient to use subsets to formulate this rule.

Definition 1.2

Suppose that both $A \subseteq B$ and $B \subseteq A$, then we say A and B are *equal* and write $A = B$.

This definition is carefully worded, and it has the consequence that

$$\{3,2,1\} = \{1,2,3\} = \{1,2,3,3\} = \{1,1,1,1,2,2,3,3,3,3,3,3,3\}.$$

Thus we are deeming two sets to be equal even though they do not 'look' the same. This should not be too disturbing; we do not worry about writing $1+1=2$, even though there are three symbols to the left of the equals sign, but just one to the right. The equals sign is used with wild abandon during school days; numbers can be equal, as can line segments. Polynomials can be equal: $(x-1)(x+1) = x^2 - 1$. In computing one even sees $n = n+1$, to mean that the value of a variable n is incremented by 1. We have been cautious – and given a precise rule to decide when we are allowed to write an equals sign between a pair of sets.

1.3 Well-known Sets

Some sets are used so frequently that mathematicians have given them special names. The most basic one is the set of *natural numbers* which is the set which consists of all the positive whole numbers. This set is denoted by $\mathbb{N}$. To be explicit, we have

$$\mathbb{N} = \{1, 2, 3, 4, 5, 6, \ldots\}. \tag{1.1}$$

Some mathematicians like to include 0 in the set $\mathbb{N}$. When going to a new course, or reading a new textbook, make sure you know which convention is being used. For the purposes of this book, 0 is not a natural number. No-one has been burnt at the stake over this issue for several years.

The set $\mathbb{N}$ is our first example of an infinite set. There are various technical definitions of "infinite", but for our purposes the following explanation will suffice. A set is *finite* if you can count, and finish counting, its elements. Thus $\{1, 2, 3\}$ is a finite set. Any attempt to count the natural numbers, no matter in what order you choose to do it, will go on for ever. A set which is not finite is *infinite*. The number of (distinct) elements of a set S is called its *cardinality*, or, more casually, its *size*. This number is written $|S|$. Thus $|\{1, 2, 3\}| = 3$ and $|\mathbb{N}|$ is infinite. Notice that $|\{1, 2, 2, 2, 6\}| = 3$ – the extra 2's do not affect the cardinality. Another example of an infinite set is the set of *integers*, economically written as $\mathbb{Z}$. This set consists of all the whole numbers.

$$\mathbb{Z} = \{\ldots, -2, -1, 0, 1, 2, \ldots\}. \tag{1.2}$$

The letter $\mathbb{Z}$ may seem an odd choice; it comes from the German word for the integers, *Zahlen*. The integers come in an obvious order – the order which we happen to have used here when describing $\mathbb{Z}$. This order has the curious property that there is no first element and no last element. This ordering of $\mathbb{Z}$ is not captured in any way by set notation. We could write

$$\mathbb{Z} = \{\ldots, 2, 1, 0, -1, -2, \ldots\}, \tag{1.3}$$

or even

$$\mathbb{Z} = \{\ldots, 4325, -214256, 3, 8765328476872, \ldots\}, \tag{1.4}$$

though the last description would be less than clever. It is impossible to know how to interpret the dots in Equation (1.4). Indeed, even in Equations (1.1), (1.2) and (1.3), we are relying on our prior knowledge of what the natural numbers and integers are like in order to continue the pattern in our minds. This is clearly an unsatisfactory state of affairs. Mathematics requires precision. It is possible to formulate a definition of the natural numbers which does not include these rather dubious dot-dot-dots. We will explain more in Section 1.6. Once you have a decent definition of $\mathbb{N}$, it is then a simple matter to build $\mathbb{Z}$.

This is only a technical point and we shall not pursue it. We allow ourselves to define subsets of a set by using "properties". We may write

$$\mathbb{N} = \{x \mid x \in \mathbb{Z}, x > 0\} \text{ or } \{x : x \in \mathbb{Z}, x > 0\}. \tag{1.5}$$

You read the vertical line or colon as "such that" and the comma as "and"; thus Equation (1.5) asserts (quite correctly) that the set $\mathbb{N}$ has as its elements exactly those integers which are strictly positive. Similarly one might write

$$\{1,2,3\} = \{x \mid x \in \mathbb{Z}, 0 < x < 4\}$$

or

$$\{1,2,3\} = \{x \mid x \in \mathbb{N}, x^2 < 10\}.$$

We settle on | rather than : as our symbol for "such that".

Remark 1.1

Note that we have used a property to slim down the membership of pre-existing sets – we started with $\mathbb{N}$ and $\mathbb{Z}$, and then sorted out which of the elements passed the property test. We did **not** simply define a set by saying "the collection of all things that have this property". As we will see, that is a dangerous game.

Another celebrated set is the set of *rational numbers*, written $\mathbb{Q}$. This is easy to remember because you can think of "quotients". The rational numbers consist of those numbers which can be written as ratios of integers:

$$\mathbb{Q} = \{a/b \mid a, b \in \mathbb{Z}, b \neq 0\}.$$

Notice that $\mathbb{Z} \subseteq \mathbb{Q}$, because if $x \in \mathbb{Z}$, then $x = x/1$. If you add, subtract or multiply $q_1, q_2 \in \mathbb{Q}$, then you get more rational numbers. To be explicit, we have $q_1 + q_2, q_1 - q_2$ and $q_1 q_2 \in \mathbb{Q}$. We say that $\mathbb{Q}$ is *closed* under addition, subtraction and multiplication. The rationals are also closed under division where it is legal. Thus if $q_1, q_2 \in \mathbb{Q}$ and $q_2 \neq 0$, then $q_1/q_2 \in \mathbb{Q}$. The rationals are an example of an algebraic structure called a *field* which we will discuss in Section 3.2.

Not all of the numbers we use can be expressed as ratios of integers. Pointing out – on the fly – that we negate symbols by putting a slash through them, we express the truth that π is not rational by writing: $\pi \notin \mathbb{Q}$. When you read in a textbook that "you may assume that $\pi = 22/7$", you are being told to assume that there are fairies at the bottom of the garden. It is true that $22/7$ is a very good approximation to π, and in fact $355/113$ is an even better one. Indeed, if you specify a positive margin of error, no matter how small, say ε, it is possible to find a rational number which differs from π by less than ε. Thus π can be approximated by rational numbers as well as you like. However, π itself is not rational – something which is not that easy to prove. We will explore this idea of arbitrarily good approximations further in Chapter 6. The impatient and brave might look at Definition 6.3.

By the way, the lower case Greek ε is traditionally used to denote a small positive quantity, or more precisely, a positive quantity which becomes interesting when it happens to be small. Of course, "traditionally" means little more than "habitually" in this context, and you are not breaking any law if you use ε to mean a negative quantity, or a large positive number. However, you are likely to test your friendships if you do that.

It was Pythagoras, or at least his school, who discovered the existence of non-rational numbers – usually called *irrational numbers*. They did this by showing that $\sqrt{2}$ was not rational. This result is not hard, and we will give a proof in Proposition 2.6. The Pythagorean cult was a mystic secret society, and it had most peculiar rules of behaviour. There is a highly amusing chapter on these eccentric scholars in Russell's *History of Western Philosophy*. The existence of irrational numbers was a dark Pythagorean secret.

When you are doing practical calculations, there is often little harm in approximating π, or indeed other numbers. When doing some types of mathematics, however, this is simply not good enough. There are other familiar numbers which are not rational; you will, no doubt, be acquainted with e, the base of natural logarithms, a number which is approximately 2.7183. This number is also not rational.

Those numbers which can be approximated arbitrarily well by rational numbers are called *real numbers*. This is not a very satisfactory definition, but it will have to do at this stage. Rational numbers can be approximated fantastically well by rational numbers (no margin of error is necessary at all!). We write $\mathbb{R}$ for the set of real numbers and observe that $\mathbb{Q} \subseteq \mathbb{R}$. Notice that π, e and $\sqrt{2}$ are all real numbers which are not rational. The real numbers, like the rational numbers, form a field. Any number which has a finite decimal expansion is automatically rational. For example,

$$311.324536547 = 311324536547/1000000000.$$

That argument generalizes easily of course. There are also numbers with infinite decimal expansions which are rational – for example, $1/3$. Computers, being only finite in size, cannot store an irrational number by listing its decimal expansion. Either the number can be approximated by a rational number, or it can be handled symbolically – much in the way that we write $\sqrt{2}$ without worrying about its decimal expansion. All we usually need to know about $\sqrt{2}$ is that $(\sqrt{2})^2 = 2$. The representation of irrational numbers in this way is used in computer algebra. This symbolic technique is a relatively recent development and, at present, most scientific calculations are performed using rational approximations. It has therefore been very important to analyze the size of the errors being introduced, and to devise schemes of calculation which minimize the eventual error that is delivered. The subject that does this is called numerical analysis.

1.4 Rationals, Reals and Pictures

We leave the world of mathematics for a moment. Let us try to construct a mental picture of the real numbers. Imagine an infinite straight line in the plane, and calibrate it by marking in the integers – in the usual order – spaced at equal distances along the line in both directions. Now imagine refining this calibration, marking in all the rational numbers. Infinitely many points on the line will now be mentally marked, indeed, that is true even if you look at any segment of the line of (non-zero) finite length. Nonetheless, π, e, $\sqrt{2}$, and infinitely many other numbers will not be marked. One can think of the rational numbers as fine mist sprayed on the line.

There are infinitely many irrational numbers, and, in a sense, there are even more irrational numbers than rational numbers. It may seem peculiar to say that one infinite set has more elements than another, but a perfectly satisfactory way of doing this was worked out by a mathematician called Georg Cantor. The irrational numbers can also be thought of as a mist sprayed on the line, but to take account of the fact that there are, in some sense, more of them, you may care to think of the mist as being thicker in this case. Note that if $a, b \in \mathbb{R}$ and $a < b$, then we can find $c \in \mathbb{Q}$ and $d \in \mathbb{R}$ but $d \notin \mathbb{Q}$ with $a < c, d < b$. This notation is supposed to mean that both c and d are between a and b. In particular, between every two distinct rational numbers is an irrational number, and between every two distinct irrational numbers is a rational number. Thus the rationals and irrationals are completely jumbled. We will prove this now, on the assumption that $\sqrt{2}$ is irrational (see Proposition 2.6). In fact we don't really need to use $\sqrt{2}$ – any positive irrational number will do equally well. However, it is particularly easy to prove that $\sqrt{2}$ is irrational.

Proposition 1.1

Suppose that a, b are real numbers, and that $a < b$. It follows that there is a rational number x in the range $a < x < b$ and an irrational number y in the range $a < y < b$.

Proof

Choose $N \in \mathbb{N}$ sufficiently large that $1/N < b - a$. Choose the least integer z such that $bN \leq z$ and so $b \leq z/N$. Thus $z - 1 < bN$, so $(z-1)/N < b$. However

$$a = b - (b - a) < b - 1/N \leq z/N - 1/N = (z-1)/N.$$

We conclude that $a < (z-1)/N < b$. However, $x = (z-1)/N$ is rational so we are half way home.

We now seek an irrational number y in the range $a < y < b$. It suffices to find such a y in the range $x < y < b$. Choose $M \in \mathbb{N}$ sufficiently large that $\sqrt{2} < M(b - x)$, so $\sqrt{2}/M < b - x$, and let $y = x + \sqrt{2}/M$. The choice of M ensures that $x < x + \sqrt{2}/M = y < x + b - x = b$, and y is not rational, else $\sqrt{2} = M(y - x)$ would be rational.

□

(The symbol □ is used to denote the end of a proof.)

Perhaps you have studied complex numbers. If so, this paragraph ought to make sense. If not, skip it. We will be studying complex numbers properly in Chapter 3. Observe that $i = \sqrt{-1}$ is not a real number. We define the set of *complex numbers* as

$$\mathbb{C} = \{a + bi \mid a, b \in \mathbb{R}\}.$$

The appropriate mental image of the set $\mathbb{C}$ is known as the *Argand diagram* and consists of identifying the points in an infinite plane with the complex numbers.

Finally, there is one other set which is so important that it merits a special name. This is the set with no elements at all. It is written $\emptyset$ and is called the *empty set* or *null set*. Thus $\emptyset = \{\ \}$. Notice that if A is any set at all, then $\emptyset \subseteq A$. In particular, we have $\emptyset \subseteq \mathbb{R}$. Can you see it sitting there in the real line?

The empty set is remarkable in many ways, not the least of which is the opportunity it gives for very precise reasoning. We have used the term "the empty set", when, from a sufficiently bureaucratic point of view, we should really have defined "an empty set", and then proved that if you have two sets, each with no elements, then they are equal. To do this we need to use carefully the precise definitions of "subset" and "equal" to be found in Definition 1.1

and Definition 1.2. The reader is invited to do this while listening to the sound of one hand clapping. Reasoning about the non-existent elements of the empty set may seem a little like fraud. However, it is legitimate, and we will pin down exactly what reasoning is allowed in Section 1.11.

1.5 Set Operations

We can make new sets from old. The most well-known such operations are union and intersection. Suppose that A and B are sets, we define their *union* $A \cup B$ as

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

This is the mathematical usage of the word "or" – it is not exclusive. We have $x \in A \cup B$ even if x is an element of both A and B. In ordinary English usage, some people use the word "or" in a slightly different way. Such folk might say to a small child "you can have an ice-cream or a bar of chocolate", meaning that the options were one or the other but not both. This is perfectly acceptable English, but, in keeping with their compulsion to be precise, mathematicians have eschewed ambiguity by opting for a specific meaning of the word "or" – the inclusive one.

Another way of making new sets from old is intersection. If A and B are sets, we define their *intersection*, $A \cap B$, as

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

Remember that the comma is read as 'and' in this context. Special terminology is used to describe the situation when A and B are two sets such that $A \cap B = \emptyset$. We say A and B are *disjoint*. We are all familiar with certain laws which, say, the integers obey:

$$x + y = y + x \text{ for all integers } x \text{ and } y.$$

This is called the commutative law of addition – and it has analogues which hold in the theory of sets:

$$\begin{aligned} A \cup B &= B \cup A \text{ for all sets } A \text{ and } B, \text{ and} \qquad (1.6) \\ A \cap B &= B \cap A \text{ for all sets } A \text{ and } B. \qquad (1.7) \end{aligned}$$

We term these (respectively) the *commutative law of union* and the *commutative law of intersection.*

These laws are not mere assertions. They can be deduced in a few lines from the definitions of union and intersection. Let us prove that Equation (1.6) is valid.

Suppose that $x \in A \cup B$, then $x \in A$ or $x \in B$, so $x \in B \cup A$. Thus $A \cup B \subseteq B \cup A$. Suppose that $y \in B \cup A$, then $y \in B$ or $y \in A$, so $y \in A \cup B$. Thus $B \cup A \subseteq A \cup B$. Putting our two conclusions together, we deduce $A \cup B = B \cup A$ thanks to Definition 1.2.

That was easy. Both union and intersection satisfy ***associative laws***. Recall that for the integers, the associative law of multiplication is

$$(x \cdot y) \cdot z = x \cdot (y \cdot z) \text{ for all } x, y, z \in \mathbb{Z}.$$

The corresponding set-theoretic associative laws are

$$(A \cup B) \cup C = A \cup (B \cup C) \text{ for all sets } A, B \text{ and } C \tag{1.8}$$

and

$$(A \cap B) \cap C = A \cap (B \cap C) \text{ for all sets } A, B \text{ and } C. \tag{1.9}$$

The diligent reader is urged to write out proofs of Equations (1.8) and (1.9).

It is fairly obvious (but that is not a proof!) that an extended union (or intersection) is independent of the bracketing. We will prove this later (Proposition 2.4).

For example, for any sets A, B, C and D, we have

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

We may write $A \cup B \cup C \cup D$ unambiguously. Until you have read the proof of the legitimacy of this practice, or better still supplied one yourself, you should feel guilty when doing this.

There are laws which show how intersection and union interact; these laws are analogous to the distributive law

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

They are, for all sets A, B and C,

$$A \cap (B \cup C) = (A \cap B) \cup (A \cap C) \tag{1.10}$$

and

$$A \cup (B \cap C) = (A \cup B) \cap (A \cup C). \tag{1.11}$$

Notice that intersection and union both distribute over one another – this is a happy state of affairs, and is much cleaner than what happens with addition and multiplication of integers. Addition does not distribute over multiplication in that context. Anyone who thinks that

$$x + (y \cdot z) = (x + y) \cdot (x + z) \text{ for all } x, y, z \in \mathbb{Z}$$

is a law of the integers is rather wasting their time reading this book.

Equation (1.10) is called the ***distributive law of intersection over union***, and Equation (1.11) is called the ***distributive law of union over intersection.*** We prove Equation (1.10).

Suppose that $x \in A \cap (B \cup C)$, then $x \in A$ and either $x \in B$ or $x \in C$. Thus either $x \in A \cap B$ or $x \in A \cap C$. Thus $x \in (A \cap B) \cup (A \cap C)$, so $A \cap (B \cup C) \subseteq (A \cap B) \cup (A \cap C)$.

Conversely, suppose that $y \in (A \cap B) \cup (A \cap C)$. This forces $y \in A$. Similarly, we know that $y \in B$ or $y \in C$ so $y \in B \cup C$. Therefore $y \in A \cap (B \cup C)$. We deduce that $(A \cap B) \cup (A \cap C) \subseteq A \cap (B \cup C)$.

Putting the two conclusions together yields the desired result.

The empty set enjoys some special properties. Suppose that A is any set, then we have:

$$A \cup \emptyset = A \text{ and } A \cap \emptyset = \emptyset.$$

We now introduce the notion of the *difference* of two sets. Suppose that A and B are two sets, then we write $A \setminus B$ for $\{x \mid x \in A, x \notin B\}$. Notice that $A \setminus B$ and $B \setminus A$ are disjoint.

Suppose that all our discussion takes place about sets which are all subsets of a given fixed set called the universe U. For example, we might be discussing subsets of the integers, so it would make sense to put $U = \mathbb{Z}$. If A is a set (so $A \subseteq U$) we put $A' = U \setminus A$. The set A' has as elements exactly those elements of U which are not in A. The set A' is called the *complement* of A. Clearly $A'' = A$.

Observe that for any such A we have $A \cup A' = U$ and $A \cap A' = \emptyset$. The following rules, known as *De Morgan's laws*, are not quite so obvious:

$$\begin{aligned} (A \cup B)' &= A' \cap B' \text{ and} && (1.12) \\ (A \cap B)' &= A' \cup B'. && (1.13) \end{aligned}$$

Both equations hold for any $A, B \subset U$.

Complementation has no meaning unless there is a universe U. Let us illustrate De Morgan's laws. Suppose that $U = \mathbb{Z}, A = \mathbb{N}$ and $B = \{0\}$. The first law is valid because both sides are $\{x \mid -x \in \mathbb{N}\}$ (the set of negative integers). The second law is valid because both sides are $\mathbb{Z}$.

Warning: Alternative notations to $\cap$ and $\cup$ exist. You may see $\wedge$ and $\vee$ instead, which may still be pronounced *intersection* and *union*, but are sometimes pronounced as *meet* and *join*. Another variation which you may find is that some people put a horizontal bar $\overline{S}$ over the set S to denote its complement, but this is not common, since this notation is the usual one when describing a *closure* – terminology which arises in analysis and topology.

1.6 Sets of Sets

It is possible to have sets whose elements are other sets. For example, let

$$T = \{\{1,2\},\{123\},\{1,2,3\},1,2,3\}.$$

The set T has six elements. Its elements are 1, 2, 3 which are all numbers, and three elements which are not numbers at all, but other sets; $\{1,2\}$, $\{123\}$ and $\{1,2,3\}$. These latter elements are sets consisting of two, one and three elements respectively. Notice that 1 and $\{1\}$ are not the same thing at all. The first is a number, the second is a set with one element. Similarly we can consider yet another set called $\{\{1\}\}$. This is a set with one element, that element being $\{1\}$.

If we start with $\emptyset$, we can form $\{\emptyset\}$, which is not the same thing. The set $\emptyset$ has no elements, whereas $\{\emptyset\}$ has one element. On the other hand $\{\emptyset,\{\emptyset\}\}$ has two elements: $\emptyset$ and $\{\emptyset\}$.

Remark 1.2

Where do babies come from? has initiated a difficult conversation or two down the years. *Where do numbers come from?* is at least as dangerous. The Peano Axioms [1] are the currently fashionable method for making the natural numbers. We skimp heavily on technical detail here, but basically we can implement Peano's scheme in a way suggested by von Neumann by saying that $0, 1, 2, 3,$ etc. are shorthand for

$$\emptyset, \qquad \{\emptyset\}, \qquad \{\emptyset,\{\emptyset\}\}, \qquad \{\emptyset,\{\emptyset\},\{\emptyset,\{\emptyset\}\}\},$$

and so on. Notice that this sequence of sets is defined by each term being the set whose elements are its predecessors. What von Neumann is doing is, in a sense, making something from nothing. This has the entirely accidental consequence that $0 \in 1 \in 2 \in 3$ and so on. This is silly, or more accurately, unnecessary. What we have here is simply an artifact of the manufacturing process, and has nothing to do with the usual properties of the natural numbers. It is as if every infinite plane was imprinted with ©*Euclid 303 BC, made in Alexandria* in letters so small that no-one has ever noticed it. This artificiality is worth it though, since it avoids having to say numbers are "obvious" or "exist in a mental reality" or the need to make some other vulnerable assertion. The philosophers know this ground much better than most mathematicians, and so, rather than let them humiliate us, we use the von Neumann-Peano machine

[1] The Peano Axioms were invented by Dedekind

to reduce everything to pure technique. That way the philosophers don't get a look in.

You need to be warned about the construction, since otherwise you might come across a book where you find $\emptyset = 0$, which is true in the world of Peano and von Neumann, but is totally confusing. Having noted this important contribution to keeping philosophy at bay, we will quietly forget the Peano-von Neumann construction, and, making the unusual journey from experience to innocence, elect to think of 0, 1, 2, 3 (and so on) as atoms, and not as having some internal structure by which they are mysteriously related via set membership.

The upshot of the preceding two paragraphs is that I am encouraging you to think about numbers the way you almost certainly did before you read the preceding two paragraphs, but to be circumspect about when you admit this fact. Thus we decide that $\{\emptyset\} = 1$ is wrong because 1 is not a set but $\{\emptyset\}$ is one. We also decide that $1 \in 2$ is nonsense because 2 isn't a set so can't have elements. *Read this remark* $\{\emptyset, \{\emptyset\}, \{\emptyset, \{\emptyset\}\}\}$ *times.*

EXERCISES

1.1 Let $A = \{1,2,3\}$, $B = \{1,2\}$, $C = \{1,3\}$, $D = \{2,3\}$, $E = \{1\}$, $F = \{2\}$, $G = \{3\}$, $H = \emptyset$. Simplify the following expressions; in each case the answer should be one of these sets.

(a) $A \cap B$

(b) $A \cup C$

(c) $A \cap (B \cap C)$

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

(e) $A \setminus B$

(f) $C \setminus A$

(g) $(D \setminus F) \cup (F \setminus D)$

(h) $G \setminus A$

(j) $A \cup ((B \setminus C) \setminus F)$

(k) $H \cup H$

(l) $A \cap A$

(m) $((B \cup C) \cap C) \cup H.$

1.2 In this question, each part contains a description of three sets. In every case two of the sets are the same, and one is different. Write down the number of the set which is different. When we say two sets are the same, we mean that they are equal. See the definition of equality of sets in Section 1.2.

(a) (i) $\emptyset$ (ii) $\{\ \}$ (iii) $\{\emptyset\}$

(b) (i) $\{x \mid x \in \mathbb{Z},\ 0 < x < 1\}$ (ii) $\{y \mid y \in \mathbb{Z},\ 0 < y < 1\}$ (iii) $\{z \mid z \subseteq \{\emptyset\}\}$

(c) (i) $\{x \mid x \in \mathbb{N},\ 1 \leq x < 8\}$ (ii) $\{r \mid r \in \mathbb{Z},\ 1 \leq r^2 < 64\}$ (iii) $\{\zeta \mid \zeta \in \mathbb{Z},\ 1 \leq \zeta^3 < 512\}$

(d) (i) $\{\lambda \mid \lambda \in \mathbb{Z},\ \lambda \geq 0\}$ (ii) $\mathbb{N}$ (iii) $\{\nu \mid \nu \in \mathbb{N}\}$

(e) (i) $\emptyset \cup \emptyset$ (ii) $\{\emptyset,\ \emptyset\}$ (iii) $\emptyset \setminus \emptyset$

In part (f), the symbols π and e have their usual meanings, connected with mensuration of the circle and logarithmic growth respectively (mensuration means measuring, but sounds considerably more impressive).

(f) (i) $\{\rho \mid \rho \in \mathbb{N}, e < \rho < \pi\}$ (ii) $\{\sigma \mid \sigma \in \mathbb{N}, -\pi < \sigma < -e\}$ (iii) $\{\gamma \mid \gamma \in \mathbb{N}, \pi < \gamma < e\}$

In part (g), the *Universe* is $\mathbb{N}$, and so the notion of *complementation* makes sense.

(g) (i) $\{\beta \mid \beta \subseteq \mathbb{N},\ |\beta'| < \infty\}$ (ii) $\{\mu \mid \mu \subseteq \mathbb{N},\ \mu \text{ is infinite}\}$ (iii) $\{\nu \mid \nu \subseteq \mathbb{N},\ \nu'' \text{ is infinite}\}$

(h) (i) $\mathbb{N} \cup \{\tau \mid -\tau \in \mathbb{N}\}$ (ii) $\mathbb{Z} \setminus \{0\}$ (iii) $\{\alpha \mid \text{there is } A \subseteq \mathbb{N} \text{ such that } \alpha \in A\}$

In parts (j) and (k), the symbol I_r^s (temporarily) denotes the set of real numbers $\{\eta \mid r \leq \eta < s\}$ where r and s are themselves real numbers. In part (k) we introduce notation for the intersection of a collection of sets. This usage is entirely analogous to the notation Σ used to specify a sum; Σ means *add this lot up*, whereas $\bigcap$ means *intersect this lot.* A similar notation makes sense for any associative and commutative operation.

(j) (i) $I_0^1 \cap I_{-1}^0$ (ii) $I_1^0 \cap I_0^{-1}$ (iii) $\{0\}$

(k) (i) $\bigcap_{\lambda>0} I_0^\lambda$ (ii) $\bigcap_{\lambda\geq 0} I_0^\lambda$ (iii) $\bigcap_{\lambda<0} I_0^\lambda$

(l) (i) $\mathbb{Q} \cap \mathbb{Z}$ (ii) $\mathbb{Q} \cap \mathbb{N}$ (iii) $\mathbb{R} \cap \mathbb{N}$

(m) (i) $(\emptyset \cup \{\emptyset\}) \cup \{0\}$ (ii) $\{\emptyset, \{\emptyset\}, 0\}$ (iii) $\{\emptyset,\ 0\}$.

1.7 Paradox

At the start of this book, in the second sentence of Section 1.1, we used the words "roughly speaking, a set is". We have been a little too naïve. We have never properly defined a set. This can be done, but would take us too far afield. Our approach – regarding sets as collections of arbitrary objects (either atoms or other sets) – has its limitations. The greatest fear of mathematicians is paradox, and careless attempted formation of sets can lead to this.

Remark 1.3

We have, so far, used properties only to describe subsets of a given set. We have not written $S = \{x \mid P(x)\}$; in words, we have never said S is the set of elements x which enjoy the property P. We have always written $S = \{x \mid x \in A, P(x)\}$. We have only used properties to define subsets of a pre-existing set A. If one does that, there is no problem with set theory. The set $\{x \mid x \in \mathbb{Z},\ x \text{ is a perfect square}\}$ is perfectly harmless. The innocent user of sets is rather unlikely to stumble upon a paradox – you really have to go looking for trouble if you want to find it (in this instance). Nonetheless, the danger is there, and you must know how to guard against it.

If we were recklessly to allow any property to define a set, we would be in deep trouble – as the following argument of Bertrand Russell shows. Now, by the way, is a good time to concentrate. A stimulant of some sort may be in order.

Let $S = \{x \mid x \text{ is a set }\}$. This is the set whose elements are all possible sets. Thus $S \in S$. There are other sets which are not members of themselves, for example $\{1\} \notin \{1\}$. We define $T = \{x \mid x \in S, x \notin x\}$. Now suppose that $T \notin T$, then from the definition of T it follows that $T \in T$. This is impossible, since we cannot have both $T \in T$ and $T \notin T$. Thus we are led to conclude that $T \in T$. Now examine the definition of T again. It now follows that $T \notin T$. We have the same problem as before.

Paradox, chaos and confusion reign. The reasoning is sound; one escape route is to prevent S from being a set in the first place. This in turn will undermine the legitimacy of the definition of T. We can then serenely walk on, having disposed of the troublesome T – it is not a set so the paradox does not matter. Nonetheless, Russell's paradox is very disturbing. The fact that we can dodge it mathematically is all very well, but it is surely rather surprising that a primitive notion such as "a collection of objects" can lead us so quickly into the quagmire of unreason. Russell's argument, by the way, was first formulated in a letter to the German logician Frege. It knocked the foundations out from under Frege's very substantial book, which was about to appear. That is how to make someone's day.

Russell's argument is analogous to ***the barber's paradox***. Consider a village with one adult male barber, and ponder upon the assertion – "The barber shaves every man in the village who does not shave himself, and does not shave anyone else". Now ask yourself the question – who shaves the barber?

Returning to the issue at hand, our objective is to stop S from being a set. We accomplish this by *not allowing* sets to be defined by arbitrary properties. The constructions outlined in this book will not lead to disaster. We do not want bizarre sets B with the feature that $B \in B$. If you ever see a set even thinking of having this property, deport it. If you want to read an account of what exactly one is (and is not) allowed to do when constructing sets, have a look at *Naïve Set Theory* by Paul Halmos (published by Springer-Verlag).

The sets we have been discussing, $\mathbb{N}$, $\mathbb{Z}$, $\mathbb{Q}$, $\mathbb{R}$, $\mathbb{C}$, and their subsets do not cause us difficulties, nor do any sets that we construct from them in sensible ways.

Let us be clear. You *can* use a property to define a subset of a pre-existing set. For example $\{x \mid x \in \mathbb{N},\ x \text{ is a perfect square}\}$. Here we have a pre-existing set $\mathbb{N}$ and we focus on its elements which enjoy a property. That is allowed. However, we ban $\{y \mid y \text{ is a set}\}$ since this is an (illegal) attempt to define a set using a property alone. There is nothing wrong with the property though; we might consider $\{x \mid x \in \mathbb{N}, x \text{ is a set}\}$ which, given that Messrs Peano and von Neumann have been shown the door, and no philosophers are looking, is the empty set. On the other hand, at inter-faculty parties you can argue that it is $\mathbb{N}$.

1.8 Set-theoretic Constructions

Despite the warnings in the previous section, there are some set-theoretic constructions which do not lead to difficulties, and we permit ourselves to use them. Suppose that A is a set. We allow ourselves to form a new set called the *power set* of A. This set is written $P(A)$, and has as its elements exactly the subsets of the original set. In symbols we have

$$P(A) = \{x \mid x \subseteq A\}.$$

Thus if $S = \{1,2,3\}$, then

$$P(S) = \{\emptyset, \{1\}, \{2\}, \{3\}, \{1,2\}, \{1,3\}, \{2,3\}, \{1,2,3\}\}.$$

If A is a finite set, then clearly $P(A)$ is finite. If A is an infinite set, then $P(A)$ is infinite.

Another construction we often need is that of the *Cartesian product* of two (or more) sets. Suppose that A and B are sets, then we form a new set called $A \times B$ which consists of "ordered pairs", the first element drawn from A, the second from B. In symbols we write

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

A purist reader might immediately object. So far we have just used properties and set membership to define new sets from old. Now a new notion – that of an ordered pair – has suddenly been slipped in. What on earth is an ordered pair? There is an artificial set-theoretic dodge to avoid the problem. We just regard (a,b) as notation for the set $\{a, \{a,b\}\}$, and then the problem goes away. Rather than chase this hare, we will adopt the naïve view and say that (a,b) is a symbol denoting a pair of elements a and b, that it can so happen that $a = b$, and that $(a,b) = (c,d)$ if and only if (i.e. exactly when) both $a = c$ and $b = d$.

One can continue the process indefinitely. If A, B, and C are sets we can form $(A \times B) \times C$ and $A \times (B \times C)$. In fact these sets are not formally the same thing, but they are usually identified in the obvious way; we think of either of them as the set of ordered triples

$$\{(a,b,c) \mid a \in A, b \in B, c \in C\}.$$

In a similar way we can form the Cartesian product of any finite number of sets. There is even a way of forming the Cartesian product of infinitely many sets.

Suppose that A is a set. We refer to $A \times A$ as the *Cartesian square* of A, and allow the notation A^2 for this set. More generally, we call a Cartesian product with n factors $A \times A \times ... \times A$ a *Cartesian power* of A, and write the set as A^n. We use ordinary "x, y" co-ordinates to study the geometry of the plane via algebra and calculus. In order to do this we identify the points of the plane with the Cartesian square of $\mathbb{R}$, usually called $\mathbb{R}^2$.

1.9 Notation

There has already been one notational crisis mentioned in the text. The (fashion) question as to whether or not 0 should be considered to be a natural number has not been resolved. Our convention, which is by no means universal, is that $0 \notin \mathbb{N}$.

There are other issues that deserve (very brief) attention. The subset notation is problematic. We have been using $\subseteq$ to denote "is a subset of". There are people who use $\subset$ in this way too. The trouble is that there are yet other people lurking in the shadows who read $A \subset B$ to mean that A is a subset of B but that $A \neq B$. The jargon for this is that A is a *proper* subset of B. This is by analogy with how we use $<$ and $\leq$. In the opinion of the author, this nice distinction between $\subset$ and $\subseteq$ just causes trouble, since you often do not know to which convention the writer adheres.

Another low-level controversy is the notation for set difference; we have written $A \setminus B$ for $\{x \mid x \in A, x \notin B\}$. There is a school of thought that this should instead be written $A - B$. Watch out for this – especially in computer graphics for example.

1.10 Venn Diagrams

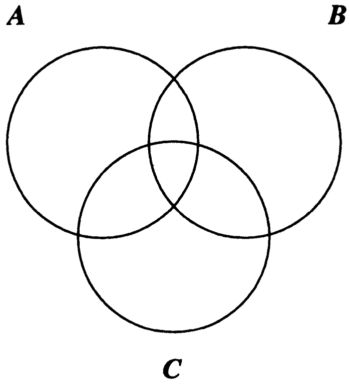

Fig. 1.1. Venn diagram for three sets

I assume that the reader knows what a Venn diagram is. Venn diagrams are very good aids to thinking. They constitute a picture of what is going on – if correctly drawn. If there are more than three sets involved, this gets very tricky. If there are at most three sets involved, all is well. In mathematics you sometimes need to think about more than three sets at once, and for this reason you will eventually need to learn to live without Venn diagrams.

EXERCISES

1.3 (a) Use A, B, C together with $\cap$ and complementation to describe eight sets corresponding to the regions of Figure 1.1. Note the existence of the region outside the interlocking circles, and that we suppose that there is a universe U.

(b) Do the same as in part (a), but use A, B, C together with $\cup$ and complementation.

1.4 (a) Prove that $\{1,2,3\} = \{1,1,2,3\}$.

(b) Recall the notation for power sets; the power set of a set A is written $P(A)$. How many elements are there in the set $P(\{1,2,3\})$?

(c) Determine the cardinality of $P(\emptyset)$.

(d) Determine the cardinality of $P(P(\emptyset))$.

(e) Determine the cardinality of $P(P(P(P(P(\emptyset)))))$.

1.5 (a) Draw Venn diagrams illustrating the truth of De Morgan's laws (1.12) and (1.13).

(b) Prove De Morgan's laws. No Venn diagrams allowed!

(c) Find two subsets A, B of the natural numbers $\mathbb{N}$ such that both $A \setminus B$ and $B \setminus A$ are infinite sets.

1.6 (Harder) Let $A = \{1, 2, 3, \ldots, n\}$. What is the cardinality of each of the following sets?

(a) $\{(a, S) \mid a \in S, S \in P(A)\}$

(b) $\{(a, S) \mid a \in A, a \notin S, S \in P(A)\}$

(c) $\{(S, T) \mid S \in P(A), T \in P(A), S \cap T = \emptyset\}$

(d) $\{(S, T) \mid S \in P(A), T \in P(A), S \cup T = A\}$.

1.11 Quantifiers and Negation

There are two symbols called *quantifiers* which you come across all the time in mathematics. They are $\forall$ and $\exists$. You have a little latitude in how you read them.

You pronounce $\forall$ as *for each*, *for every* or *for all*, and for $\exists$ you say *there exist*, *there exists*, *there is* or *there are*, depending on the context. Consider the following two statements.

$$\forall x \in \mathbb{Z}\, \exists y \in \mathbb{Z} \text{ such that } x < y. \tag{1.14}$$

$$\exists y \in \mathbb{Z} \text{ such that } \forall x \in \mathbb{Z}\, x < y. \tag{1.15}$$

The "such that" phrases are just padding to make the English read better. In fact you can take $\exists x$ to mean "there exists x such that", and then there is no real need to put these English words into the mathematical expression. It is quite a good idea to put the non-quantifier part of the expression in round brackets, otherwise the stream of symbols can be a bit overwhelming. Our expressions (1.14) and (1.15) become

$$\forall x \in \mathbb{Z}\, \exists y \in \mathbb{Z}(x < y) \text{ and} \tag{1.16}$$

$$\exists y \in \mathbb{Z}\, \forall x \in \mathbb{Z}(x < y). \tag{1.17}$$

This is a case where we read from left to right (as in written English), so Equations (1.14) and (1.16) correctly assert that there is no largest integer. On the other hand, Equations (1.15) and (1.17) incorrectly assert that there is a largest integer. Viewed purely symbolically, this tells us that $\exists$ and $\forall$ do not commute.

Suppose that P is a proposition, let $\neg P$ denote "not P", a statement which is true if and only if P false. We say that we *negate* P when we place the symbol $\neg$ before it. Thus for real numbers x and y, we have $\neg(x < y)$ is logically equivalent to $x \not< y$ (often written $y \leq x$). The words *logically equivalent* have a technical meaning, but please read them informally. There is a nice interaction between quantifiers and negation. If P is a proposition, let $\neg P$ denote its negation. The algebra of quantifiers is as follows.

$$\neg\forall x(P) \text{ is logically equivalent to } \exists x(\neg P)$$

and similarly

$$\neg\exists x(P) \text{ is logically equivalent to } \forall x(\neg P).$$

Thus pushing a $\neg$ through a quantifier changes $\forall$ to $\exists$ and changes $\exists$ to $\forall$. We use this information to turn the falsehood (1.17) into a truth. According to our rules

$$\neg(\exists y \in \mathbb{Z}\, \forall x \in \mathbb{Z}\, (x < y))$$

is logically equivalent to

$$\forall y \in \mathbb{Z}\, \exists x \in \mathbb{Z}\, (y \leq x).$$

which asserts that for every integer y there is an integer x which is at least as big as y. Now let us negate the truthful statement (1.16) and confidently expect a falsehood to emerge. The negation is

$$\exists x \in \mathbb{Z} \forall y \in \mathbb{Z}\, (y \leq x),$$

which asserts that there is a largest integer. This sort of reasoning really comes into its own when you have a string of several quantifiers, one after the other, and you want to negate the statement. There is no need to think about what the statement *means*. You can negate it just by pushing symbols around.

Also notice that membership of sets over which quantified variables are ranging is unchanged by the process of passing a negation through a quantifier. The (false) assertion that there is a fixed integer which is less than every real number is written

$$\exists x \in \mathbb{Z}\ \forall y \in \mathbb{R}(x < y)$$

and when negated becomes the true statement

$$\forall x \in \mathbb{Z}\ \exists y \in \mathbb{R}(y \leq x),$$

which asserts that given any integer x, there is a real number y such that $y \leq x$.

Particular care should be taken when reasoning about the empty set. Any statement about the members of the empty set is true. For example, and readers of a nervous disposition are warned that they may find this disturbing, the statements

$$\forall x \in \emptyset\, (x \text{ is a banana})$$

and

$$\forall x \in \emptyset\, (x \text{ is not a banana})$$

are both true statements. We say that both statements are *vacuously true*. The informal reasoning goes as follows. The first statement is true because there is no element of the empty set which is not a banana. The second statement is true because there is no element of the empty set which is a banana.

It should be admitted that reasoning about non-existent elements may seem rather brave. The justification is that when $P(x)$ is some proposition concerning $x \in S$, we want $\forall x \in S\, (P(x))$ to have exactly the same meaning as $\neg\exists x \in S\, \neg(P(x))$. When S happens to be empty, irrespective of $P(x)$, The statement $\neg\exists x \in \emptyset\, \neg(P(x))$ has to be true because it begins $\neg\exists x \in \emptyset$. Note that any statement which asserts the existence of an element of the empty set is false.

1.12 Informal Description of Maps

Sets by themselves are not very interesting. In order to explore mathematics, we need to be able to leap about from set to set. The device that is used to do this is called a *map*. It is also called a *mapping*, and also a *function*. We will later give a completely formal definition of a map, but in the first instance we outline a useful and rather relaxed way of thinking about maps. This way of thinking about maps is not actually harmful, but since we shall introduce ideas which are not set-theoretic, a mathematical zealot (gongs sound, prayer wheels spin, mantras drone) will find this approach distasteful. We embark on this route for two reasons. First because this is the way almost all mathematicians think about maps, and second because maps are rather easier to understand if one is prepared to slum it. We will indicate how we could have kept the faith in Section 1.15.

Informally then, a map is a *rule*. You feed in a piece of data, and the rule responds with an item of output data. The rule can be thought of as living in a black box with separate input and output devices attached. Maps are utterly reliable. If you feed in identical pieces of data on two separate occasions, the output response will be the same in both instances. The jargon for this is to say that they are *deterministic*. The rule does not toss a coin to decide how to respond to a piece of input, nor does it alter its behaviour over time. The item being fed to the rule is traditionally called the *argument* (one can imagine an indignant item of input data objecting to its fate!).

Definition 1.3

The set A (the possible inputs) is called the *domain* of the map. The set B (the set from which the outputs will be drawn) is called the *codomain* of the map. We say that the mapping is from A to B.

A mapping might be specified by the following statement: "Only feed me data from the set $\mathbb{Z}$; output is guaranteed to be from the set $\mathbb{N}$." The map therefore has domain $\mathbb{Z}$ and codomain $\mathbb{N}$. We write

$$f : \mathbb{Z} \to \mathbb{N};\ \ f : x \mapsto 1 + x^2 \ \forall x \in \mathbb{Z}.$$

This tells us everything we could wish to know. It translates into English as follows:

The name of the map is f (the colon following f tells us that the name has finished). The symbols $\mathbb{Z} \to \mathbb{N}$ tell us that the domain of f is $\mathbb{Z}$ and the codomain of f is $\mathbb{N}$. The semicolon tells us to expect either a formula or some other recipe for describing which inputs will yield which outputs. The final piece of information is that if the integer x is fed in to the black box, the output will be $1 + x^2$. We abbreviate the domain of f to $\mathrm{dom}(f)$ and the codomain to $\mathrm{cod}(f)$.

The notation $f : x \mapsto 1 + x^2 \ \forall x \in \mathbb{Z}$ has the virtue of complete clarity, but is perhaps a little cumbersome. It is important to note the tail on the second arrow – we use $\to$ between the domain and codomain, but $\mapsto$ in the recipe section. You can pronounce $\to$ as "maps to" and $\mapsto$ as "goes to" or "is sent to".

We allow ourselves to denote the output of the function by $f(x)$ when the input is x. The description could therefore read

$$f : \mathbb{Z} \to \mathbb{N};\ f(x) = 1 + x^2 \ \forall x \in \mathbb{Z}.$$

There is one type of map which is so important that we single it out for special attention.

Definition 1.4

Given any set A we can always form the identity map from A to A. This is the map $\mathrm{Id}_A : A \to A;\ \mathrm{Id}_A : a \mapsto a \ \forall a \in A$.

We list below various other possible descriptions of maps:

$$a : \mathbb{N} \to \mathbb{N};\ a : x \mapsto 1 \ \forall x \in \mathbb{N}, \tag{1.18}$$

$$b : \mathbb{R} \to \mathbb{R};\ b : r \mapsto \sqrt{|r|} \ \forall r \in \mathbb{R}, \tag{1.19}$$

$$c : \{1,2,3\} \to \{1,2,3,4\};\ 1 \mapsto 1,\ 2 \mapsto 2,\ 3 \mapsto 4, \tag{1.20}$$

and

$$d : \mathbb{Z} \to \mathbb{N} \cup \{0\};\ f(x) = x \ \forall x \geq 0,\ f(x) = -x \ \forall x < 0. \tag{1.21}$$

The map (1.21) is of course the "modulus function", a map from $\mathbb{Z}$ to $\mathbb{N} \cup \{0\}$. Notice that the rule need not be simple. Here is a rather peculiar map:

$$f : \mathbb{R} \to \mathbb{Q};\ f(x) = x \text{ if } x \in \mathbb{Q},\ f(x) = 0 \text{ if } x \notin \mathbb{Q}. \tag{1.22}$$

1.13 Injective, Surjective and Bijective Maps

Suppose that $f : A \to B$ is a map. We define the *image* of f to be

$$\mathrm{im}(f) = \{x |\ x \in B,\ x = f(a) \text{ for some } a \in A\}.$$

This could equally well be written

$$\mathrm{im}(f) = \{f(a) |\ a \in A\}.$$

Thus $\mathrm{im}(f) \subseteq \mathrm{cod}(f)$. Note that we do not require equality here.

It is of course rather interesting when $\mathrm{im}(f) = \mathrm{cod}(f)$; in this circumstance we say that the map is *onto* or *surjective* or *epic*. One can only apologize for the multiplicity of jargon. It all depends on whether you prefer your English roots Anglo-Saxon, Latin or Hellenic. We shall use the term *surjective* in this book. We shall also write that f is a *surjection*. The map (1.21) was surjective, as was the map (1.22), and identity maps are always surjective.

Another way that a map can stand out from the crowd is if different inputs always give different outputs. More formally, let us suppose that $f: A \to B$ has the property that if $a, a' \in A$ and $a \neq a'$, then $f(a) \neq f(a')$. We say (here we go again) f is *1-1* or *injective* or *monic*. Once again, we plump for the Latin option. Thus we refer to *injective* maps and to *injections*. Another way of saying that f is injective is to observe that for all $a, a' \in A$, if $f(a) = f(a')$, then $a = a'$. The map (1.20) is an injection, as are all identity maps.

Suppose that $f: A \to B$ is both injective and surjective, then we say that it is *bijective* or a *bijection*. Those who prefer an earthier tongue may call it a *1-1 correspondence*.

We say two maps f and g are *equal* and write $f = g$ if and only if three conditions are satisfied:

$$\begin{aligned} \mathrm{dom}(f) &= \mathrm{dom}(g), \\ \mathrm{cod}(f) &= \mathrm{cod}(g), \\ \forall x \in \mathrm{dom}(f),\ f(x) &= g(x). \end{aligned}$$

Notice that the map

$$\alpha:\ \mathbb{Z} \to \mathbb{N} \cup \{0\};\ a(x) = x^2\ \forall x \in \mathbb{Z}$$

and the map

$$\beta:\ \mathbb{Z} \to \mathbb{Z};\ b(x) = x^2\ \forall x \in \mathbb{Z}$$

are not equal even though they are very similar. They have different codomains.

1.14 Composition of Maps

Suppose that $f: A \to B$ and $g: B \to C$, then we can define a map called $g \circ f$ called the composition of f with g. This map is described formally as follows:

$$g \circ f:\ A \to C;\ g \circ f:\ x \mapsto g(f(x))\ \forall x \in A.$$

Informally, you feed x into f, take the output and feed it straight into g, and finally gather the output.

Notice that we can form $g \circ f$ if and only if (i.e exactly when) $\mathrm{cod}(f) = \mathrm{dom}(g)$. We do not allow ourselves to be seduced into permitting map composition when $\mathrm{cod}(f)$ is a proper subset of $\mathrm{dom}(g)$ – tempting though it is. Some writers are not so strict.

Example 1.1

Suppose that

$$f : \mathbb{Z} \to \mathbb{Z};\ f(x) = x - 1\ \forall x \in \mathbb{Z}$$

and

$$g : \mathbb{Z} \to \mathbb{N} \cup \{0\};\ g(x) = x^2\ \forall x \in \mathbb{Z},$$

then

$$g \circ f : \mathbb{Z} \to \mathbb{N} \cup \{0\};\ (g \circ f)(x) = x^2 - 2x + 1\ \forall x \in \mathbb{Z}.$$

In cultures where people write horizontally from left to right, you might think it perfectly natural that f composed with g would be written $f \circ g$ – first apply f and then apply g. In fact we should write $(x)f$ for the output of the map f when x was fed in. Happily there is such a world of reason, justice and harmony. It is called ***abstract algebra***, and in that better place, maps are often written on the right – so the composition of maps can be written sensibly. Alas, the rest of mathematics is inhabited by curious people who insist on writing their maps on the left. Folk have been writing $\sin(x)$ for too long to be able to cope with writing $(x)\sin$. The author is, as always, neutral on this issue. Nonetheless, having seen colleagues dusting down the thumb-screws, we shall write maps on the left (under protest) most of the time.

However, there are times when, for good or historical reasons, it makes sense to write maps on the right. The exclamation mark indicating the application of the factorial function is usually written on the right, and permutations are best written on the right, as we will see. Other more exotic notation has become familiar through usage. For example, to apply the squaring function one usually writes a small 2 to the right of the argument but raised above the line of print. We apply the modulus function by writing vertical lines on *both sides* of the argument. To apply the square root function we draw a piece-wise linear ideogram draped over the top and left of the argument: $\sqrt{x}$.

It is high time we proved something.

Proposition 1.2

(a) The composition of two injective maps is injective.

(b) The composition of two surjective maps is surjective.

Proof

Suppose that $f: A \to B$ and $g: B \to C$ are the maps concerned.

(a) We assume f and g are injective. Suppose that $a, a' \in A$ and $a \neq a'$. The map f is injective so $f(a) \neq f(a')$. The map g is also injective so $g(f(a)) \neq g(f(a'))$. Thus $g \circ f$ is injective.

(b) We now assume f and g are surjective. Choose any $c \in C$. The map g is surjective so there exists $b \in B$ such that $g(b) = c$. Also the map f is surjective so there exists $a \in A$ such that $f(a) = b$. Thus $(g \circ f)(a) = g(f(a)) = c$, so $g \circ f$ is surjective.

□

Corollary 1.1

The composition of two bijective maps is bijective.

A new word has crept in – *corollary*. A corollary is a result that you get for free, or almost for free, from the previous result. If you cannot see why a corollary follows from its preceding result, either the writer is mistaken, or you are being dense.

Now suppose that we have three maps

$$f: A \to B, \quad g: B \to C, \quad h: C \to D.$$

Arguably the single most important equation in this book is

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

for each of these maps sends each $x \in A$ to $h(g(f(x))) \in D$. We say that composition of maps is ***associative***. See Proposition 2.4, where you will find a formal justification of the fact that brackets can always be omitted from any repeated application of an associative operation.

In particular, if we have a map $k: A \to A$ and a natural number n, then the map

$$\underbrace{k \circ k \circ \cdots \circ k}_{n \text{ copies of } k}: A \to A$$

is unambiguously defined, so it makes sense to call this map k^n. You have to be careful though; if A has a multiplicative structure, there is an excellent *confusion opportunity* for the unwary student, as the next example shows.

Example 1.2

Consider the map $\sin: \mathbb{R} \to \mathbb{R}$. In the notation we have just constructed, $\sin^2 x = \sin(\sin x)$, whereas conventional usage is that $\sin^2 x = (\sin x)^2$. Now

$$\sin(\sin -\pi/2) = \sin(-1) < 0$$

whereas $(\sin(-\pi/2))^2$ is the square of a real number and so is non-negative. Any lingering flicker of hope that these two functions are the same has just been extinguished, and we have a genuine problem. There is no easy way out; you just have to be careful, and to make sure you understand what you are reading or writing.

EXERCISES

Recall that we make the convention that $0 \notin \mathbb{N}$.

1.7 Let $f: \mathbb{Z} \to \mathbb{Z}$ be defined by $f(x) = x^2 \, \forall x \in \mathbb{Z}$ and let $g: \mathbb{Z} \to \mathbb{Z}$ be defined by $g(x) = x + 1 \, \forall x \in \mathbb{Z}$.

(a) Give the formulas which define the maps $f \circ g$ and $g \circ f$, carefully distinguishing which is which.

(b) Let $n \in \mathbb{N}$. Give the formulas which define the maps f^n and g^n.

(c) Which of the maps f, g, f^2 and g^2 are bijections?

1.8 (a) Define a map from the natural numbers to themselves which is injective but not surjective (i.e. you must name the map, and explain what value is obtained when the map is applied to each natural number).

(b) Define a map from the natural numbers to themselves which is surjective but not injective.

(c) Define two maps α and β which are both bijections from the natural numbers to themselves but which enjoy the property that $\alpha \circ \beta \neq \beta \circ \alpha$.

1.9 (a) Define a bijective map from $\mathbb{N}$ to $\mathbb{Z}$.

(b) Define a surjective map from $\mathbb{Z}$ to $\mathbb{Z}$ which is not a bijection.

1.10 (a) Suppose that α, β and γ are all maps from a set A to itself, and that α is a bijection. Furthermore, suppose that $\alpha \circ \beta = \alpha \circ \gamma$. Does it follow that $\beta = \gamma$? Justify your answer.

(b) Suppose that A is a set, and both f and g are bijections from A to A such that $f^2 = g^2$. Does it follow that $f = g$? Justify your answer.

(c) Suppose that the conditions of part (b) hold but that also $f^3 = g^3$. Now does it follow that $f = g$? Justify your answer.

1.11 Let $S = \{1, 2, 3, \ldots n\}$.

(a) How many maps are there from S to S?

(b) How many surjective maps are there from S to S?

(c) How many injective maps are there from S to S?

(d) How many bijective maps are there from S to S?

1.12 True or false? Justify your answer.

(a) If $f : A \to B$ is bijective, then it is surjective.

(b) If $f : A \to A$ is injective but not surjective, then A is an infinite set.

(c) If $f : A \to A$ is such that f^2 is bijective, then f is bijective.

(d) If $f : A \to B$ is surjective and B is finite, then A is finite.

(e) If $f : A \to B$ is surjective and B is infinite, then A is infinite.

(f) If $f : A \to B$ is injective and B is infinite, then A is infinite.

(g) If $f : A \to B$ is injective and B is finite, then A is finite.

(h) If $f : A \to B$ is surjective and A is finite, then B is finite.

(j) If $f : A \to B$ is surjective and A is infinite, then B is infinite.

(k) If $f : A \to B$ is injective and A is infinite, then B is finite.

(l) If $f : A \to B$ is injective and A is finite, then B is finite.

1.13 (a) Suppose that f, g, h are all maps from a set U to itself. Suppose that $f \circ g = g \circ f$ and that $f \circ h = h \circ f$. Prove that $f \circ (g \circ h) = (g \circ h) \circ f$.

(b) Show by example that subject to the hypotheses of part (a), one cannot deduce that $g \circ h = h \circ g$.

(c) Let C be a non-empty set and suppose that $\alpha : C \to C$ has the property that whenever $\beta : C \to C$ then $\alpha \circ \beta = \beta \circ \alpha$. Prove that $\alpha = \mathrm{Id}_C$. *(Hint: Ασσυμϵ φορ κοντραδικτιον θατ $\alpha \neq Id_C$; if you don't get the point about κοντραδικτιον, look ahead to the νϵξτ χαπτϵρ).*

(d) Let D be a set and suppose that $\gamma : D \to D$ is a bijection which has the property that whenever $\delta : D \to D$ is a bijection, then $\gamma \circ \delta = \delta \circ \gamma$. If $\gamma \neq \mathrm{Id}_D$, what can be said about $|D|$?

1.14 Let W denote the set of finite strings of lower case Roman letters (i.e. the set of all words that could be written in that alphabet). Thus *dog* $\in W$ and *wetquytriut* $\in W$. Define three maps from W to W as follows. $f_1 : W \to W$ concatenates a word with itself, i.e. $f_1(\lambda) = \lambda\lambda \ \forall \lambda \in W$. For example $f_1(do) = dodo$. The map f_2 removes the first letter of a word if there is one – so $f_2(qbath) = bath$ but f_2 applied to the empty word is the empty word. The map f_3 is like f_2 except that it removes the last letter of a word if there is one.

The problem is to decide whether or not, starting with the word

$$abcdefghijklmnopqrstuvwxyz,$$

one can successively apply these maps (in a judiciously chosen order) to obtain the word

$$zyxwvutsrqponmlkjihgfedcba.$$

Justify your answer.

1.15 In this question we discuss a map $f : A \to B$.

(a) Suppose that there is a map $g : B \to A$ such that $f \circ g = \mathrm{Id}_B$. Prove that f is surjective.

(b) Suppose that there is a map $h : B \to A$ such that $h \circ f = \mathrm{Id}_A$. Prove that f is injective.

(c) Suppose now that the hypotheses of parts (a) and (b) hold simultaneously. Prove that f is bijective and that $g = h$.

1.15 Graphs and Respectability Reclaimed

Definition 1.5

Suppose that $f: A \to B$ is a map. The graph of the map f is the subset of $A \times B$ defined by

$$\mathrm{graph}(f) = \{(a, f(a)) |\ a \in A\} = \{(a, b) |\ b = f(a), a \in A\}.$$

Suppose that $g : \mathbb{R} \to \mathbb{R}$, then graph$(g)$ is a subset of $\mathbb{R} \times \mathbb{R} = \mathbb{R}^2$ and corresponds exactly to the casual use of the term graph when $\mathbb{R}^2$ is identified with the plane in the usual way, as in Figure 1.2. Meanwhile, back at the general case, notice that graph(f) is not an arbitrary subset of $A \times B$. Let $S = \mathrm{graph}(f)$. It has two special properties.

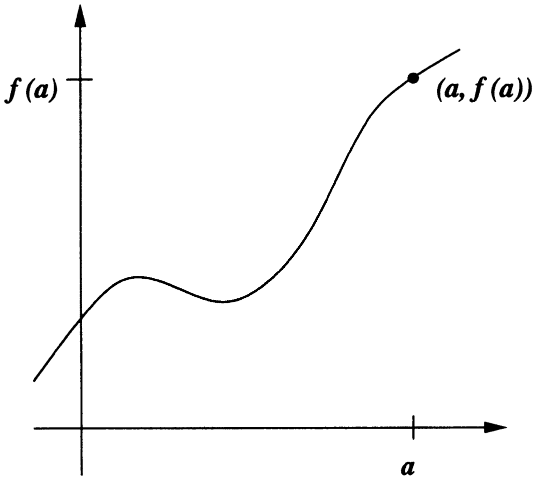

Fig. 1.2. $A = B = \mathbb{R}$ gives the familiar picture

(i) Given any $a \in A$, $\exists b \in B$ such that $(a, b) \in S$.

(ii) If (a, b), $(a, b') \in S$, then $b = b'$.

This gives us the appropriate hint about how maps should have been defined in the first place. All that nonsense about black boxes was of course designed to appeal to persons with a limited attention span, their minds addled by television soap-operas and junk-food. One simply defines a *graph* in $A \times B$ to be a subset S of $A \times B$ which has properties (i) and (ii). If S is a *graph*, then it gives rise to a map $f_S : A \to B$ as follows; for any $a \in A$, we write $f_S(a)$ for the unique element of B such that $(a, f_S(a)) \in S$. No "rules" are required. This way we can build the notion of a map straight out of set-theory. In all honesty though, most people do find that thinking of maps as rules is the easy thing to do.

1.16 Characterizing Bijections

Suppose that $\alpha : X \to Y$ is a bijection. We can define a map $\beta : Y \to X$ as follows: for each $y \in Y$, $\beta(y)$ is the unique element $x \in X$ such that $\alpha(x) = y$. In order for β to be properly defined it is crucial that α is a bijection. The fact that $\text{im}(\alpha) = Y$ ensures that $\text{dom}(\beta) = Y$, and the injectivity of α gives us the recipe for constructing β. Notice that $\beta \circ \alpha = \text{Id}_X$ and $\alpha \circ \beta = \text{Id}_Y$. The map β is the inverse of α, and we can write $\beta = \alpha^{-1}$.

We can characterize bijections using maps in the following way:

Proposition 1.3

Let $\alpha : X \to Y$ be a map. The following are equivalent:

(a) The map α is a bijection.

(b) There is $\beta : Y \to X$ such that $\beta \circ \alpha = \text{Id}_X$ and $\alpha \circ \beta = \text{Id}_Y$.

Proof

The preceding discussion ensures that $(a) \Rightarrow (b)$ (the symbol $\Rightarrow$ means *implies*, it has a rather infectious quality and should be used with extreme restraint – we will discuss this matter further in Section 2.6).

The heart of the theorem remains to do; we must show $(b) \Rightarrow (a)$. First we shall show that α is surjective. Suppose that $y \in Y$, then $\alpha(\beta(y)) = \text{Id}_Y(y) = y$ so $y \in \text{im}(\alpha)$. Thus $\text{im}(\alpha) = Y$ and α is surjective. It remains to prove that α is injective. Suppose that $x, x' \in X$ and $\alpha(x) = \alpha(x')$, then $\beta(\alpha(x)) = \beta(\alpha(x'))$ so $\text{Id}_X(x) = \text{Id}_X(x')$ and finally $x = x'$. Thus α is injective.

□

Corollary 1.2

The map β having property (b) is a bijection, and is the unique map rendering (b) valid. We say that β is the *inverse map* to α, and write β as α^{-1}. An extremely rich source of error is provided by the chance to write that α^{-1} is a map when α is not bijective.

1.17 Sets of Maps

We can form sets whose elements are maps. For example, suppose that A and B are sets, we might consider

$$\mathrm{map}_{A,B} = \{\alpha \mid \alpha : A \to B\}. \tag{1.23}$$

This set has various subsets, for example

$$\mathrm{inj}_{A,B} = \{\gamma \mid \gamma \in \mathrm{map}_{A,B},\ \gamma \text{ is injective }\}$$

and

$$\mathrm{sur}_{A,B} = \{\gamma \mid \gamma \in \mathrm{map}_{A,B},\ \gamma \text{ is surjective }\}.$$

We give the obvious meaning to $\mathrm{bij}_{A,B}$ and observe that

$$\mathrm{bij}_{A,B} = \mathrm{inj}_{A,B} \cap \mathrm{sur}_{A,B}.$$

Notice that $\mathrm{sur}_{A,B} = \emptyset$ if $|A| < |B|$, and that $\mathrm{inj}_{A,B} = \emptyset$ if $|A| > |B|$. This squares nicely with the fact that $\mathrm{bij}_{A,B} = \emptyset$ unless $|A| = |B|$.

1.18 Relations

We have studied two of the types of creature in the mathematical jungle, sets and functions. Next we study a third type, *relations*. Relations are things like $=$, $<$, $\subseteq$. That isn't a definition, it's a bit of waffle. When you want to discuss a function (perhaps f) you need to start off with two sets, the domain and the codomain (perhaps A and B respectively). Relations are a bit easier, you need only start with a single set (say S), and then you talk about *a relation* $\bowtie$ on S. The idea is that some elements of S may be related, but others may not. If s, $t \in S$ are related we write $s \bowtie t$. If not we write $s \not\bowtie t$.

An example of a relation is $<$ on the set $\mathbb{Z}$. Here we write $a < b$ exactly when $b - a \in \mathbb{N}$. Notice that $1 < 2$ but $2 \not< 1$.

Now, just as functions can have special properties which make them interesting and useful (injectivity, surjectivity, bijectivity) there are some fairly natural properties that a relation might have. A relation $\bowtie$ on a set S may or may not have some or all of the following properties.

(R) $\bowtie$ is *reflexive* if $s \bowtie s$ for every $s \in S$.

(S) $\bowtie$ is *symmetric* if whenever $s \bowtie t$, then $t \bowtie s$.

(T) $\bowtie$ is *transitive* if whenever both $s \bowtie t$ and $t \bowtie u$, then $s \bowtie u$.

An easy way to remember these names is to observe that we have listed the three properties in order of increasing complexity; the first involves only one element at a time, the second is about two elements and the third is about three. Simultaneously, the initial letters of the properties come successively in the alphabet. We look at some examples, some informal and some mathematical.

Let P be the set of all people and write $p \sharp q$ when $p, q \in P$ and p is an ancestor of q. This relation is neither reflexive nor symmetric, yet it is transitive.

Again let P be the set of all people and write $p \heartsuit q$ to mean that p loves q. Now, while it is certainly the case that $p \heartsuit p$ for some $p \in P$, it is definitely not the case for every $p \in P$, at least I hope not. Thus $\heartsuit$ is not reflexive. Similarly, $\heartsuit$ is not symmetric (a fact crucial to the lyricists of popular songs). Finally, $\heartsuit$ is not transitive (no doubt the reader can supply an example from personal experience).

Now for some mathematical examples. Consider the following relations on $\mathbb{N}$ with their usual meanings: $=, <, >, \leq$ and $\geq$. The first one satisfies all three conditions, but the other four do not. Please check to see which condition(s) each of them fails.

Let L be the set of (infinite) straight lines in the plane. The relation $\|$ satisfies all three conditions ($\theta \parallel \psi$ means that the lines θ and ψ are parallel).

Let T be the set of all triangles in the plane. The relation "is congruent to", is often written $\cong$ and satisfies all three conditions.

Inspired by these examples, I hope that you are thinking that *a relation satisfying reflexivity, symmetry and transitivity is really something like equality, subject to local conditions.*

Definition 1.6

A relation $\bowtie$ on a set S is called an *equivalence relation* exactly when it is simultaneously reflexive, symmetric and transitive.

Equivalence relations are incredibly important, and we will be studying them at greater length later (see Definition 1.10 and Proposition 1.4).

Definition 1.7

A relation $\bowtie$ on a set S is called antisymmetric if whenever $a, b \in S$ and $a \bowtie b$, then $b \not\bowtie a$.

An example of an antisymmetric relation is $<$ on $\mathbb{N}$. One could go on and on. If a relation occurs naturally in mathematics, you should look to see what (if any) special properties it has. If it has none, or at any rate you can't find any, then you are going to have a hard time proving much about it.

So far this Section has been a cheat, because we have not given a proper definition of a relation (just as Section 1.12 on maps began on shaky ground with all that persuasive but slightly doubtful stuff about *rules*). The formal definition may look a bit weird, but if you think about it, it is utterly clear.

Definition 1.8

A relation R on a set S is a subset of $S \times S$.

Instead of $(s, t) \in R$ we often write sRt. This is so-called *infix* notation where the relation is placed in between the elements.

This looks a little crazy at first. For the relation equality on the set of natural numbers we have that $=$ is $\{(a, a) \mid a \in \mathbb{N}\}$ and using infix notation we can write $a = b$ instead of $(a, b) \in \{(a, a) \mid a \in \mathbb{N}\}$. This may seem a bit unnatural (you probably didn't think of $=$ as a set before) but you have to admit it is crisp.

This also gives us a way of deciding if two relations are the same. It is clear. Two relations on a set S are the same exactly when they are the same subset of $S \times S$.

Definition 1.9

Let S be a set. A *partition* of S is a collection C of subsets of S with the following three properties.

(a) $X \neq \emptyset$ whenever $X \in C$.

(b) If $X, Y \in C$ and $X \neq Y$, then $X \cap Y = \emptyset$.

(c) The union of all the elements of the partition is S; in symbols we have

$$\bigcup_{X \in C} X = S.$$

Thus a partition of S is a collection C of non-empty subsets of S which are pairwise disjoint (i.e. any two distinct ones are disjoint), and the union of all these subsets of S is S. In crude terms, you have taken the elements of S and sorted them into various non-overlapping subsets. Gender partitions humanity. The infinite plane can be partitioned into infinitely many parallel lines.

This is a very important idea, and equivalence relations are the technical gadgets which enable us to reason about partitions.

Definition 1.10

Suppose that $\sim$ is an equivalence relation on a set S. For each $x \in S$ let

$$[x] = \{y \mid y \in S,\ x \sim y\} \subseteq S.$$

These subsets $[x]$ of S are called ***equivalence classes***, and $[x]$ is called the equivalence class of x.

Here is the result which ties things together.

Proposition 1.4

Using the notation of Definition 1.10, the sets $[x]$ (as x varies over S) form a partition of S.

Proof (tough but worth it)

We need to check that the three conditions of Definition 1.9 are all satisfied.

(a) Consider an arbitrary $[y]$ for $y \in S$. Now $y \sim y$ so $y \in [y]$ and therefore $[y] \neq \emptyset$.

(b) We twist things round. It suffices to show that if $[y] \cap [z] \neq \emptyset$, then $[y] = [z]$. We will assume $[y] \cap [z] \neq \emptyset$, and deduce that $[z] \subseteq [y]$.

Since $[y] \cap [z]$ is non-empty we may choose $x \in [y] \cap [z]$. Now $y \sim x$ and $z \sim x$. Use the symmetric law to spin the second of these facts so $x \sim z$. Now we know both $y \sim x$ and $x \sim z$ so by the transitive law we can deduce that $y \sim z$.

Now choose an arbitrary element $a \in [z]$. Now $y \sim z$ and $z \sim a$ so by transitivity we have $y \sim a$. Thus $a \in [y]$. This shows that $[z] \subseteq [y]$.

An entirely similar argument yields the reverse inclusion, so $[z] = [y]$.

(c) The union of subsets of S is a subset of S. Thus all we need to show is that if $x \in S$, then there is an equivalence class which contains the element x. However, $x \sim x$ by the reflexive law, so $x \in [x]$ and we are done.

□

In fact one can reverse the argument of Proposition 1.4 and show that a partition C of a set S gives rise to an equivalence relation of S. Just put $x \sim y$ exactly when there is $X \in C$ (i.e. a set in the given partition of S) with the property that both $x \in X$ and $y \in X$. The diligent reader will check that this is an equivalence relation on S. The *very* diligent reader will do a little more. We now have two procedures: one is a way of making a partition from an equivalence relation, the other is a way of making an equivalence relation from a partition. The enthusiast should show that these two procedures are mutually inverse. That is to say, if you do one, and then the other, make sure that you end up where you started. This is really two problems, because you can start with either procedure!

Finally, when you have a partition of a set S, it is often useful to have one representative from each set comprising the partition. A set of these representatives is called a *transversal* for the partition. If humanity is partitioned by gender, any set consisting of one male and one female is a transversal for the partition. We view a plane as a set, the elements of which are the geometric points comprising the plane. The plane can be partitioned into parallel straight lines, and then any straight line which is skew to the lines in the partition will suffice as a transversal. Such a skew straight line will intersect each of the family of parallel lines in exactly one point, as required. One could think of crazy transversals, where a point is selected from each of the family of parallel lines in some fashion, but our transversal has the virtue of being geometrically pleasant.

EXERCISES

1.16 Let $A = \{1, 2, \ldots, n\}$.

(a) How many relations are there on the set A?

(b) How many reflexive relations are there on the set A?

(c) How many symmetric relations are there on the set A?

(d) How many relations are there on the set A which are both reflexive and symmetric?

1.17 Let $F = \mathbb{Z} \times \mathbb{N}$.

(a) Define a relation $\sim$ on F by $(a,b) \sim (c,d)$ if and only if $ad - bc = 0$. Show that $\sim$ is an equivalence relation.

(b) Suppose that $(a_1, b_1) \sim (a_2, b_2)$ and $(c_1, d_1) \sim (c_2, d_2)$. Prove that $(a_1 d_1 + c_1 b_1, b_1 d_1) \sim (a_2 d_2 + c_2 b_2, b_2 d_2)$ and moreover that $(a_1 c_1, b_1 d_1) \sim (a_2 c_2, b_2 d_2)$.

(c) What might a rational person make of this construction?

1.18 Exhibit relations on $\mathbb{Z}$ which demonstrate all 2^3 possibilities of being reflexive or not, symmetric or not, and transitive or not.

1.19 *What, if anything, is wrong with this "proof"?*
Theorem: If a relation $\sim$ on a set A is both symmetric and transitive, then it must also be reflexive.
Proof: Suppose that $a \in A$. Choose any $b \in A$ such that $a \sim b$. Use the symmetric property to deduce that $b \sim a$. Now both $a \sim b$ and $b \sim a$ so by transitivity we deduce that $a \sim a$. However, a was an arbitrary element of A so the relation must be reflexive.

1.20 Suppose that R is a relation on the set of real numbers $\mathbb{R}$. We write $a \sim b$ if and only if $(a,b) \in R$. Consider the set R; now $R \subseteq \mathbb{R}^2$ so we may give a geometric interpretation of R as a subset of the plane. What can you say about the subset R of the plane in the event that

(a) R is reflexive,

(b) R is symmetric,

(c) R is both reflexive and symmetric?

1.21 (a) How many equivalence relations are there on a set of size 3?

(b) How many equivalence relations are there on a set of size 4?

1.22 (a) Let $f : A \to B$ be a function. Define a relation $\sim$ on A via $x \sim y$ if and only if $f(x) = f(y)$. Determine (with justification) whether or not $\sim$ is an equivalence relation.

(b) We define a relation on $\mathbb{R}^2$ $(= \mathbb{R} \times \mathbb{R})$ by $(a,b) \sim (c,d)$ if and only if $(c-a, d-b) \in \mathbb{Z}^2$. Prove that $\sim$ is an equivalence relation. Identifying $\mathbb{R}^2$ with the plane in the usual way, describe the most natural transversal for $\sim$ which you can find. What, if anything, has this question to do with doughnuts?

(c) We define a relation on $\mathbb{R}^3 \setminus \{(0,0,0)\}$ by $(a,b,c) \sim (d,e,f)$ if and only if there is $\lambda \in \mathbb{R}$ with $\lambda > 0$ and $a = \lambda d$, $b = \lambda e$ and

$c = \lambda f$. Show that $\sim$ is an equivalence relation. Identifying $\mathbb{R}^3$ with three-dimensional space in the usual way, find a transversal for this equivalence relation which is geometrically pleasant.

1.19 Intervals

Definition 1.11

A *real interval* (or just an *interval*) is a subset I of the real numbers with the property that if $x \in \mathbb{R}$ and $a, b \in I$ are such that $a < x < b$, then $x \in I$.

We establish some notation. Suppose that $a, b \in \mathbb{R}$.

Definition 1.12

(a) The *open interval* $(a, b) = \{x \mid x \in \mathbb{R},\ a < x < b\}$.

(b) The *closed interval* $[a, b] = \{x \mid x \in \mathbb{R},\ a \leq x \leq b\}$.

(c) The *half open interval* $[a, b) = \{x \mid x \in \mathbb{R},\ a \leq x < b\}$.

(d) The *half open interval* $(a, b] = \{x \mid x \in \mathbb{R},\ a < x \leq b\}$.

Of course people with a less sunny disposition are welcome to refer to the intervals (c) and (d) as being half closed. There are language vandals who even call these intervals *clopen*. Please don't.

You can introduce artificial symbols ∞ and $-\infty$ deemed to be respectively greater than and less than all real numbers. Thus the positive reals can be written as $(0, \infty)$, the non-positive reals as $(-\infty, 0]$, the non-negative reals as $[0, \infty)$ and $\mathbb{R}$ as $(-\infty, \infty)$. Some people actually adjoin ∞ and $-\infty$ to $\mathbb{R}$ to form an extended real number system. This sort of behaviour is fairly common among geometers, topologists and the like, but you would be unlikely to catch an algebraist doing it (they are very sensitive about $0 \times \infty$).

Of course, there is a *confusion opportunity* in the notation for intervals. Open intervals (a, b) look just like elements of $\mathbb{R}^2$. Be careful!

EXERCISES

1.23 (a) Show that the empty set is an interval.

(b) Show that $\{1\}$ is a closed interval.

(c) Show that $\{1\}$ is not an open interval.

1.24 (a) Prove that the intersection of two intervals is an interval.

(b) Prove that the intersection of two open intervals is an open interval.

(c) Prove that the intersection of two closed intervals is a closed interval.

1.25 [Harder] In this question, union and intersection are not necessarily of two sets, but rather of arbitrary collections of sets.

(a) Show that the closed interval $[0, 1]$ is expressible as the intersection of infinitely many open intervals.

(b) Can the closed interval $[0, 1]$ be expressed as the union of infinitely many open intervals? Justify your answer.

1.26 *Once upon a time a six-year-old child sat in class while the teacher explained that if you took a ruler, and cut it in two equal pieces, each part would be the same length. When the lesson was over, the child went to see the teacher, to explain that you couldn't cut a piece of wood into two equal pieces, because the middle point would have to be attached to one piece but not the other. Then one broken piece would have two ends but the other would only have one because its end was missing. The teacher was very patient, and tried to explain again.* Discuss this using the interval notation.

2
Proof

This chapter is deliberately short, and contains no exercises. This is because it briefly outlines techniques which will be used throughout the rest of the book.

2.1 Induction

We have already met the set of natural numbers $\mathbb{N} = \{1, 2, 3, \ldots\}$, and raised the vexed question of the dot-dot-dots. We all carry around a mental image of these numbers going on for ever – there being no last natural number. If you want to play a game of "who can think of the biggest number", the person who goes last will always win. There is a simple procedure for the last player to win this game – just add one to the largest number mentioned by any other player.

Such a strategy will work until you play the game against some clown who says "banana" (or more likely "infinity"). This person may assert that banana is a natural number bigger than all the other natural numbers. There are two obvious responses to this:

(i) *Aggressive and pompous:* you lie. There is no such thing. Go and study some lesser subject.

(ii) *Awkward customer:* you are wrong there. The natural number orange is even bigger than banana.

As we mentioned before, mathematicians abhor ambiguity. They have decided to agree upon two things. First that there is no largest natural number, and

G. Smith, *Introductory Mathematics: Algebra and Analysis*

second that every natural number other than 1 is the successor of exactly one other natural number as you count $1, 2, 3, 4, 5, 6, \ldots$, and that 1 is not the successor of a natural number.

This does two things. It ensures that the set $\mathbb{N}$ is infinite, and it also guarantees that if anyone mentions a natural number, we can count, starting at 1, and reach that natural number after a finite amount of time.

This is neat in two ways. First, it clears up any doubt as to what is and what is not a natural number. Second, it gives us a way of proving things about the natural numbers. The mathematical way of crystallizing this discussion is to say that we have adopted the *axiom of mathematical induction*. This axiom gives us a way of proving things about properties of the natural numbers. We do not have to go through a case-by-case analysis; we build a sort of mathematical "proving machine" which does all the work for us. An axiom is something we do not have to prove. The axiom of mathematical induction is part of the definition of the natural numbers.

When we spot a pattern in mathematics, we cannot simply say – "that is a *law of mathematics*". Perhaps the pattern is more complex than we at first suppose. The axiom of mathematical induction gives us a way of taking a pattern that we think describes something in mathematics, and then proving that we are right. This method is not available in science, and that is why scientific knowledge has a different status. Scientific knowledge is provisional; mathematical knowledge is not. The only possibility for error in mathematics is human fallibility in the course of calculation or proof. For this reason, to keep mathematics as certain as possible, great attention is paid to standards of proof. Enough of this discussion – we now investigate the method of mathematical induction.

Axiom of (Simple) Induction

Let $P(n)$ be a proposition about the natural number n. Suppose that we can show two things:

(i) $P(1)$ is true.

(ii) For each $r \in \mathbb{N}$, whenever $P(r)$ is true, then $P(r+1)$ is true.

We can then deduce $P(n)$ is true for each natural number n.

The proposition $P(n)$ is known in the trade as the *inductive hypothesis*. Let us immediately see this axiom in action.

Proposition 2.1

For each natural number n, the sum $1 + 2 + \cdots + n$ $(= \sum_{i=1}^{n} i)$ is equal to $n(n+1)/2$.

Proof

$P(n)$ is the proposition that the particular natural number n has the property that $1 + \cdots + n = n(n+1)/2$.

(i) $P(1)$ asserts that the sum of the natural numbers, starting at 1 and going up to 1 is 1. This is true (starting the induction is almost always easy, but you must never forget to do this part).

(ii) Suppose that for some natural number r, $P(r)$ is true. We must show that it follows that $P(r+1)$ is true. Now,

$$1 + 2 + \cdots + r + (r+1) = (1 + 2 + \cdots + r) + (r+1).$$

We know that $1 + 2 + \cdots + r = r(r+1)/2$ so

$$1 + 2 + \cdots + (r+1) = r(r+1)/2 + (r+1) = (r+1)(r+2)/2.$$

which asserts that proposition $P(r+1)$ is true.

By the axiom of mathematical induction the result is proved.

□

That was all rather formal. As we look at more examples, the expositions of the proofs will gradually become less stiff, but will still (one hopes) be clear. Some readers will grasp the idea very quickly, but others may feel that we are in some way cheating – "you are assuming the answer to prove the result" is a common *cri de coeur*. This complaint is not justified, but it easy to see why it happens. We have used an auxiliary symbol r in the course of the above proof. In real life, many writers do not bother to do this. They use n. The point is that the statement is to be proved for all n, but the "counter" r (sometimes confusingly called n) is a particular natural number. If one uses n as a counter, the first line of (ii) becomes "Suppose that for some natural number n, $P(n)$ is true". Perhaps that is still not too bad, because the emphasis is still that for a *particular* number n the proposition $P(n)$ holds. However, someone who "knows what they are doing" might easily write the first line of (ii) as "Suppose that $P(n)$ is true". That is the sort of thing which can cause panic.

If you are completely at home with induction arguments, such confusing casual phrasing is (alas) acceptable. Too many mathematicians do it for there to be any hope of change in the near future. Nonetheless, this is sloppy practice, and an auxiliary variable such as r is a relatively small price to pay for clarity. It all depends on the target readership of course. Writing for experienced mathematicians one can afford to be very loose. One often sees statements in the literature such as "this proposition follows by induction", and the argument, being routine, is completely omitted. A sophisticated reader will then mentally check that the argument is genuinely routine in her head. On the other hand, undergraduates write most of their mathematics to convince their tutors that they (the undergraduates) understand what is going on. If you are at the stage of your career when the tutor might have grounds to doubt that you completely understand induction, you must write out the argument in all its gory detail.

There is another point worth making. In order to prove Proposition 2.1 we somehow had to have prior knowledge that the appropriate formula was $n(n+1)/2$. It would be pretty easy to guess that formula by doing a few experiments. It might not be so easy to spot the formula for the sum of the first n perfect squares – that is $n(n+1)(2n+1)/6$ by the way. Induction does have this limitation – you must intelligently guess what you should try to prove.

Let us now do another example. We discussed the power set of a set in Section 1.8. Recall that if A is a set we let $P(A) = \{x \mid x \subseteq A\}$ and call this the power set of A. Look back and see that when A has three elements then the power set of A has eight elements. The general formula is easy. If A is a finite set with n elements, then $P(A)$ has 2^n elements. You can check the plausibility of this assertion by looking at a few examples. Evidence, however, is not proof. We shall now banish doubts.

Proposition 2.2

If A is a non-empty finite set of cardinality n, then $|P(A)| = 2^n$.

Proof

For each natural number r, let $Q(r)$ denote the statement that the result holds for sets of cardinality r.

(i) The statement clearly holds for a set with one element.

(ii) We deem the statement to hold for the natural number r (this is the *inductive hypothesis*). Suppose that $|A| = r + 1$. We seek to show that $|P(A)| = 2^{r+1}$. Select $B \subseteq A$ such that $|B| = r$ so $A = B \cup \{a\}$ and $a \notin B$. We now count the subsets of A. The number of subsets of A which do not contain a is exactly the same as the number of subsets of B. We know this number is 2^r (by inductive hypothesis). Any other subset of A must contain the element a, and will be expressible as $S \cup \{a\}$ where $S \subset B$. Conversely, any subset S of B will determine a subset of A containing a via $S \cup \{a\}$. Thus the number of subsets of A which contain a is the same as the number of subsets of B, i.e. 2^r.

Now the number of subsets of A is therefore $2^r + 2^r = 2^{r+1}$. This is precisely the content of $Q(r+1)$ so we are finished by induction.

□

Notice that this proof has been finished with the words "by induction". This is a very acceptable variant of the more formal "by the axiom of mathematical induction". This formula for the size of a power set even holds when A is empty. We could easily have incorporated this into our statement and proof by allowing 0 to be a natural number. It is wise to be flexible in these things.

Another way of thinking about induction is by analogy with an infinite row of standing dominoes. You show it is possible to knock the first one over, and you also demonstrate that if the r^{th} domino falls over, then it will collide with domino $r+1$ and knock that over too. If you succeed in demonstrating these facts, then you know you can knock down the entire row with a flick of a finger.

2.2 Complete Induction

A great fuss is often made about something called *complete induction*, and the difference between complete induction and *simple* induction – which is what we have been looking at just now. There is nothing substantial going on here. If you can prove a proposition by one method then you can prove it by the other. Though it does not merit such exalted status, we will treat it as an axiom (reluctantly).

Axiom of Complete Induction

Let $Q(n)$ be a proposition about the natural number n. Suppose that we can show two things:

(i) $Q(1)$ is true.

(ii) Whenever $Q(x)$ is true for all natural numbers x less than r, then $Q(r)$ is true.

We can deduce that $Q(n)$ is true for each natural number n. In fact, and whisper this softly, part (i) is not strictly necessary. This is because applying part (ii) when $r = 1$ (and a little vacuous reasoning) gives a proof that $Q(1)$ holds. However, a beginner would be well advised not to rely on vacuous reasoning, and checking that $Q(1)$ holds is unlikely to cause much pain.

Before explaining why this is not a genuinely new axiom, we shall demonstrate the axiom of complete induction in action.

We all know what prime numbers are. They are the set of natural numbers which have no proper factors.

$$\mathfrak{P} = \{2, 3, 5, 7, 11, 13, 17, 19, 23, 29, 31, 37, 41, \ldots\}.$$

For good reasons, we place a special ban on the number 1. The number 1 is not a prime. By the way, that letter is a Gothic P; astonishingly enough, there is no standard notation for the set of all prime numbers, so we create our own.

Proposition 2.3

Each natural number n is the product of prime numbers. (We deem 1 to be the product of no prime numbers!)

Proof

We prove this by complete induction. (We are getting more relaxed in our levels of formality. The slippers are now on.)

(i) The induction starts using the special status of 1.

(ii) Assume that we have proved the proposition in the cases $x < r$. Now consider the case $x = r > 1$. If $r \in \mathfrak{P}$, then the proposition holds. If, on the other hand, $r \notin \mathfrak{P}$, then $\exists u, v \in \mathbb{N}$, $u \neq 1 \neq v$ such that $r = u \cdot v$. Now $u, v < r$ so (complete induction rearing its head here) each of u and v is a product of primes (because z is a product of primes $\forall z \in \mathbb{N}$ with $z < r$). However $r = u \cdot v$ so r is a product of prime numbers. By (complete) induction, the proposition is proved.

□

Let us now tackle something a little more ambitious. We asserted in Section 1.5 that in the presence of the associative law we could dispense with brackets. We shall now prove this. A *binary operation* on a set A is just a way of multiplying elements of A together to yield elements of A. The word "binary" is used because you are multiplying *two* things together. A more sophisticated point of view is that a binary operation on A is a map from $A \times A$ to A. For economy, we will write the product if $x, y \in A$ as $x \cdot y$. This abstract setting includes the integers under addition or multiplication, the collection of all subsets of the set U under either union or intersection, and much else besides.

Proposition 2.4

Suppose that the set A is equipped with an associative binary operation denoted by a dot. If $x_1, x_2, \ldots, x_n \in A$, then the value of the product

$$x_1 \cdot x_2 \cdot x_3 \cdot \cdots \cdot x_n \tag{2.1}$$

is independent of the order in which the multiplications (operations) are performed (i.e. how we bracket this expression).

In such a bracketed expression, the last multiplication to be performed must be indicated by a dot sitting between x_i and x_{i+1} for a particular value of i. We say that the expression ***breaks*** at i. For example: $((x_1 \cdot x_2) \cdot x_3) \cdot (x_4 \cdot x_5))$ breaks at 3. We have not yet begun the proof of Proposition 2.4. If we had done so, the word "proof" would have occurred on the left-hand side of the page. We are still just having a chat. When a theorem or proposition needs a long or complicated argument to prove, it is a form of cruelty simply to write the thing out in one go. The poor reader is faced with the prospect of wading through the entire argument, hoping against hope that life's little intrusions (coffee is ready, building is on fire, etc.) will not interrupt her train of thought two lines from the end.

It is also a foolish thing to do from the point of view of the writer. If the proof is in one piece, it may be difficult to check that it is correct. If you do find a flaw, then it may be difficult to see what can be salvaged from the proof. For these reasons, it has become regarded as very classy to break up proofs into mind-sized pieces which take little time to read or check, and which stand alone as true propositions. These minitheorems are often not really of any interest in their own right, but only insofar as they contribute to the proof of the theorem. For this reason we do not dignify them with the hallowed title *theorem* (usually reserved for exceptionally important results), or even *proposition* (which is used for a result of interest in its own right). Instead we call them *lemmas*. Nonetheless, a good lemma may be re-used, and acquire fame and stature in its own right; lifted, as it were, from the chorus line to the limelight. You may sometimes see a phrase like "by Bobker's lemma, blah-blah is true". You are clearly supposed to know Bobker's lemma. Bobker's lemma is an internationally famous result, and only its name betrays its humble origins.

Either Bobker didn't realize how good the result was when he proved it, or more likely, he was being modest.

In many respects a lemma is rather like a *subroutine* or *function* in computer science.

Now we address ourselves to the proof of Proposition 2.4. It has been a fair time since we stated it, so now is a good time to read the statement again – and to refresh your memory about the notation "break".

As the first stage of our proof, we obtain a lemma (and we will deliberately be a little sloppy in the course of the argument – if you can still follow it easily, all is well).

Lemma 2.1

Any product of elements of A,

$$x_1 \cdot x_2 \cdot \cdots \cdot x_n, \tag{2.2}$$

no matter how it is bracketed, is equal to a product which breaks at 1.

Proof

Let the value of the product (2.2) be X. We proceed by induction on r, where the original product (2.1) breaks at r. If $r = 1$ we are done already. If $r > 1$, then we suppose that our product is $p.q$ where x_r is the last letter of p and x_{r+1} is the first letter of q, and both p and q are products in their own right. Now p breaks at j and of course $j < r$ so p is equal to a product breaking at 1 (by complete induction) so $X = (x_1 \cdot s) \cdot q$; we can now deploy the associative law to obtain $X = x_1 \cdot (s \cdot q)$, and the product on the right breaks at 1 as required.

□

Proof (of Proposition 2.4)

By Lemma 2.1

$$X = x_1 \cdot p \text{ where } p \text{ is a product in its own right.} \tag{2.3}$$

We now prove (by induction on n) that any product X may be bracketed to the right; that is $X = x_1(x_2(\ldots(x_n)\ldots))$.

If $n = 1$ (or 2) the result is trivially true. Suppose that the proposition is true for $n = r$, and consider the case $n = r + 1$. By Equation (2.3) $X = x_1 \cdot p$ and p contains r terms so the result holds for p. Therefore the result holds for X and we are done.

Finally we observe that any X can be rebracketed to the right without changing the value of the product. The value of X is therefore independent of the bracketing.

□

We asserted earlier that there was not really much difference between (simple) induction and complete induction. We now justify this remark. Suppose that you have a proof a proposition $P(n)$ by simple induction. You can turn it into a proof by complete induction by replacing the final phrase "by induction the proof is complete" with the phrase "by complete induction the proof is complete". The proof will remain valid. You are guilty of overkill of course.

The other way round requires a ruse. Suppose that you have a proposition $P(n)$ which you prove true for all $n \in \mathbb{N}$ by complete induction. Consider the proposition $Q(n)$.

$$Q(n) : \text{“The proposition } P(x) \text{ is true } \forall x \in \mathbb{N},\ 1 \leq x \leq n\text{”}$$

Your proof of "$P(n)$ is true $\forall n \in \mathbb{N}$" will become a proof that "$Q(n)$ is true $\forall n \in \mathbb{N}$" by replacing $P(n)$ by $Q(n)$ throughout the proof, and deleting the word "complete" before all references to induction. Now if $Q(n)$ is true for all natural numbers n, then certainly so too is $P(n)$.

2.3 Counter-examples and Contradictions

Suppose that we consider the proposition: "Every natural number is the sum of three perfect squares". You can mentally check a few instances:

$$\begin{aligned} 1 &= 1^2 + 0^2 + 0^2 \\ 2 &= 1^2 + 1^2 + 0^2 \\ 3 &= 1^2 + 1^2 + 1^2 \\ 4 &= 2^2 + 0^2 + 0^2 \end{aligned}$$

Everything seems to be going well. An optimist might stop looking at examples and try to prove the proposition. This would be a waste of time, because, as you may check, 7 is not the sum of three perfect squares. We say that 7 is a *counter-example* to the proposition. Counter-examples are wonderful things. They are sometimes easy to find, so can take much less effort than using intelligence, ingenuity and skill to prove things true. Suppose that someone claims that he or she has proved a marvellous theorem – and has a beautiful 250-page proof (this happens in real life). As a mathematician you have a duty to establish whether or not the proof is correct. If the proposition is true, then you may have a lot of work on your hands – checking every line of the proof. If however, the proposition is false, you can avoid finding the flaw in the proof. Even if you did find the flaw, it would only establish that the *proof* was wrong. You would still wonder if it might be the case that the theorem was correct, but that a different proof was needed. Much easier is to find a counter-example to the alleged theorem. This is usually fast, it increases the sum of human knowledge (this proposition is false), and shows that there must exist a fault in the proof being offered. There is no need to find the flaw. The author will usually run away and find it for himself.

This very efficient, but smacks of intellectual vandalism. However, there is a very constructive use for the *idea* of a counter-example.

There is another method of proof which is actually equivalent to induction, but sometimes yields a slicker proof; that is *proof by minimal counter-example*. Suppose that you wish to prove that some proposition $P(n)$ is true for all natural numbers n. What you do is to use a "contradiction" to accomplish the proof. You suppose (for the moment, and hoping to be shown that you are wrong) that there exists a natural number m for which $P(m)$ is false. There must therefore exist a smallest natural number r such that $P(r)$ is false. You then trade off the consequences of $P(r)$ being false but $P(x)$ being true whenever x is a natural number smaller than r. If you manage to show that this situation is impossible, you can deduce that there is no smallest natural number r for which $P(r)$ is false, and so that there are no natural numbers m for which $P(m)$ is false. Thus $P(n)$ is true for all natural numbers n.

Let us see this in action. We shall reprove Proposition 2.3.

Proposition 2.5

This is just Proposition 2.3 (revisited). Each natural number n is the product of prime numbers. (We deem 1 to be the product of no prime numbers!)

Proof

Suppose (for contradiction) that the proposition is false, and let x be a minimal counter-example. Now, $x \notin \mathfrak{P}$ (otherwise it would not be a counter-example), and of course $x > 1$. Thus $x = y \cdot z$ for $y, z \in \mathbb{N}$ and $y, z < x$. Now, y and z are products of primes (because x is the *smallest* counter-example). Thus x is a product of primes. However, x is not a product of primes by hypothesis. We have a contradiction. Therefore x does not exist and the proposition is proved.

□

One must admit that the argument is very similar to proof by induction. Some mathematicians do not like proofs by contradiction. Reasoning about a non-existent counter-example gives them the philosophical wobblies. If you are one of these people, you will often find that you can reformulate a "contradiction" argument in a more straightforward way.

The following challenge was set as problem 6 of the finals of the International Mathematical Olympiad 1988 held in Canberra, Australia. This competition can be thought of as the world championship finals for secondary school mathematicians. This was certainly the most difficult question in the competition, only 11 out of 268 competitors managed it completely. The solution that follows was not the one that the organizers were expecting, but was found under exam conditions by a Bulgarian competitor named Emanuel Atanasov. In order to appreciate the magnitude of this achievement, you must realize that there was no hint that this question was amenable to attack by induction or minimal counter-example. It could have been complex numbers, or geometry, or indeed any type of mathematics that was appropriate. This competitor both realized what method would be likely to succeed, and produced an argument half the length of the "official" answer. You may care to cover up the proof and have a go yourself. You have the advantage over Atanasov that you know which type of argument is necessary. If you do find a solution, see if it is as elegant as the following *tour de force*.

Problem 2.1

Suppose that $(a^2+b^2)/(1+ab) = t \in \mathbb{N}$ for some natural numbers a and b. Show that t is a perfect square.

Solution 2.1

We fix t and choose b minimal and $a \geq b > 0$ such that the equation is satisfied. Thus a is a root of the quadratic polynomial

$$x^2 - tbx + b^2 - t. \qquad (2.4)$$

Let the other root be c. Now, $a + c = tb$ so $c \in \mathbb{Z}$. Substitute c into Equation (2.4) and rearrange to obtain $t(1+bc) = c^2+b^2 > 0$ so $bc > -1$ and $c \geq 0$. The product of the roots of our quadratic is $ac = b^2 - t < b^2$ so $b > c \geq 0$.

Thus $(b^2 + c^2)/(1 + bc)$ violates the minimality of the choice of b unless $c = 0$. Thus 0 is a root of Equation (2.4) and so $t = b^2$ is a perfect square.

To phrase this argument in terms of *minimal counter-example* one observes that we could have said – choose $b \in \mathbb{N}$ minimal so that $a \geq$

$b > 0$ and $(a^2 + b^2)/(1 + ab)$ is a natural number which is not a square. The above proof then produces the necessary contradiction.

2.4 Method of Descent

The *method of descent* is closely associated with the mathematician Fermat. You may well have heard of him because of the celebrated result known as Fermat's last theorem. This is a misnomer of course, since, no proof was found *and published* until 1995.

Theorem 2.1 (Fermat's Last Theorem (Wiles and Taylor–Wiles))

There do not exist $x, y, z, n \in \mathbb{N}$, with $n \geq 3$ such that $x^n + y^n = z^n$.

Fermat claimed to have found a proof, but no trace of it remains, and no-one else produced valid proof until 1995. It is just possible that Fermat did have a proof, but the balance of mathematical opinion (for what that is worth) is that he had a flawed argument. It is of course possible that Fermat, who was not stupid, had an idea which no-one has had since. It is not conceivable that Fermat had a proof which was anything like the modern argument – relying as it does on sophisticated technical machinery. It would be rather like speculating that Archimedes had secretly invented the laser.

There is a special term for a result that we believe to be true, but for which we cannot find a proof. Such propositions are called *conjectures* – we have already used the term once in the previous paragraph. Some conjectures have been standing open for hundreds of years – as Fermat's conjecture did. Others have a much shorter life (a couple of seconds is the average, and fortunately most of them don't reach as far as the vocal chords).

Flavour A

Suppose that you want to prove that a proposition $P(n)$ is true for all natural numbers n. You suppose (for contradiction) that there exists $k \in \mathbb{N}$, $k > 1$ such that $P(k)$ is false. You then show that it follows that there must be some $k' \in \mathbb{N}$, $k' < k$ such that $P(k')$ is false. By repeated application of your argument, it follows that $P(1)$ must be false. You now examine $P(1)$. If $P(1)$ is true, you have a contradiction, and $P(n)$ must be true for all $n \in \mathbb{N}$.

Flavour A is equivalent to proof by minimal counter-example. Flavour B is in some ways more exciting.

Flavour B

Suppose that you have propositions $P(n)$ concerning each natural number n, and that you want to show that $P(1)$ is true. First, find $t \in \mathbb{N}$ such that $P(t)$ is true. Second, you find a proof that if for some $s \in \mathbb{N}$, $s > 1$, $P(s)$ is true then $\exists s' \in \mathbb{N}$, $s' < s$ such that $P(s')$ is also true. By repeated application of your argument you can conclude that $P(1)$ is true.

We asserted in Section 1.3 that the number $\sqrt{2}$ was not rational. We shall demonstrate this fact by using Flavour A of the method of descent.

Proposition 2.6 (Pythagoras)

The number $\sqrt{2}$ is not rational – i.e. there do not exist natural numbers a and b such that $a^2 = 2b^2$.

Proof

Suppose that $a, b \in \mathbb{N}$ and $a^2 = 2b^2$. Now a^2 is even, so a is even. Let $a = 2b'$ where $b' \in \mathbb{N}$. Thus $4b'^2 = 2b^2$ so $b^2 = 2b'^2$. Let $a' = b$, then $a'^2 = 2b'^2$ and $1 \leq a' < a$. By the method of descent (applied to a) there must exist $d \in \mathbb{N}$ such that $1^2 = 2d^2$. This is nonsense, so the natural numbers a and b could not have existed in the first place, so $\sqrt{2}$ is irrational.

□

If you want to see how the proof of Proposition 2.6 ties in with the description of Flavour A, consider the following proposition:

$$P(n): \text{ For } n \in \mathbb{N} \ \nexists m \in \mathbb{N} \text{ such that } n^2 = 2m^2.$$

We illustrate Flavour B by giving a sketch proof of a famous result of Fermat. All the technical details are omitted so that you can concentrate on the structure of the proof. This result is certainly striking enough to be called a "theorem".

Theorem 2.2 (Fermat – two squares)

Let p be a prime number, and suppose that p leaves remainder 1 when divided by 4, then $p = x^2 + y^2$ for some natural numbers x and y.

Proof (sketch)

First find $n, a, b \in \mathbb{N}$ such that $np = a^2 + b^2$ (a proof that this can always be done is not deep – though it would be a remarkable first-year undergraduate who could discover it unaided). If $n = 1$ you have finished, so assume $n > 1$. Next, show that you can find $n', a', b' \in \mathbb{N}$, $n' < n$ such that $n'.p = (a')^2 + (b')^2$.

Fermat's method of descent yields a proof that the required natural numbers x and y do exist.

□

As before, we point out the relevant family of propositions which ties the proof of Theorem 2.2 to the method Flavour B:

$$P(n) : \exists x, y \in \mathbb{N} \text{ such that } np = x^2 + y^2$$

The remarkable thing about Fermat's method of descent (Flavour B) is that it is often *constructive*. It certainly is in this instance. Once you have got your hands on a, b and n, there is a recipe for producing a', b' and n' using simple formulas. You repeat the procedure to actually construct x and y. The sensible way to do this sort of thing is on a computer of course. We not only know that x and y exist, we can actually find them. In the case of Fermat's two squares theorem, this is not that wonderful. A crude computer search will find x and y. The theorem simply guarantees that the search will succeed. Nonetheless, the proof method does actually find x and y by successively "improving" a first guess at a solution – a and b. This illustrates a technique of which computer scientists are very fond – for they usually need to be able to construct solutions.

Some mathematicians (especially of the computing variety) do not really care for proofs which are not constructive. They take the view that if you cannot actually "calculate an answer", then you are whistling in the wind.

Other mathematicians (especially of the analysis variety – *analysis* is calculus made interesting) disagree completely. They often take the view that a proof of the existence of something is quite sufficient, thank you very much. Only a grubby calculator would want to actually lay her hands on the relevant numbers.

Fortunately, the planet is big enough for both views.

2.5 Style

In the course of this book, we repeatedly give proofs of results. Perhaps the time has come to discuss what *proof* is about. A mathematical proof is an utterly convincing argument that some proposition is true. An argument which is merely persuasive or fairly convincing simply will not do. It should not only be logically perfect – it should also be clear.

Clarity is a matter of style. It takes time and trouble to learn how to write mathematics beautifully. Remember that you are writing for other people, and that your work will be enjoyed only if your exposition gives pleasure to the reader. It is perfectly possible to write correct mathematics in an ugly way, but it is horribly antisocial. The "other" person most likely to read your work is yourself, but not your current self. It is yourself six months or a year older. By that time you will most likely have forgotten details which are obvious to you now. In this respect you will be a different person. Write for that person now, and you will be grateful to yourself later. The same remarks apply to computer programming. Comment the code!

First state exactly what you are trying to prove. Give your statement a label to distinguish it from other material – a label such as *Theorem, Lemma, Result* or *Proposition*. When writing casually, a good label is *Claim*. If you are not sure that the result is true, then call it a *Conjecture*. Thus, on the back of the proverbial envelope, you might write:

Conjecture 2.1

If n is a natural number, then there is a prime number p such that $n \leq p \leq 2n$.

In fact this conjecture is true. After stating the proposition you wish to prove, write the word *Proof* boldly. Then make your utterly convincing submission, and finish it off with an end-of-proof symbol – anything that clearly marks the end of the proof will do. Some people use a couple of slashes, and others prefer Q.E.D. (*Quod Erat Demonstrandum*).

In the course of your proof, decorate the lines of mathematics with well-chosen English words and phrases to render the meaning as transparent as possible. Useful pointers include *but, conversely, it follows that, hence, however, since, so, then, therefore, we see that, whence, whereas* and many other friends and relatives.

2.6 Implication

Some arguments or proofs take a linear form – you deduce each line from its immediate predecessor. A legitimate (though very dull) tactic is to prepend to each line the word "therefore" or its symbol ($\therefore$), or alternatively append to each line the slightly dilatory "since" or its symbol ($\because$). Try to introduce a little variety into your exposition – pepper the page with appropriate conjunctions.

The word "implies" is very widely misused – even by mathematics undergraduates. If you cannot use it selectively and correctly, perhaps the best stratagem might be to banish it from your vocabulary completely. Even worse is the widely used implication symbol ⇒ . There is little in life more unpleasant than a page of purported mathematics in which each line begins with ⇒ . Some undergraduates just dump it down on the page as if to say "I am going to start another line now. I wonder if the reader is clever enough to work out if and how this line is related to the other lines on this page".

The implication symbol ⇒ is a fine thing in the proper context. It is a symbol borrowed from mathematical logic in fact. If you use it at all, use it sparingly. There is one circumstance when this symbol really is rather useful. We will come to that shortly.

2.7 Double Implication

Suppose that A and B are two propositions. You may be asked to show that A is true if and only if B is true. This means that you will almost certainly have to do two pieces of work. You must demonstrate both of the following propositions.

1 If A is true, then B is true.

2 If B is true, then A is true.

The phrase *if and only if* is used so frequently that a notational convention has been adopted which saves time, trees and ink. We write *iff* to mean "if and only if". A symbolic synonym for *iff* is ⇔ (implies and is implied by). There is nothing wrong with ⇔, save that it can encourage the wanton use of ⇒ . Another useful synonym for iff which looks less like a spelling error is *exactly when.*

When doing the two parts of an "if and only if" proof, it is polite to tell the reader which part you are doing first and which second. This is an opportunity to use ⇒ and ⇐ (is implied by) to rather good effect. Perhaps this is best illustrated by an example:

First a quick definition. If n and d are natural numbers and there exists a third natural number q such that $n = qd$, then we say that d is a *divisor* of n. This is written $d \mid n$.

Proposition 2.7

Suppose that n is a natural number. The number of distinct divisors of n is odd if and only if n is a perfect square.

Remark 2.1

In terms of the general set-up described above, A is the proposition "the number of distinct divisors of n is odd"; the proposition B is that "n is a perfect square". We must show that if A is true (for some particular n), then B must be true. Then we will have done exactly half the problem. To finish we must show that if B is true (for some particular natural number n), then A must be true. Thus our plan is first to show $A \Rightarrow B$ and then $A \Leftarrow B$. We can use these implication symbols to inform the reader what is going on.

Proof

$\Rightarrow$) If $d \mid n$, then let $d' = n/d \in \mathbb{N}$. Notice that $dd' = n$ so d' is also a divisor of n. Also observe that $d'' = d$ so the divisors of n occur in pairs – except possibly in the case $d = d'$. We are assuming that the number of distinct divisors of n is odd, so there must exist a divisor d of n such that $d = d'$. Thus $n = dd' = d^2$ is a perfect square.

$\Leftarrow$) Now we assume that n is a perfect square. We use the notation outlined in the first half of the proof. The number of divisors of n with the property that $d \neq d'$ is even, so it suffices to show that the number of divisors d such that $d = d'$ is odd. In fact a stronger statement is true. The number of distinct divisors d of n such that $d = d'$ is exactly one. Certainly there is such a divisor f (since n is a perfect square). Suppose g were a rival divisor such that $g = g'$. Then $n = ff' = gg' = f^2 = g^2$. Thus $f^2 = g^2$ so $f^2 - g^2 = 0$ and thus $(f - g)(f + g) = 0$. Now $f + g$ is strictly positive so $f - g = 0$ and $f = g$. Therefore f is the unique divisor of n with the property that $f = f'$. Thus the number of divisors of n must be odd.

Thus the number of divisors of n is odd if and only if n is a perfect square.

□

2.8 The Master Plan

When you are planning to write down a proof that is more than a few lines long, it is a good idea to spell out the plan of the proof. A short comment such as "we will prove this proposition by simple induction on x", or perhaps "we will use a contradiction argument to establish this proposition. We therefore suppose that the proposition is false, and produce a contradiction as follows"– or more succinctly "Assume, for contradiction, that the result is false".

3
Complex Numbers and Related Functions

3.1 Motivation

Think back to when you first started doing algebra. No doubt you were taxed with problems such as "find x such that $3x = 12$". When you had learned to deal with those, the teacher may have produced more tricky ones such as $2x = 3$ or even $2x + 4 = 0$. You can't solve either of these equations using the natural numbers, and if either equation is supposed to tell you the number of people in the room, you know there is a mistake somewhere. To solve these equations you need more extensive number systems, involving fractions in the first case, and negative numbers in the second. These equations are all *linear equations*, in other words, the graphs of the functions defined by the left-hand sides of these equations are straight lines. The algebra which grows out of the study of these equations is called *linear algebra.*

In this way we are driven to work in $\mathbb{Q}$ rather than $\mathbb{N}$ when doing linear algebra. Any equation of the form $ax + b = 0$ with $a, b \in \mathbb{Q}$ and $a \neq 0$ has exactly one rational solution $-b/a$. If a, b are restricted to be in $\mathbb{N}$ (or $\mathbb{Z}$), and we look for a solution in the same set, we find a complete mess; sometimes there is a solution, sometimes not.

So, the rationals are a great place to work in if you are doing linear algebra. Also $\mathbb{R}$ is just as good. There are intermediate sets which are also perfectly satisfactory places to do linear algebra; for example $F = \{a + b\sqrt{2} \mid a, b \in \mathbb{Q}\}$ is fine; you should check that F is closed under addition, subtraction, multiplication and division by a non-zero quantity. There are infinitely many sets

G. Smith, *Introductory Mathematics: Algebra and Analysis*

intermediate between $\mathbb{Q}$ and $\mathbb{R}$ which are good places to do linear algebra; can you find another one?

Anyway, let's make life more difficult, and worry about quadratic equations of the form $ax^2 + bx + c = 0$ where the coefficients a, b, c are in some specified set – say $\mathbb{R}$. As you should know, if $a \neq 0$ this equation has solutions of the form $\frac{-b \pm \sqrt{b^2-4ac}}{2a}$ provided $\sqrt{b^2 - 4ac}$ exists. In fact these are the only possible solutions, so there are either two solutions, or a repeated solution (when $b^2 - 4ac = 0$) or no solutions when $b^2 - 4ac < 0$. This is all a bit untidy. There is not much you can do about $x^2 - 2x + 1 = (x - 1)^2 = 0$ having a repeated solution, but it seems a bit awkward that a quadratic equation will sometimes have two solutions and sometimes none.

The solutions of an equation are sometimes called its roots. There is a strong relationship between functions (i.e. maps) and equations. For example, suppose that you are told to find all the real roots of $2x^2 - 5x + 1 = 0$; this problem can be cast in the language of maps. Consider $f : \mathbb{R} \to \mathbb{R}$ defined by $f(x) = 2x^2 - 5x + 1 \ \forall x \in \mathbb{R}$. The solutions of the equation are the elements of $\{\alpha \mid \alpha \in \mathbb{R},\ f(\alpha) = 0\}$.

Just in case you don't know it, here is a proof that the roots of $ax^2 + bx + c = 0$ are given by the usual formula.

Proposition 3.1

Consider the equation

$$ax^2 + bx + c = 0 \text{ where } a, b, c \in \mathbb{R} \text{ and } a \neq 0. \tag{3.1}$$

This equation has no roots in $\mathbb{R}$ if $b^2 - 4ac < 0$. The real roots are given by the formula

$$\frac{-b \pm \sqrt{b^2 - 4ac}}{2a}$$

if $b^2 - 4ac \geq 0$.

Proof

Assume (possibly, but not necessarily, for contradiction) that $\alpha \in \mathbb{R}$ is a root of Equation (3.1) so $a\alpha^2 + b\alpha + c = 0$ and $a \neq 0$. Divide through by a and rearrange slightly to see that

$$\alpha^2 + \frac{b}{a}\alpha = -\frac{c}{a}.$$

Now add $\left(\frac{b}{2a}\right)^2$ to each side to get

$$\alpha^2 + \frac{b}{a}\alpha + \left(\frac{b}{2a}\right)^2 = \left(\frac{b}{2a}\right)^2 - \frac{c}{a}.$$

Tidy up to obtain

$$\left(\alpha+\frac{b}{2a}\right)^2=\frac{b^2-4ac}{4a^2}.$$

If the right hand side is negative we have a contradiction, since the square of a real number is not negative. In this event the initial assumption that there is a real root α must have been nonsense so there are no real roots.

On the other hand, if the right hand side is positive we have

$$\left(\alpha+\frac{b}{2a}\right)^2=\left(\frac{\sqrt{b^2-4ac}}{2a}\right)^2.$$

Now in $\mathbb{R}$ we have $u^2=v^2$ if and only if $u=\pm v$ because $u^2-v^2=(u-v)(u+v)$ and the product of two reals is 0 exactly when at least one of the quantities being multiplied is 0.

Thus $(\alpha+\frac{b}{2a})=\frac{\pm\sqrt{b^2-4ac}}{2a}$. We have proved that if $\alpha\in\mathbb{R}$ is a root of Equation (3.1), then

$$\alpha=\frac{-b\pm\sqrt{b^2-4ac}}{2a}$$

and $b^2-4ac\geq 0$.

This is not quite the end of the story. We also need to show that $\frac{-b\pm\sqrt{b^2-4ac}}{2a}$ is a real solution of Equation (3.1) provided $b^2-4ac\geq 0$. This can be done in two ways. Either you can retrace the argument in reverse (which is probably the quickest thing to do) or the devoted reader can simply substitute the alleged roots into Equation (3.1) and verify that $a\alpha^2+b\alpha+c=0$.

□

The fuss in the last paragraph when we insisted on verifying that $a\alpha^2+b\alpha+c=0$ may have come as a surprise. We will now try to convince you that it was necessary.

Example 3.1

You are asked to find all $x\in\mathbb{R}$ such that $|x|=-1$. The correct answer is that no real number has modulus -1. However, we may reason like this. Suppose that $\alpha\in\mathbb{R}$ and $|\alpha|=-1$. Take the modulus of each side so $||\alpha||=|-1|$. Now $||\alpha||=|\alpha|$ and $|-1|=1$ so $|\alpha|=1$. It follows that $\alpha=1$ or $\alpha=-1$. This might cause an eyebrow to quiver, but there is no fault in the reasoning. The point is that the original assumption that there was a real number α satisfying the equation was false. From a false premise, you can deduce anything at all, and in particular you can deduce that $\alpha=1$ or -1. You pick up the fact that the assumption was wrong by substituting 1 and -1 into the equation and finding that it is broken.

False assumptions are not the only way you can generate trouble.

Example 3.2

You are asked to find all $x \in \mathbb{R}$ such that $x - 1 = 0$. Suppose that $\alpha \in \mathbb{R}$ is a solution of this equation. Now, the more astute reader may have already observed that this equation has a real solution, so perhaps we can avoid the unpleasantness of the preceding example. We know that $\alpha - 1 = 0$. Multiply both sides by $\alpha + 1$ to yield $\alpha^2 - 1 = 0$. Add one to each side so $\alpha^2 = 1$ and thus α is either 1 or -1. At this point you must not fall into the trap of thinking that -1 is a solution of the equation. Our argument shows that if α is a real root, then α is 1 or -1. This is right, since the only root is 1 and it is true that $1 = 1$ or $1 = -1$. That is how "or" works.

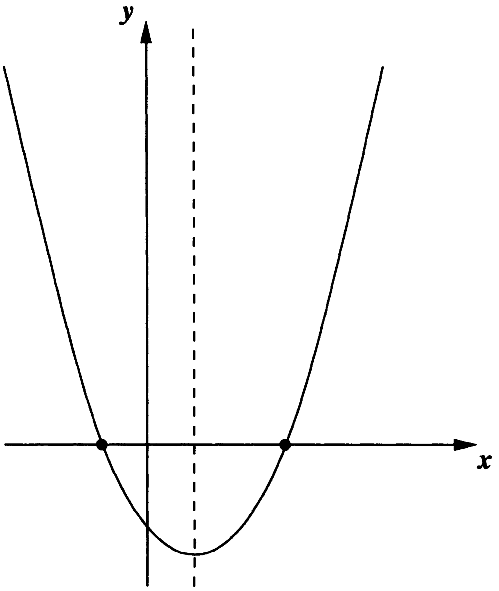

Fig. 3.1. Nose-down parabolic graph of a quadratic polynomial with two roots

Now we address the issue of finding a geometric interpretation of solving a quadratic equation. Suppose $a \neq 0$ and $f : \mathbb{R} \to \mathbb{R}$ is defined by $f(x) = ax^2 + bx + c \; \forall x \in \mathbb{R}$. The graph of this function is $\{(x, f(x)) \mid x \in \mathbb{R}\}$ to which, thanks to Descartes, you can give a geometric interpretation as a subset of the x, y plane. The picture of the graph (what you might call the graph itself) is a parabola, with axis of symmetry parallel to the y-axis, as shown in Figure 3.1. This parabola will be nose down or nose up as a is positive or negative. The equation $ax^2 + bc + c = 0$ will have a solution α exactly when $(\alpha, 0)$ is in the graph of the function. In the picture, $ax^2 + bx + c = 0$ has a solution precisely when the parabola meets the x-axis. This gives you two solutions or no solutions in general, but you can have a repeated solution if the parabola just kisses (is tangent to) the x-axis.

Returning to the algebraic view, the obstruction to using the formula to solve a quadratic is that you can be asked to extract the square root of a negative real number. The way forward is to extend our number system yet again in such a way that we can do this. This may seem a little uncomfortable at first: personally I can recall feeling very suspicious about negative numbers and their arithmetic. It was all very well being told to think of negative quantities as debts, but how can you multiply two debts together to get a credit? The point is that it is intellectually crippling to give symbols real-world interpretations all the time. It may be useful to invest symbols with meaning temporarily as a psychological prop, or to help you gain inspiration about how to solve a mathematical problem, but ultimately the symbols are symbols and nothing else.

A teacher who tries to explain that a piece of mathematics is easy because x is *really* temperature and y is *really* height above sea-level and so the equation *means* something or other may be very effective at getting across an idea. However, this is not mathematics. The negative numbers are symbols, nothing more, and we endow them with multiplication, division, addition and subtraction as we see fit. We do it in such a way that the laws of algebra (associativity, distributivity, etc.) continue to be valid in $\mathbb{Z}$ just as they were in $\mathbb{N}$. When we try to extend $\mathbb{R}$ to build a larger system where you can extract a square root of anything you fancy, our obligation is to do it in such a way that the algebraic properties of $\mathbb{R}$ continue to be valid in the new world. That way we can carry on doing mathematics.

Of course, there is a strong association *in our minds* between the real numbers and a powerful image, the infinite line. We should not be surprised if a sufficiently sweet generalization of $\mathbb{R}$ has a fine geometric interpretation. That is written with the benefit of hindsight of course. In every case I have ever seen, a good geometric interpretation of symbols has been a useful inspiration, and notwithstanding the previous paragraph, it seems to be a good strategy to keep track of any relevant geometry. I have no idea whether this is intrinsic to thought, or because vision is by far the most developed of human senses.

3.2 Creating the Complex Numbers

Recall from Chapter 1 that $\mathbb{R}^2$ consists of ordered pairs of real numbers, and that two ordered pairs are equal if and only if their corresponding entries are equal. We endow $\mathbb{R}^2$ with addition *co-ordinatewise*, by

$$(a,b)+(c,d)=(a+c,b+d)\ \forall a,b,c,d\in\mathbb{R}.$$

It easy to check that this operation is associative and commutative. The element $(0,0)$ acts as an additive identity element and the additive inverse of (a,b) is $(-a,-b)$. Now, more ambitiously, give $\mathbb{R}^2$ a multiplication via

$$(a,b)\star(c,d)=(ac-bd,ad+bc)\ \forall a,b,c,d\in\mathbb{R}.$$

This is a commutative operation since $(c,d)\star(a,b)=(ca-db,da+cb)=(ac-bd,ad+bc)$. The element $(1,0)$ acts as a multiplicative identity. It is also true that this multiplication is associative. We shall stay honest and check this; for all $a,b,c,d,e,f\in\mathbb{R}$ we have

$$\begin{aligned}
&((a,b)\star(c,d))\star(e,f)=(ac-bd,ad+bc)\star(e,f)\\
&=((ac-bd)e-(ad+bc)f,(ac-bd)f+(ad+bc)e)\\
&=(a(ce-df)-b(cf+de),a(cf+de)+b(ce-df))\\
&=(a,b)\star(ce-df,cf+de)=(a,b)\star((c,d)\star(e,f)).
\end{aligned}$$

Finally, note that we have a multiplicative inverse for $(a,b)\neq(0,0)$. It is $(\frac{a}{a^2+b^2},\frac{-b}{a^2+b^2})$ because

$$(a,b)\cdot\left(\frac{a}{a^2+b^2},\frac{-b}{a^2+b^2}\right)=\left(\frac{a^2+b^2}{a^2+b^2},\frac{-ab+ba}{a^2+b^2}\right)=(1,0).$$

This multiplication is thus a splendid operation. Ordinary multiplication of real numbers interacts with addition via the distributive law: $\forall a,\ b,\ c\in\mathbb{R}$ we have $a\cdot(b+c)=(a\cdot b)+(a\cdot c)$, or just $a\cdot b+a\cdot c$ subject to the usual conventions about priority of operations. We now check that our addition and multiplication on $\mathbb{R}^2$ is just as good.

$$\begin{aligned}
&(a,b)\star((c,d)+(e,f))=(a,b)\star(c+e,d+f)\\
&=(ac+ae-bd-bf,ad+af+bc+be)\\
&=(ac-bd,ad+bc)+(ae-bf,af+be)\\
&=(a,b)\star(c,d)+(a,b)\star(e,f)\ \forall a,b,c,d,e,f\in\mathbb{R}.
\end{aligned}$$

Consider the map $\theta:\mathbb{R}\to\mathbb{R}^2$ defined by $\theta(r)=(r,0)$. This is an injective map, and it preserves all the algebraic features of $\mathbb{R}$. In fact you can think of $\mathbb{R}$

as simply acquiring a bit of decoration; a left bracket in front, and a comma, a zero and a right bracket behind. We'll use a bold font $\mathbf{x}$ as shorthand for $(x, 0)$ and use juxtaposition or a dot to denote multiplication of these symbols in bold print. We will also use $+$ to denote addition of ordered pairs. We will use $\mathbf{R}$ to denote $\{(x, 0) \mid x \in \mathbb{R}\}$. Using our definitions of addition and multiplication in $\mathbb{R}^2$ it is still the case that $\mathbf{2} + \mathbf{3} = \mathbf{5}$ and $\mathbf{6} \cdot \mathbf{7} = \mathbf{42}$ because $(2,0)+(3,0) = (5,0)$ and $(6,0) \cdot (7,0) = (6 \cdot 7 - 0 \cdot 0, 6 \cdot 0 + 0 \cdot 7) = (42, 0)$. Our set $\mathbf{R}$ is just a copy of $\mathbb{R}$ and is every bit as good as $\mathbb{R}$.

Let $\mathbf{i} = (0,1)$, then $\mathbf{i}^2 = \mathbf{i} \cdot \mathbf{i} = (0 \cdot 0 - 1 \cdot 1, 0 \cdot 1 + 0 \cdot 1) = (-1, 0) = \mathbf{-1}$.

Notice that every element of $\mathbb{R}^2$ can be expressed in terms of bold symbols by $(a, b) = (a, 0) + (0, b) = (a, 0) + (0, 1) \star (b, 0) = \mathbf{a} + \mathbf{ib}$.

Definition 3.1

The set $\mathbb{C}$ of *complex numbers* is the set $\mathbb{R}^2$ of ordered pairs endowed with addition and multiplication as above.

We have the usual geometric picture of $\mathbb{R}^2$ where two axes are drawn at right angles in the plane, and points are associated to pairs (x, y). In this case however, $(x, y) = (x, 0) + (0, 1) \star (y, 0) = \mathbf{x} + \mathbf{iy}$.

Definition 3.2

The Argand diagram (Figure 3.2) is the usual x, y plane corresponding to $\mathbb{R}^2$, except that each point is labelled with a complex number $\mathbf{a} + \mathbf{ib}$ rather than the ordered pair (a, b).

Thus complex numbers are things that look like $\mathbf{x} + \mathbf{iy}$, where $\mathbf{x}, \mathbf{y} \in \mathbf{R}$. Notice that $\mathbf{R} \subseteq \mathbb{C}$. The addition and multiplication work like this. Suppose that $\nu = \mathbf{3} - \mathbf{2i}$ and $\mu = \mathbf{3/2} + \mathbf{i}$, then $\nu + \mu = \mathbf{9/2} - \mathbf{i}$ and

$$\nu \cdot \mu = \mathbf{3(3/2) + 3i + (-2i)(3/2) + (-2i)i = 9/2 + 2 + (3 - 3)i = 13/2}.$$

In other words, you can add and multiply these complex numbers by thinking of $\mathbf{3} - \mathbf{2i}$ and $\mathbf{3/2} + \mathbf{i}$ as linear polynomial expressions in the variable $\mathbf{i}$, but with the extra proviso that you replace $\mathbf{i}^2$ by $\mathbf{-1}$ whenever it occurs. This always works, and applies to subtraction and division as well.

We don't need our old real numbers any more. The new copy $\mathbf{R}$ has all the properties of the old real numbers, with the bonus that it is a subset of $\mathbb{C}$. We now discard for ever our old real numbers, and use the ones in bold print instead. Having done that, there is no need to use bold type any more, so $\mathbf{i}$ gets written as i and $\mathbf{3\frac{3}{4}}$ as $3\frac{3}{4}$. Since the symbol $\mathbb{R}$ is now redundant, it

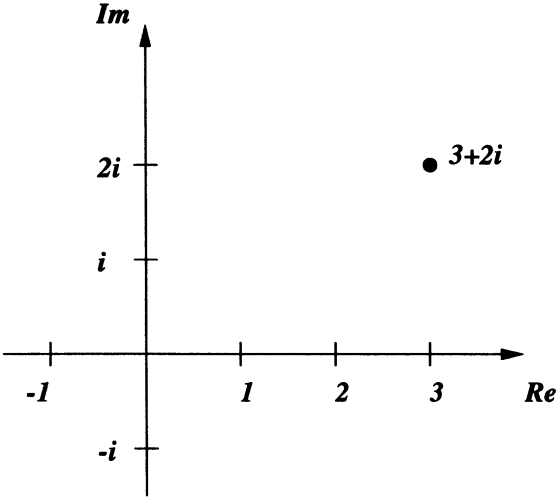

Fig. 3.2. The Argand diagram

is harmless to recycle it and write $\mathbb{R}$ instead of $\mathbf{R}$. Now $\mathbb{R} \subseteq \mathbb{C}$. So, we have gone to considerable trouble to construct $\mathbb{C}$ properly. Having laid these secure foundations, we will be able to take a confident and relaxed attitude to $\mathbb{C}$. We will not have to think of $\mathbb{C}$ as consisting of ordered pairs of real numbers; the elements of $\mathbb{C}$ are just expressions of the form $a + bi$ with $a, b \in \mathbb{R}$, and we have simple rules for doing algebra with these symbols.

Definition 3.3

If $a, b \in \mathbb{R}$, then the real and imaginary parts of $z = a + ib \ \in \mathbb{C}$ are a and b respectively. We write $\text{Re}(z) = a$ and $\text{Im}(z) = b$. We say that the complex number z is real if $\text{Im}(z) = 0$, and that z is *purely imaginary* if $\text{Re}(z) = 0$.

Note the slightly strange fact that the imaginary part of a complex number is always real. Also observe that the imaginary part of a real number is 0. If you see a phrase such as "Consider the complex number $a + ib$" the author often means to imply that $a, b \in \mathbb{R}$, but that may accidentally be omitted in the rush, especially in a lecture course. If this is not specified, there is the danger that a and b might be complex numbers which are not real. As long as you are alert to the possibility of ambiguity, it should be possible to work out the intended meaning from the context.

If we are going to do mathematics with $\mathbb{C}$ we had better make sure which laws of algebra it satisfies. In fact it is an example of a *field.*

Definition 3.4

A set F is called a *field* if it is endowed with binary operations $+$ and $*$, and contains elements 0 and 1 so that the following axioms are all satisfied.

The Field Axioms

$$
\begin{gathered}
a+b=b+a \ \forall a,b\in F \text{ (addition is commutative)}\\
a+(b+c)=(a+b)+c \ \forall a,b,c\in F \text{ (addition is associative)}\\
\exists 0\in F \text{ such that } a+0=a \ \forall a\in F \text{ (0 is an additive identity)}\\
\forall a\in F\ \exists -a\in F \text{ such that } -a+a=0 \text{ (there are additive inverses)}\\
a*b=b*a \ \forall a,b\in F \text{ (multiplication is commutative)}\\
a*(b*c)=(a*b)*c \ \forall a,b,c\in F \text{ (multiplication is associative)}\\
\exists 1\in F \text{ such that } a*1=a \ \forall a\in F \text{ (1 is a multiplicative identity)}\\
\forall a\in F\setminus\{0\}\,\exists a^{-1}\in F \text{ such that } a^{-1}*a=1 \text{ (multiplicative inverses)}\\
a*(b+c)=a*b+a*c \ \forall a,b,c\in F \text{ (distributive law)}
\end{gathered}
$$

and, last but not least, $0\neq 1$.

We have given the usual priority to $*$ over $+$ to lose a few brackets, and from now on, $*$ will be replaced by juxtaposition, × or a central dot.

The pattern is clear. The first four and second four axioms tell us that addition and multiplication are good operations. The distributive axiom tells us that they interact well, and the final axiom is a purely technical one to preclude the unwanted case that $F=\{0\}$. We now have have a formal characterization of the golden lands ("good places to do linear algebra") mentioned in Section 3.1.

Notice that $\mathbb{Q}$ forms a field, but $\mathbb{Z}$ does not (since 2 has no multiplicative inverse in $\mathbb{Z}$). We are a little hazy about the real numbers $\mathbb{R}$, but let us agree that they satisfy the field axioms. On that assumption, $\mathbb{C}$ is also a field – we have checked almost all of the details already.

There is a considerable advantage in using axiom systems to study mathematics. Not only do they clarify ideas, but they also permit the following strategy. Try to prove a theorem about an arbitrary system satisfying a given collection of axioms. If you succeed, then the theorem becomes true of all those systems at once – including systems which no-one has ever invented or considered. The following result is well known for the rational numbers and indeed for $\mathbb{R}$. However, if you have never worked with $\mathbb{C}$ before it may not be obvious to you that it works for the complex numbers too, but $\mathbb{C}$ is a field so it does.

Proposition 3.2

Let F be a field.

(i) For all $f \in F$ we have $f \cdot 0 = 0$.

(ii) If $h, k \in F$ and $hk = 0$, then $h = 0$ or $k = 0$.

Proof

(i) For any f we have

$$f \cdot 0 = f \cdot (0 + 0) \text{ by definition of } 0$$

so

$$f \cdot 0 = f \cdot 0 + f \cdot 0 \text{ by distributivity.}$$

Now add the additive inverse of $f \cdot 0$ to each side and then use additive associativity to yield

$$-(f \cdot 0) + f \cdot 0 = -(f \cdot 0) + (f \cdot 0 + f \cdot 0) = (-(f \cdot 0) + f \cdot 0) + f \cdot 0.$$

The definitions of an additive inverse and 0 ensure that

$$0 = 0 + f \cdot 0 = f \cdot 0$$

and we are done.

(ii) Suppose that $h, k \in F$ and $hk = 0$. Either $h = 0$ (and we are done) or there exists $h^{-1} \in F$ such that $h^{-1}h = 1$. Now $h^{-1}(hk) = h^{-1}0$ so by associativity we have $h^{-1}0 = (h^{-1}h)k = 1k = k$. Thus we can deduce $k = 0$ because of (i).

□

The algebraic argument that we used to find the roots of a quadratic equation is valid for any field (except for those fields where $1 + 1 = 0$; we won't be working with these monsters, but they do exist, and division by $2 = 1 + 1$ is impossible in such fields). In fields where $1 + 1 \neq 0$ the delicate moment is where, if you recall, you have to deduce that $u = \pm v$ from $u^2 = v^2$. The point is that $u^2 - v^2 = 0$ so $(u - v)(u + v) = 0$. Now you use Proposition 3.2 (ii) to see that either $u - v = 0$ or $u + v = 0$, from which it follows that $u = \pm v$.

In order to be able to solve all quadratic equations with coefficients in $\mathbb{R}$ we need to be able to extract square roots of negative numbers. If you want to solve quadratic equations where the coefficients are in $\mathbb{C}$, then you will need to be able to extract square roots of complex numbers. We shall now see that this is always possible.

Proposition 3.3

Suppose that $a, b \in \mathbb{R}$. The square roots of $a + ib$ are $\pm(x + iy)$ where

$$x = \sqrt{\frac{a + \sqrt{a^2 + b^2}}{2}}, \ y = \text{sign(b)}\sqrt{\frac{-a + \sqrt{a^2 + b^2}}{2}}$$

and $\text{sign}(b)$ is -1 if b is negative and otherwise is $+1$.

Proof

The result is straightforward when $b = 0$, and we leave that case to the reader. Thus we assume that $b \neq 0$. Suppose that $\zeta, \eta \in \mathbb{C}$ with $\zeta = a+ib$ and $\eta = x+iy$ where $a, b, x, y \in \mathbb{R}$ and $\zeta = \eta^2$, so $\text{Re}(\zeta) = \text{Re}(\eta^2)$ and $\text{Im}(\zeta) = \text{Im}(\eta^2)$. Thus $a = x^2 - y^2$ and $b = 2xy$. We eliminate y between these equations by observing that $4ax^2 + b^2 = 4x^4 - 4x^2y^2 + 4x^2y^2 = 4x^4$. Thus

$$4x^4 - 4ax^2 - b^2 = 4(x^2)^2 - 4ax^2 - b^2 = 0.$$

Although this is a quartic equation, it is actually a quadratic equation in the variable x^2 so

$$x^2 = \frac{4a \pm \sqrt{16a^2 + 16b^2}}{8} = \frac{a \pm \sqrt{a^2 + b^2}}{2}.$$

We are looking for $x \in \mathbb{R}$ so $x^2 \geq 0$. One of the two candidate values for x^2 looks doubtful. Remember that a and b are real, so $|a| = \sqrt{a^2} < \sqrt{a^2 + b^2}$ since $b \neq 0$. Thus $a - \sqrt{a^2 + b^2} < a - |a| \leq 0$ and so

$$\frac{a - \sqrt{a^2 + b^2}}{2}$$

is not the square of a real number. Thus we have eliminated one possibility and are left with the other which is

$$x^2 = \frac{a + \sqrt{a^2 + b^2}}{2}$$

and extracting square roots yields

$$x = \pm\sqrt{\frac{a + \sqrt{a^2 + b^2}}{2}}.$$

Now $x^2 - y^2 = a$ so we have

$$y^2 = x^2 - a = \frac{-a + \sqrt{a^2 + b^2}}{2}$$

and so

$$y = \pm\sqrt{\frac{-a + \sqrt{a^2 + b^2}}{2}}.$$

Suppose that $b > 0$, then you want $b = 2xy > 0$ and this can be accomplished by choosing the signs consistently; both + or both −. Conversely if b is negative the signs must be chosen opposite in order to obtain $2xy < 0$.

We now have the candidate square roots mentioned in the proposition. By reversing the reasoning, or by directly verifying that $(x + iy)^2 = a + ib$, we are done.

□

We are now in a happy state. Just as any linear equation with real coefficients can be solved with a real solution, any quadratic equation with complex coefficients can be solved using the complex numbers. At this point, the reader with a good mathematical imagination might be thinking that in order to solve cubics and then quartic and higher degree equations, we will need to construct larger and larger number systems, progressively more exotic generalizations of $\mathbb{R}$ and $\mathbb{C}$. Happily or unhappily (it depends upon your taste), this doesn't happen.

A polynomial equation of degree n with coefficients in $\mathbb{C}$ is one of the form

$$a_n x^n + a_{n-1}x^{n-1} + \ldots + a_0 = 0$$

with $a_i \in \mathbb{C}\ \forall i \in \{0, \ldots, n\}$, x being an unknown and $a_n \neq 0$.

Such an equation has exactly n solutions (possibly allowing for repetitions) in $\mathbb{C}$. In this sense, $\mathbb{C}$ is the end of the road. This result is called the *Fundamental Theorem of Algebra* and we will not prove it here; it is too deep. However, it is not beyond mortal range, and can be proved using a variety of techniques. It was first proved by the great Hanoverian mathematician Carl Friederich Gauss.

EXERCISES

3.1 Put the following expressions into the shape $a + ib$ with $a, b \in \mathbb{R}$.

(a) i^3

(b) $(1 - i)i$

(c) $(1+i)^2$

(d) $(1+i)^8$

(e) $(1-i)^8$

(f) $(1-2i)(3-4i)-i^3$

(g) $(\frac{-1+i\sqrt{3}}{2})^2$

(h) $(\frac{-1+i\sqrt{3}}{2})^3$

(j) $(\frac{-1+i\sqrt{3}}{2})^{999}$

(k) $\frac{1}{1+i}$ (Hint: *Μυλτιπλι τοπ ανδ βοττομ βι ονϵ μινυσ i.*)

(l) $\frac{1}{1+i}+\frac{1}{1-i}$ (Hint: *Σϵϵ πρϵφιουσ ιντ.*)

(m) $\frac{1-i}{2-3i}$ (Hint: *Ασ βϵφορϵ.*)

3.2 Find all complex numbers satisfying each of the following equations in the unknown x.

(a) $x^2 = -1$

(b) $x^2 = -2$

(c) $x^2 = i$

(d) $x^2 + x + 1$

(e) $x^3 - 1 = 0$

(f) $x^2 + 4ix + 5 = 0$

(g) $x^2 + 4ix - 5 = 0$

3.3 Use the field axioms (and Proposition 3.2) to prove each of the following statements about a field F. Justify every step of your argument by appealing to an axiom or a previously proved result.

(a) $(-1)(-1) = 1$ (Hint: *Σταρτ οφφ υζινγ νϵγατιφ ονϵ πλυσ ονϵ ϵκαλσ ζϵρο.*)

(b) $\forall f \in F$ we have $-f = (-1)f$.

(c) $\forall f, g \in F$ we have $(-f)\cdot g = -(fg)$.

(d) $\forall f, g \in F$ we have $(-f)\cdot(-g) = fg$.

(e) If $f, g \in F$ and $fg = g$, then either $g = 0$ or $f = 1$.

(f) If $f, g, h \in F$ and $fg = hg$, then either $g = 0$ or $f = h$.

3.3 A Geometric Interpretation

Let's see what all this means geometrically. First we go back to $\mathbb{R}$, and think about the real line. Addition has an easy interpretation. If you add 5 to a real number, the answer is the number situated distance 5 to the right on the real line. You can think of adding 5 as a map. We call it add_5 or more formally $\text{add}_5 : \mathbb{R} \to \mathbb{R}$ defined by $\text{add}_5(x) = x + 5\ \forall x \in \mathbb{R}$.

The map add_5 is a bijection, and its inverse is the map add_{-5}. Using this notation, add_0 would be the identity map from $\mathbb{R}$ to $\mathbb{R}$. One very special feature of add_5 is that it preserves distances. The distance between x and y is the same as between $x+5$ and $y+5$ because $|x-y| = |(x+5)-(y+5)|$. The same is true for add_r whenever r is a real number – it always preserves distances. Think of the real line as having a house at each real number, and each house having a single occupant. The map add_5 tells everyone to move to the house situated distance 5 to the right. When this is done, the people find that in their new homes, the neighbours seem rather familiar. The fact that Vin Luthra used to live a distance π from Pete Whitelock is still true after the application of add_5. Maps of the form add_r simply translate people (or numbers) in a rigid way.

EXERCISES

3.4 (a) Let $f : \mathbb{R} \to \mathbb{R}$ be a map which preserves distances. Prove that there are numbers $a \in \{1, -1\}$ and $b \in \mathbb{R}$ such that $f(x) = ax + b\ \forall x \in \mathbb{R}$. Conversely, show that any map defined by a formula of this type will preserve distances.

(b) What happens if you replace $\mathbb{R}$ by $\mathbb{Q}$ or $\mathbb{Z}$ in part (a)?

3.5 Try to classify (list or describe) all maps from $\mathbb{R}^2$ to $\mathbb{R}^2$ which fix the origin and preserve distances.

The jargon for a distance-preserving bijection is that it is an *isometry*. Next consider multiplication. Fix a real number, perhaps $2 \in \mathbb{R}$ and consider the effect of multiplying by 2. The map $\text{mul}_2 : \mathbb{R} \to \mathbb{R}$ is defined by $\text{mul}_2(x) = 2x\ \forall x \in \mathbb{R}$. In terms of the house, the person in the house at x is told to move to the house at $2x$. This certainly does not preserve all distances. For example, $|7 - 3| = 4$ but $|14 - 6| = 8$. The effect of this map is to stretch numbers (or people) apart. This map is a bijection, and its inverse $\text{mul}_{1/2}$ shrinks the distances between numbers (or people). Notice that $\text{mul}_1 = \text{add}_0$. Also mul_0 is not a bijection, but rather a sort of black hole which sucks everything to 0.

We want a similar geometric picture for addition and multiplication in the

Argand diagram. This is no exotic 17-dimensional confection, but simply the ordinary plane labelled with complex numbers in the obvious way.

Addition in $\mathbb{C}$ has a very straightforward interpretation in the Argand diagram. If you are familiar with vectors you will see that addition of complex numbers z_1 and z_2 to get $z_1 + z_2$ corresponds to vector addition. We will go into that topic in great detail in Chapter 4.

This time you need to imagine that there is a house at each point of the Argand diagram. Choose and fix a complex number; perhaps $1 - i$. Consider the map $\mathrm{add}_{1-i} : \mathbb{C} \to \mathbb{C}$. This is defined by $\mathrm{add}_{1-i}(z) = z+1-i \; \forall z \in \mathbb{C}$. What happens is that all the points of the plane are subject to a translation. Thinking of our plane as being inhabited, adding $1-i$ tells the occupants to move distance 1 to the right and distance 1 down. The net effect is that everyone moves a distance of $\sqrt{2}$ in a south-easterly direction. Notice that the distance between complex numbers p, q is unchanged if you add the same complex number to each of them. The story about people having the same neighbours holds good.

Multiplying complex numbers by a fixed positive real number simply pushes them directly away from or towards the origin, as the real number is bigger than or less than 1. Multiplication by a fixed negative real number sends points to the other side of the origin, and as before multiplication by 0 has a dramatic astronomical analogy. Multiplication by a non-real complex number is not nearly so obvious. For example, what does the map mul_{1-i} do? It maps 0 to 0, and 1 to $1 - i$, and you can do experiments to try to guess the geometric rule. Rather than reading the explanation immediately, you are asked to break off from the text and do some experiments.

In order to understand the geometry of complex multiplication, we will develop some machinery. In particular, the notion of distance in the Argand diagram ought to be captured in $\mathbb{C}$ itself.

Definition 3.5

If $z = a + ib \in \mathbb{C}$ with $a, b \in \mathbb{R}$, then the *modulus* of z is $|z| = \sqrt{a^2 + b^2}$. Geometrically this is the distance from 0 to z in the Argand diagram, thanks to a theorem of Pythagoras.

Observe that this definition is consistent with the use of $|x|$ for $x \in \mathbb{R}$. Also notice that if $u, v \in \mathbb{C}$, then the distance from u to v in the Argand diagram is $|u - v|$. This is extremely important. Please draw a picture to convince yourself that we have captured the notion of distance properly.

Definition 3.6

If $z = a + ib \in \mathbb{C}$ with $a, b \in \mathbb{R}$, then the *complex conjugate* of z is $\overline{z} = a - ib$.

As usual it is interesting to look at the geometric interpretation of what we have just defined. Complex conjugation is a map from $\mathbb{C}$ to $\mathbb{C}$ which corresponds to reflection in the real axis in the Argand diagram. In terms of our analogy, complex conjugation amounts to a home exchange, where northerners and southerners swap homes, and the unadventurous folk on the real line swap homes with themselves. Notice that if $z = a + ib \in \mathbb{C}$ with $a, b \in \mathbb{R}$, then $z\overline{z} = a^2 + b^2 = |z|^2$ is the square of the distance of z from the origin in the Argand diagram.

Proposition 3.4

(i) $\overline{z_1 + z_2} = \overline{z}_1 + \overline{z}_2\ \forall z_1, z_2 \in \mathbb{C}$.

(ii) $\overline{z} = z$ if and only if $z \in \mathbb{R}$.

(iii) $\overline{z} = -z$ if and only if z is purely imaginary.

(iv) $z + \overline{z} \in \mathbb{R}\ \forall z \in \mathbb{C}$.

(v) $z - \overline{z}$ is purely imaginary $\forall z \in \mathbb{C}$.

(vi) $\overline{z_1 z_2} = \overline{z}_1\ \overline{z}_2\ \forall z_1, z_2 \in \mathbb{C}$.

(vii) $|z_1 z_2| = |z_1| \cdot |z_2|\ \forall z_1, z_2 \in \mathbb{C}$.

(viii) $|z_1 + z_2| \leq |z_1| + |z_2|\ \forall z_1, z_2 \in \mathbb{C}$ (the triangle inequality).

Proof

(i) – (vi) are straightforward application of the definitions. Each one takes a line or two to justify, and the reader is strongly urged to do that now.

Part (vii) is more interesting. The naive solution might be to let $z_1 = a + ib$ and $z_2 = c + id$ with $a, b, c, d \in \mathbb{R}$. Work out both sides and notice that you get the same answer. That is far too much effort. Here is a neat way to do it.

Observe that $|z_1 z_2|^2 = z_1 z_2 \overline{z_1 z_2}$ and in turn this is $z_1 z_2 \overline{z_1}\ \overline{z_2}$. Now use the commutativity of multiplication to recast our expression as $z_1 \overline{z_1} z_2 \overline{z_2}$ which is $|z_1|^2 |z_2|^2$. Take the square roots of these real numbers and we are done.

Part (viii) has the interesting geometrical interpretation that the length of the longest side of a triangle cannot exceed the sum of the lengths of the other two sides. We give an algebraic proof. Suppose that $\alpha \in \mathbb{C}$. Notice that

$\alpha - \overline{\alpha}$ is purely imaginary so $(\alpha - \overline{\alpha})^2$ is a non-positive real number. Thus $(\alpha - \overline{\alpha})^2 + 4\alpha\overline{\alpha} \leq 4\alpha\overline{\alpha}$. We deduce that $(\alpha + \overline{\alpha})^2 \leq 4|\alpha|^2$. It follows that

$$\alpha + \overline{\alpha} \leq 2|\alpha|.$$

Now put $\alpha = z_1\overline{z}_2$ so our inequality reads

$$z_1\overline{z}_2 + \overline{z}_1 z_2 \leq 2|z_1\overline{z}_2| = 2|z_1||z_2|.$$

Add $z_1\overline{z}_1 + z_2\overline{z}_2$ to each side to obtain that $(z_1 + z_2)(\overline{z}_1 + \overline{z}_2) \leq |z_1|^2 + 2|z_1||z_2| + |z_2|^2$ or in other words

$$|z_1 + z_2|^2 \leq (|z_1| + |z_2|)^2.$$

The result now follows.

□

It is instructive to check the validity of each of the eight parts of the Proposition when it so happens that $z, z_1, z_2 \in \mathbb{R}$.

If $z \in \mathbb{C}$ and $z \neq 0$ let $\hat{z} = z/|z|$ so $|\hat{z}| = 1$ by a geometrical argument or because $\frac{a^2+b^2}{a^2+b^2} = 1$. Multiplication by $z = |z|\hat{z}$ can be accomplished in two stages: first multiply by $|z|$ and then by $\hat{z}$. We know the geometric interpretation of multiplication by a real number, so the problem reduces to understanding multiplication by a complex number u of modulus 1.

Let the paths round the unit circle from 1 to u have lengths $\theta + 2k\pi$ (where $k \in \mathbb{Z}$) measured in an anticlockwise direction. We allow negative lengths because of clockwise paths, and infinitely many lengths because you can go round the circle as many times as you wish in either direction before stopping at u.

Definition 3.7

In this notation, we define the argument of u to be $\arg(u) = \{\theta + 2k\pi \mid k \in \mathbb{Z}\}$. The unique element of $\arg(u)$ in the interval $(-\pi, \pi]$ is called the principal argument of u, and is written capitalized as $\operatorname{Arg}(u)$. If you don't remember interval notation, look back to Section 1.19.

If $z \in \mathbb{C}\setminus\{0\}$, then we have $z = |z|\hat{z}$ for a unique $\hat{z}$ and notice that $|\hat{z}| = 1$. We define $\arg(z)$ and $\operatorname{Arg}(z)$ to be $\arg(\hat{z})$ and $\operatorname{Arg}(\hat{z})$ respectively. The length θ is also the size of the angle shown in Figure 3.3 measured in radians.

We now have the language to address our geometric problem. The map from $\mathbb{C}$ to $\mathbb{C}$ defined by multiplication by u of modulus 1 corresponds to a rotation about the origin through $\operatorname{Arg}(u)$ – and we will have to justify this bold claim. Multiplication by 1 is a trivial rotation through $0 = \operatorname{Arg}(1)$. Multiplication by i

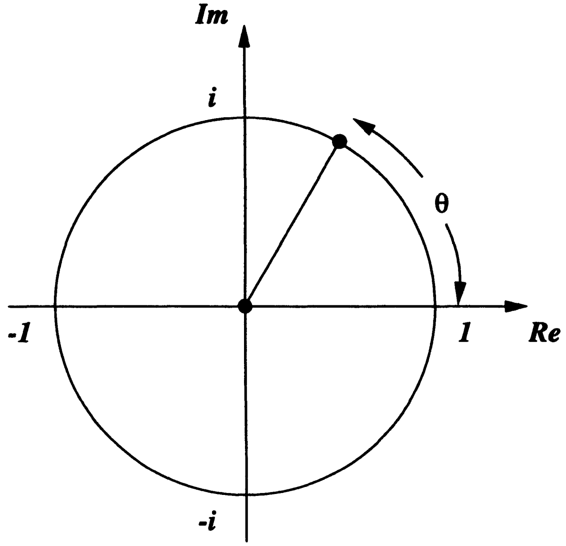

Fig. 3.3. Angle measure defined by arc length

rotates a point about the origin anti-clockwise through $\pi/2$; to see this simply inspect the effect of multiplying $x + iy$ by i as $x + iy$ lives in each of the four quadrants in turn. However, $\mathrm{Arg}(i) = \pi/2$ so all is well so far. We have verified this assertion about u inducing a rotation through $\mathrm{Arg}(u)$ when u happens to be 1 or i, but that is just two cases, and there are infinitely many other cases to worry about.

The key is provided by the distributive law of complex multiplication. Suppose that $u \in \mathbb{C}$ is such that $|u| = 1$. We consider the map $\mathrm{mul}_u : \mathbb{C} \to \mathbb{C}$. We claim that it is an isometry. Consider an arbitrary pair $\alpha, \beta \in \mathbb{C}$. We must compare $|\alpha - \beta|$ with $|\mathrm{mul}_u(\alpha) - \mathrm{mul}_u(\beta)|$. However, the distributive law tells us that $u(\alpha - \beta) = u\alpha - u\beta$ and taking moduli we have

$$|\mathrm{mul}_u(\alpha) - \mathrm{mul}_u(\beta)| = |u||\alpha - \beta| = |\alpha - \beta|$$

because $|u| = 1$. Thus multiplication by u preserves distances.

We want to show that multiplication by u is a rotation through $\mathrm{Arg}(u)$ radians in the Argand diagram. Now $\arg(ui) = \{\mathrm{Arg}(u) + \pi/2 + 2k\pi \mid k \in \mathbb{Z}\}$ and $\arg(i) = \{\pi/2 + 2k\pi \mid k \in \mathbb{Z}\}$. Of course $|ui| = |u||i| = 1$. This is good news, since it means that multiplication by u rotates i anti-clockwise through $\mathrm{Arg}(u)$.

We finish off the proof using geometrical arguments. Suppose that z is any complex number, then z determines a triple $(|z - 1|, |z - 0|, |z - i|)$. These are the three distances from z to 1, 0 and i respectively. The important point is that no other $z \in \mathbb{C}$ yields the same three numbers in that order. This is a nice geometrical exercise in drawing circles, and the reader should fill in the details. It amounts to showing that three circles whose centres are not concurrent cannot meet in more than one point.

Now, $(|uz-u|, |uz-0|, |uz-ui|) = (|z-1|, |z-0|, |z-i|)$ since multiplication by u preserves distances. The unique point w satisfying the condition $(|w-u|, |w-0|, |w-ui|) = (|z-1|, |z-0|, |z-i|)$ has modulus $|z|$ and $\arg(w) = \{\operatorname{Arg}(u) + \operatorname{Arg}(z) + 2k\pi \mid k \in \mathbb{Z}\}$, again by a geometrical argument. The idea is that the triangle with vertices at 1, i and 0 is rotated about 0 through $\operatorname{Arg}(u)$, and therefore so is the quadrilateral with vertices at 1, i, 0 and z.

Finally, we conclude that multiplication by u has the effect of rotating points in the Argand diagram about the origin through $\operatorname{Arg}(u)$.

EXERCISES

3.6 Suppose that $a, b \in \mathbb{C}$.

(a) Show that $|a+b|^2 = |a|^2 + |b|^2 + a\bar{b} + \bar{a}b$.

(b) Show that $|a+b|^2 + |a-b|^2 = 2|a|^2 + 2|b|^2$.

(c) Give an elegant geometrical interpretation of part (b) involving a parallelogram, and the lengths of its sides and diagonals.

3.7 Suppose that $z = c + id \in \mathbb{C}$ with $c, d \in \mathbb{R}$. Recall that $|z|^2 = z\bar{z} = c^2 + d^2$.

(a) Suppose that m and n are natural numbers and each of them is the sum of two perfect squares. Prove that the natural number mn is also the sum of two perfect squares. (Definition: A perfect square is the square of an integer.)

(b) Express 97,000,097 as the sum of two perfect squares without the aid of a calculator.

(c) Suppose that $a, b, c, d \in \mathbb{R}$. Prove that

$$ac + bd \leq \sqrt{a^2 + b^2}\sqrt{c^2 + d^2}.$$

3.4 Sine, Cosine and Polar Form

We now define the functions sine and cosine, so the reader should temporarily forget everything hitherto known about these functions. Take any $\theta \in \mathbb{R}$ and let u be the unique complex number such that $|u| = 1$ and $\theta \in \arg(u)$. We define maps sin: $\mathbb{R} \to \mathbb{R}$ and cos: $\mathbb{R} \to \mathbb{R}$. In the notation we have just set up, $u = \cos\theta + i\sin\theta$. Notice that this definition is consistent with the definitions of cosine and sine of acute angles using right-angled triangles. For acute angles the length of the circular path from 1 to u is same thing as the angle (or more precisely the measure of the angle) in radians. This is the reason why radians are a great way to measure angles.

Of course if θ_1 and θ_2 differ by an integer multiple of 2π, then $\sin\theta_1 = \sin\theta_2$ and $\cos\theta_1 = \cos\theta_2$. The triangular definitions of sine and cosine were fine for angles in the interval $[0, \pi/2)$ radians, but for larger angles that business about the angles of a triangle summing to π is a bit of a problem, as are negative angles. You may have seen the following formulas before, but it is possible that the only proofs you have seen assumed that $\alpha, \beta, \alpha + \beta \in [0, \pi/2)$.

Proposition 3.5

Suppose that $\alpha, \beta \in \mathbb{R}$, then

(i) $\cos(\alpha + \beta) = \cos\alpha\cos\beta - \sin\alpha\sin\beta$, and

(ii) $\sin(\alpha + \beta) = \sin\alpha\cos\beta + \sin\beta\cos\alpha$.

Proof

Since sine and cosine have been defined geometrically, we need a geometric proof. Let $u = \cos\alpha + i\sin\alpha$ and $v = \cos\beta + i\sin\beta$. Recall that α and β are the lengths of paths from the origin to u and v respectively round the unit circle. The geometric effect of multiplying by u and then by v is to rotate the Argand diagram about the origin through α and then through β. The combined effect will be to rotate it through $\alpha + \beta$. The only complex number of modulus 1 which contains $\alpha + \beta$ in its set of arguments is $\cos(\alpha + \beta) + i\sin(\alpha + \beta)$. Thus we may equate real and imaginary parts of

$$\cos(\alpha + \beta) + i\sin(\alpha + \beta) = (\cos\alpha + i\sin\alpha)(\cos\beta + i\sin\beta).$$

□

There are various well-known (or ought to be well-known) trigonometric formulas which follow from Proposition 3.5. Among them is

$$\tan(\alpha+\beta) = \frac{\sin(\alpha+\beta)}{\cos(\alpha+\beta)} = \frac{\sin\alpha\cos\beta+\sin\beta\cos\alpha}{\cos\alpha\cos\beta-\sin\alpha\sin\beta} = \frac{\tan\ \alpha+\tan\beta}{1-\tan\alpha\tan\beta}.$$

Perhaps the most celebrated variation on this theme is the result often known as De Moivre's theorem.

Proposition 3.6 (De Moivre)

Suppose that $\theta \in \mathbb{R}$ and $n \in \mathbb{N}$, then

$$(\cos\theta + i\sin\theta)^n = \cos n\theta + i\sin n\theta.$$

Proof

We prove this result by induction. The case $n = 1$ is trivially true. Assume that the result holds when $n = m \in \mathbb{N}$ and try to deduce that the result holds when $n = m + 1$.

Now

$$\begin{aligned}(\cos\theta + i\sin\theta)^{m+1} &= (\cos\theta + i\sin\theta)^m(\cos\theta + i\sin\theta)\\ &= (\cos m\theta + i\sin m\theta)(\cos\theta + i\sin\theta)\end{aligned}$$

by inductive hypothesis. Next use Proposition 3.5, so

$$\begin{aligned}(\cos\theta + i\sin\theta)^n &= (\cos m\theta\cos\theta - \sin m\theta\sin\theta) + i(\sin m\theta\cos\theta + \sin\theta\cos m\theta)\\ &= \cos(m+1)\theta + i\sin(m+1)\theta\end{aligned}$$

as required. Thus by mathematical induction the proof is complete.

□

Definition 3.8

Suppose that $z \in \mathbb{C}$ and $z \neq 0$. Let $r = |z|$ and choose $\theta \in \arg(z)$. The *polar form* of z is its expression as $r(\cos\ \theta + i\sin\ \theta)$.

Note that r is uniquely determined by z but that there is ambiguity in the choice of θ since you may add or subtract an arbitrary integer multiple of 2π.

If you want to multiply or divide non-zero complex numbers, it is often best to put them in polar form $r(\cos\ \alpha + i\sin\ \alpha)$ and $s(\cos\ \beta + i\sin\ \beta)$ so that their product is $rs(\cos(\alpha+\beta) + i\sin(\alpha+\beta))$, as shown in Figure 3.4 and their quotient is $(r/s)(\cos(\alpha-\beta) + i\sin(\alpha-\beta))$.

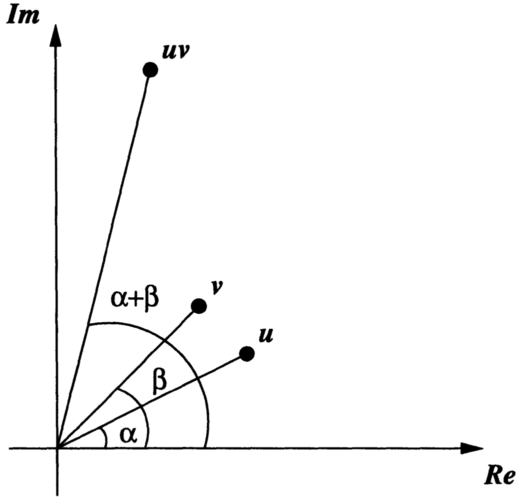

Fig. 3.4. Geometry of complex multiplication. Note that $|\mathbf{u}||\mathbf{v}| = |\mathbf{uv}|$

Note that the inverse of $r(\cos\theta + i\sin\theta)$ is $(1/r)(\cos(-\theta) + i\sin(-\theta))$. However, from the geometry of the Argand diagram we see that $\cos(-\theta) = \cos\theta$ and $\sin(-\theta) = -\sin\theta$ for all $\theta \in \mathbb{R}$. Thus this inverse can also be written as $r^{-1}(\cos\theta - i\sin\theta)$. Thus complex conjugates and inverses are related. However, this is clear because $z\overline{z} = |z|^2$ so when $z \neq 0$ we have $z^{-1} = |z|^{-2}\overline{z}$.

The fact that the functions sine and cosine repeat every 2π is very important.

Definition 3.9

Suppose that $f : \mathbb{R} \to X$ where X is any set. We say f is *periodic* if there exists a positive $p \in \mathbb{R}$ such that $f(x + p) = f(x)$ $\forall x \in \mathbb{R}$, and p is called a period of the function. If f is periodic, and there exists a smallest period, the smallest period is called the fundamental period.

We leave it as an easy exercise to show that sin: $\mathbb{R} \to \mathbb{R}$ and cos: $\mathbb{R} \to \mathbb{R}$ have fundamental period 2π (actually people are sloppy about this, and might easily say "have period 2π").

Notice that $\sin(x + \pi) = -\sin x$ and $\cos(x + \pi) = -\cos x$ for all $x \in \mathbb{R}$. Thus $\tan x = \sin x/\cos x$ has period π, and in fact this is its shortest period. There is a minor problem with tan because it blows up (jargon - it has a *singularity*) at odd integer multiples of $\pi/2$ so it is not really a function from $\mathbb{R}$

to $\mathbb{R}$; an easy solution is to invent a meaningless symbol such as ∞ and decide that at odd integer multiples of π the value of tan will be ∞ rather than the undefined ratio formerly suggested. That way tan becomes a function again, and tan: $\mathbb{R} \to \mathbb{R} \cup \{\infty\}$. Alternatively you can eject odd integer multiples of $\pi/2$ from the domain of tan.

Definition 3.10

Let $f : \mathbb{R} \to \mathbb{R}$ (or $\mathbb{C}$) be a map. We say f is *even* if $f(x) = f(-x)$ for every $x \in \mathbb{R}$. We say f is *odd* if $f(x) = -f(-x)$ for every $x \in \mathbb{R}$.

Thus cosine is an even function but sine is an odd function.

EXERCISES

3.8 For all real θ it happens to be true that $\cos 4\theta = 8\cos^4\theta - 8\cos^2\theta + 1$.

(a) Verify this formula in the cases when $\theta = 0, \pi/2, \pi/3$ and π.

(b) Prove the formula is valid for all $\theta \in \mathbb{R}$ by considering the instance

$$\cos 4\theta + i\sin 4\theta = (\cos\theta + i\sin\theta)^4$$

of De Moivre's theorem.

3.9 The "integer part" of a real number x is written $\lfloor x \rfloor$, and is defined to be the largest integer not greater than x. Define $\gamma : \mathbb{R} \to \mathbb{R}$ by $\gamma(x) = x - \lfloor x \rfloor$.

(a) Sketch a graph of this function.

(b) Find all of the periods of γ.

(c) Does γ have a fundamental period, and if so what is it?

3.10 Consider the function $\Xi : \mathbb{R} \to \mathbb{R}$ defined by $\Xi(r) = 1$ if $r \in \mathbb{Q}$ but $\Xi(r) = 0$ if $r \notin \mathbb{Q}$.

(a) Show that Ξ is a periodic function.

(b) Find all of the periods of Ξ.

(c) Does Ξ have a fundamental period, and if so what is it?

3.11 Consider a function $f : \mathbb{R} \to \mathbb{R}$.

(a) Show that if f is both even and odd, then it is the constant function which always takes the value 0.

(b) Show that the function defined by the formula $e(x) = (f(x) + f(-x))/2$ is an even function.

(c) Show that f may be written as a sum of an even function and an odd function.

(d) Show that f may be written as a sum of an even function and an odd function in only one way. (Thus we may talk about the *even part* and the *odd part* of a function such as f.)

(e) Discuss your answers in the case that f is an even function.

(f) Suppose that $p : \mathbb{R} \to \mathbb{R}$ is defined by $p(x) = 2x^5 - x^4 + 3x^2 + \sin(-x^2) + \cos(x^3)$ $\forall x \in \mathbb{R}$. Describe the even and odd parts of p.

3.5 *e*

We now introduce the number e, which is a number of mysterious mathematical ubiquity fourth only to $0, 1$ and π. This number is about 2.718, and has the property that $\frac{d}{dx}e^x = e^x$. Here x is supposed to be a real variable and it is time to worry a little. What is a^b when $a, b \in \mathbb{R}$? To concentrate our minds, what is e^π? You can try to work it out, but you will only be using approximations to e and π. The trouble with e and π is that they are real numbers but not rational numbers, and since we haven't got a proper definition of real numbers, the ice is a little thin. However, with a little goodwill perhaps we can overcome this difficulty.

We will regard it as given that for every positive $r \in \mathbb{R}$ and any $n \in \mathbb{N}$, there exists a unique positive $s \in \mathbb{R}$ such that $s^n = r$.

So, what could we be driving at when writing a^b with $a, b \in \mathbb{R}$? Well, even if b isn't rational, it can be approximated arbitrarily well by rationals. For example, $22/7$ is a good stab at π and $355/113$ is even better. Thus if you want to work out e^π, then it should be close to $(272/100)^{22/7}$ and closer yet to $(271828/100000)^{355/113}$. The latter is the real 113th root of $(271828/100000)^{355}$. As you refine the approximations to e and π the value of the approximate exponentiation settles down towards a number. That number is e^π. That is a bit fluffy, and all sorts of questions arise about what a good approximation actually is, and what "settles down towards" means, but I hope the reader will go along with this for now. These issues will be explored from Chapter 6 onwards.

So, with a fair wind we know what the expression e^x means for real x. If we write $e^{1/2}$ we definitely mean the positive real number whose square is e. We haven't given a formal construction of the calculus either, but it may help to see how to get at e from a calculus perspective, and to see why $d/dx(e^x) = e^x$.

We first investigate the natural logarithm function log: $(0, \infty) \to \mathbb{R}$ defined by $\log(x) = \int_1^x 1/t\,dt$ (see Figure 3.5). This function is sometimes written ln or $\log_e$ to emphasize its connection with the yet undefined number e. Suppose that $y_1, y_2 > 0$, then

$$\begin{aligned}\log(y_1y_2) &= \int_1^{y_1y_2} \frac{1}{x}dx = \int_1^{y_1} \frac{1}{x}dx + \int_{y_1}^{y_1y_2} \frac{1}{x}dx\\ &= \log(y_1) + \int_1^{y_2} \frac{1}{v}dv.\end{aligned}$$

In the second integral we have made the change of variable $v = x/y_1$. Of course, the name of the variable of integration is irrelevant so

$$\log(y_1y_2) = \int_1^{y_1} \frac{1}{x}dx + \int_1^{y_2} \frac{1}{x}dx = \log(y_1) + \log(y_2) \tag{3.2}$$

which is a wonderful formula and holds for all $y_1, y_2 > 0$.

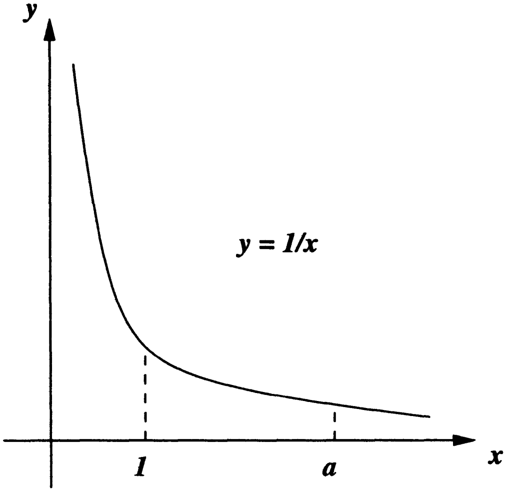

Fig. 3.5. $\log a = \int_1^a 1/x\,dx$

Notice that $\log(1) = 0$. Now log is a strictly increasing function of x (because of its definition as an area under the graph of a positive function (see

Figure 3.5), and so is injective. By induction on n it follows from Equation (3.2) that $\log(nx) = n\log(x)\ \forall n \in \mathbb{N}$. Since $\log(2) > 0$ this means that for natural numbers m the quantity $\log(2^m) = m \log(2)$ can be made arbitrarily large by choosing bigger and bigger m. For $x > 0$ we have

$$0 = \log(1) = \log(x \cdot (1/x)) = \log(1/x) + \log(x)$$

so $\log(1/x) = -\log(x)$. This means that for sufficiently large numbers x we have that $\log(1/x)$ will assume arbitrarily large negative values.

Thus log is a continuous function (there are no jumps in its values) which assumes arbitrarily large positive and negative values, and therefore log takes all real values. It is therefore surjective. We already knew it was injective so it is bijective, and has an inverse which we call $\exp:\mathbb{R} \to (0, \infty)$.

Definition 3.11

The number e is the unique natural number with the property that $\log(e) = 1$, or equivalently $e = \exp(1)$.

If you want to see e concretely, observe that it is alive and well in a picture of the integral

$$\int_1^e \frac{1}{t} dt = 1.$$

Next we study the function exp. In fact $\exp(x)$ will turn out to be the same thing as e^x, but we don't know that yet!

Let $x_1 = \log(y_1)$ and $x_2 = \log(y_2)$ and substitute into Equation (3.2) to get $\log(y_1 y_2) = x_1 + x_2$. Apply the map exp to each side so $y_1 y_2 = \exp(x_1 + x_2)$. However, since log and exp are inverse maps $y_1 = \exp(x_1)$ and $y_2 = \exp(x_2)$. We conclude that

$$\exp(x_1 + x_2) = \exp(x_1) \cdot \exp(x_2)\ \forall x_1, x_2 \in \mathbb{R}.$$

It follows by induction that $\exp(n) = \exp(1)^n = e^n$ for all natural numbers n. Also when $n \in \mathbb{N}$ we have $\exp(-n)\exp(n) = \exp(0) = 1$, so $\exp(-n) = e^{-n}$. Thus $e^z = \exp(1)^z = \exp(z)$ whenever $z \in \mathbb{Z}$. That is good progress, but we need equality whenever $z \in \mathbb{R}$. We next examine $\exp(1/n)$ when $n \in \mathbb{N}$. Now $\exp(1/n)$ is real and positive, and $(\exp(1/n))^n = \exp(n/n) = \exp(1) = e$. Thus $\exp(1/n)$ must be the unique real n-th root of e. Next consider the rational m/n with $m \in \mathbb{Z}$ and $n \in \mathbb{N}$. Now $\exp(m/n) = \exp(1/n)^m = \exp(1)^{m/n} = e^{m/n}$. We now know that the functions defined by the formulas $\exp(x)$ and e^x coincide at all rational values of x. However, these functions are continuous, and all real numbers may be approximated by rational numbers, so $\exp(x)$ and e^x define the same function from $\mathbb{R}$ to $(0, \infty)$.

The celebrated property of e is that $d/dx(e^x) = e^x$. This follows because $e^x = \exp(x)$. Put $y = \exp(x)$ so $\log(y) = x$. Differentiating we obtain that $1/y \cdot dy/dx = 1$ so $dy/dx = y$. In other words

$$\frac{d}{dx}\exp(x) = \exp(x)$$

and we are done.

We next want to define e^z for an arbitrary $z \in \mathbb{C}$, and we need to make sure that we select a definition which is consistent with the important properties of the function given by e^x when $x \in \mathbb{R}$.

Definition 3.12

Suppose that $z = u + i\theta$ where $u, \theta \in \mathbb{R}$. We write e^z to mean $e^u \cdot (\cos\theta + i\sin\theta)$.

This looks rather brave. Our definition of e^z when $z \in \mathbb{C}$ is consistent with this when z happens to be real since $\cos 0 + i\sin 0 = 1 + 0 = 1$. In fact there are very compelling reasons for defining e^z for $z \in \mathbb{C}$ as we have done. If you go on to study complex analysis you will learn theorems which tell you that this is the best possible definition of e^z. Notice that when $z = i\theta$ is purely imaginary we have $e^{i\theta} = \cos\theta + i\sin\theta$.

A minor fuss is traditional. Our definition of e^z yields the celebrated equation $e^{\pi i} = \cos\pi + i\sin\pi = -1$ or equivalently $e^{\pi i} + 1 = 0$, an equation relating the seven most interesting objects of human thought. Given the way we have developed the subject this a really a bit of a cheat, since we have cooked the definition to make it happen. If you construct this subject by another route, it is possible to make this equation seem considerably more dramatic. From our point of view, the marvellous thing is that this definition of e^z proves so attractive. We immediately supply solid algebraic evidence of this.

Proposition 3.7

Suppose that $z_1 = u_1 + i\theta_1$ and $z_2 = u_2 + i\theta_2$ with $u_1, u_2, \theta_1, \theta_2 \in \mathbb{R}$. Moreover suppose that $k \in \mathbb{Z}$. The following equations are valid.

(i) $e^{z_1}e^{z_2} = e^{z_1+z_2}$

(ii) $(e^{z_1})^k = e^{kz_1}$.

Proof

(i)

$$\begin{aligned} e^{z_1}e^{z_2} &= e^{u_1}(\cos\theta_1 + i\sin\theta_1)\cdot e^{u_2}(\cos\theta_2 + i\sin\theta_2) \\ &= e^{u_1}e^{u_2}(\cos\theta_1 + i\sin\theta_1)\cdot(\cos\theta_2 + i\sin\theta_2) \\ &= e^{u_1+u_2}(\cos(\theta_1+\theta_2) + i\sin(\theta_1+\theta_2)). \\ &= e^{u_1+u_2}e^{i(\theta_1+\theta_2)} = e^{z_1+z_2}. \end{aligned}$$

(ii) This is little more than De Moivre's theorem for k positive. For $k = 0$ the result is trivial, and the case when $k < 0$ follows from the case $k > 0$ since $(e^z)^{-1} = e^{-z}\forall z \in \mathbb{C}$.

□

Polar Form Revisited

We know that for real θ we have $e^{i\theta} = \cos\theta + i\sin\theta$; it follows that the polar form of a non-zero complex number

$$r(\cos\theta + i\sin\theta)$$

simplifies to $re^{i\theta}$. We have $(re^{i\theta})^{-1} = (r^{-1})e^{-i\theta}$ and

$$re^{i\theta}\cdot se^{i\psi} = (rs)e^{i(\theta+\psi)}.$$

It follows that the rule for division is

$$(re^{i\theta})/(se^{i\psi}) = (rs^{-1})e^{i(\theta-\psi)}.$$

We will return to this topic later when we discuss finding roots of complex numbers.

Since $e^{i\theta} = \cos\theta + i\sin\theta$ we have $e^{-i\theta} = \cos(-\theta) + i\sin(-\theta) = \cos\theta - i\sin\theta$. Thus $e^{i\theta} + e^{-i\theta} = 2\cos\theta$ and $e^{i\theta} - e^{-i\theta} = 2i\sin\theta$. Rearranging these expressions we discover the important facts that

$$\cos\theta = \frac{e^{i\theta} + e^{-i\theta}}{2} \text{ and } \sin\theta = \frac{e^{i\theta} - e^{-i\theta}}{2i}.$$

These handy formulas facilitate the derivation of multiple angle formulas. For example we have

$$\cos(2\theta) + 1 = \frac{e^{i2\theta} + e^{-i2\theta}}{2} + 1$$

$$= \frac{e^{i2\theta} + 2 + e^{-i2\theta}}{2} = \frac{e^{i2\theta} + 2e^{i\theta}e^{-i\theta} + e^{-i2\theta}}{2}$$
$$= \frac{\left(e^{i\theta} + e^{-i\theta}\right)^2}{2} = 2\left(\frac{e^{i\theta} + e^{-i\theta}}{2}\right)^2 = 2\cos^2\theta.$$

We recover the standard result that for any $\theta \in \mathbb{R}$ we have $\cos 2\theta = 2\cos^2\theta - 1$.

This characterization of sine and cosine in terms of $e^{i\theta}$ gives us a wonderful opportunity to define $\cos : \mathbb{C} \to \mathbb{C}$ and $\sin : \mathbb{C} \to \mathbb{C}$ in such a way as to extend the usual real functions. We simply put $\cos z = (e^{iz} + e^{-iz})/2$ and $\sin z = (e^{iz} - e^{-iz})/2i$.

EXERCISES

3.12 Prove that $\cos 4\theta = 8\cos^4\theta - 8\cos^2\theta + 1$ for all real θ by using the expression we have developed for $\cos\psi$ in terms of $e^{i\psi}$.

3.13 Find all $z \in \mathbb{C}$ such that $\sin z = 2$.

3.6 Hyperbolic Sine and Hyperbolic Cosine

Inspired by the relationship between $e^{i\theta}$, $\sin\theta$ and $\cos\theta$, we examine what happens if you replace $e^{i\theta}$ by e^{θ} for $\theta \in \mathbb{R}$.

Given any function $f : \mathbb{R} \to \mathbb{C}$ we can express it as a sum of an even function $e_f(x)$ and an odd function $o_f(x)$ by putting $e_f(x) = (f(x) + f(-x))/2$ and $o_f(x) = (f(x) - f(-x))/2$. Cosine and sine are the even and odd parts of the function given by the formula $e^{i\theta}$.

We define $\cosh t$ and $\sinh t$ to be the even and odd parts of the function defined by the formula e^t. Thus

$$\cosh t = \frac{e^t + e^{-t}}{2} \text{ and } \sinh t = \frac{e^t - e^{-t}}{2}.$$

Notice that

$$\cosh^2 t - \sinh^2 t = \left(\frac{e^t + e^{-t}}{2}\right)^2 - \left(\frac{e^t - e^{-t}}{2}\right)^2 = \frac{2}{4} - \frac{-2}{4} = 1$$

so $\cosh^2 t - \sinh^2 t = 1$, a formula which is chillingly similar to Pythagoras's theorem which asserts that $\cos^2 t + \sin^2 t = 1$.

If you consider the points in $\mathbb{R}^2$ of the form $(\cos t, \sin t)$ as t varies over $\mathbb{R}$ you obtain the unit circle (traced over and over again). On the other hand, the points $(\cosh t, \sinh t)$ describe one of the two branches of the hyperbola consisting of points (x, y) satisfying $x^2 - y^2 = 1$. The other branch is described by $(-\cosh t, \sinh t)$ as t ranges over $\mathbb{R}$. Hence the name *hyperbolic functions* for $\cosh t$, $\sinh t$, and their kindred functions such as the hyperbolic tangent, cotangent, secant and cosecant defined by

$$\tanh t = \frac{\sinh t}{\cosh t}, \ \coth t = \frac{1}{\tanh t},$$

$$\operatorname{sech} t = \frac{1}{\cosh t}, \ \operatorname{cosech} t = \frac{1}{\sinh t}.$$

How you pronounce these is a matter of personal taste, but I use the speech of a stage drunk; *shine, coshine, thangent, cothangent, shecant* and *coshecant* – with written abbreviations *sinh, cosh, tanh, coth, sech* and *cosech.*

If you know something about the geometric curves known as conic sections, you might suppose that there would be a class of elliptic functions corresponding to a parameterization of an ellipse, or a class of parabolic functions corresponding to a parabola. This is not the case. There are things called elliptic functions but they are something completely different. If you want to parameterize an ellipse, you can do it using *circular functions.* The circular functions are just the functions which arise naturally when studying the circle: sine, cosine, tangent, cotangent, secant and cosecant. For example, as t varies over $\mathbb{R}$, the points of the form $(2\cos t, 3\sin t)$ trace an ellipse of points in $\mathbb{R}^2$ satisfying $9x^2 + 4y^2 = 36$. To parameterize a parabola is even easier, for example $\{(t, t^2) \mid t \in \mathbb{R}\}$ does the job.

Notice that

$$\frac{d}{dx}\cosh x = \frac{d}{dx}\left(\frac{e^x + e^{-x}}{2}\right) = \frac{e^x - e^{-x}}{2} = \sinh x$$

and similarly $\frac{d}{dx}\sinh x = \cosh x$. Combining this with $\cosh^2 x - \sinh^2 x = 1$, yields the interesting result that

$$\frac{d}{dx}\sinh x = \sqrt{1 + \sinh^2 x}.$$

Thus $\sinh x = \int_0^x \sqrt{1 + \sinh^2 t}\, dt$. We can twist this around, since it is easy to show that $\sinh : \mathbb{R} \to \mathbb{R}$ is an injective function, and so has an inverse function arcsinh. Now, just as arcsin leads to a standard integral, so does arcsinh. Here we go.

Suppose that $y = \operatorname{arcsinh} x$ so $\sinh y = x$. Differentiate with respect to x to obtain that $\cosh y \frac{dy}{dx} = 1$. Thus

$$\frac{dy}{dx} = \frac{1}{\cosh y} = \frac{1}{\sqrt{1 + \sinh^2 y}} = \frac{1}{\sqrt{1 + x^2}}.$$

Thus

$$\operatorname{arcsinh} t = \int_0^t \frac{1}{\sqrt{1+x^2}}dx.$$

Perhaps this reminds you of

$$\arcsin t = \int_0^t \frac{1}{\sqrt{1-x^2}}dx,$$

which is a more feeble formula because it is only valid for $t \in (-1,1)$. The robust formula for $\operatorname{arcsinh} t$ is valid for all real t. Why is one so much better than the other? Well, the simple explanation is that $\sinh(x)$ assumes all real values but $\sin x$ takes values only in $[-1,1]$. However, it is interesting to examine the integrands (the formulas nestling after the integral signs and before the dx). The expression $\sqrt{1+x^2}$ never vanishes when $x \in \mathbb{R}$ and, since we are integrating along a portion of the real axis, this is good news. However, the (dodgy) function defined by $\frac{1}{\sqrt{1-x^2}}$ isn't quite a function from $\mathbb{R}$ to $\mathbb{R}$ at all. It has a singularity when $x = \pm 1$ because $\sqrt{1-x^2}$ vanishes there. What is happening is this; the formula $\arcsin t = \int_0^t \frac{1}{\sqrt{1-x^2}}dx$ is valid when $t = 0$, and for small t. It does its best to be true for all real t but when t reaches a singularity at ± 1 all is lost.

What we have been discussing is not a special situation. The study of functions which are nice in most places but have singularities is central to the study of calculus in the context of $\mathbb{C}$.

On a less philosophical note, we observe that we are only touching upon hyperbolic functions and their associated inverse functions. Whenever there is a valid formula involving circular functions, there will be an analogous one involving hyperbolic functions. The reader who needs drill should find a calculus text which contains a substantial section on hyperbolic functions, and do the exercises. We append a couple below to get you started.

EXERCISES

3.14 Express the functions $\sinh(2x)$, $\cosh(2x)$, and $\tanh(2x)$ in terms of $\sinh(x)$, $\cosh(x)$, and $\tanh(x)$ to get formulas which remind you of the ones for the corresponding circular functions.

3.15 Show that $\tanh(x) \in (-1,1)\, \forall x \in \mathbb{R}$.

3.7 Integration Tricks

This section does not develop the theory of complex numbers, but rather illustrates how they can be used to evaluate integrals. What you see here is the protrusion of an iceberg. The reasoning here is not supposed to be completely rigorous – but rather we introduce a technique which is very useful and can be properly justified.

Suppose that $f : \mathbb{R} \to \mathbb{C}$ and suppose that the value of the function at the point $x \in \mathbb{R}$ is written $f(x)$. Though we have not formally developed the calculus, we sketch how to proceed. Write $f(x) = u(x) + iv(x)$ where u and v are real-valued functions of a real variable x. Define $\frac{df}{dx} = \frac{du}{dx} + i\frac{dv}{dx}$, so differentiation of real and imaginary parts is done separately. Equally well, integration is defined by

$$\int_a^b f(t)dt = \int_a^b u(t)dt + i\int_a^b v(t)dt.$$

Just as in the real case, differentiation and integration are inverse procedures. It turns out that differentiation and integration formulas valid for real constants are equally good when complex constants are involved.

Recall that whenever m is a real constant

$$\frac{d}{dx}e^{mx} = me^{mx}.$$

Consider the function $f : \mathbb{R} \to \mathbb{C}$ defined by $f(\theta) = e^{i\theta} = \cos\theta + i\sin\theta$. Now, differentiation with respect to θ can be shown to respect addition and multiplication by constants (even complex ones – have faith).

$$\begin{aligned}\frac{d}{d\theta}e^{i\theta} &= \frac{d}{d\theta}(\cos\theta + i\sin\theta) = \frac{d}{d\theta}\cos\theta + i\frac{d}{d\theta}\sin\theta \\ &= -\sin\theta + i\cos\theta = i(\cos\theta + i\sin\theta) = ie^{i\theta}.\end{aligned}$$

This is exactly as we would wish, and incidentally is yet more evidence that our definition of e^z for $z \in \mathbb{C}$ was sensible.

We now demonstrate how powerful this technique can be for evaluating integrals.

Example 3.3

Suppose that you are confronted with $I = \int_0^\pi e^x \sin(x)dx$. Now, a charming method of evaluation is to integrate by parts twice (do it!), but instead we could observe that $\sin x$ is the imaginary part of e^{ix}. Thus

$$I = \operatorname{Im}(\int_0^\pi e^{x(1+i)}dx) = \operatorname{Im}\left(\left[\frac{e^{x(1+i)}}{1+i}\right]_0^\pi\right) = \operatorname{Im}\left(\frac{e^\pi e^{\pi i} - 1}{1+i}\right)$$

$$= \operatorname{Im}\left(\frac{(-e^{\pi}-1)(1-i)}{(1+i)(1-i)}\right) = \frac{e^{\pi}+1}{2}.$$

3.8 Extracting Roots and Raising to Powers

If you work exclusively in $\mathbb{R}$, then you can't find square roots of negative numbers. If $x > 0$, then there are exactly two real numbers which square to y, the positive one called $x^{1/2}$ and the negative one $-x^{1/2}$. There is only one number which has 0 as its square, and that is 0.

Any $z \in \mathbb{C}\setminus\{0\}$ can be written in polar form as $re^{i\theta}$ for any $\theta \in \arg(z)$. For each natural number n we seek all $\eta \in \mathbb{C}$ such that $\eta^n = z = re^{i\theta}$. Now, any such η must be non-zero. Also $|\eta|^n = r$ so $|\eta| = r^{1/n}$. Choosing $\psi \in \arg(\eta)$ we find that $\eta^n = z$ exactly when $re^{i\theta} = (r^{1/n}e^{i\psi})^n = re^{in\psi}$ which happens if and only if $\theta - n\psi$ is an integer multiple of 2π.

Thus $\eta = r^{1/n}e^{i\psi}$ will be an n-th root of z if and only if $\psi = (\theta + 2k\pi)/n$ for some integer k. Now although there are infinitely many elements of this set, there are exactly n possible values of η. The reason is that values of ψ give rise to the same η exactly when they differ by an integral multiple of $2\pi/n$.

Thus the n-th roots of z are $r^{1/n}e^{i(\theta+2k\pi)/n}$ for $k = 0, 1, \ldots, n-1$. Any non-zero complex number therefore has exactly n different n-th roots.

As usual we seek a geometric interpretation in the Argand diagram. The n numbers $r^{1/n}e^{i(\theta+2k\pi)/n}$ $(k = 0, 1, \ldots, n-1)$ all have modulus $r^{1/n}$ so they are on a circle centred at the origin of radius $r^{1/n}$. Their arguments are evenly spaced (with a common difference of $2\pi/n$) so they form the vertices of a regular n-gon. This determines the positions of these points up to a rotation about the origin. The last uncertainty is removed when we observe that $r^{1/n}e^{i\theta/n}$ is a vertex.

We are going to give a definition of w^z where $w, z \in \mathbb{C}$ and w is not 0. Now, we already have a definition of e^z where e is the base of natural logarithms. Our definition must be consistent with that.

Write w in its infinitely many polar forms as $e^u e^{i(\theta+2k\pi)}$ as k ranges over $\mathbb{Z}$. We try (unsuccessfully) to define w^z to be $e^{uz} \cdot e^{i(\theta+2k\pi)z}$ $\forall k \in \mathbb{Z}$; this usually has infinitely many values, rather than just one value, which is what we have in mind. However, when $z \in \mathbb{Z}$, then $e^{2k\pi iz} = 1$ and this first attempt at a definition works because then it defines just one value for w^z.

The way out in general is to *insist* that $k = 0$ and $\theta \in (-\pi, \pi]$. This unnatural intrusion will cause a few problems for us, but we have no choice.

First the good news.

Proposition 3.8

Suppose that $w \in \mathbb{C}$ and $w \neq 0$.

(i) $w^{z_1+z_2} = w^{z_1} \cdot w^{z_2}\ \forall z_1, z_2 \in \mathbb{C}$.

(ii) $(w^z)^k = w^{kz}\ \forall z \in \mathbb{C},\ \forall k \in \mathbb{Z}$.

Proof

(i) Put $w = e^u e^{i\theta}$ in polar form with the restriction on θ. Now $w^{z_1+z_2} = e^{u(z_1+z_2)}e^{i\theta(z_1+z_2)}$ using Proposition 3.7 (i) we have

$$w^{z_1+z_2} = e^{uz_1}e^{i\theta z_1}e^{uz_2}e^{i\theta z_2} = w^{z_1} \cdot w^{z_2}.$$

(ii) Now we use part (ii) of Proposition 3.7 and

$$(w^z)^k = (e^{uz}e^{i\theta z})^k = e^{ukz}e^{ikz\theta} = (e^u e^{i\theta})^{kz} = w^{kz}.$$

□

Now the bad news. The problem arises when you study $(w_1w_2)^z$. In general this is not the same thing as $w_1{}^z w_2{}^z$, and, as we should expect, our intrusion on the value of θ in the definition of w^z has caused pain.

EXERCISES

3.16 Suppose that $z \in \mathbb{C}$ has the property that for all $w_1, w_2 \in \mathbb{C}$ we have $(w_1w_2)^z = w_1{}^z w_2{}^z$. What can you say about z?

3.9 Logarithm

There is an obvious connection between logarithms and raising numbers to a power in the context of the reals. We have already had a minor problem with complex exponentiation. No doubt this means there will be problems with complex logarithms.

Consider the function exp: $\mathbb{R} \to (0, \infty)$ defined by $\exp(x) = e^x$. Recall that this is a bijective function and its inverse map is called the (natural) logarithm, and is written log, $\log_e$ or ln.

Now, if $z = u + iv$ with $u, v \in \mathbb{R}$, then $e^z = e^u e^{iv}$. We can recycle the notation to define a map exp: $\mathbb{C} \to \mathbb{C}^* = \mathbb{C} \setminus \{0\}$ by $\exp(z) = e^z\ \forall z \in \mathbb{C}$. This is certainly a surjective function since any non-zero complex number can be written as $e^u e^{i\theta}$ for some $u, \theta \in \mathbb{R}$.

Suppose that $z' = u' + iv'$ with $u', v' \in \mathbb{R}$, then in order to examine the injectivity (or otherwise) of this complex version of exp, we consider the equation $e^z = e^{z'}$ or equivalently the pair of equations $e^u = e^{u'}$ and $e^{iv} = e^{iv'}$. The injectivity of exponentiation of real numbers yields the promising start that $u = u'$. However, $e^{iv} = e^{iv'}$ holds exactly when v and v' differ by an integer multiple of 2π. In other words, $e^z = e^{z'}$ if and only if $z - z'$ is an integer multiple of $2\pi i$.

We can render exp a bijective function by restricting its domain to an infinitely wide horizontal strip in the Argand diagram of height 2π containing one but not the other of its boundaries. For example $-\pi < \mathrm{Im}(z) \leq \pi$.

If $z \in \mathbb{C}^*$ we define (capital L) Log z to be the unique $w \in \{\alpha \mid \alpha \in \mathbb{C},\ \mathrm{Im}(\alpha) \in [-\pi, \pi)\}$ with the property that $e^w = z$. Now $w = \log(r) + i\mathrm{Arg}(z)$. By construction $\exp(\mathrm{Log}(z)) = z$ for all $z \in \mathbb{C}^*$. However, it is not in general the case that $\mathrm{Log}(\exp(w)) = w$ – but the discrepancy must be an integer multiple of $2\pi i$.

Here is a concrete example of the problem. Let $x = y = e^{2\pi/3}$ so $xy = e^{4\pi/3} = e^{-2\pi i/3}$. Now, using the definition of Log we see that $\mathrm{Log}(x) + \mathrm{Log}(y) = 2\pi i/3 + 2\pi i/3 = 4\pi i/3$. However, $\mathrm{Log}(xy) = -2\pi i/3$ (and not $4\pi i/3$).

If you want to rescue the situation, you can define $\log(z)$ to be a set, as opposed to $\mathrm{Log}(z)$ which is a particular number, just as $\arg(z)$ is a set but $\mathrm{Arg}(z)$ is a number. For a non-zero complex number z you would need to put $\log(z) = \{\log(|z|) + i\psi \mid \psi \in \arg(z)\}$. Note that the log on the right hand side of the equation is an uncontroversial logarithm of a positive real number; log on the left is being defined.

This rescues the situation, provided you are prepared to add sets elementwise. In the case which caused us a problem $\log(x) = \log(y) = \{2\pi/3 + 2k\pi \mid k \in \mathbb{Z}\}$ and $\log(xy) = \{-4\pi/3 + 2l\pi \mid l \in \mathbb{Z}\}$. Now we define the sum of two sets in the obvious way as

$$\log(x) + \log(y) = \{a + b \mid a \in \log(x),\ b \in \log(y)\}$$

and we see that

$$\log(xy) = \log(x) + \log(y).$$

Order has been restored, but at the price of defining $\log(x)$ to be a set, or if you prefer a "many-valued function".

As z varies continuously the set $\log(z)$ will vary in a continuous way, which is more than can be said for $\mathrm{Log}(z)$ which has horrible jumps though $2\pi i$ every time $\mathrm{Im}(z)$ passes through an odd multiple of π.

Finally, note the connection between complex logarithms and exponentiation. For each $w \in \mathbb{C}^*$ and $z \in \mathbb{C}$ we have $w^z = e^{z\text{Log}(w)}$. Observe that the intrusion to define Log matches exactly the intrusion which selected w^z from the available candidates.

3.10 Power Series

This section is not tightly reasoned, and is to let you see ahead. You may have come across power series before. An example of such a thing is

$$1 + x + x^2 + x^3 + \dots.$$

You can think of a power series as a purely formal object, just decoration on the page. Alternatively, you can think of x as coming from a domain, and then the power series is an attempt to define a function. There is a problem though, as illustrated by our power series. When $x = 0$ it is uncontroversial to assert that the sum of the series is 1, though you do have to go along with the idea that the sum of infinitely many zeros is zero. When $x = 1/2$, the series is $1 + \frac{1}{2} + \frac{1}{4} + \frac{1}{8} + \dots$. Now, we will not address the problem of giving a formal definition of an infinite sum until Chapter 6. However, in this case it is fairly clear what is happening. The sum of the first n terms is $2 - (\frac{1}{2})^{n-1}$ and as n gets larger and larger the contribution of $(\frac{1}{2})^{n-1}$ will get smaller and smaller. It is therefore not unreasonable to assert that the sum of the series is 2 when $x = \frac{1}{2}$. When $x = 1$, the infinite sum is not co-operating at all. When $x = -1$, the sum is $1 - 1 + 1 - 1 + 1 - 1 + \dots$. This is quite interesting; if you look at the *partial sums* (the sequence whose terms are the sum of the first n terms of the sum) you get $1, 0, 1, 0, 1, 0, \dots$. In a sense, $\frac{1}{2}$ is a reasonable compromise value for this sum, but using the most common definitions this infinite sum will not exist.

All this is quite a mess. For some values of x the series appears to have meaning. In fact this happens for real x in the range $x \in (-1, 1)$.

Now consider the infinite sum

$$1 + \frac{z}{1!} + \frac{z^2}{2!} + \frac{z^3}{3!} + \dots.$$

It will turn out that this sum has a sensible meaning for all $z \in \mathbb{C}$ and that it coincides with our function e^z. Pretend to be brave, and try differentiating the power series term by term. Is that allowed? Well that is a serious question, but just go ahead. Also ignore the concern that z is a complex variable not a real one. Do it quietly while nobody is looking.

You should find that resolute differentiation yields that the derivative of the power series is itself. That is very encouraging of course, since we claimed that the power series defines e^z. Using $\cos\theta = \frac{e^{i\theta}+e^{-i\theta}}{2}$, and $\sin\theta = \frac{e^{i\theta}-e^{-i\theta}}{2i}$ and putting trust in the way power series might add and subtract you should discover that

$$\cos\theta = 1 - \frac{\theta^2}{2!} + \frac{\theta^4}{4!} - \ldots = \sum_{n=0}^{\infty} \frac{\theta^{2n}(-1)^n}{(2n)!}$$

and

$$\sin\theta = \theta - \frac{\theta^3}{3!} + \frac{\theta^5}{5!} - \ldots = \sum_{n=0}^{\infty} \frac{\theta^{2n+1}(-1)^n}{(2n+1)!}.$$

These power series for sine and cosine are very well behaved (the jargon is that they *converge*) for all $\theta \in \mathbb{R}$, and even for all $\theta \in \mathbb{C}$. Moreover, it turns out that these series sum to the values of $\cos\theta$ and $\sin\theta$.

In fact everything we did can be properly justified, but there are plenty of similar looking cases where simple ignorant manipulation leads to wrong answers. There is clearly much work to be done here, and we will start to address it in Chapters 6 and 7 of this book. To understand this material properly is to command subjects called mathematical analysis and complex analysis.

4
Vectors and Matrices

This topic is very substantial, but we will try to give the spirit of this important area in one short chapter.

4.1 Row Vectors

The notion of a vector is geometrically inspired. A geometric vector in our 3-dimensional world (or the 2-dimensional world of this page) is an arrow i.e. a straight line segment of finite length, with orientation. A geometric vector has direction but not position. Thus parallel geometric vectors of the same length, pointing in the same direction, are deemed to be equal.

The method of adding geometric vectors is "nose to tail" as in Figure 4.1. We need to liberate ourselves from the tyranny of pictures for at least two reasons. First, because they can be difficult to draw, and second because we must escape from the 2- and 3-dimensional prisons in which our geometric imaginations are trapped. We do this by capturing the geometric notion of a vector in purely algebraic terms.

To this end we define a *row vector* to be an element of $\mathbb{R}^n$. Thus a row vector is a finite sequence $(x_1, x_2, \ldots, x_n)$ of real numbers. Sequences of the same length can by added (or subtracted) co-ordinatewise. Thus

$$(x_1, \ldots, x_n) + (y_1, \ldots, y_n) = (x_1 + y_1, \ldots, x_n + y_n) \tag{4.1}$$

G. Smith, *Introductory Mathematics: Algebra and Analysis*
© Springer-Verlag London Limited 1998

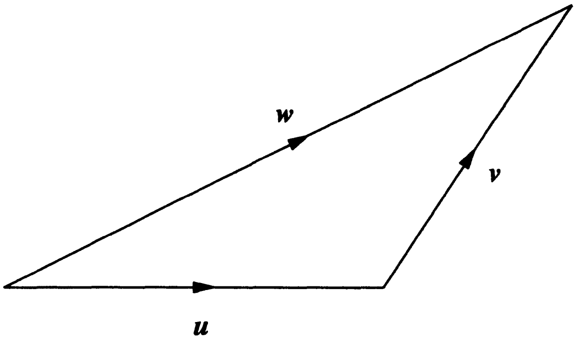

Fig. 4.1. Geometric vector addition: $\mathbf{u} + \mathbf{v} = \mathbf{w}$

and

$$(x_1, \ldots, x_n) - (y_1, \ldots, y_n) = (x_1 - y_1, \ldots, x_n - y_n).$$

There is also a way of multiplying a row vector by a real number λ according to the recipe

$$\lambda \cdot (x_1, x_2, \ldots, x_n) = (\lambda x_1, \lambda x_2, \ldots, \lambda x_n). \tag{4.2}$$

We now show that this algebraic version of the theory of vectors captures the geometry correctly. We work in 3-dimensional space but 2- or 1-dimensional space would do just as well. Set up a co-ordinate system with an origin and mutually perpendicular axes. Calibrate the axes – which is a way of saying that you will regard each axis as a copy of the real line with 0 at the origin. Take any geometric vector $\mathbf{v}$ (we will write all vectors in bold type) and translate it until its tail is parked at the origin. Here *translate* means that you must not change the length, direction or orientation of the geometric vector when you move its tail to the origin. The co-ordinates of the tip of the vector are now at (x_1, x_2, x_3). This sets up a bijection between geometric vectors and ordered triples of real numbers (i.e. $\mathbb{R}^3$). It is easy to check that the addition of row vectors exactly captures addition of geometric vectors. Thus you can add geometric vectors the geometric way, and then read off the row vector equivalent – or alternatively take the two geometric vectors to be added, turn them into elements of $\mathbb{R}^3$ by the specified procedure, and then add the row vectors using Equation (4.1). It doesn't matter which you do, you get the same answer.

In this context, the real numbers are called *scalars* because of their rôle in Equation (4.2). The reason is that multiplication by a scalar quantity *scales* the length of the vector, without changing its direction. The orientation will reverse if you multiply by a negative real number. If you multiply any row vector by 0 you will obtain the zero vector $\mathbf{0} = (0, 0, \ldots, 0)$ which acts as an additive identity for the vectors. The scalars could come from any field, for example from $\mathbb{Q}$ or $\mathbb{C}$. We will stick to the concrete case where the field is $\mathbb{R}$ to fix ideas.

4.2 Higher Dimensions

We have now found a route to higher dimensional worlds. We think of $\mathbb{R}^n$ as being n-dimensional space. You use the algebraic definitions of vector addition and scalar multiplication to do geometry in a place which my mind (and probably yours) cannot envisage pictorially.

Do you have to write elements of $\mathbb{R}^n$ as rows? Well no, of course not. You can write them as columns, but $\begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}$ uses up a lot of space in cultures where people normally write from side to side. Even so, it is sometimes very useful to use these column vectors. It is obvious how to modify our definitions of addition and scalar multiplication so that they apply to column vectors. However, we don't want to mix things up, so we forbid adding a row vector to a column vector (unless $n = 1$ of course, when the two notions coincide).

If the number n is not prime, you also have the option of writing your vectors as rectangular arrays. For example, an element of $\mathbb{R}^6$ might be written as $(x_1, x_2, x_3, x_4, x_5, x_6)$, or as

$$\begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \\ x_6 \end{pmatrix}, \text{ or } \begin{pmatrix} x_1 & x_2 & x_3 \\ x_4 & x_5 & x_6 \end{pmatrix}, \text{ or } \begin{pmatrix} x_1 & x_2 \\ x_3 & x_4 \\ x_5 & x_6 \end{pmatrix}.$$

One can define addition and scalar multiplication in the usual entry-by-entry fashion. Each rectangular array of numbers is called a *matrix*. Even row vectors and column vectors are really matrices. They are merely rather short or thin (respectively). We say that we have an n by m matrix if there are n rows and m columns, and so our matrix is really an element of $\mathbb{R}^{nm}$. Note that scalar multiplication by λ has the effect of multiplying each matrix entry by λ.

4.3 Vector Laws

We list various algebraic laws which vectors enjoy (whether they are geometric vectors, row vectors, column vectors or matrices). In what follows, all vectors must be of the same type (e.g. row vectors of the form $(x_1, \ldots, x_7)$). We will use bold lower case letters as names for vectors, and lower case Greek letters as names for scalars (i.e. real numbers).

$$\begin{aligned}
\forall \mathbf{u},\ \mathbf{v} \text{ we have } \mathbf{u}+\mathbf{v} &= \mathbf{v}+\mathbf{u} \\
\forall \mathbf{u},\ \mathbf{v},\ \mathbf{w} \text{ we have } (\mathbf{u}+\mathbf{v})+\mathbf{w} &= \mathbf{u}+(\mathbf{v}+\mathbf{w}) \\
\exists \mathbf{0} \text{ such that } \forall \mathbf{u} \text{ we have } \mathbf{0}+\mathbf{u} &= \mathbf{u} \\
\forall \mathbf{u}\ \exists -\mathbf{u} \text{ such that } \mathbf{u}+-\mathbf{u} &= \mathbf{0} \\
\forall \mathbf{u},\ \forall \lambda, \mu \in \mathbb{R} \text{ we have } (\lambda+\mu)\mathbf{u} &= (\lambda\mathbf{u})+(\mu\mathbf{u}) \\
\forall \mathbf{u},\ \mathbf{v},\ \forall \lambda \in \mathbb{R} \text{ we have } \lambda(\mathbf{u}+\mathbf{v}) &= (\lambda\mathbf{u})+(\lambda\mathbf{v}) \\
\forall \mathbf{u}\ 1\cdot\mathbf{u} &= \mathbf{u} \\
\forall \lambda, \mu \in \mathbb{R},\ \forall \mathbf{u} \text{ we have } (\lambda\mu)\mathbf{u} &= \lambda(\mu\mathbf{u})
\end{aligned}$$

Remark 4.1

From a more sophisticated point of view, we can take the laws and turn them into axioms (and the scalars can be drawn from any field). An *abstract vector space* will be a set V (of vectors) equipped with addition and scalar multiplication so that these axioms are satisfied. However, at this stage it is probably best to stay firmly grounded in $\mathbb{R}^n$.

4.4 Lengths and Angles

Let us suppose that we are working with $\mathbb{R}^n$, written as row vectors.

Definition 4.1

The *modulus* or *length* of the vector $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ is the real number

$$||\mathbf{x}|| = \sqrt{\sum_{i=1}^{n} x_i^2} = \sqrt{x_1^2 + x_2^2 + \ldots + x_n^2}.$$

This notion of length coincides with the geometric definition of length when $n = 1$. It also coincides when $n = 2$ by Pythagoras's theorem. One can use

Pythagoras's theorem twice to show that it also works when $n = 3$. After that, there is no pictorial definition of length, but our definition becomes a platform on which higher dimensional Euclidean geometry can be built.

There is a notion of multiplying two vectors together to obtain a scalar. This product suffers from notational proliferation. It is variously called the dot product, scalar product or inner product, and the notation for the product of $\mathbf{x}$, $\mathbf{y} \in \mathbb{R}^n$ can be $\langle \mathbf{x}, \mathbf{y} \rangle$, $(\mathbf{x}, \mathbf{y})$, $\mathbf{x} \cdot \mathbf{y}$ or $\mathbf{x}.\mathbf{y}$. We plump for the following notation.

Definition 4.2

Suppose that $\mathbf{x}$, $\mathbf{y} \in \mathbb{R}^n$. We define the scalar product of these two vectors by

$$\langle \mathbf{x}, \mathbf{y} \rangle = \sum_{i=1}^{n} x_i y_i = x_1 y_1 + x_2 y_2 + \ldots + x_n y_n. \tag{4.3}$$

There are some purely formal consequences of this definition which will come in handy from time to time. Each equation holds for all row vectors $\mathbf{x}, \mathbf{y}$ and $\mathbf{z}$, and for all scalars λ.

$$||\mathbf{x}||^2 = \langle \mathbf{x}, \mathbf{x} \rangle \tag{4.4}$$

$$||\lambda \mathbf{x}|| = |\lambda| \cdot ||\mathbf{x}|| \tag{4.5}$$

$$\langle \mathbf{x}, \mathbf{y} \rangle = \langle \mathbf{y}, \mathbf{x} \rangle \tag{4.6}$$

$$\langle \mathbf{x} + \mathbf{y}, \mathbf{z} \rangle = \langle \mathbf{x}, \mathbf{z} \rangle + \langle \mathbf{y}, \mathbf{z} \rangle \tag{4.7}$$

$$\langle \mathbf{x}, \mathbf{y} + \mathbf{z} \rangle = \langle \mathbf{x}, \mathbf{y} \rangle + \langle \mathbf{x}, \mathbf{z} \rangle \tag{4.8}$$

$$\lambda \langle \mathbf{x}, \mathbf{y} \rangle = \langle \lambda \mathbf{x}, \mathbf{y} \rangle \tag{4.9}$$

$$\lambda \langle \mathbf{x}, \mathbf{y} \rangle = \langle \mathbf{x}, \lambda \mathbf{y} \rangle \tag{4.10}$$

This notation $\langle \mathbf{x}, \mathbf{y} \rangle$ therefore obeys some nice algebraic laws, so we should be able to do mathematics with it. There is also a geometric interpretation when $n \leq 3$. This is perhaps best seen when $n = 2$. Suppose that $\mathbf{x}, \mathbf{y} \in \mathbb{R}^2$ are *unit vectors*, i.e. assume that $||\mathbf{x}|| = ||\mathbf{y}|| = 1$. Thus $x_1 = \cos \psi$ and $x_2 = \sin \psi$. Similarly $\mathbf{y} = (y_1, y_2) = (\cos \varphi, \sin \varphi)$. Here ψ and φ are the angles between $(1, 0)$ and the corresponding unit vectors, measured in an anti-clockwise direction. In fact you are at liberty to add or subtract multiples of 2π to or from these angles; it does not matter. Now

$$\langle \mathbf{x}, \mathbf{y} \rangle = \cos \psi \cos \varphi + \sin \psi \sin \varphi = \cos(\psi - \varphi). \tag{4.11}$$

This gives the proof of most of

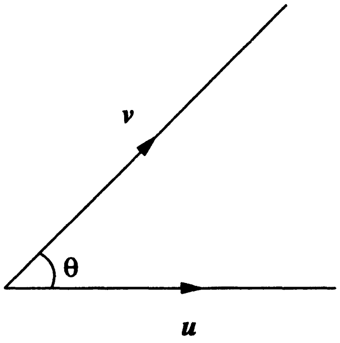

Fig. 4.2. The angle between the vectors is θ

Proposition 4.1

Suppose that $\mathbf{u}, \mathbf{v} \in \mathbb{R}^2$, then $\langle \mathbf{u}, \mathbf{v} \rangle = ||\mathbf{u}|| \cdot ||\mathbf{v}|| \cdot \cos\theta$ where θ is the angle between the vectors $\mathbf{u}$ and $\mathbf{v}$.

Proof

Our discussion shows that the result is true provided $\mathbf{u}$ and $\mathbf{v}$ are of length 1. However, $\mathbf{u} = ||\mathbf{u}|| \cdot \hat{\mathbf{u}}$ and $\mathbf{v} = ||\mathbf{v}|| \cdot \hat{\mathbf{v}}$ with $\hat{\mathbf{u}}$ and $\hat{\mathbf{v}}$ both unit vectors. Thus $\langle \mathbf{u}, \mathbf{v} \rangle = ||\mathbf{u}|| \cdot ||\mathbf{v}|| \cdot \langle \hat{\mathbf{u}}, \hat{\mathbf{v}} \rangle = ||\mathbf{u}|| \cdot ||\mathbf{v}|| \cdot \cos\theta$ as required.

□

We could look at it another way, using the happy chance that we have a natural identification between $\mathbb{R}^2$ and $\mathbb{C}$, the set of complex numbers. The heart of the proof is that the scalar product of two vectors of unit length is the cosine of the angle between them. Thought of as a complex number, the unit vector $\mathbf{u}$ corresponds to $e^{i\beta}$ and the unit vector $\mathbf{v}$ corresponds to $e^{i\alpha}$ as shown in Figure 4.3.

Now

$$\langle \mathbf{u}, \mathbf{v} \rangle = \langle (u_1, u_2), (v_1, v_2) \rangle = u_1 v_1 + u_2 v_2.$$

This is the real part of $(u_1 + iu_2)(v_1 - iv_2)$ and so the real part of $e^{i\alpha} e^{-i\beta} = e^{i(\alpha-\beta)}$. This is, of course, the cosine of the angle between $\mathbf{u}$ and $\mathbf{v}$.

In three dimensions it is not quite so obvious that Proposition 4.1 still holds, but in fact it does. The reader with a flair for 3-dimensional trigonometry might like to fill in the details.

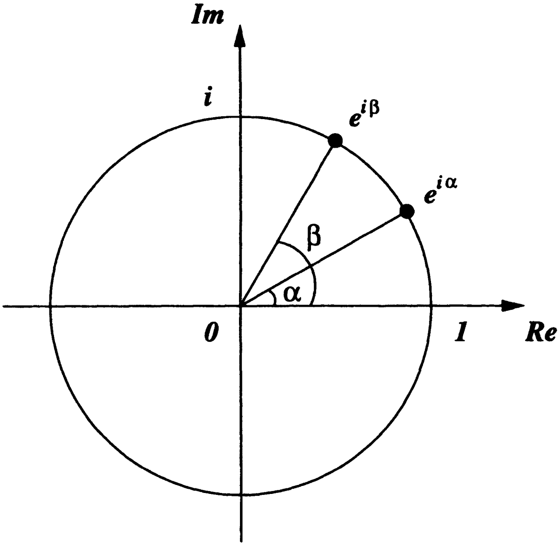

Fig. 4.3. A dot product in the Argand diagram

EXERCISES

These exercises were already mentioned in the text.

4.1 Verify Equations (4.5) to (4.10).

4.2 Show that Proposition 4.1 holds in $\mathbb{R}^3$.

We would like to *define* the angle between non-zero vectors $\mathbf{x}$, $\mathbf{y} \in \mathbb{R}^n$ to be that angle θ in the range $0 \le \theta \le \pi$ such that

$$\cos\theta = \frac{\langle \mathbf{x}, \mathbf{y} \rangle}{||\mathbf{x}|| \cdot ||\mathbf{y}||}.$$

However, that will only work if

$$-1 \le \frac{\langle \mathbf{x}, \mathbf{y} \rangle}{||\mathbf{x}|| \cdot ||\mathbf{y}||} \le 1,$$

but we are in luck. A celebrated inequality comes to our aid.

Proposition 4.2 (Cauchy, Schwarz)

Suppose that $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$, then

$$|\langle \mathbf{x}, \mathbf{y} \rangle| \le ||\mathbf{x}|| \cdot ||\mathbf{y}||.$$

Proof

The result is clear if either $\mathbf{x} = \mathbf{0}$ or $\mathbf{y} = \mathbf{0}$, so we may assume $\mathbf{x} \neq \mathbf{0} \neq \mathbf{y}$. We have $||\mathbf{x} + \lambda\mathbf{y}||^2 \geq 0$ for all values of λ, since it the square of a real number. Thus $\langle \mathbf{x} + \lambda\mathbf{y}, \mathbf{x} + \lambda\mathbf{y} \rangle \geq 0$. Expand this out using various properties of the scalar product to obtain

$$\lambda^2 ||\mathbf{y}||^2 + 2\lambda \langle \mathbf{x}, \mathbf{y} \rangle + ||\mathbf{x}||^2 \geq 0. \tag{4.12}$$

The left hand side of this inequality is a quadratic polynomial in λ with positive coefficient of λ^2. Its graph is therefore a nose-down parabola. This polynomial cannot have two distinct roots, else any value of λ between the roots would violate inequality (4.12).

Since it has at most one real root, we deduce that its discriminant ("$b^2 - 4ac$") is non-positive. Thus

$$4\langle \mathbf{x}, \mathbf{y} \rangle^2 - 4||\mathbf{x}||^2 ||\mathbf{y}||^2 \leq 0.$$

Divide by 4, rearrange to put $||\mathbf{x}||^2 ||\mathbf{y}||^2$ on the right and take the (non-negative) square root to obtain $|\langle \mathbf{x}, \mathbf{y} \rangle| \leq ||\mathbf{x}|| \cdot ||\mathbf{y}||$.

□

Thus we can assign meaning to "angle" in higher dimensions. Now we can deduce the following important result.

Proposition 4.3 (triangle inequality for row vectors)

If $\mathbf{x},\ \mathbf{y} \in \mathbb{R}^n$, then

$$||\mathbf{x} + \mathbf{y}|| \leq ||\mathbf{x}|| + ||\mathbf{y}||. \tag{4.13}$$

Note: This is the n-dimensional generalization of the assertion that each side of a triangle has length less than or equal to the sum of the lengths of the other two sides.

Proof

For any vectors $\mathbf{x}$ and $\mathbf{y}$ we have

$$\begin{aligned} ||\mathbf{x} + \mathbf{y}||^2 &= \langle \mathbf{x} + \mathbf{y}, \mathbf{x} + \mathbf{y} \rangle = ||\mathbf{x}||^2 + 2\langle \mathbf{x}, \mathbf{y} \rangle + ||\mathbf{y}||^2 \\ &\leq ||\mathbf{x}||^2 + 2|\langle \mathbf{x}, \mathbf{y} \rangle| + ||\mathbf{y}||^2 \leq ||\mathbf{x}||^2 + 2||\mathbf{x}|| \cdot ||\mathbf{y}|| + ||\mathbf{y}||^2. \end{aligned}$$

The final inequality uses the Cauchy–Schwarz result. Now take square roots and we are done.

□

4.5 Position Vectors

Although geometric vectors do not have position, you can elect to nail the tail down to an origin, and the arrow then does point to a specific place. Such a vector is called a *position vector.*

Example 4.1

Find the equation of the straight line in the plane going through the point with co-ordinates $(1,2)$ and the point with co-ordinates $(3,4)$.

Happily, the position vector of the first point is the vector $\mathbf{a} = (1,2)$ and the position vector of the second point is $\mathbf{b} = (3,4)$ (see Figure 4.4). The vector you need to add to $\mathbf{a}$ to get $\mathbf{b}$ is $\mathbf{b} - \mathbf{a}$, a vector which is the direction of the line under discussion. Points P on this line have position vector $\mathbf{r} = \mathbf{a} + \lambda(\mathbf{b} - \mathbf{a})$. When $\lambda = 0, \mathbf{r} = \mathbf{a}$. When $\lambda = 1, \mathbf{r} = \mathbf{b}$. When $0 < \lambda < 1$ the position vector $\mathbf{r}$ points to places on the line strictly between $(1,2)$ and $(3,4)$.

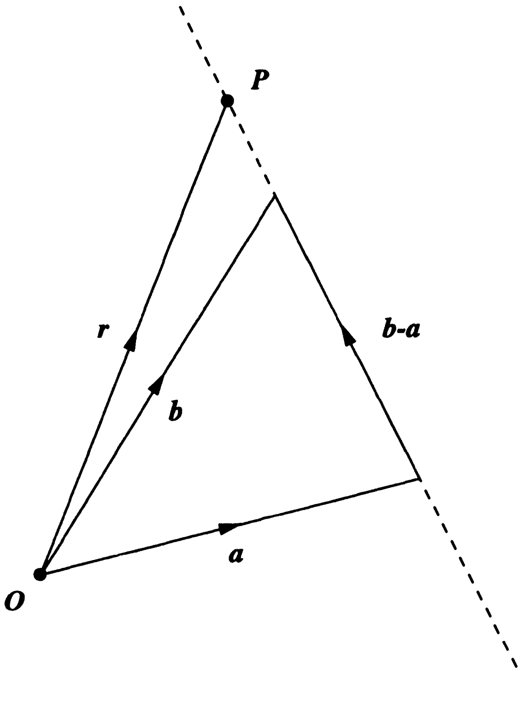

Fig. 4.4. Line through points with position vectors **a** and **b**

One very attractive way to define a line in the plane which does not pass though the origin is as follows. See Figure (4.5) Let the nearest point in the plane to the origin be C. Let the position vector of C be $\mathbf{c}$. Now a point with position vector $\mathbf{r}$ is on the line exactly when $\mathbf{r} - \mathbf{c}$ is perpendicular to $\mathbf{c}$. This is given by the simple condition

$$\langle \mathbf{r} - \mathbf{c}, \mathbf{c} \rangle = 0.$$

Now, this can be rearranged into the equivalent condition $\langle \mathbf{r}, \mathbf{c} \rangle = \langle \mathbf{c}, \mathbf{c} \rangle$ or, if you prefer, $\langle \mathbf{r}, \mathbf{c} \rangle = ||\mathbf{c}||^2$.

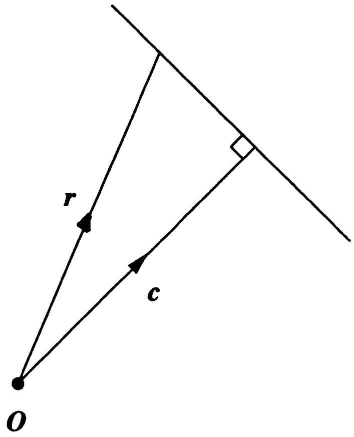

Fig. 4.5. Line (or plane) perpendicular to $\mathbf{c}$

In $\mathbb{R}^3$ we can use exactly the same construction to write down the equation of a plane not passing through the origin. The equation of a circle in $\mathbb{R}^2$ is quite straightforward. If its centre has position vector $\mathbf{a}$ and the radius of the circle is c, then the equation of the circle is $||\mathbf{r} - \mathbf{a}|| = c$. The same equation in $\mathbb{R}^3$ defines a sphere. The inequality $||\mathbf{r} - \mathbf{a}|| \le c$ defines a disk in $\mathbb{R}^2$ and a ball in $\mathbb{R}^3$.

4.6 Matrix Operations

We need a good way of describing any rectangular matrix. We will start off with a concrete example. Suppose

$$M = \begin{pmatrix} 1 & 1 & 2 \\ -1 & 7 & 2 \end{pmatrix}.$$

We give the entries names in a systematic way, so

$$M = \begin{pmatrix} m_{11} & m_{12} & m_{13} \\ m_{21} & m_{22} & m_{23} \end{pmatrix}.$$

Thus $m_{11} = 1$, $m_{21} = -1$ and so on. This is still cumbersome, since you might be dealing with a 100 by 100 matrix. The neat way to write M in terms of its entries is (m_{ij}). This means that the entry in row i and column j is m_{ij} whenever i is a legal row number, and j a legal column number. The row numbers always run from top to bottom and the column numbers from left to right.

Using this notation, we can write down how to add matrices together very cleanly. Suppose that $A = (a_{ij})$ and $B = (b_{ij})$ are both n by m matrices. Let $(c_{ij}) = C = A + B$ be defined by $c_{ij} = a_{ij} + b_{ij}$ for each legal i and j.

There is a way of multiplying matrices which allows the multiplication of matrices of various shapes, but the shapes must be compatible. Suppose that $A = (a_{ij})$ and $B = (b_{ij})$ are matrices, where A has shape n by m, and B has shape m by p. Their product AB is a matrix $C = (c_{ij})$ of shape n by p. Its entries are given by the formula

$$c_{ij} = \sum_{k=1}^{m} a_{ik}b_{kj}$$

which probably looks a bit intimidating if you have not seen it before. We expand it out and discover that

$$c_{ij} = a_{i1}b_{1j} + a_{i2}b_{2j} + \ldots + a_{im}b_{mj}.$$

Thus to calculate the entry in row i and column j of the product, use ingredients from the i-th row of the first matrix and the j-th column of the second matrix. What you do is to take the j-th column of the second matrix, and turn it round so it becomes a row vector, and then form the scalar product with the i-th row of the first matrix. Concretely, we have

$$\begin{pmatrix} 1 & 1 & 2 \\ -1 & 7 & 2 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 2 & 3 \end{pmatrix} = \begin{pmatrix} 5 & 7 \\ 3 & 13 \end{pmatrix},$$

because

$$\begin{aligned} \langle(1,1,2),(1,0,2)\rangle &= 5, \\ \langle(1,1,2),(0,1,3)\rangle &= 7, \\ \langle(-1,7,2),(1,0,2)\rangle &= 3, \\ \langle(-1,7,2),(0,1,3)\rangle &= 13. \end{aligned}$$

We need a more formal way of saying "turn round" a column vector to make it a row vector.

Definition 4.3

Let $A = (a_{ij})$ be an n by m matrix. The *transpose* of A, written A^T is the m by n matrix whose entry in the i-th row and j-th column is a_{ji}.

Suppose that the rows of A are the row vectors $\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_n$ reading from top to bottom, and the columns of B are the column vectors $\mathbf{b}_1^T, \mathbf{b}_2^T, \ldots, \mathbf{b}_p^T$. Our vectors are row vectors unless we state otherwise, so each $\mathbf{b}_j$ is a row vector and each $\mathbf{b}_j^T$ is a column vector. The entry in the i-th row and j-th column of AB is the scalar product $\langle \mathbf{a}_i, \mathbf{b}_j \rangle$. Thus multiplication of an n by m matrix by an m by p matrix can be performed by working out np scalar products.

4.7 Laws of Matrix Algebra

In what follows A, B and C are supposed to be matrices of such shapes that the operations of matrix addition and matrix multiplication mentioned in the equations are allowed. Subject to this condition, the following equations always hold.

$$A(B + C) = (AB) + (AC) \tag{4.14}$$

$$(A + B)C = (AB) + (AC) \tag{4.15}$$

$$(AB)C = A(BC) \tag{4.16}$$

Equations (4.14) and (4.15) are routine consequences of the definitions of matrix addition and matrix multiplication. However, Equation (4.16) is perhaps more mysterious. It is not clear at first why this recipe for multiplication should yield an associative operation. One can look at a few examples, and you are urged to do so. However, the fact that it works out in a few special cases is not a proof that it will always work.

Proposition 4.4

Suppose that A, B and C are matrices. We use the usual notation $A = (a_{ij})$ and so on. We also suppose that A has r columns and B has r rows, and that B has s columns and C has s rows. It follows that $A(BC) = (AB)C$.

Proof

Both products are matrices of the same shape, the number of rows being the number of rows of A, and the number of columns being the number of columns of C.

We write out formulas for x_{ij} and y_{ij}, the typical entries of the two respective matrix products

$$x_{ij} = \sum_{u=1}^{r} a_{iu} \left(\sum_{v=1}^{s} b_{uv} c_{vj} \right)$$

and

$$y_{ij} = \sum_{v=1}^{s} \left(\sum_{u=1}^{r} a_{iu} b_{uv} \right) c_{vj}.$$

Both of these expressions are different ways of writing the sum of all possible terms of the form $a_{iu}b_{uv}c_{vj}$ where i and j are fixed but u and v take all possible legal values. There are r possible values of u and s possible values of v so there are rs terms to be summed.

□

The proof is complete, but we will use a concrete example just to make sure the point is understood.

Example 4.2

Suppose that

$$A = \begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \end{pmatrix}, B = \begin{pmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \\ b_{31} & b_{32} \end{pmatrix}, C = \begin{pmatrix} c_{11} & c_{12} \\ c_{21} & c_{22} \end{pmatrix}.$$

We look at corresponding entries of the 2 by 2 matrices $A(BC)$ and $(AB)C$. The entry in row i and column j of the first is

$$\begin{aligned}
\sum_{u=1}^{3} a_{iu}(\sum_{v=1}^{2} b_{uv}c_{vj}) &= \sum_{u=1}^{3} a_{iu}(b_{u1}c_{1j} + b_{u2}c_{2j}) \\
&= a_{i1}(b_{11}c_{1j} + b_{12}c_{2j}) \\
&\quad + a_{i2}(b_{21}c_{1j} + b_{22}c_{2j}) \\
&\quad + a_{i3}(b_{31}c_{1j} + b_{32}c_{2j}) \\
&= a_{i1}b_{11}c_{1j} + a_{i1}b_{12}c_{2j} \\
&\quad + a_{i2}b_{21}c_{1j} + a_{i2}b_{22}c_{2j} \\
&\quad + a_{i3}b_{31}c_{1j} + a_{i3}b_{32}c_{2j}
\end{aligned}$$

whereas the entry in row i and column j of the second is

$$\begin{aligned}\sum_{v=1}^{2}(\sum_{u=1}^{3} a_{iu}b_{uv})c_{vj} &= \sum_{v=1}^{2}(a_{i1}b_{1v} + a_{i2}b_{2v} + a_{i3}b_{3v})c_{vj} \\ &= (a_{i1}b_{11} + a_{i2}b_{21} + a_{i3}b_{31})c_{1j} \\ &\quad +(a_{i1}b_{12} + a_{i2}b_{22} + a_{i3}b_{32})c_{2j} \\ &= a_{i1}b_{11}c_{1j} + a_{i2}b_{21}c_{1j} + a_{i3}b_{31}c_{1j} \\ &\quad +a_{i1}b_{12}c_{2j} + a_{i2}b_{22}c_{2j} + a_{i3}b_{32}c_{2j}.\end{aligned}$$

Thanks to the commutative law of addition, we are done.

4.8 Identity Matrices and Inverses

First a piece of useful notation. The *Kronecker delta* is the symbol δ_{ij} where $\delta_{ij} = 1$ if $i = j$, and otherwise $\delta_{ij} = 0$.

Definition 4.4

The n by n *identity matrix* is $I_n = (\delta_{ij})$.

Thus $I_1 = (1)$,

$$I_2 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \quad I_3 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \text{ and so on.}$$

The reason we call each of these matrices an identity matrix is the following result.

Proposition 4.5

Let X be an n by m matrix. It follows that $I_nX = X = XI_m$.

Proof

The entry in the i-th row and j-th column of I_nX is $\sum_k \delta_{ik}x_{kj} = x_{ij}$ so $I_nX = X$. The proof that $XI_m = X$ is similar.

□

Matrices which are square play a special rôle in the theory. For fixed n the set of n by n matrices is closed under multiplication (i.e. the product of two n by n matrices is an n by n matrix). Another reason is that I_n plays the rôle of a two-sided multiplicative identity, just like 1 does in $\mathbb{Z}$, $\mathbb{Q}$, $\mathbb{R}$ and $\mathbb{C}$.

Definition 4.5

Suppose that X is an n by n matrix. An inverse matrix Y for X is an n by n matrix such that $XY = I_n$.

Now for a little peering into the future; the following assertions are correct, but we will not prove them for a few pages yet. If such an inverse matrix Y exists, then it is unique (i.e. it is the only matrix which will do the job). Since there is no ambiguity, we may write Y as X^{-1}. Moreover (and this is not at all obvious), X^{-1} is also the unique matrix such that $X^{-1}X = I_n$.

The 1 by 1 matrices are just elements of $\mathbb{R}$ with brackets as adornments. Addition and multiplication are just as in $\mathbb{R}$. The matrix (c) is invertible, with inverse $(1/c)$ unless $c = 0$. If course (0) has no multiplicative inverse.

Now step up to the 2 by 2 case. Just as before, the zero matrix $0_2 = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}$ fails to be invertible for obvious reasons. However, this time there are matrices other than 0_2 which are not invertible. It works like this. Suppose

$$X = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$$

Let $\Delta = ad - bc$. If $\Delta = 0$, then the matrix is not invertible. If $\Delta \neq 0$, then X is invertible and

$$X^{-1} = \begin{pmatrix} d/\Delta & -b/\Delta \\ -c/\Delta & a/\Delta \end{pmatrix}.$$

Definition 4.6

The quantity Δ is called the *determinant* of the 2 by 2 matrix X, and we write it as $\det(X)$ or $|X|$.

There is an analogous notion of determinant for n by n matrices. We will explore this later.

Suppose that you want to study a system of two linear equations in two unknowns. We fix our notation:

$$\begin{aligned} aX + bY &= r, \\ cX + dY &= s. \end{aligned}$$

We simplify these two equations by writing down one matrix equation. Let

$$A = \begin{pmatrix} a & b \\ c & d \end{pmatrix}, \ \mathbf{z} = \begin{pmatrix} X \\ Y \end{pmatrix} \text{ and } \mathbf{c} = \begin{pmatrix} r \\ s \end{pmatrix}.$$

The relevant matrix equation is $A\mathbf{z} = \mathbf{c}$. Provided you know A^{-1} you can solve for $\mathbf{z}$ immediately because

$$\mathbf{z} = I_2\mathbf{z} = (A^{-1}A)\mathbf{z} = A^{-1}(A\mathbf{z}) = A^{-1}\mathbf{c}.$$

Notice the crucial use of the associativity of matrix multiplication in that analysis. Thus being able to calculate the inverse of a square matrix is intimately related to solving systems of linear equations.

Example 4.3

Consider the simultaneous linear equations $2x + y = 1$ and $3x - y = 0$. Let

$$A = \begin{pmatrix} 2 & 1 \\ 3 & -1 \end{pmatrix} \text{ so } \begin{pmatrix} x \\ y \end{pmatrix} = A^{-1}\begin{pmatrix} 1 \\ 0 \end{pmatrix} = \begin{pmatrix} \frac{1}{5} & \frac{1}{5} \\ \frac{3}{5} & -\frac{2}{5} \end{pmatrix}\begin{pmatrix} 1 \\ 0 \end{pmatrix} = \begin{pmatrix} \frac{1}{5} \\ \frac{3}{5} \end{pmatrix}.$$

Notice that if you vary the constant terms in these linear equations, you can find the new solutions immediately since you already have A^{-1} to hand. When dealing with larger systems of linear equations, working out A^{-1} can be a big problem. There are formulas for 3 by 3, 4 by 4 and so on, but it gets very complicated very quickly.

4.9 Determinants

We first give a purely mechanical definition of determinants, and then afterwards we will explain more about what is going on.

Definition 4.7

Suppose that $A = (a_{ij})$ is an n by n matrix. We define the determinant $|A|$ (also written $\det(A)$ or $|a_{ij}|$) inductively. Thus we assume that we know how to calculate a determinant $|X|$ where X is an $n-1$ by $n-1$ matrix, and start the induction off by defining the determinant of a 1 by 1 matrix (a) to be a.

Pick any row of A, say the i-th row. We will work out $|A|$ by using the entries of the i-th row of A and the determinants of some $n-1$ by $n-1$ matrices. For each entry a_{ij} in the i-th row, let A_{ij} be the $n-1$ by $n-1$ matrix obtained by striking out the i-th row and j-th column of A. Let $c_{ij} = (-1)^{i+j}|A_{ij}|$. Now let $|A| = \sum_{j=1}^{n} a_{ij}c_{ij}$.

The stunning thing about this definition is that you are allowed to choose the row you work with. In fact you could use a column instead; it doesn't matter, you will always get the same answer. This may seem slightly magical – but the reason everything works is just that multiplication distributes over addition in the real numbers.

Remark 4.2

A matrix obtained by striking out the same number of rows and columns of a matrix is called a *minor* of the matrix. Our our case, we are only striking out one row and one column at a time. The quantities c_{ij} mentioned above, obtained by adjusting the signs of the determinants of $n-1$ by $n-1$ minors, are called *cofactors* of the corresponding entries a_{ij} of A.

Let us look at a specific example.

Example 4.4

Suppose that

$$A = \begin{pmatrix} 1 & 0 & 2 \\ -1 & 1 & 1 \\ 0 & 1 & 0 \end{pmatrix}.$$

We expand using the third row.

$$|A| = 0 \cdot \begin{vmatrix} 0 & 2 \\ 1 & 1 \end{vmatrix} + 1(-1) \cdot \begin{vmatrix} 1 & 2 \\ -1 & 1 \end{vmatrix} + 0 \cdot \begin{vmatrix} 1 & 0 \\ -1 & 1 \end{vmatrix} = -3.$$

Now expand using the first column.

$$|A| = 1 \cdot \begin{vmatrix} 1 & 1 \\ 1 & 0 \end{vmatrix} + 1 \cdot \begin{vmatrix} 0 & 2 \\ 1 & 0 \end{vmatrix} + 0 \cdot \begin{vmatrix} 0 & 2 \\ 1 & 1 \end{vmatrix} = -3.$$

Now we will try to demystify this. We work with the 3 by 3 case $X = (x_{ij})$ for clarity. No matter how you work out $|X|$ you get

$$x_{11}x_{22}x_{33} - x_{11}x_{23}x_{32} + x_{12}x_{23}x_{31} - x_{12}x_{21}x_{33} + x_{13}x_{21}x_{32} - x_{13}x_{22}x_{31}.$$

Perhaps you can see the pattern now. The determinant consists of a sum of products (sometimes multiplied by -1). The products range over all possible ways of multiplying n matrix entries which have the property that no two are in the same row or column. There are $n!$ such products. Half the terms are multiplied by -1 and half are not.

We need to think about even and odd permutations in order to see where these minus signs should go. There is much material on permutations in Chapter 5, but we only need a small amount of the theory of permutations here, so you shouldn't need to look ahead. We think of the subscripts in a given product as defining a permutation of $\{1, 2, 3\}$. Thus $x_{11}x_{22}x_{33}$ defines the identity permutation but $x_{11}x_{23}x_{32}$ defines the permutation which fixes 1 (because of x_{11}), sends 2 to 3 (because of x_{23}) and sends 3 to 2 (because of x_{32}). In the order we have written the products, the permutations which arise are Id, $(2,3), (1,2,3), (1,2), (1,3,2)$ and $(1,3)$. We hope the notation is obvious, so $(1,2,3)$ is the permutation which sends 1 to 2, 2 to 3 and (wrapping round) 3 to 1. On the other hand $(2,3)$ fixes 1, and swaps 2 and 3. A permutation which swaps two things and fixes everything else is called a *transposition*. It turns out that every permutation is a product of transpositions (see Section 5.4). For example $(1,2,3)$ can be obtained by first applying (1,2) and then (1,3). The odd permutations are those which can be written as a product of an odd number of transpositions, and the even permutations are those which can be written as a product of an even number of transpositions. It is a slightly mysterious fact that there are no permutations which are both even and odd. To see this, consider a polynomial p in n commuting variables $x_1, \ldots, x_n$. Explicitly

$$p = \prod_{i<j}(x_i - x_j).$$

We need only be worrying about $n \geq 2$. When $n = 2$ our polynomial p is just $x_1 - x_2$, and when $n = 3$ it is $(x_1 - x_2)(x_1 - x_3)(x_2 - x_3)$ and so on. Now suppose that σ is a permutation of $\{1, 2, \ldots, n\}$. If you apply σ to the subscripts of p, the resulting polynomial is $\pm p$. Moreover, the effect of applying σ_1 then σ_2 is the same as composing the permutations (maps) σ_1 with σ_2 and then applying the result to the subscripts. A careful examination (do it!) shows that a transposition changes the sign of p. It therefore follows that any σ which changes the sign of p is the product of an odd number of transpositions, and that any σ which does not change the sign of p is the product of an even number of transpositions.

Back to the issue at hand: determinants. Each of our six products gives rise to a permutation, and you have to insert a factor of -1 when the permutation is odd.

Perhaps you can now see why you can use any row or column to expand a determinant. It is because of the distributive law. For example, in

$$x_{11}x_{22}x_{33} - x_{11}x_{23}x_{32} + x_{12}x_{23}x_{31} - x_{12}x_{21}x_{33} + x_{13}x_{21}x_{32} - x_{13}x_{22}x_{31},$$

you might focus on the second column of the matrix. Each of the six products forming the sum contains exactly one entry from the second column. You see what multiplies each one of these second column entries and tidy up. You get

$$-x_{12}(x_{21}x_{33} - x_{23}x_{31}) + x_{22}(x_{11}x_{33} - x_{13}x_{31}) - x_{32}(x_{11}x_{23} - x_{13}x_{21}).$$

To remember the signs you have to use, just think of the pattern below, starting with a + in the top left corner.

$$\begin{array}{ccccccc} + & - & + & - & + & - & \cdots \\ - & + & - & + & - & + & \cdots \\ + & - & + & - & + & - & \cdots \\ - & + & - & + & - & + & \cdots \\ + & - & + & - & + & - & \cdots \\ - & + & - & + & - & + & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \end{array}$$

An immediate consequence of the fact that you can expand a determinant using any row or column is that the determinant is a "linear function" of the rows, and of the columns, of a matrix. To be explicit, suppose that $\mathbf{a_1}, \ldots, \mathbf{a_n} \in \mathbb{R}^n$ then we can use these row vectors to form an n by n matrix with these rows. This matrix is $A = \begin{pmatrix} \mathbf{a_1} \\ \vdots \\ \mathbf{a_i} \\ \vdots \\ \mathbf{a_n} \end{pmatrix}$ and its determinant is $A = \begin{vmatrix} \mathbf{a_1} \\ \vdots \\ \mathbf{a_i} \\ \vdots \\ \mathbf{a_n} \end{vmatrix}$. Now if $\mathbf{a_i} = \mathbf{b_i} + \mathbf{c_i}$, then

$$\begin{vmatrix} \mathbf{a_1} \\ \vdots \\ \mathbf{a_i} \\ \vdots \\ \mathbf{a_n} \end{vmatrix} = \begin{vmatrix} \mathbf{a_1} \\ \vdots \\ \mathbf{b_i} \\ \vdots \\ \mathbf{a_n} \end{vmatrix} + \begin{vmatrix} \mathbf{a_1} \\ \vdots \\ \mathbf{c_i} \\ \vdots \\ \mathbf{a_n} \end{vmatrix}.$$

Of course i is arbitrary, and the same argument works for columns. The same remarks apply to the observation that if $\lambda \in \mathbb{R}$, then

$$\begin{vmatrix} \mathbf{a_1} \\ \vdots \\ \lambda\mathbf{a_i} \\ \vdots \\ \mathbf{a_n} \end{vmatrix} = \lambda \begin{vmatrix} \mathbf{a_1} \\ \vdots \\ \mathbf{a_i} \\ \vdots \\ \mathbf{a_n} \end{vmatrix}.$$

Something which is a little more tricky to see is that if you swap two rows or two columns of a square matrix, then in the expanded determinant, every single permutation gets multiplied by a transposition. Thus all the odd permutations become even and vice versa, so the determinant is multiplied by -1. Now if two rows are equal (or two columns are equal), swapping them simultaneously does nothing at all, and also multiplies the determinant by -1. The only number which is unchanged by multiplication by -1 is 0 so the determinant vanishes. Finally, if you add any multiple of any row to any other row, or any multiple of any column to any column, you do not change the determinant. This is because

$$\begin{vmatrix} \mathbf{a}_1 \\ \vdots \\ \mathbf{a}_i \\ \vdots \\ \mathbf{a}_j \\ \vdots \\ \mathbf{a}_n \end{vmatrix} = \begin{vmatrix} \mathbf{a}_1 \\ \vdots \\ \mathbf{a}_i \\ \vdots \\ \mathbf{a}_j \\ \vdots \\ \mathbf{a}_n \end{vmatrix} + \lambda \begin{vmatrix} \mathbf{a}_1 \\ \vdots \\ \mathbf{a}_i \\ \vdots \\ \mathbf{a}_i \\ \vdots \\ \mathbf{a}_n \end{vmatrix} = \begin{vmatrix} \mathbf{a}_1 \\ \vdots \\ \mathbf{a}_i \\ \vdots \\ \mathbf{a}_j \\ \vdots \\ \mathbf{a}_n \end{vmatrix} + \begin{vmatrix} \mathbf{a}_1 \\ \vdots \\ \mathbf{a}_i \\ \vdots \\ \lambda\mathbf{a}_i \\ \vdots \\ \mathbf{a}_n \end{vmatrix} = \begin{vmatrix} \mathbf{a}_1 \\ \vdots \\ \mathbf{a}_i \\ \vdots \\ \mathbf{a}_j + \lambda\mathbf{a}_i \\ \vdots \\ \mathbf{a}_n \end{vmatrix} .$$

Remark 4.3

Let us summarize some of the properties of the determinant.

(a) It is a function $\mathbb{R}^{n^2} \to \mathbb{R}$.

(b) You can view det as a function

$$\mathbb{R}^n \times \mathbb{R}^n \times \cdots \times \mathbb{R}^n \to \mathbb{R},$$

with the factors in the Cartesian product coming from the rows of the matrix. In this way, you can think of det as a function with n arguments, each argument being an element of $\mathbb{R}^n$.

(c) The map is *multilinear* in each of the n row vector variables. In other words, if you choose i in the range $1 \leq i \leq n$ and keep all arguments fixed except the variable in position i, you obtain a linear map from $\mathbb{R}^n$ to $\mathbb{R}$. To be explicit, if we fix all rows except one, determinant defines a function α from $\mathbb{R}^n$ to $\mathbb{R}$ by varying the distinguished row. To say that α is linear is to assert that

$$\alpha(\lambda\mathbf{x} + \mu\mathbf{y}) = \lambda\alpha(\mathbf{x}) + \mu\alpha(\mathbf{y})\,\forall\lambda, \mu \in \mathbb{R},\ \forall\mathbf{x}, \mathbf{y} \in \mathbb{R}^n.$$

(d) Swapping the values of two arguments changes the sign of the value of the function.

(e) If two arguments take the same value, then the value of the function is 0.

In fact the same remarks are valid if we use the columns instead of the rows, but we do not want to concern ourselves with columns in our future applications.

Remark 4.4

In the light of these properties of determinant, we calculate $|X|$ again where $X = (x_{ij})$ is a 3 by 3 matrix. Instead of using Definition 4.7, let us see how far we get in an attempt to evaluate $|X|$ using Remark 4.3.

Think of X as a column vector, its entries being row vectors. Let the i-th row be $\mathbf{x}_i$. Now let $\mathbf{e}_i$ be the i-th row of I_n, so that $\mathbf{x}_i = \sum_j x_{ij}\mathbf{e}_j$. We now use property (c) in Remark 4.3 to find that $|X|$ is

$$x_{11}x_{22}x_{33}\begin{vmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{vmatrix} + x_{11}x_{23}x_{32}\begin{vmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{vmatrix}$$

$$+x_{12}x_{23}x_{31}\begin{vmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{vmatrix} + x_{12}x_{21}x_{33}\begin{vmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{vmatrix}$$

$$+x_{13}x_{21}x_{32}\begin{vmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{vmatrix} + x_{13}x_{22}x_{31}\begin{vmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{vmatrix}.$$

In fact there are terms missing from this expression, for example terms such as

$$+x_{13}x_{23}x_{31}\begin{vmatrix} 0 & 0 & 1 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{vmatrix}$$

should really be mentioned, but thanks to Property (e) of Remark 4.3 we know that the determinant of a matrix with a repeated row must be 0. So, there are these vital 6 determinants consisting of zeros and ones, and if we know their values, then we know $|X|$ where X is an arbitrary 3 by 3 matrix. One of these all-important 6 determinants is that of the identity matrix. The other 5 important matrices can all be obtained form the identity matrix by swapping rows. Thus, thanks to Remark 4.3 Property (d), we can calculate $|X|$ once we know $|I_3|$. Our definition of determinant ensures that all identity matrices have determinant 1. In that analysis, the fact that we were working with a 3 by 3 matrix was of no significance. The same argument would work for any square matrix. We have laid the ground for a slightly tricky argument. We seek to prove the following beautiful result.

Proposition 4.6

If X and Y are n by n matrices, then

$$\det(XY) = \det(X)\det(Y). \tag{4.17}$$

Proof (harder)

We construct a function f from $\mathbb{R}^{n^2} \to \mathbb{R}$ which satisfies all the properties listed in Remark 4.3.

Think of the n by n matrix X as being a column vector, each entry being a row vector. Thus $X = (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n)^T$ where $\mathbf{x}_i$ is the i-th row. Define the function f by

$$f(X) = |(\mathbf{x}_1 Y, \mathbf{x}_2 Y, \ldots, \mathbf{x}_n Y)^T| - |X| \cdot |Y|. \tag{4.18}$$

This looks far worse than it is. All we have done is to calculate $|(XY)| - |X||Y|$. Our reason for writing the formula in this way is that it is clear (from staring at Equation (4.18)) that our function satisfies all the conditions listed in Remark 4.3. Now we can use these properties to evaluate the function exactly as was done in Remark 4.4. It therefore turns out that the value of $f(X)$ is determined by the value of $f(I_n)$. Now $f(I_n) = |I_n Y| - |I_n| \cdot |Y| = |Y| - |Y| = 0$. Thus f vanishes at all matrices obtained by swapping the rows of I_n (repeatedly), and so $f(X) = 0$ for all matrices X.

We conclude that for all n by n matrices X and Y, we have $|XY| = |X||Y|$, or in a slightly more descriptive notation, $\det(XY) = \det(X)\det(Y)$. The proof is complete.

□

EXERCISES

4.3 Evaluate the following determinant in four different ways: $\begin{vmatrix} 1 & 2 \\ 3 & 4 \end{vmatrix}$.

4.4 Evaluate the following determinant without working hard.

$$\begin{vmatrix} 1 & 2 & 3 & 4 & 5 \\ 3 & 5 & 7 & -2 & 1 \\ 4 & 7 & 10 & 2 & 6 \\ 7 & \pi & 22 & \sqrt{13} & -10^{60} \\ -11 & 14.3 & 97 & 1 & 0 \end{vmatrix}$$

If you take the transpose of a square matrix, the value of its determinant does not change.

Proposition 4.7

Suppose that $A = (a_{ij})$ is an n by n square matrix. It follows that $|A| = |A^T|$.

Proof

$|A| = \sum_{j=1}^{n} a_{ij}c_{ij}$ is the expansion of the determinant of A using its i-th row and the relevant cofactors. It is also the expansion of the determinant of A^T using its i-th column. There is an implicit induction here, since c_{ij} is (up to sign) the determinant of an $(n-1)$ by $(n-1)$ square matrix. Thus $|A| = |A^T|$ and we are done.

□

Definition 4.8

Suppose that $A = (a_{ij})$ is an n by n square matrix. Let A^* (the *adjugate* of A) be the n by n matrix whose entry in row i and column j is $a^*_{ij} = c_{ji} = |A_{ji}|\,(-1)^{i+j}$.

Here A_{ij} is the minor mentioned in Definition 4.7. Thus you obtain the adjugate of A by replacing each entry of A by its corresponding cofactor, and then transposing the matrix. This operation is largely of theoretical interest, since you would not wish to perform it on an actual matrix unless n were very small. However, the construction is of great theoretical importance as you are about to see. We remind the reader that if $X = (x_{ij})$ is a matrix and λ is a real number, then λX has entry λx_{ij} in the i-th row and j-th column.

Proposition 4.8

Let $A = (a_{ij})$ be an n by n matrix, with adjugate matrix A^* as defined above.

(a) $AA^* = A^*A = \det A \cdot I_n$.

(b) A has a right inverse if and only if $|A| \neq 0$.

(c) If A has a right inverse X, then X is also a left inverse of A, and is also the unique inverse on either side.

Proof

(a) Let $Z = (z_{ij}) = AA^*$. The entry z_{ij} is given by the formula

$$z_{ij} = \sum_{k=1}^{n} a_{ik}a^*_{kj} = \sum_{k=1}^{n} a_{ik}c_{jk} \tag{4.19}$$

Now, if $i = j$ this is a definition of $|A|$. However, if $i \neq j$, then Equation (4.19) can be viewed as a calculation of the determinant of a matrix. That matrix is almost the same as A, except that the j-th row of A has been discarded and replaced by a duplicate of the i-th row, and the determinant is being calculated by expanding along the row which has suffered this change. However, this must yield 0 since we are calculating the determinant of a matrix with a repeated row. Thus $z_{ij} = \delta_{ij}|A|$ and $Z = |A| \cdot I_n$. If $|A| \neq 0$, then $A \cdot (1/|A|)A^* = I_n$. A similar calculation yields that $A^*A = I_n$. You find yourself manipulating columns rather than rows, and, perhaps not surprisingly, you need to use Proposition 4.7 on the way.

(b) Suppose that $|A| = 0$. Suppose, for contradiction, that there is an n by n matrix B such that $AB = I_n$. Using Proposition 4.6, we see that $0 = |A| \cdot |B| = |I_n| = 1$. This is absurd, so A has no right inverse. Conversely, if $|A| \neq 0$, then $(1/|A|)A^*$ is a right inverse of A by part (a).

(c) Suppose $AY = I_n$, so $|A| \neq 0$. Pre-multiplication by $(1/|A|)A^*$ yields that $Y = (1/|A|)A^*$. A similar argument works if $YA = I_n$.

□

In the light of Proposition 4.8, if A is an invertible n by n matrix, its unique two-sided inverse can be written A^{-1}. Observe that if B is another invertible matrix of the same shape, then AB is also invertible and $(AB)^{-1} = B^{-1}A^{-1}$. This is because

$$(AB)(B^{-1}A^{-1}) = A(BB^{-1})A^{-1} = AI_nA^{-1} = AA^{-1} = I_n.$$

The theory tells us that $B^{-1}A^{-1}$ should also be a left inverse for AB, and of course it is. Also observe that if A is an an invertible n by n matrix then A^{-1} is invertible, and $(A^{-1})^{-1} = A$.

Suppose that B is an n by n matrix and there is $k \in \mathbb{N}$ such that $B^k = 0_n$ where 0_n is the matrix of zeros. Observe that $I_n - B$ is invertible, since $I_n + B + \ldots + B^{k-1}$ is its inverse.

4.10 Geometry of Determinants

This short section is just a discussion; we will not prove the assertions made here. Suppose that X is a 2 by 2 real matrix. Define a map $m_X : \mathbb{R}^2 \to \mathbb{R}^2$ via $\mathbf{a}^T \mapsto X\mathbf{a}^T$ for every $\mathbf{a}^T \in \mathbb{R}^2$ (we are working with column vectors). Thus we multiply a column vector on the left by a square matrix to yield another column vector. Suppose that $S \subseteq \mathbb{R}^2$; define $m_X(S) = \{m_X(s) \mid s \in S\}$. This is definitely abuse of notation, but no-one is looking. So $m_X(S) \subseteq \mathbb{R}^2$. The remarkable fact about the map m_X is that it sends any region of area d to a region of area $\det(X) \cdot d$. If $\det(X) = 0$, then either $X = O_2$ and m_X sends all of $\mathbb{R}^2$ to $(0, 0)$, or $m_X \neq O_2$ and the image of m_X is a straight line through the origin. In either circumstance, all subsets of $\mathbb{R}^2$ are either shrunk to a point or flattened to become a subset of a line. Thus the area of the image must vanish, and that is the geometric reason why $\det(X) = 0$. On the other hand, if $\det(X) \neq 0$, then it turns out that m_X is a bijection, and enjoys the curious property that it multiplies area by a fixed quantity.

Notice that m_{I_2} is the identity map from $\mathbb{R}^2$ to $\mathbb{R}^2$, and that $\det(I_2) = 1$. Furthermore, notice that if X and Y are both 2 by 2 matrices, then $m_{XY} = m_X \circ m_Y$. This is because you can multiply a column vector on the left by the matrix XY by first multiplying it by Y, and then by X. If you accept the truth of the assertions about area, it follows that $\det(XY) = \det(X)\,\det(Y)$.

An entirely similar story is true in the case of n by n matrices. Thus from this geometric point of view, the fact that $\det(XY) = \det(X)\,\det(Y)$ is completely clear. The subtlety of the algebraic proof of Proposition 4.6 is not needed. Of course, you would have to go to the trouble of setting up the notion of volume in $\mathbb{R}^n$, and that is the main reason we have avoided the geometric route. Another point in favour of the algebraic proof is that it works for matrices with entries in any field.

EXERCISES

4.5 Show that if $\mathbf{a}, \mathbf{b} \in \mathbb{R}^2$ and $\lambda \in \mathbb{R}$, then we have both of the following equations.

(a) $m_X(\mathbf{a} + \mathbf{b}) = m_X(\mathbf{a}) + m_X(\mathbf{b})$.

(b) $m_X(\lambda\mathbf{a}) = \lambda m_X(\mathbf{a})$ for all $\lambda \in \mathbb{R}$.

4.11 Linear Independence

We will work with row vectors in $\mathbb{R}^n$.

Definition 4.9

We say that a finite sequence $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_m$ of vectors is *linearly dependent* if there are real numbers $\lambda_1, \lambda_2, \ldots, \lambda_m$ which are not all 0, such that

$$\sum_{i=1}^{m} \lambda_i \mathbf{v}_i = 0. \tag{4.20}$$

The reason for the terminology is this. Suppose that $\lambda_1 \neq 0$, then

$$\mathbf{v}_1 = -(1/\lambda_1)(\lambda_2 \mathbf{v}_2 + \ldots + \lambda_m \mathbf{v}_m),$$

so you can express $\mathbf{v}_1$ as a linear combination of the other vectors in the sequence; if you like, $\mathbf{v}_1$ depends on the other vectors. Of course, it needn't be $\mathbf{v}_1$ which can be expressed as a linear combination of the others. Given that our finite sequence of vectors is linearly dependent, all you know is that at least one of the vectors can be expressed as a linear combination of the others.

Definition 4.10

If a finite sequence $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_m$ of vectors is not linearly dependent, we say that it is *linearly independent.*

This a very powerful idea; if you know that $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_m$ is a linearly independent sequence of vectors, and that you have real numbers $\alpha_1, \ldots, \alpha_m$ such that $\sum_i \alpha \mathbf{v}_i = 0$, then you can immediately deduce that $\alpha_i = 0$ for every i.

Proposition 4.9

Suppose that $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_m$ is a linearly independent sequence of vectors, and that there is a vector $\mathbf{u}$ which you can write as a linear combination $\mathbf{u} = \sum_i \beta_i \mathbf{v}_i$, then this representation is unique (i.e. you can't do the same thing using other scalars).

Proof

Suppose that there is a rival expression $\mathbf{u} = \sum_i \gamma_i \mathbf{v}_i$, then subtract the two competing equations to obtain

$$\mathbf{0} = \mathbf{u} - \mathbf{u} = \sum_i (\beta_i - \gamma_i)\mathbf{v}_i.$$

Now apply the definition of linear independence to deduce that $\beta_i - \gamma_i = 0$ for every i. Thus $\beta_i = \gamma_i$ for $1 \leq i \leq m$ as required.

□

4.12 Vector Spaces

Definition 4.11

A *concrete vector space* is a subset V of $\mathbb{R}^n$ which satisfies the following three conditions.

(i) If $\mathbf{u}, \mathbf{v} \in V$, then $\mathbf{u} + \mathbf{v} \in V$.

(ii) If $\mathbf{u} \in V$ and $\lambda \in \mathbb{R}$, then $\lambda \mathbf{u} \in V$.

(iii) $\mathbf{0} \in V$.

Of course, there is an entirely analogous notion of a geometric vector space (this being a subset of the 3-dimensional space of geometric vectors which obeys these same axioms). Moreover, the only such subsets are $\{\mathbf{0}\}$, straight lines through the origin, planes through the origin and all of 3-dimensional Euclidean space. Moreover, moving up the ladder to get a better view, a subset of an abstract vector space which satisfies our axioms will itself be an abstract vector space; this is simply a matter of checking definitions.

Definition 4.12

Suppose that we have a finite sequence $\mathbf{v}_1, \mathbf{v}_2 \ldots, \mathbf{v}_t \in V$ of vectors in a concrete vector space V. We define the *span* of this sequence to be the set of vectors

$$\langle \mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_t \rangle = \{\sum_{i=1}^{t} \lambda_i \mathbf{v}_i \mid \lambda_i \in \mathbb{R} \forall i = 1, \ldots, t\}.$$

The most important sequences of linearly independent vectors in a subspace V are those which also span V. Such a sequence is called a ***basis*** of V. In view of Proposition 4.9, every element of V will be uniquely representable as a linear combination of basis elements.

It is true that (a) every concrete vector space has a basis containing finitely many vectors, and (b) any two bases of a concrete vector space are sequences of the same length. The length of a sequence forming a basis is therefore an important number; it is called the ***dimension*** of the subspace. Working in $\mathbb{R}^3$ this terminology agrees with the casual idea of dimension. Note that we have not proved these statements about dimension – but have merely asserted them.

We give a proof of (b).

Theorem 4.1

Any two bases of a concrete vector space are sequences of the same length.

Proof

Suppose that V is a concrete vector space with bases $\mathbf{u}_1, \ldots, \mathbf{u}_r$ and $\mathbf{v}_1, \ldots, \mathbf{v}_s$. We express the terms of each basis as linear combinations of the other basis via $\mathbf{u}_i = \sum_j a_{ij}\mathbf{v}_j$ and $\mathbf{v}_i = \sum_j b_{ij}\mathbf{u}_j$ for every legal i, and the sums are taken over every legal j. Thus $A = (a_{ij})$ is an r by s matrix and $B = (b_{ij})$ is an s by r matrix. Assume, for contradiction, that $r > s$. Now AB expresses each $\mathbf{u}_i$ in terms of $\mathbf{u}_1, \ldots, \mathbf{u}_r$, so $AB = I_r$. Pad the right of A with 0's, and the bottom of B with λ's (where $\lambda \in \mathbb{R}$) to obtain square matrices $\overline{A}$ and $\overline{B}(\lambda)$. Moreover, those 0's ensure that $\overline{A} \cdot \overline{B}(\lambda) = I_r$. However, λ is arbitrary, so $\overline{A}$ has infinitely many, and so more than one, right inverse. This contradicts Proposition 4.8. We deal with the case $r < s$ similarly, so $r = s$ as required.

□

For a full development of this theory, the reader should consult a textbook on linear algebra.

As we have already mentioned, when viewed geometrically, the subspaces of $\mathbb{R}^3$ are exactly its subsets of the following types: $\{\mathbf{0}\}$, straight lines through the origin, planes through the origin and the whole of $\mathbb{R}^3$. These subspaces have dimension 0, 1, 2, and 3 respectively. To get that the $\{\mathbf{0}\}$ is 0-dimensional requires a little faith, but that is the best way to define it since statements of theorems then work out cleanly.

EXERCISES

4.6 Prove the following. The rows of the n by n identity matrix are linearly independent and span $\mathbb{R}^n$.

4.7 Any subsequence of a linearly independent sequence of vectors is linearly independent (a subsequence of a sequence of vectors is obtained by omitting some vectors).

4.13 Transposition

Definition 4.13

A square matrix X is *symmetric* if $X = X^T$.

Definition 4.14

A square matrix X is *alternating* or *antisymmetric* if $X = -X^T$.

The entries on the leading diagonal of an alternating matrix must vanish.

Proposition 4.10

Suppose that A and B are square matrices of the same size. It follows that both

(i) $(A+B)^T = A^T + B^T$ and

(ii) $(AB)^T = B^T A^T$.

Proof

(i) is a triviality and we omit the proof.

(ii) Let $A = (a_{ij})$, $B = (b_{ij})$, $X = (x_{ij}) = A^T$ and $Y = (y_{ij}) = B^T$. The entry in the i-th row and j-th column of $(AB)^T$ is the entry in row j and column i of AB. This is $\sum_k a_{jk}b_{ki}$. On the other hand, the entry in the i-th row and j-th column of $B^T A^T$ is $\sum_k y_{ik}x_{kj} = \sum_k b_{ki}a_{jk} = \sum_k a_{jk}b_{ki}$. □

Definition 4.15

A square matrix X is said to be *orthogonal* if $XX^T = I$ where I is the appropriate identity matrix.

Here is an example of a family of interesting orthogonal matrices:

$$\left\{ \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix} \mid \theta \in \mathbb{R} \right\}.$$

Proposition 4.11

For a fixed natural number n, the n by n orthogonal matrices enjoy the following properties.

(i) If X is orthogonal, then $|X| = \pm 1$.

(ii) If X, Y are orthogonal, then XY is orthogonal.

(iii) The identity matrix I_n is orthogonal.

(iv) Orthogonal matrices are invertible, and their inverses are orthogonal matrices.

Proof

(i) $XX^T = I_n$ so $|XX^T| = 1$ and therefore $|X||X^T| = 1$ by Proposition 4.6. Also $|X^T| = |X|$ by Proposition 4.7 so $|X|^2 = 1$ and we are done.

(ii) $XY(XY)^T = XYY^TX^T = X(YY^T)X^T = XI_nX^T = XX^T = I_n$.

(iii) $I_nI_n^T = I_nI_n = I_n$.

(iv) $XX^T = 1$ so X is invertible, and its inverse is X^T. Now $X^T(X^T)^T = X^TX = X^{-1}X = I_n$ so X^T is an orthogonal matrix.

When you have understood Chapter 5, Proposition 4.11 will assume more meaning.

5
Group Theory

5.1 Permutations

We now discuss permutations, and rather than use the language of functions developed in Chapter 1, we will first try to get the idea across without using fancy notation.

We are all familiar with the idea of swapping things round. Suppose that you have three boxes labelled 1, 2 and 3. Put three different items into the boxes, perhaps an apple in box 1, a banana in box 2 and a cantaloup in box 3.

Consider the operation of swapping the contents of boxes 1 and 2. We need a snappy name for this operation. Let's be original and call it x. We let y denote the operation of swapping the contents of box 2 and box 3. Now look at the effect on the fruit of first doing x and then y. The apple will end up in box 3, the banana in box 1 and (by elimination or thinking) the cantaloup in box 2. Each operation of rearranging the contents of the boxes is called a *permutation*.

We must also decide how to write the operation that consists of first doing x then doing y. There are two obvious choices; $x \cdot y$ or $y \cdot x$. Partly because we read from left to right, and partly because it turns out to make life simpler later, we will write "first x, then y" as $x \cdot y$. The little central dot is supposed to remind you of multiplication, but what is going on is really composition of maps. The way in which we combine permutations guarantees that this operation will be associative; this also clear from the "composition of maps" perspective; if you can't see it, look ahead to Section 5.3.

Rewind time back to the start, so that the apple is parked in box 1 once more. This time consider the effect of first doing y then x on the apple. Doing y has no effect on the contents of box 1, and the turmoil in the other boxes leaves the apple undisturbed. Next the application of x moves the apple to box 2. In summary, "x then y" leaves the apple in box 3 but "y then x" puts the apple in box 2. We have not considered the effects on the banana and the cantaloup, but we do know that "x then y" does not yield the same rearrangement as "y then x".

Definition 5.1

Suppose that u and v are permutations of the same collection of objects. We write $u = v$ if u and v have exactly the same effect on each item being permuted.

Thus it follows that $x \cdot y \neq y \cdot x$.

In a cunning notational manoeuvre that the reader may have sensed imminent, we shall replace each fruit by the capitalized initial letter of its name.

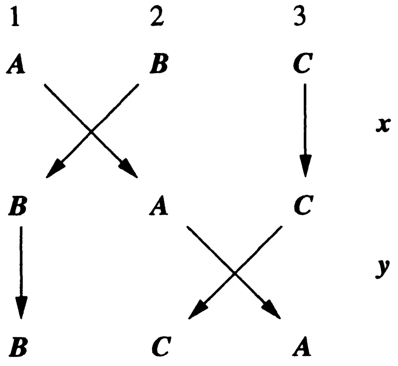

Fig. 5.1. x then y

If you compare the effects of $x \cdot y$ and $y \cdot x$ on each one of A, B or C you find that you get different answers. See Figures 5.1 and 5.2. Now compare x and $y \cdot x$. We examine the effect on A. Well, in each case A ends up in box 2, so as far as A is concerned, there is no difference between x and $y \cdot x$. Any temptation to write $x = y \cdot x$ is removed if we consider their effects on B. Please check that x and $x \cdot x \cdot x$ have the same effect on each of A, B and C so we are allowed to write $x = x \cdot x \cdot x$. This last expression is a bit cumbersome – we shorten it to x^3, so $x = x^3$. Notice that associativity is not really an issue here. We say that $x \cdot y$ is the ***product*** of x and y and that we have ***multiplied*** x by y.

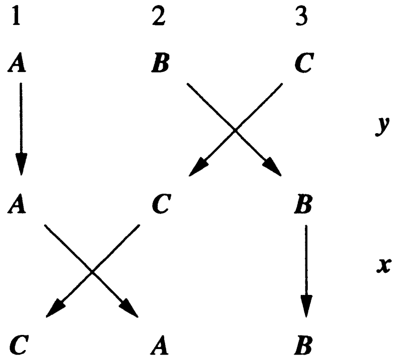

Fig. 5.2. y then x

We now introduce an even slicker way of describing permutations, and to give ourselves a bit more room, let us suppose that we have six boxes labelled 1 through 6, containing six objects labelled A to F. Consider the permutation p which leaves box 1 alone, puts the contents of box 2 in box 4, the contents of box 3 in box 6, the contents of box 4 in box 5, the contents of box 5 in box 2 and the contents of box 6 in box 3.

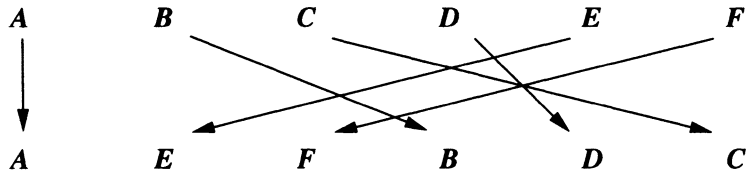

Fig. 5.3. Six objects being permuted

This isn't a very economical way of describing this permutation. Here is the slick way: $(2,4,5)(3,6)$. You read this notation like this; the fact that 1 is not mentioned means that you leave the contents of box 1 alone. The contents of box 2 gets put in box 4 because 4 is on the right of 2, and the contents of box 4 gets put in box 5 for the same reason. We have a little problem with the contents of box 5 – but the game is to regard 2, 3 and 5 as being arranged in a circle – if you like 2 *wraps around* and is deemed to be immediately to the right of 5 for these purposes. So far we know what happens to the contents of boxes 1, 2, 4 and 5. The final part tells us that the contents of boxes 3 and 6 are swapped because 6 is written to the right of 3, and in the "wrap around" sense, 3 is also to the right of 6.

Definition 5.2

A *cycle* is a permutation which is of the form $(*, *, \ldots, *)$. If a cycle has n entries it is called an n-cycle. Two cycles are *disjoint* if they have no entry in common.

Notice that every permutation can be written as a product of pairwise disjoint cycles, and that disjoint cycles commute. The first fact follows from our new procedure used to write down permutations, and the second is clear.

By the way, sometimes it is convenient to put the unmentioned numbers back in the notation, so $(2,4,5)(3,6)$ could be written $(1)(2,4,5)(3,6)$ and the initial (1) serves to remind you that the contents of box 1 must be left alone.

Notice that $(4,5,2)(6,3)$ is exactly the same permutation as $(5,2,4)(3,6)$ and it is also the same as $(6,3)(4,5,2)$. If you want to write your notation in a standard form, then the trick is to write each cycle with the smallest number at the beginning, and then to sort the cycles according to their first entry. If you do that in this case, you write the cycles as $(2,4,5)$ and $(3,6)$ and since 2 is smaller than 3 you write the permutation as $(2,4,5)(3,6)$. Notice, by the way, that both $(2,4,5)$ and $(3,6)$ considered as permutations themselves are both written in standard form, and their product is obtained by erasing the multiplication sign, since $(2,4,5)\cdot(3,6) = (2,4,5)(3,6)$. If you do the multiplication the other way round you have $(3,6)\cdot(2,4,5) = (3,6)(2,4,5) = (2,4,5)(3,6)$.

The other great advantage of this way of writing permutations is that it makes it very easy to work out the result of first doing one then another. Also notice that the Roman letters have completely disappeared from the notation. This is not that surprising really, since it doesn't really matter what the objects in the boxes are, as long as they are different.

Now we will demonstrate that combining permutations is very easy in our notation. Let $p = (2,4,5)(3,6)$ and $q = (1,3,4)(2,6)$. We will show that $p \cdot q = (1,3,2)(4,5,6)$. You write p and q next to one another like this

$$(2,4,5)(3,6)(1,3,4)(2,6)$$

and then run your eye from left to right. Start of with the contents of box 1. The initial segment $(2,4,5)(3,6)$ tells you to leave the contents of box 1 alone, but life starts to get more interesting when you reach $(1,3,4)$. This tells you to put the contents of box 1 into box 3. In future we shall be looking for instructions telling us what to do with the contents of box 3, since that is where the original contents of box 1 now resides. Read on. We have the final instruction $(2,6)$, which tells us to leave the contents of box 3 alone. The upshot of this paragraph is that the contents of box 1 is now in box 3. Exciting isn't it? Well, of course not, but it is easy. You can begin to write down the answer $(1,3$ – but you don't yet know what will happen to the contents of box 3.

Now let us find out. Start again with $(2,4,5)(3,6)(1,3,4)(2,6)$ and read from left to right, looking for instructions on what to do with the original contents of box 3. The first interest is when you get to $(3,6)$ which tells you to put the contents of box 3 into box 6. Now you refocus your interest on the contents of box 6. The next instruction $(1,3,4)$ tells you to leave it alone but the final one $(2,6)$ tells you to put it in box 2. Thus the original contents of box 3 ends up in box 2 (after a brief and irrelevant sojourn in box 6). We can write more of the answer now. It is $(1,3,2$ – and we are getting there. Now start again with $(2,4,5)(3,6)(1,3,4)(2,6)$ and focus on the original contents of box 2. It immediately is put in box 4 by $(2,4,5)$ and is moved on to box 1 by $(1,3,4)$. The upshot is that we are able to close off our partial answer $(1,3,2)$ like that. It remains to see what happens to the original contents of boxes 4, 5 and 6. First focus on box 4. The first instruction $(2,4,5)$ sends it to box 5 and the contents of box 5 is never subsequently moved. Thus our partial answer is $(1,3,2)(4,5$ – nearly there. Now the original contents of box 5 doesn't have too many choices in life. It must end up in box 4 or box 6, since we have established that there are going to be other objects parked in boxes 1, 2, 3 and 5 when this is all over. Let's find out which. The original contents of box 5 gets put into box 2 by $(2,4,5)$ and is unmoved for a while until we reach $(2,6)$ and it is put in box 6. Our partial answer is now $(1,3,2)(4,5,6$. Now you can calculate yourself that the original contents of box 6 will end up in box 4, or you can just observe that it must end up in box 4 since all the other boxes will be occupied. In any event our answer is $(1,3,2)(4,5,6)$ and that took far longer to explain than to do.

EXERCISES

In these exercises, each permutation is deemed to be a permutation of the contents of five boxes.

5.1 Simplify the following products of permutations.

(a) $(1,3,5,4,2)\cdot(1,3,2,5,4)$

(b) $(1,2,3)\cdot(2,4)(1,3,5)$

(c) $(1,3,4)(2,5)\cdot(2,3,4)(1,5)$

(d) $(1,2)(3,4)\cdot(1,2)$.

5.2 Find a permutation λ with the property that $\lambda\cdot(1,2,3)(4,5) = (1,3,5,4,2)$.

5.3 Find a permutation ρ with the property that

$$(1,2,3)(4,5)\cdot\rho=(1,3,5,4,2).$$

5.4 Find a permutation α with the property that

$$\alpha\cdot(1,2,3)=(1,2,3)\cdot\alpha.$$

There is a special permutation which is so dull that it deserves a distinguished name. This is the *identity permutation* of the contents of the boxes which leaves everything where it is. In our slick notation it is written as *nothing at all!* This is elegant but drastic. On the other hand, writing it as a product of 1-cycles is too cumbersome; if you have five boxes the identity permutation could be written as $(1)(2)(3)(4)(5)$. We write it as id, an obvious abbreviation.

Notice that no matter what the permutation x, we have $x\cdot\mathrm{id}=x=\mathrm{id}\cdot x$.

5.2 Inverse Permutations

The inverse of a permutation can be worked out with a video camera. Record a permutation taking place, then play back the recording in reverse. The inverse of $(1,2,3)(4,5)$ sends the contents of box 1 to box 3 (where it came from via $(1,2,3)(4,5)$), the contents of box 3 to box 2, and so on.

We write the inverse of the permutation π as π^{-1}, and then $\pi\cdot\pi^{-1}=\mathrm{id}=\pi^{-1}\cdot\pi$.

The easy way to work out the inverse of a permutation is to write the whole thing backwards – write the cycles in reverse order and reverse the contents of each cycle – and then tidy up into standard form if you want to. In this case the inverse of $(1,2,3)(4,5)$ is $(5,4)(3,2,1)$ and in standard form this is $(1,3,2)(4,5)$. This even works if you haven't simplified a product; for example, suppose that $\theta=(1,2,3)(2,3)(1,2,3)$. You can simplify θ to get $\theta=(2,3)$ and so $\theta^{-1}=(3,2)=(2,3)$. On the other hand you can work this out by writing the original expression for θ backwards, and then simplifying; in this case $\theta^{-1}=(3,2,1)(3,2)(3,2,1)=(2,3)$ which is the same answer, as promised.

We make the convention that $\pi^0=\mathrm{id}$ for all permutations π. We want to give meaning to π^{-t} where $t\in\mathbb{N}$. There are two candidates for π^{-t}; it could be $(\pi^t)^{-1}$ or $(\pi^{-1})^t$. Fortunately these last two permutations coincide, since they are each the inverse permutation of π^t so we can unambiguously assign π^{-t} to either of them.

The usual algebra of exponents works perfectly well. If $\alpha,\beta\in\mathbb{Z}$ and π is a permutation then $(\pi^\alpha)^\beta=\pi^{(\alpha\beta)}$ and we often write the latter as $\pi^{\alpha\beta}$ in the usual way. Similarly $\pi^{\alpha+\beta}=\pi^{(\alpha+\beta)}=\pi^\alpha\cdot\pi^\beta$.

EXERCISES

5.5 Work out the inverse of each of the following permutations.

(a) $(1,2,5)(2,3)$

(b) $(1,2,5)\cdot(2,3)$

(c) $(1,2,3,4)^5$

(d) id

(e) $(1,2,3,4)\cdot(2,3,4,5)\cdot(1,3,5)$

5.6 (A little harder). Find a permutation ψ such that $\psi^{-1}\cdot(1,2,3,4)\cdot\psi = (1,3,2,4)$.

Now let's see how we could have expressed all this slickly using sets and functions. Suppose that we want to describe the permutation $(1,2)(3,4,5)$ as a map $\pi : \{1,2,3,4,5\} \to \{1,2,3,4,5\}$. Let $\pi : 1 \mapsto 2,\ 2 \mapsto 1,\ 3 \mapsto 4,\ 4 \mapsto 5,$ and $5 \mapsto 3$. Since we are composing permutations from left to right it makes sense to write π on the right of its argument, so $(1)\pi = 2$, etc.

So, what is it about this map $\pi : \{1,2,3,4,5\} \to \{1,2,3,4,5\}$ which makes it a permutation? It should now be clear; it is the fact that it is a bijection. A permutation is just a bijection from $\{1,2,\ldots,n\}$ to itself.

5.3 The Algebra of Permutations

We have a way of multiplying permutations together, and we can invert them. We have briefly mentioned the ***associative law*** for permutation multiplication. We remind you of the issue here. The symbols α, β and γ denote permutations, and we want to work out $\alpha\cdot\beta\cdot\gamma$. Someone who is very officious might say "Hold on a minute, what do you mean by $\alpha\cdot\beta\cdot\gamma$ – it could be $(\alpha\cdot\beta)\cdot\gamma$ or $\alpha\cdot(\beta\cdot\gamma)$ and you haven't said which so I am totally confused". Well, it's time to get unconfused. The point is that $(\alpha\cdot\beta)\cdot\gamma = \alpha\cdot(\beta\cdot\gamma)$ so it doesn't matter. Both sides mean do α, then β, then γ, they just tell you to do that in slightly different ways.

However, we should be able to see this another way now. Permutations are really maps (bijections in fact but that is irrelevant). Permutations are composed as maps, and map composition is associative. Thus permutation composition is associative. There will come a time when you will want to simply omit brackets on the ground that they don't matter. Fair enough, but not yet please.

Another rather obvious but important property of the multiplication of permutations is that it is a *closed operation*, in other words, when you do one permutation then another, the result amounts to doing a single permutation, and not something which is not a permutation, like a waltz for example. This is clear if you think in terms of rearrangements. Rearranging the contents of the boxes, and then rearranging the contents again could have been accomplished by a single rearrangement.

There is the distinguished *identity permutation* which is just the identity map from the underlying set to itself.

Each permutation has a *two-sided inverse*. This is just the inverse map which exists since permutations are bijections.

These four algebraic properties (closure, associativity, identity, inverses) form the basis of what will become group theory.

Definition 5.3

A *permutation group* S is a non-empty collection of permutations (of distinct objects in n boxes) which is closed under multiplication and inversion.

Note that it follows that the relevant identity permutation is in the permutation group.

Definition 5.4

The *order* of a permutation group S is simply $|S|$. The *degree* of a permutation group is the number of boxes involved.

The set of all permutations of the n boxes is called the symmetric (permutation) group of degree n, and we shall write it as S_n. It is not hard to see that $|S_n| = n!$. The permutation group S_1 is not worth writing home about, and $S_2 = \{\text{id}, (1,2)\}$ doesn't stir the blood either. We pick the smallest interesting value of n, which is of course 3, and try to understand everything there is to be known about S_3.

This little central dot $\cdot$ can get a bit irritating. From now on we will usually just omit it.

First of all, we give short names to the $3! = 6$ elements of this set. One is called id already, so is well equipped. Let $\alpha = (1,2)$, $\beta = (2,3)$, $\gamma = (1,3)$, $\delta = (1,2,3)$ and $\varepsilon = (1,3,2)$. We will make a table which gives all possible products. Notice, for example, that $\alpha\gamma = (1,2)(1,3) = (1,2,3) = \delta$. Thus where the row labelled α meets the column labelled γ we write the letter δ. The only way you can get this wrong is to get rows and columns confused. So don't.

S_3	id	α	β	γ	δ	ε
id	id	α	β	γ	δ	ε
α	α	id	ε	δ	γ	β
β	β	δ	id	ε	α	γ
γ	γ	ε	δ	id	β	α
δ	δ	β	γ	α	ε	id
ε	ε	γ	α	β	id	δ

EXERCISES

Simplify the following expressions, using the notation given for the permutations in S_3. In each case your answer should be a single Greek letter.

5.7 (a) $\alpha(\beta\gamma)$

(b) $(\alpha\beta)\gamma$

(c) α^2

(d) α^3

(e) δ^2

(f) δ^3

(g) $(\alpha\beta\varepsilon\gamma\delta)^6$

(h) $(\beta\alpha\beta\varepsilon)^{216}$

5.4 The Order of a Permutation

In the group S_3 discussed in Section 5.3 we have

$$\mathrm{id}^1 = \alpha^2 = \beta^2 = \gamma^2 = \delta^3 = \varepsilon^3 = \mathrm{id}.$$

By inspection of the multiplication table of S_3 we find that the exponents given are the smallest positive natural numbers such that raising each given permutation to that power yields the identity element.

Definition 5.5

Suppose that $\pi \in S_n$ is a permutation. The *order* of π, written $o(\pi)$, is the smallest natural number n such that $\pi^n =$ id.

Now, this looks slightly worrying – might it not be the case that no positive power of some particularly perverse permutation is the identity? In fact that cannot happen. Also, the word "order" is getting heavy use, but in fact this is a good idea, as we will see in Section 5.9.

Theorem 5.1

Every permutation $\pi \in S_n$ has an order $o(\pi)$.

Proof

First consider a cycle $\gamma = (*, *, *, \ldots, *)$ containing m entries. Now $\gamma^m =$ id and no smaller positive exponent has that property so $o(\gamma)$ exists and is m. It is easy to show that if $u \in \mathbb{Z}$, then there are $q, r \in \mathbb{Z}$ with $0 \leq r < m$ such that $u = qm + r$. Now $\gamma^u = (\gamma^m)^q\gamma^r = \gamma^r$ so $\gamma^u =$ id if and only if m divides u. From the theory, any permutation π can be expressed as a product of disjoint cycles γ_i so $\pi = \prod_{i=1}^t \gamma_i = \gamma_1 \ldots \gamma_t$ and as we have already observed, disjoint cycles commute. Raising such a product to a positive power u is very easy:

$$\pi^u = (\gamma_1 \ldots \gamma_t)^u = \gamma_1^u \ldots \gamma_t^u.$$

Since different γ_i move different numbers, we have that $\pi^u =$ id if and only if $\gamma_i^u =$ id for each $i = 1, \ldots, t$. Each γ_i is a cycle of order m_i. Now $\pi^u =$ id if and only if u is a multiple of each m_i. The order of π therefore exists and is the lowest common multiple of the lengths of the disjoint cycles comprising π.

□

For example, the order of $(1,2)(3,4,5)(6,7,8,9)$ is the lowest common multiple of 2, 3 and 4 and so is 12. Notice, by the way, that a power of a cycle need not be a cycle, since $(1,2,3,4)^2 = (1,3)(2,4)$.

Definition 5.6

A cycle of length 2 is called a *transposition*.

EXERCISES

5.8 Which permutations are positive powers of cycles?

5.9 In each case, find the order of the given permutation.

(a) (1,2)(3,4)

(b) (1,2,3)(4,5,6)

(c) (1,2)(1,3)

(d) (1,2,3)(1,3,2)

(e) (1,2,3)(2,3,4)(3,4,5)(4,5,1)(1,4,5)

(f) (1,2)(1,3)(1,4)(1,5).

5.10 (a) A permutation of order 2 must be transposition. True or false?

(b) Show that any cycle is a product of transpositions.

(c) Show that any permutation is a product of transpositions.

5.5 Permutation Groups

Recall from Definition 5.3 that a *permutation group of degree n* is a non-empty subset S of S_n which is closed under multiplication and inversion. In symbols we could write $S \neq \emptyset$, and if $x, y \in S$, then $xy \in S$ and $x^{-1} \in S$.

When someone defines a concept abstractly, it often helps to have concrete examples in mind, at least initially. There are two obvious examples of permutation groups of degree n. They are S_n itself, and its subset $\{\mathrm{id}\}$.

However, there are examples which sit between these two extremes. Let $\tau = (1,2,3,4)(5,6) \in S_6$ and let T be the set of all powers of τ, so $T = \{\tau^z \mid z \in \mathbb{Z}\}$. Now T is visibly closed under multiplication and inversion, and contains the identity element τ^0 so T is itself a permutation group.

Similarly, the set of all powers of an arbitrary permutation is a permutation group. There are interesting groups which do not arise in this way. For example the four elements id, λ, μ, ν of S_4 form a group where $\lambda = (1,2)(3,4)$, $\mu = (1,3)(2,4)$, and $\nu = (1,4)(2,3)$. The group $K = \{\mathrm{id}, \lambda, \mu, \nu\}$ has the property that all of its elements commute with one another.

To see it arise geometrically, take a rectangular piece of plain paper without any holes in it, and which is not square. Label the four corners 1,2,3,4 clockwise, and label the back of the paper with the same numbers so that a given corner has the same label on both sides. If you cannot tell clockwise from anti-clockwise, it is your lucky day.

Now imagine that the universe is completely empty apart from this piece of paper. Now remove the piece of paper from the universe, leaving (obviously) a piece-of-paper shaped hole in the void. Now put the paper back in the hole in any way it will fit. Now, there are exactly four ways to do this, two with the paper turned over and two with it the same way up. These four moves correspond to id, λ, μ, and ν when you keep track of the labels on the corners. Don't let life pass you by; do it! The four group elements are the four symmetries of the piece of paper.

EXERCISES

5.11 Repeat the procedure with a square piece of paper. How big is the group that you get? What are its elements (as permutations)?

5.12 Repeat the procedure with a regular pentagonal piece of paper. How big is the group that you get? What are its elements (as permutations)?

5.13 Repeat the procedure with a triangular piece of paper. In each of the following cases, how big is the group that you get? What are its elements (as permutations)?

(a) The triangle is equilateral.

(b) The triangle is isosceles but not equilateral.

(c) The triangle is scalene.

Confession time: We defined a permutation group to be a group consisting of permutations in S_n for some $n \in \mathbb{N}$. The fact that the labels on the boxes were $1, 2, \ldots, n$ *was not very important.* You could use the elements of any finite set of appropriate size. You can always rename the set as $1, 2, 3, \ldots n$ of course.

5.6 Abstract Groups

The great thing about permutation groups is that the elements are really concrete things. You know how to multiply them and that is that. Now we will explore the idea of an ***abstract*** group – which is one level of abstraction up from permutation groups. Abstract groups are usually just called groups.

This is how to make an abstract group. You need a set, often called G, and a binary operation $\star$ on G. A binary operation on a set G is a rule which allows you to "multiply" any two elements a, b of the set G together to produce something called $a \star b$. The fancy view is that it is a map $\mu : G \times G \to H$ for some set H, and $a \star b$ is another name for $\mu((a, b))$ – or if you are less fussy, $\mu(a, b)$. Moreover, H is a set sufficiently big to hold each $a \star b$, and may have other stuff in it too. Of course these two perspectives amount to the same thing. After about 10 minutes, it is usual to get fed up with writing $a \star b$ and write ab instead.

As you might have guessed, there are four axioms (rules) which this binary operation $\star$ must satisfy: *closure, associativity, identity* and *inverses.* Such is the passion for habit of our species that these axioms are invariably listed in that order.

Definition 5.7

A set G equipped with a binary operation $\star$ is called a *group* if all of the following axioms are satisfied;

$$\forall a, b \in G \quad a \star b \in G. \text{ (closure)}$$

$$\forall a, b, c \in G \ (a \star b) \star c = a \star (b \star c). \text{ (associativity)}$$

$$\exists e \in G \ \forall a \in G \quad e \star a = a = a \star e. \text{ (identity)}$$

$$\forall a \in G \ \exists b \in G \ a \star b = e = b \star a. \text{ (inverses)}$$

Let elaborate on these axioms. First a translation into English. We put them in words.

closure: For every $a, b \in G$ their product $a \star b$ is in G.

associativity: For every $a, b, c \in G$ it follows that $(a \star b) \star c = a \star (b \star c)$.

identity: There is $e \in G$ such that $e \star a = a = a \star e$ whenever $a \in G$.

inverses: Given any $a \in G$ there is $b \in G$ such that $a \star b = e = b \star a$.

Now, any permutation group satisfies these axioms, so anything you can prove true about a group must be true about a permutation group. However, the set $\mathbb{Q}^*$ of non-zero rational numbers under multiplication also satisfies the axioms, and that isn't a permutation group (or rather it isn't usually thought of as a permutation group). In this case $e = 1 \in \mathbb{Q}^*$ and if $q = r/s \in \mathbb{Q}^*$ with $r, s \in \mathbb{Z}$ then $q^{-1} = s/r \in \mathbb{Q}^*$ as required. Groups arise all over mathematics. For example, Proposition 4.11 shows that for each $n \in \mathbb{N}$ the n by n orthogonal matrices form a group under matrix multiplication.

You can add extra axioms, which corresponds to looking at a more special type of group. Perhaps the most obvious one is to append the axiom of commutativity:

$$\textit{commutative: } a \star b = b \star a \ \forall a, b \in G.$$

A group which satisfies this extra axiom is called a commutative group (a bit of a surprise there) or, more usually, an *abelian* group. Now this concept is named in honour of the Norwegian mathematician N. H. Abel. You might expect that as a proper adjective, it should begin with a capital letter. In fact common usage has it with a small letter, a tribute to the importance of the concept. When we use the unadorned term group we do not specify whether or not the group is abelian.

Another point: if $A, B \subseteq G$ then let $AB = \{a{\star}b \mid a, b \in G\}$. Thus subsets of G may be multiplied to give subsets of G. The associativity of $\star$ is transmitted and multiplication of subsets is also associative since

$$(A \star B) \star C = \{(a \star b) \star c \mid a \in A,\ b \in B,\ c \in C\}$$
$$= \{a \star (b \star c) \mid a \in A,\ b \in B, c \in C\} = A \star (B \star C).$$

Abuses of notation: In the event that a subset is a singleton set, we will happily abuse notation to write x instead of $\{x\}$. If $A \subseteq G$ and $x \in G$ we write Ax rather than the correct but cumbersome $A\{x\}$. This is also particularly important as far as the identity element e of a group G is concerned. We are sometimes led to think about the group $\{e\}$, a so-called trivial group. It gets tiresome to write the brackets all the time, and this subset of G is often written just as e.

Flexibility of notation: The identity element of a group G, or the set which contains it, is often written e_G to emphasize the group in question. If the group operation is written as multiplication or juxtaposition then we can use 1_G or even, provided the group is either clear or irrelevant, simply 1. When the operation is commutative, groups are sometimes written using additive notation, in which case it makes sense to use 0_G or just 0.

We are now going to develop some group theory. Notice that we know absolutely nothing about the set G, except that it is non-empty – this is true because $1 \in G$. We are now going to omit the stars. We make an important definition for later use.

Definition 5.8

Let G be a group and suppose that $g \in G$. The *order* of G is $|G|$. The *order* of g is the smallest natural number $\mathrm{o}(g)$ (if any) with the property that $g^{\,\mathrm{o}(g)} = 1$. If no such natural number exists, we write $\mathrm{o}(g) = \infty$, and say that g has infinite order.

So, now we will build up some theory from the axioms. What follows will become very familiar to those readers who go on to specialize in Pure Mathematics. The relentless application of logic gradually builds a better and better understanding of the consequences of a set of axioms.

Proposition 5.1 (The cancellation law)

Let G be a group and suppose that $x, y, z \in G$ are such that either (a) $xy = xz$ or (b) $yx = zx$, then $y = z$.

Proof

Suppose that (a) holds. By *inverses* there is $w \in G$ such that $wx = 1$. We are given that $xy = xz$. Pre-multiply by w so we learn that $w(xy) = w(xz)$. Use the associative axiom on each side so $(wx)y = (wx)z$. By the choice of w we have $1y = 1z$. By the definition of 1 we have $y = z$ and we are done. Part (b) is entirely similar.

□

Notice that we have justified every step of the argument. This was easy to do because we had no knowledge about G, and couldn't make the error of slipping in some unmentioned and possibly flawed "knowledge" about G. If you are dealing with $\mathbb{Z}$ or $\mathbb{Q}$ this is much harder to do. We get so used to thinking about integers and rationals, and acquire such strong intuition concerning their structure, that it is quite tricky to learn to argue logically from the axioms. It is the same problem that self-taught guitarists or two-finger typists face when they decide to take proper lessons. You have to throw away reasonably effective techniques in order to learn very good ones, and this will slow you down for a while. It is, of course, worth the trouble.

Proposition 5.2

Let G be a group and suppose that $h \in G$ has the property that there exists $g \in G$ with either (a) $gh = h$ or (b) $hg = h$, then $g = e$.

Proof

(a) We have $h = eh$ by definition of e so $gh = eh$. Now apply the cancellation law to obtain $g = e$ as required. Part (b) is entirely similar.

□

Now think about Proposition 5.2. It says that if an element g masquerades as a multiplicative identity, then it is the particular element e mentioned in the group axioms.

Corollary 5.1

(a) A group G contains a *unique* element e which acts as a multiplicative identity on both sides.
(b) In the inverse axiom there is no freedom in the choice of b, since if $ab = e = a\widehat{b}$ then $b = \widehat{b}$ by the cancellation law.

Both corollaries are very important. We now know that given any $a \in G$ there is a *unique* b such that $ab = e = ba$. More formally we can define a function $i : G \to G$ by letting $i(a)$ be the unique two-sided inverse of a.

Proposition 5.3

The map i which we have just defined is a bijection.

Proof

First for surjectivity. Suppose that $g \in G$. We need to show that $g = i(h)$ for some $h \in G$. Now $gi(g) = e = i(g)g$ so g is the (unique) two-sided inverse of $i(g)$ so $g = i(i(g))$. Put $h = i(g)$ and we have $g = i(h)$, so i is surjective.

Now for injectivity. We need to show that if $i(x) = i(y)$, then $x = y$. Now we multiply the equation $i(x) = i(y)$ on the left by x. Thus $xi(x) = xi(y)$. Now $xi(x) = e$ so $xi(y) = e = yi(y)$. The cancellation law yields that $x = y$ and we are done.

□

So inversion simply permutes the elements of G. Let's see this in concrete terms using a particular group. If we had been given that G was finite, then we could have quit the proof half-way through, since a surjective map from a finite set to itself must also be injective.

An alternative mature proof to Proposition 5.3 is to observe that $i \circ i = \text{id}$ so i has a two-sided inverse and so must be a bijection, using Proposition 1.3. This is serene, perhaps a little too serene at this stage.

Example 5.1

In S_3 we have $i(\mathrm{id}) =$ id, $i(\alpha) = \alpha$, $i(\beta) = \beta$, $i(\gamma) = \gamma$, $i(\delta) = \varepsilon$, and $i(\varepsilon) = \delta$. The map $i : G \to G$ has served its purpose, but it is time to move on. The usual name for $i(g)$ is g^{-1}, and from now on that is the notation we will use.

Remark 5.1

For $g \in G$ and $n \in \mathbb{N}$ we have $(g^n)^{-1} = (g^{-1})^n$. Each of them is the unique multiplicative inverse of g^n. Also we put $g^0 = 1_G$.

Proposition 5.4

Let G be a group and suppose that $a, b \in G$ and $\alpha, \beta \in \mathbb{Z}$, then (a) $(ab)^{-1} = b^{-1}a^{-1}$, (b) $a^\alpha a^\beta = a^{\alpha+\beta}$ and (c) $(a^\alpha)^\beta = a^{(\alpha\beta)}$.

Proof

(a) This part is a triviality since

$$(ab)(b^{-1}a^{-1}) = a(bb^{-1})a^{-1} = aea^{-1} = aa^{-1} = e.$$

Notice the use of associativity. Similarly $(b^{-1}a^{-1})(ab) = e$. Parts (b) and (c) are easy exercises. Please fill in the details.

□

The apparently dull parts (b) and (c) of Proposition 5.4 are the reason why writing g^{-1} for $i(g)$ was a good move. The familiar laws of exponentiation can be wheeled out and used without error.

Proposition 5.5

Let G be a group and fix some $h \in G$. Now suppose that $g \in G$, then there exist elements $x, y \in G$ such that $g = xh = hy$.

Proof

Let $x = gh^{-1}$ and $y = h^{-1}g$.

□

Let's pause for a moment and try to take in Proposition 5.5. Let G be a group and choose $h \in G$. Multiplication by h defines a map from G to G, but you have to choose whether you will multiply on the right or left. Let's multiply by h on the left, so we have a map $\lambda : G \to G$ defined by $\lambda(g) = hg$. Proposition 5.1 tells you that this map is injective and Proposition 5.5 tells you it is surjective. Thus multiplication by h on the left is a bijection from G to G, a permutation of G if you like. Of course you could say the same about the map defined by multiplying by h on the right, and it would make sense to call that map ρ.

These bijections λ and ρ are manufactured from the ingredient $h \in G$. If you choose a different h you will build different λ and ρ, so it makes sense to emphasize the dependence on the particular $h \in G$ by calling the maps λ_h and ρ_h.

Example 5.2

Let $G = S_3$, and choose the element δ to create the maps λ_δ and ρ_δ. Using the multiplication table in Section 5.3 the map λ_δ is defined by $\lambda_\delta(\text{id}) = \delta$, $\lambda_\delta(\alpha) = \beta$, $\lambda_\delta(\beta) = \gamma$, $\lambda_\delta(\gamma) = \alpha$, $\lambda_\delta(\delta) = \varepsilon$ and $\lambda_\delta(\varepsilon) = \text{id}$.

EXERCISES

5.14 In the notation of Example 5.2, describe the map ρ_δ.

5.15 Return to the general case where G is an arbitrary group and $x, y \in G$. Show that the maps λ_x and ρ_y commute.

5.7 Subgroups

Definition 5.9

Let G be a group. A *subgroup* H of G is a subset $H \subseteq G$ with the property that the binary operation of G, restricted to H, renders H a group.

Example 5.3

(a) Let G be a group. G is a subgroup of itself.

(b) Let G be a group with identity element e. The set $\{e\}$ is a subgroup –

abusively called e.

(c) The non-zero real numbers form a group under multiplication. The non-zero rational numbers form a subgroup.

In order to check that a subset H of a group G is actually a subgroup, you needn't check all four of the group axioms, since associativity is automatic. The other three axioms can be rolled up into two things to check, one of which is usually easy to do, but unfortunately easy to forget.

Proposition 5.6

Let G be a group. A subset H of G is a subgroup of G if and only if both (i) $H \neq \emptyset$ and (ii) $\forall h, k \in H$ we have $hk^{-1} \in H$.

Proof

If H is a subgroup of G then obviously (i) and (ii) are satisfied.

Conversely, suppose that (i) and (ii) hold. Choose $a \in H \neq \emptyset$ (by (i)), then by (ii) $aa^{-1} = e \in H$ so H contains the identity element of G. Now suppose that $c \in H$, then by (ii) we have $ec^{-1} = c^{-1} \in H$ so H is closed under inversion. Finally, if $f, g \in H$, then $f, g^{-1} \in H$ so $f(g^{-1})^{-1} = fg \in H$ and H is closed under multiplication and we are done.

□

That is a great way to check that a subset is a subgroup, but you must not overlook the unfortunately unmemorable non-emptiness condition.

5.8 Cosets

Let G be a group and suppose that H is a subgroup of G. Now suppose that $x \in G$. Consider Hx. Well, the first thing you have to do is to work out what it might mean. Look away from the page and think about it now. Let's see if you guessed (or remembered!) correctly. The subgroup H is a subset of G, and Hx is what you get when you multiply all the elements of H by x on the right. More formally

$$Hx = \{hx \mid h \in H\}.$$

Notice that if $h_1, h_2 \in H$ and $h_1x = h_2x$ then $h_1 = h_2$ by the cancellation property. This means that multiplying the elements of H on the right by x

doesn't cause any of them to "stick together". In serene form, multiplication on the right by x induces an injection from H to Hx. This is obviously also an onto map, and so is a bijection. That is a particularly interesting fact in the event that H happens to be finite of size m, since it follows that H and Hx both have size m. There is an inverse bijection induced from right multiplication by x^{-1}.

A set of the form Hx is called a *right coset* of H in G. Now H was a group in its own right, but Hx need not be a group – but it is a subset of G of course.

Let's look at an example from permutation group theory. Let $S = S_4$ be the symmetric group, and let K be the so-called Klein 4-group, so $K = \{\mathrm{id}, \lambda, \mu, \nu\}$ where $\lambda = (1,2)(3,4)$, $\mu = (1,3)(2,4)$, and $\nu = (1,4)(2,3)$. The diligent reader will check that the six cosets K, $K(1,2)$, $K(1,3)$, $K(1,4)$, $K(1,2,3)$, $K(1,3,2)$ are distinct, pairwise disjoint, all have size 4 and have union the whole of S. Moreover, if $\eta \in S$ then $K\eta$ is one of the six cosets in our list (this can be hard work, unless you think about it of course).

This is so important; let's say it again a slightly different way. We see that if you look at the 24 right cosets Kx in turn as x ranges over S, you do not get 24 different sets. You get just 6 different cosets, each occurring 4 times. These 6 cosets are disjoint from one another (different ones don't overlap) and their union is S. The six cosets therefore partition S into 6 sets of size 4, which is consistent with $|S| = 24$.

In this example then, the right cosets of K in S have these very special properties. Our next task is to show that this was no accident; that it was not really anything to do with the particular groups in question, but that we have been looking at the footprint of a general theorem about groups.

Suppose that G is a group and H is a subgroup. It follows from the cancellation law (Proposition 5.1) that right multiplication by x induces a bijection from H to Hx. The right coset Hx is just H whenever $x \in H$; this follows because right multiplication by x is then a bijection from H to itself. On the other hand if $x \notin H$ then $H \neq Hx$. This is because $x = ex \in Hx$ but $x \notin H$. That is worth knowing, but it doesn't help us very much.

If $g \in G$ then $g = eg \in Hg$ so every element of G is in at least one right coset of H in G. Of course, all the right cosets are subsets of G, so the union of all the cosets is therefore G.

We need to sort out the business that distinct cosets are disjoint. The neat way to do this is take an arbitrary pair of cosets Hx and Hy, assume that they are not disjoint, and then try to deduce that $Hx = Hy$.

Now $Hx = Hy$ if and only if $Hxy^{-1} = H$, and the latter condition holds if and only if $xy^{-1} \in H$.

Since $Hx \cap Hy \neq \emptyset$ it follows that there is $z \in Hx \cap Hy$. Thus $zx^{-1}, zy^{-1} \in H$. The previous paragraph comes to our rescue and we deduce both $Hz = Hx$

and $Hz = Hy$. It follows that $Hx = Hy$. Thus the right cosets of H in G form a partition of G.

Definition 5.10

The number of right cosets of H in G is called the *index* of H in G and is written $|G : H|$.

If $|G|$ is finite then both $|H|$ and $|G : H|$ will be finite and, since G is the disjoint union of $|G : H|$ sets each size $|H|$, we have proved a celebrated result.

Theorem 5.2 (Lagrange)

Let G be a finite group and H a subgroup of G. It follows that $|G| = |G : H||H|$.

Continue to suppose, for the moment, that G is finite. Lagrange's theorem tells us that $|G : H|$ is $|G|$ divided by $|H|$, so $|H|$ divides $|G|$ exactly – just as 4 divided 24 in our example. This tells us something very strong: a finite group can only have subgroups of certain sizes. A group of order a prime p can only have subgroups of order 1 or p.

All this was with right cosets. It doesn't take a lot of imagination to see that we could have done the whole thing with left cosets instead, and obtained a *left index* instead of what we called the index but really should be the *right index*. However, since both indices are $|G|/|H|$, they coincide.

Well, that's all very well for finite groups, but what happens if G is infinite? Could it be that there are finitely many right cosets of H in G but infinitely many left cosets? Perhaps as bad, could it be that there are finitely many of each type of coset but the numbers differ? No. There is a bijection between the set of right cosets of H in G and the set of left cosets of H in G, as you will see in a forthcoming exercise.

EXERCISES

5.16 Suppose that X and Y are subgroups of a finite group G. Let $XY = \{xy \mid x \in X, y \in Y\}$. In this question, the early parts help you with the later parts.

(a) Suppose that $a, b, c, d \in G$ and $ab = cd$. Prove that there is a unique $h \in G$ such that $c = ah$ and $d = h^{-1}b$.

(b) Suppose that $x_1, x_2 \in X$ and $y_1, y_2 \in Y$, Show that $x_1 y_1 = x_2 y_2$ if and only if there is $a \in X \cap Y$ such that $x_2 = x_1 a$ and $y_2 = a^{-1} y_1$.

(c) Show that $X \cap Y$ is a subgroup of G.

(d) Show that
$$|XY| = \frac{|X| \cdot |Y|}{|X \cap Y|}.$$

(e) Now suppose that $|G| = 256$. Suppose that $|X| = |Y| = 32$. Show that $|X \cap Y| \geq 4$, and that there are at most four possible values of $|X \cap Y|$.

5.17 (a) Suppose that H is a subgroup of a group G, and that U is a right coset of H in G. Prove that $\{u^{-1} \mid u \in U\}$ is a left coset of H in G.

(b) Show that there is a bijection between the right cosets of H in G and the left cosets of H in G. You are not allowed to assume that G is a finite group.

5.9 Cyclic Groups

Definition 5.11

Let G be a group and suppose that $g \in G$. Let $\langle g \rangle$ be $\{g^i \mid i \in \mathbb{Z}\}$, the set of powers of g. Notice that $\langle g \rangle$ is a subgroup of G, called the cyclic subgroup generated by g. If $G = \langle g \rangle$ we say that G is *cyclic.*

Recall the notation $\mathrm{o}(g)$ for the order of g, explained in Definition 5.8. This terminology is quite sensible, as we see next, because the order of $g \in G$ coincides with $|\langle g \rangle|$, the order (size) of the cyclic group generated by g.

Proposition 5.7

Let C be a cyclic group generated by c. Either C is infinite, in which case the elements c^i $(i \in \mathbb{Z})$ are distinct, or $|C| = n < \infty$. In the latter case $C = \{\mathrm{id}, c, c^2, \ldots, c^{n-1}\}$, and $n = o(c)$.

Proof

Suppose that for distinct $i, j \in \mathbb{Z}$ we have $c^i = c^j$. We may assume $i < j$ and choose such a pair i and j so that $n = j - i$ is minimal. Now $c^n = 1$ and $c^0, \ldots, c^{n-1}$ are distinct by choice of i and j. Finally, if $k \in \mathbb{Z}$, then there are $q, r \in \mathbb{Z}$ such that $k = qn + r$ and $0 \leq r < n$. It follows that $c^k = (c^n)^q c^r = c^r \in \{c^0, \ldots, c^{n-1}\}$. Thus $C = \{c^0, \ldots, c^{n-1}\}$ and we are done.

□

The following beautiful result tells us that the structure of finite cyclic groups is very easy to understand.

Proposition 5.8

Let C be a finite cyclic group of order n, then C has exactly one subgroup of every given order dividing n.

Proof

Suppose that c is a generator of C so the elements c^i for $0 \leq i < n$ are distinct by Proposition 5.7. If d divides n, then the cyclic group generated by $c^{n/d}$ has order d and so there is at least one subgroup of order d.

Now suppose that H is a subgroup of order d, and notice that every element of H has order dividing d by Lagrange's theorem. Thus if $h = c^j \in H$ then n divides dj and thus n/d divides j. It follows that h is a power of $c^{n/d}$ and so $H \subseteq \langle c^{n/d} \rangle$. However, both these sets have size d and so $H = \langle c^{n/d} \rangle$ and uniqueness is established.

□

Corollary 5.2

Every subgroup of a finite cyclic group is cyclic.

In fact the finiteness assumption is not necessary, as the reader sensible enough to tackle Exercise 5.18 will discover.

EXERCISES

5.18 Let $C = \langle c \rangle$ be an infinite cyclic group, and let H be a subgroup of C. Prove that H must be cyclic. (*Hint: We may assume that H* ισ νον − τριφιαλ. *Next show that H contains a* ποσιτιφ παυερ οφ *c* ανδ φοκυσ ον θε σμαλλεστ ονε.)

5.19 Prove that a cyclic group is abelian.

5.20 Let C_n be a cyclic group of order n. Find a formula for the number of elements $c \in C_n$ such that $C_n = \langle c \rangle$. (Hint: πριμε φακτοριζασον.)

5.10 Isomorphism

Consider the integers under addition, a group usually called $\mathbb{Z}$. This is an infinite group, and its binary operation $\star$ is invariably written as $+$ and its identity element is written 0. Now consider a second group; this is the cyclic group $T = \langle 10 \rangle$ consisting of all integer powers of 10, a subgroup of the multiplicative group of non-zero rationals $\mathbb{Q}^*$. In T the identity element is 1.

These groups are not the same; their underlying sets are different, and their operations $\times$ and $+$ are different, but even so, there is a certain "sameness" about them. That sameness is, informally, isomorphism – they have the same structure, which is what isomorphism means.

Consider the map $\varphi : \mathbb{Z} \to T$ defined by $\varphi(z) = 10^z \, \forall z \in \mathbb{Z}$. It is easy to check that this map is a bijection. Now, this bijection is compatible with the two group structures in the sense that

$$\varphi(n+m) = \varphi(n) \times \varphi(m) \ \forall n, m \in \mathbb{Z}.$$

Thus you can combine n and m using the binary operation of $\mathbb{Z}$ and then apply the map φ, or, starting over, you can first apply the map φ to each of m and n individually and then combine the resulting elements of T using the binary operation of that group. *It doesn't matter which you do* because you get the same answer either way. Notice that the inverse map φ^{-1} from T to $\mathbb{Z}$, sometimes called $\log_{10}$, is also a bijection and

$$\log_{10}(10^i 10^j) = \log_{10}(10^i) + \log_{10}(10^j) \ \forall i, j \in \mathbb{Z}$$

because both sides are just $i + j$. We now see that all φ does is to rename the elements of $\mathbb{Z}$; if you want to do a calculation in $\mathbb{Z}$ you can do it there, or use

φ to transport everything into T, do the calculation there, and use φ^{-1} (also known as $\log_{10}$) to drag yourself back into $\mathbb{Z}$.

In ancient times before the invention of the pocket calculator, our ancestors used the map $\log_{10}$ to multiply decimals together. This function was called *logarithm.* To work out 3.14159×2.1828 they would look up $\log_{10}(3.14159)$ and $\log_{10}(2.1828)$ in tables. They would then add these decimal quantities, because adding numbers is easy – much easier than multiplying them. From the isomorphism property, this sum is $\log_{10}(3.14159 \times 2.1828)$. These primitive creatures would then use tables again to find out what number has the specified logarithm and that is the answer (subject to minor rounding errors). These days you rarely see $\log_{10}$ mentioned because $\log_e$ has better mathematical properties; the only advantage of $\log_{10}$ concerned bijections between digits and digits (in not so different senses).

Another two group isomorphisms can be manufactured from complex conjugation. Let $c : \mathbb{C} \to \mathbb{C}$ be the map $z \mapsto \overline{z}$ that sends each complex number to its complex conjugate. This is an isomorphism from the additive group of $\mathbb{C}$ to itself. The fact that this map is bijective is clear, and the group structure is preserved because $\overline{z_1 + z_2} = \overline{z}_1 + \overline{z}_2$ for every $z_1, z_2 \in \mathbb{C}$.

Now puncture $\mathbb{C}$ by removing 0 to form $\mathbb{C}^*$, a group under complex multiplication. Note that we have to remove 0 since it has no multiplicative inverse. Again complex conjugation induces an isomorphism from $\mathbb{C}^*$ to $\mathbb{C}^*$, the structural point being that $\overline{z_1 z_2} = \overline{z}_1 \overline{z}_2$ for every $z_1, z_2 \in \mathbb{C}^*$.

Now for a geometrical example. We consider maps from $\mathbb{R}^2$ to itself. These are $\mathrm{I}_{\mathbb{R}^2}$ the identity map, and three other maps $L, M, N : \mathbb{R}^2 \to \mathbb{R}^2$ defined as follows. These maps send an arbitrary point (x, y) to $(-x, y)$, $(x, -y)$ and $(-x, -y)$ respectively. The four maps form a group $V = \{\mathrm{I}_{\mathbb{R}^2}, L, M, N\}$. The reader is urged to verify that this group is isomorphic to the Klein 4-group of Section 5.8, where the non-identity elements match up literally. In fact you can match up the non-identity elements in any way you like, since in each group the product of any pair of distinct ones is the third, and each has order 2.

So, let's think about isomorphism from a group G to a group H for a moment. In terms of multiplication tables, the fact that the map is an isomorphism means that you can take a multiplication table of G, and systematically replace the entries in the table and the labels of the rows and columns using the given map, and obtain *a correct multiplication table* of H. It is a bit like writing out the table of G again, but using green ink the second time.

If a property of a group G is mathematically interesting (e.g. it is abelian) then it will be preserved by isomorphism, i.e. an isomorphic copy of an abelian group will be abelian. If a property of G is not of mathematical interest – for example, the fact that you first thought about the group on a Tuesday – then isomorphism may not preserve it.

Intuitively, it is clear that the composition of two isomorphisms is an isomorphism, and that the inverse of an isomorphism is an isomorphism. We show the first result here, and ask you to produce the second in Exercise 5.21. The self-isomorphisms of a group G with itself are particularly interesting, and they form a group under composition of maps. This new group is called the group of automorphisms of G. We do not pursue this topic in this book. You might like to work out the automorphism groups of the additive groups of $\mathbb{Z}$ and $\mathbb{Q}$.

Proposition 5.9

Suppose that G, H and J are groups and that $\psi : G \to H$ and $\varphi : H \to J$ are isomorphisms. It follows that $\varphi\psi : G \to J$ is an isomorphism (the juxtaposition of φ and ψ denotes composition of maps).

Proof

The composition of two bijections is a bijection by Corollary 1.1, so $\varphi\psi$ is bijective. Now for structure; let the operations of G, H and J be $\star$, $\circ$ and $*$ respectively. Suppose that $x, y \in G$, then

$$\begin{aligned}(\varphi\psi)(x \star y) &= \varphi(\psi(x \star y)) \text{ (by definition of composition)}\\ &= \varphi(\psi(x) \circ \psi(y)) \text{ (since } \psi \text{ preserves structure)}\\ &= \varphi(\psi(x)) * \varphi(\psi(y)) \text{ (since } \varphi \text{ preserves structure).}\end{aligned}$$

Thus $\varphi\psi$ is a bijective structure-preserving map and so is an isomorphism.

□

EXERCISES

5.21 Let $\zeta : G \to H$ be an isomorphism of groups. Consider the map $\zeta^{-1} : H \to G$; show that it is an isomorphism of groups.

5.22 Consider the map $\tau : \mathbb{Z} \to \mathbb{Z}$ defined by $\tau(x) = -x \, \forall x \in \mathbb{Z}$. Show that τ is an isomorphism from the additive group of $\mathbb{Z}$ to itself.

5.23 Let P denote the group of strictly positive reals under multiplication. Consider the map $\sigma : P \to P$ defined by $\sigma(r) = r^2$. Is this an isomorphism? What happens if you replace P by U, the group of positive rationals under multiplication? Justify your answers to each question.

5.11 Homomorphism

A homomorphism $\theta : G \to H$ from a group G to a group H is a map which preserves structure in the same sense that an isomorphism does, but we do not insist that θ be a bijection. Thus an isomorphism is a special kind of homomorphism, and theorems proved about homomorphisms will be true of isomorphisms.

Proposition 5.10

Let $\theta : G \to H$ be a homomorphism of groups and $x \in G$. It follows that $\theta(e_G) = e_H$ and $\theta(x^{-1}) = \theta(x)^{-1}$.

Proof

$\theta(e_G)\theta(e_G) = \theta(e_G^2) = \theta(e_G)$ so by the cancellation law in H we have $\theta(e_G) = e_H$ and we are half way there. Next observe that

$$e_H = \theta(e_G) = \theta(xx^{-1}) = \theta(x)\theta(x^{-1})$$

so $\theta(x^{-1})$ is the multiplicative inverse of $\theta(x)$ as required.

□

Definition 5.12

Let $\eta : G \to H$ be a homomorphism of groups. The *kernel* of η, written $\mathrm{Ker}(\eta)$, is $\{g \mid g \in G, \eta(g) = e_H\}$ where e_H is the identity element of H. More predictably, $\mathrm{Im}(\eta) = \{\eta(x) \mid x \in G\}$ is called the image of G under η.

In that definition, η was a homomorphism, a map that preserves structure. Thus there is reason to think that subsets defined in natural ways in terms of η might be subgroups.

Proposition 5.11

In the notation of the previous definition, both $\mathrm{Ker}(\eta)$ and $\mathrm{Im}(\eta)$ are groups, and are subgroups of G and H respectively.

Proof

From Proposition 5.10, $e_G \in \text{Ker}(\eta) \neq \emptyset$ so we have done the forgettable bit. Next suppose that $x, y \in \text{Ker}(\eta)$. Again by Proposition 5.10, $\eta(y^{-1}) = e_H$ so

$$\eta(xy^{-1}) = \eta(x)\eta(y^{-1}) = e_H^2 = e_H$$

so $xy^{-1} \in \text{Ker}(\eta)$ as required.

Now for $\text{Im}(\eta)$. From Proposition 5.10 we have $e_H = \eta(e_G) \in \text{Im}(\eta)$ so that set is not empty and we are on our way. Next suppose that $\eta(x)$, $\eta(y) \in \text{Im}(\eta)$ then

$$\eta(x)\eta(y)^{-1} = \eta(x)\eta(y^{-1}) = \eta(xy^{-1}) \in \text{Im}(\eta)$$

so all is well.

□

Example 5.4

Let $\psi : \mathbb{R} \to \mathbb{C}^*$ be the map defined by $\psi(\theta) = e^{2\pi i\theta}$. Recall that $\mathbb{C}^*$ denotes the group of non-zero complex numbers under multiplication. The reader can easily verify that ψ is a group homomorphism. The image is $\{z \mid z \in \mathbb{C}, |z| = 1\}$, the set of complex numbers of unit modulus. This is a subgroup of $\mathbb{C}^*$. Also $\text{Ker}(\psi) = \mathbb{Z}$ which is a subgroup of the additive group of reals. Thus no error is exposed.

The reason that the kernel of a group homomorphism $\mu : G \to H$ is so important is that it detects injectivity, in the sense that $|\text{Ker}(\mu)| = 1$ if and only if μ is injective. If μ is injective then obviously $|\text{Ker}(\mu)| = 1$. It is less obvious that the converse holds. Suppose that $\mu(x) \in \text{Im}(\mu)$ then $\mu(x) = \mu(y)$ if and only if $\mu(x)\mu(y)^{-1} = 1_H$. In turn this happens if and only if $\mu(xy^{-1}) = 1_H$ i.e. $xy^{-1} \in \text{Ker}(\mu)$. Thus $\mu(x) = \mu(y)$ if and only if $x = ky$ for some $k \in \text{Ker}(\mu)$, in other words if and only if $y \in \text{Ker}(\mu)x$.

Let $K = \text{Ker}(\mu)$. The coset Kx is in bijective correspondence with K. Now, μ is injective if and only if at most one element of G maps to each element of H, and this happens exactly when $K = \{e_G\}$.

6
Sequences and Series

6.1 Denary and Decimal Sequences

There is a wonderful little book by J. A. Green entitled *Sequences and Series.* It is admirably short, unfashionably old and surely out of print. However, your local library may well have a battered copy. If so, I urge you to read it.

We usually think about numbers one at a time. In particular, a lot of attention is sometimes given to 1729. If you are sufficiently open minded, you may think of 1729 not as a natural number, but as a piece of notation, consisting of a sequence of four symbols $1,\ 7,\ 2,\ 9$. The denary (base 10) notation allows us to use finite sequences of symbols $0,\ 1,\ 2,\ \dots, 9$ to represent all the natural numbers. This is a very clever move, since otherwise we would have to think up infinitely many names for all of the natural numbers. There are well-established procedures for doing arithmetic when natural numbers are represented in this familiar notation, even in the absence of a pocket calculator. For example, if you have two natural numbers x and y you may wish to multiply them. You describe x and y in denary notation and invoke the procedures. Thus in order to multiply 37 by 121 it is not necessary to know your 37 times table. You invoke the method drummed into you as a child (I hope) to conclude that the product, written in denary, is 4477. In fact, in order to multiply any pair of natural numbers, the largest multiplicative fact you need to know is that $9 \times 9 = 81$. You also need to be able to do some addition of course.

One could use this arithmetical procedure as the *definition* of the product 37×121, but that would be a little silly. After all, this arithmetical procedure

G. Smith, *Introductory Mathematics: Algebra and Analysis*
© Springer-Verlag London Limited 1998

hides all the interesting mathematical properties of multiplication. In particular, it is not at all clear that $37 \times 121 = 121 \times 37$ from this point of view. Indeed, it seems almost magical that the two calculations yield the same result. A more appealing definition of multiplication might involve looking at the area of a box which is x units wide by y units high. If you don't want to get involved with areas, you can simply make rectangles of dots, and then count them. From this point of view, the commutative law of multiplication is clear, since the area of the rectangle (or the number of dots in the rectangle) is unchanged if you turn it round. On the other hand, this procedure with rectangles is feeble from a computational point of view. Imagine working out 37×121 in this fashion. It would involve doing all that drawing, and then counting from 1 to 4477.

Following the success of denary notation, we can easily embellish it a little with decimal points and minus signs to make a way of representing real numbers. Before we start worrying about the decimal representation of π, we first look at less exalted numbers. For example, $2/5$ is written as $0{\cdot}4$ and $-11/8$ is written as -1.375. Again there are simple, mechanical procedures for doing arithmetic with these decimal expressions. One of the most minor sources of international tension is the trichotomy exemplified by 3.142, 3·142 and 3,142. You can make a case for the first and the second representations being 426. The third representation can be worrying too, since in some places it means *a number close to* π, and in other parts of the world it means the smallest natural number more than one thousand times bigger than π. Different cultures struggle to exert dominance by the choice of the decimal punctuation mark.

Let us be clear about what a finite decimal expansion means. For example,

$$12.345 = 1 \times 10^1 + 2 \times 10^0 + 3 \times 10^{-1} + 4 \times 10^{-2} + 5 \times 10^{-3}.$$

Any natural number has a denary representation which is really a sum of expressions, each of which is a digit in the range 0 to 9 multiplied by a non-negative power of 10. A decimal representation is a more general beast, and you permit negative powers of 10.

We soon get into trouble with decimal notation, because $1/3$ does not admit a finite decimal expansion. The number 0.3 is too small, but 0.4 is too big. in fact we find ourselves wanting to represent $1/3$ as $0.3333333\ldots$, economically written as $0.\overline{3}$. In a sense this isn't too bad, since at least the terms of the decimal expansion are predictable. However, there is something a little tricky going on here. As a sum $0.\overline{3}$ is

$$3 \times 10^{-1} + 3 \times 10^{-2} + 3 \times 10^{-3} + \ldots.$$

Now, I hope that you feel uncomfortable about infinite sums. Adding up two numbers is fine, and adding three numbers is also fine because you can break it up into two sums of two numbers:

$$a + b + c = (a + b) + c.$$

Notice that the associative law of addition means that you don't have to worry about brackets. By induction, there is no problem about adding up m numbers where m is a natural number. The difficulty arises when you try to add up infinitely many numbers. Let us skip to the answer: yes, you can do it sometimes, but you have to be careful, because sometimes you cannot. Troublesome sums include

$$1 + 1 + 1 + \ldots$$

which has the merit of being absurd. No-one would dream of thinking that you could add up infinitely many 1's and get an answer which is a number. More refined trouble arises from

$$1 - 1 + 1 - 1 + 1 - 1 \ldots.$$

You can make out a case for this to be 0, 1 or 1/2.

† **Warning! Flawed arguments coming.** To see this, observe that one may bracket the expression as

$$(1 - 1) + (1 - 1) + (1 - 1) + \ldots$$

or

$$1 - (1 - 1) - (1 - 1) - (1 - 1) - \ldots,$$

and these sloppy arguments support the first two answers.

Trying to get the answer 1/2 requires a bit more sophistry, though it is very instructive. Recall the following fact from school days:

$$(1 - x)(1 + x + x^2 + x^3 + \ldots + x^m) = 1 - x^{m+1}.$$

Now assume $|x| < 1$ and let m wander off towards infinity. Notice that x^{m+1} will become vanishingly small, so we have $(1 - x)(1 + x + x^2 + x^3 + \ldots) = 1$. Now rearrange this as

$$1 + x + x^2 + x^3 + \ldots = \frac{1}{1 - x}.$$

This equation is only valid for $|x| < 1$, so we cannot just plug in $x = -1$ to get the third answer. However, the equation is valid for values of x in the range $-1 < x < 1$. Let x take such values, and see what happens as x approaches -1. The right hand side of the equation approaches 1/2, and the left hand side approaches $1 - 1 + 1 - 1 + 1 - 1 + \ldots$.

All of the argument since † is highly suspect. This process of taking limits, and trying to add up infinite sums is very important, and very useful. The problem is obvious though. If we just blunder around we are likely to come up with contradictory answers. It is necessary to develop some mathematics which will tell us exactly what we are and are not allowed to do with infinite processes, taking limits and so on.

6.2 The Real Numbers

The real numbers are an infinite field. They also come equipped with an ordering; an antisymmetric transitive relation ("less than") which has the property that if $x, y \in \mathbb{R}$ and $x \neq y$ then either $x < y$ or $y < x$, but not both. The algebra and the ordering interact nicely, and we list some key laws below.

These laws hold for all real numbers a, b and c.

(i) If $a < b$ and $0 < c$, then $ac < bc$.

(ii) If $a < b$ and $c < 0$, then $bc < ac$.

(iii) If $a < b$, then $a + c < b + c$.

(iv) If $a < b$ and $c < d$, then $a + c < b + d$.

All remain valid if we relax $<$ to $\leq$. There is a well-known variation on the notation: $b > a$ means $a < b$. Similarly $b \geq a$ means $a \leq b$.

The notion of the modulus of a number was introduced in Chapters 1 and 3, but that was long ago. Without apology, we address this important matter again (but not in quite the same way!).

Definition 6.1

The *modulus* $|x|$ of the real number x is x if $x \geq 0$, and is $-x$ if $x < 0$.

If you think of the real numbers as being the names of points on a number line, then $|x|$ is just the distance of x from 0. This comment explains why this is such an important idea. Notice that if x and y are any pair of real numbers, then $|x - y|$ (also known as $|y - x|$) is their distance apart.

The function $f : \mathbb{R} \to \mathbb{R}$ defined by $f(x) = |x| \forall x \in \mathbb{R}$ is not surjective, because $|x|$ is never negative. It is not injective either, because $|1| = |-1|$. If you have read this book in sequence, then you will know much about modulii (the plural of modulus). However, we can't assume that, so we distill a few important properties of the modulus function.

For every real number x we have

$$|x| = |-x|$$

$$\begin{aligned} |x| &= x \text{ or } -x \\ x &\leq |x| \end{aligned} \tag{6.1}$$

On the basis that you cannot have too much of a good thing, we will give yet another proof of the triangle inequality.

Proposition 6.1

For all real numbers x and y we have $|x+y| \leq |x| + |y|$.

Proof

For any real numbers x, we have $x \leq |x|$ and $y \leq |y|$. Now apply law (iv) from the start of this section, so $x + y \leq |x| + |y|$. Similarly we observe that $-x \leq |x|$ and $-y \leq |y|$. Now apply law (iv) again so $-(x+y) \leq |x| + |y|$. Now by Equation (6.1) either $x + y = |x+y|$ or $-(x+y) = |x+y|$. In either event, we can deduce that $|x+y| \leq |x| + |y|$.

□

Proposition 6.2

For all real numbers x and y we have $||x| - |y|| \leq |x - y|$.

Proof

$|x| = |(x-y) + y| \leq |x-y| + |y|$. Subtract $|y|$ from each side, or if you want to be really precise, add $-|y|$ to each side. Either way you obtain $|x| - |y| \leq |x-y|$. Now swap the roles of x and y in that argument, so we obtain $|y| - |x| \leq |y-x|$, but this can be recast as $-(|x| - |y|) \leq |x-y|$. Now we employ the same trick as in the preceding proof. Either $|x| - |y| = ||x| - |y||$ or $-(|x| - |y|) = ||x| - |y||$. In either event, we can deduce that $||x| - |y|| \leq |x-y|$.

□

Corollary 6.1

Both $|x+y|$ and $|x-y|$ can be bounded above and below using expressions involving $|x|$ and $|y|$ because

$$||x| - |y|| \leq |x-y| = |x + (-y)| \leq |x| + |y|$$

and

$$||x| - |y|| \leq |x - (-y)| = |x + y| \leq |x| + |y|.$$

This places both expressions in the same sandwich.

6.3 Notation for Sequences

We consider the sequence $a_1, a_2, a_3, \ldots$ where each a_i is a real number, and there is a real number a_n for each natural number n. When $n \in \mathbb{N}$, we say that a_n is the n-th term of our sequence. For example, we might consider the sequence $b_1, b_2, b_3, \ldots$ where $b_i = 1/i$ for every $i \in \mathbb{N}$. This is the sequence

$$1, 1/2, 1/3, 1/4, 1/5, \ldots.$$

We say that two sequences $(c_i)_{i=1}^{\infty}$ and $(d_i)_{i=1}^{\infty}$ are *equal* if and only if $c_i = d_i$ for every $i \in \mathbb{N}$. If this happens, we write $(c_i)_{i=1}^{\infty} = (d_i)_{i=1}^{\infty}$.

We can write our original sequence as simply as $(a_i)_{i=1}^{\infty}$, or as $(a_i)_{i \in \mathbb{N}}$. In our notation, the sequence of reciprocals of natural numbers can be written in these four ways: $(b_i)_{i=1}^{\infty}$, $(b_i)_{i \in \mathbb{N}}$, $(1/i)_{i=1}^{\infty}$, or $(1/i)_{i \in \mathbb{N}}$. The actual letter used as the "counter" doesn't matter at all. The sequence $(a_i)_{i=1}^{\infty}$ and the sequence $(a_j)_{j=1}^{\infty}$ both mean the sequence $a_1, a_2, a_3, \ldots$, as do $(a_u)_{u \in \mathbb{N}}$ and $(a_v)_{v \in \mathbb{N}}$.

Given that our sequences are all understood to be labelled by the natural numbers, we can elect to omit that fact. Thus we can write just (a_i) as a fast notation for $(a_i)_{i=1}^{\infty}$. You have to adopt the convention that the subscript i in the expression (a_i) is automatically allowed to range over the natural numbers. You must not leave off the brackets of course, since you want a_i to denote the i-th term of the sequence, and not the whole sequence as described by (a_i). Notice the following slightly unnerving consequence of our notation. It is always true that $(a_i) = (a_j)$, but $a_i = a_j$ in not necessarily the case. The statement $(a_i) = (a_j)$ is the assertion that a sequence is equal to itself, but we have elected to describe the sequence slightly differently on either side of the symbol $=$. The statement that $a_i = a_j$ asserts that the i-th term of the sequence and the j-th term of the sequence happen to have the same value.

There are special sequences which we should make note of; these are the *constant sequences*, where all terms are equal. The notation for a sequence where every term is x is (x). Thus (1) is an infinite sequence of ones, and (0) is an infinite sequence of zeros.

We can perform algebraic operations on sequences term by term. Suppose that (c_i) and (d_i) are sequences. Their *sum* is (e_i) where $e_i = c_i + d_i$ for every $i \in \mathbb{N}$. One can also write this as $(c_i) + (d_i) = (c_i + d_i)$. Similarly the *product* of (c_i) and (d_i) is (f_i) where $f_i = c_i d_i$ for every $i \in \mathbb{N}$. Equally well, we can write $(c_i)\cdot(d_i) = (c_i d_i)$. These term-by-term operation inherit various algebraic laws of $\mathbb{R}$. For example, addition is commutative and associative, the constant sequence (0) is an additive identity, and the additive inverse of (a_i) is $(-a_i)$. In the language of Chapter 5, sequences form an abelian group under addition. Note that subtraction of sequences is effected by $(c_i) - (d_i) = (c_i) + (-d_i) = (c_i - d_i)$.

The constant sequences form a copy of $\mathbb{R}$ from the algebraic point of view; $(3/7) + (4/7) = (1)$ and so on. The constant sequences therefore form a field, and regarding them as scalars, the reader can check that the set of all real sequences forms an abstract vector space because the vector laws of Section 4.3 all hold.

There is an enormous range of interesting properties which a sequence might have. In this book we will focus on certain attributes which arise naturally when you investigate limiting processes. There are other properties too though. For example, a sequence (a_i) might be *periodic* – this means that there is a natural number p such that $a_{i+p} = a_i$ for every $i \in \mathbb{N}$. On the other hand, a sequence might be *alternating*, so its entries are alternately non-positive and non-negative. We are concentrating on the case where all the entries of a sequence are real numbers. This need not be the case in general of course, you can have sequences of anything – sequences of functions, sequences of matrices, even sequences of sequences, and sequences of sequences of sequences.

We make a definition for forthcoming exercises and for other future use.

Definition 6.2

There are various types of boundedness.

(a) A sequence (a_i) or set $\{a_i \mid i \in I\}$ of real numbers is *bounded* if there is a real number M so that $|a_i| \leq M$ for every i.

(b) A sequence (a_i) or set $\{a_i \mid i \in I\}$ of real numbers is *bounded above* if there exists $a \in \mathbb{R}$ such that $a_i \leq a$ for every i.

(c) A sequence (a_i) or set $\{a_i \mid i \in I\}$ of real numbers is *bounded below* if there exists $b \in \mathbb{R}$ such that $b \leq a_i$ for every i.

EXERCISES

6.1 (a) Give an example of a sequence which has period 1.

(b) Give an example of a sequence which has period 2 but which does not have period 1.

(c) Give an example of an alternating sequence of period 2.

(d) Can you give an example of an alternating sequence of period 3?

6.2 (a) Show that the sequence $(\sin(i))$ is bounded.

(b) Give an example of a sequence which is not bounded above or below.

(c) Prove that the sequence (a_i) is bounded if and only if it is bounded above and below.

6.3 (a) Give an example of an infinite bounded subset of $\mathbb{R}$.

(b) Is the empty set a bounded subset of $\mathbb{R}$?

(c) Is the empty set an unbounded subset of $\mathbb{R}$?

(d) Is the union of two bounded subsets of $\mathbb{R}$ a bounded subset of $\mathbb{R}$? Justify your answer.

(e) Is the intersection of two bounded subsets of $\mathbb{R}$ a bounded subset of $\mathbb{R}$? Justify your answer.

6.4 (a) Give an example of a bounded sequence (a_i) where the set $\{a_i \mid i \in \mathbb{N}\}$ is infinite.

(b) Prove that if the sequence (b_i) is such that $\{b_i \mid i \in \mathbb{N}\}$ is finite, then (b_i) is bounded.

(c) Show that a bounded sequence can be the sum of two unbounded sequences.

(d) Show that the sum of two bounded sequences is bounded.

6.4 Limits of Sequences

We first explore the notion of a limit from an informal point of view, to try to see how best to capture the idea in a formal definition. Consider the sequence $(s_i) = s_1, s_2, s_3, \ldots$ where $s_i = 1 - 2^{-i}$, so

$$(s_i) = 1 - \frac{1}{2}, 1 - \frac{1}{4}, 1 - \frac{1}{8}, 1 - \frac{1}{16}, \ldots$$

It should be clear to you that as you look along this sequence, the terms get steadily closer to 1. However, you can equally well say that as you look further along the sequence, the terms get progressively closer to 37 since numbers $|37 - s_i|$ form a strictly decreasing sequence! This is an important lesson; you can get closer to something forever, and yet it may be that you never get near it. If instead we look at the sequence of numbers $|1 - s_i|$, in other words the sequence $(1/2^i)$, we find a sequence which approaches 0 very well, but never gets there. Thus the sequence (s_i) is approaching 1 very well, but it never gets there. This is another lesson; you can get as close as you like to something, but you may never reach it.

Now let us think about the sequence $t_1, t_2, t_3, \ldots$ where $t_i = 0$ when i is odd and $1/i$ when i is even. Thus the first few terms look like

$$0, 1/2, 0, 1/4, 0, 1/6, 0, 1/8, \ldots.$$

What happens is that the terms are settling towards 0 (approaching 0 in the limit) as you look along the sequence. However, it is not the case that the terms are progressively better and better approximations to 0 (unlike the sequence $s_1, s_2, s_3, \ldots$ where the terms are steady improvements towards 1). What is happening with (t_i) is that the alternate terms which are 0 are actually achieving the limit. Insinuated between them is another sequence which is also steadily closing on 0. However, the interlacing of the two sequences has the effect that as you pass from 0 to a non-zero term, your approximation to the limit deteriorates.

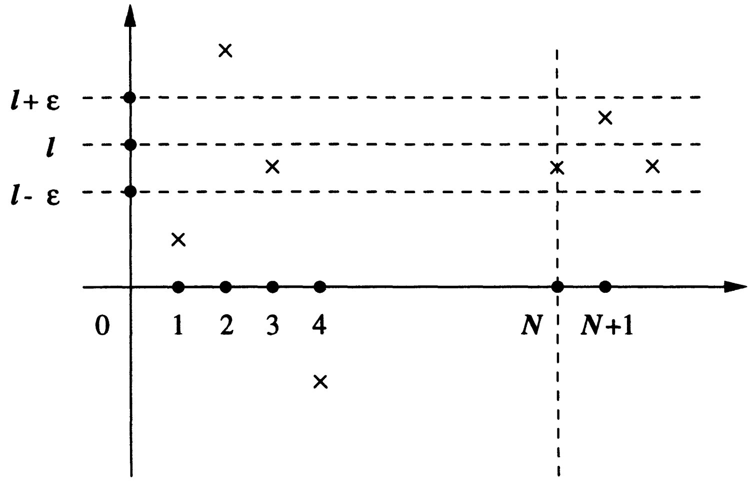

Fig. 6.1. Picture of a limit

We now give the formal definition of convergence, and of a limit of a sequence.

Definition 6.3

We say that the sequence (a_i) of real numbers converges to a limit l exactly when the following condition holds: for every $\varepsilon > 0$ there is a natural number N such that $|a_i - l| < \varepsilon$ when $i \geq N$. Here we implicitly assume that $\varepsilon \in \mathbb{R}$ and $n \in \mathbb{N}$.

In order to see how this definition works, you have to notice that N must be chosen *after* ε has been selected. The idea is that for any particular ε you can always find an N to do the job. You do not have to find a single N which will do the job for every ε simultaneously. See Figure 6.1.

Let us see what happens if you misremember the definition in the most obvious way.

Definition 6.4 (confused convergence – non-standard notation)

We say that the sequence (a_i) of real numbers confusedly converges to a limit l exactly when the following condition holds: there is a natural number N such that for every $\varepsilon > 0$ we have $|a_i - l| < \varepsilon$ when $i \geq N$.

If (a_i) is a sequence which confusedly converges to l, then $a_i = l$ for all except finitely many values of i. This means that the terms of the sequence may wander about at first, but after finitely many terms have gone by, the sequence must consist of l repeated for ever. Thus a confusedly convergent sequence must be "ultimately constant". To see this, we argue as follows.

For each $n \geq N$ we have $|a_n - l| < \varepsilon$ for every positive real number ε. In other words, $l - \varepsilon < a_n < l + \varepsilon$ for each $n \geq N$ and for each positive ε. Now, since ε may be as small as you please, it must be that a_n is neither greater than nor less than l. The only way this can happen is that $a_n = l$ whenever $n \geq N$. Of course, we have nothing to say about a_i when $i < N$, which is why the terms of the sequence may vary at first.

Now, there is nothing *wrong* with confused convergence, but it is not such a widely used concept as convergence. The reason is obvious: most people have found it too strict a definition of "settling down towards". Now we return to the proper definition of convergence, and explore some of its consequences.

Proposition 6.3

Suppose that the sequence (a_n) converges to the limit l, and that also (a_n) converges to the limit m, then $l = m$.

Remark 6.1

This means that a sequence (a_n) can converge to at most one number. Once we have proved this theorem, we can call l (or m) *the* limit of the sequence, and write $l = \lim(a_n) = \lim_{n\to\infty} a_n = \lim a_n$, all of which are common ways of writing this limit. Thus every convergent sequence converges to exactly one limit.

Proof

We will begin by using the triangle inequality of Proposition 6.1. As you will see, it is a very powerful tool when proving results concerning convergence and (in Chapter 7) continuity. Suppose that a, b and x are real numbers, then it follows that

$$|a - b| = |(a - x) + (x - b)| \leq |a - x| + |x - b|. \tag{6.2}$$

Suppose that (a_n) converges to l and m. We want to show that $l = m$. Suppose that $\varepsilon > 0$, then there exists $L, M \in \mathbb{N}$ such that $|a_i - l| < \varepsilon$ when $i \geq L$, and moreover $|a_i - m| < \varepsilon$ when $i \geq M$. Let $N = \max\{L, M\}$. When $i \geq N$ we have $|l - m| \leq |l - a_i| + |a_i - m| < 2\varepsilon$ using Equation (6.2).

Thus $0 \leq |l - m| < 2\varepsilon$ for all positive ε (no matter how small). Now we can not have $|l - m| > 0$ because then putting $\varepsilon = |l - m|/2$ violates the condition that $0 \leq |l - m| < 2\varepsilon$ for all positive ε. Thus $|l - m| = 0$. We conclude that $l = m$.

□

You can do a finite amount of damage to a convergent sequence without changing its limit. This is a loose way of saying that if you take a convergent sequence, and change a finite number of its terms to other values, then the resulting sequence is still convergent, and its limit is unchanged. We will now clarify and prove these remarks.

Proposition 6.4

Suppose that (a_i) is a sequence of real numbers converging to l.

(a) Form a new sequence (b_i) by discarding finitely many terms from the beginning of the sequence (a_i), so for some $m \in \mathbb{N} \cup \{0\}$ we have $b_i = a_{i+m} \forall i \in \mathbb{N}$. It follows that (b_i) converges to l.

(b) Start again with (a_i). Form a new sequence (c_i) by inserting finitely many extra terms at the beginning of the sequence (a_i), so for some $m \in \mathbb{N} \cup \{0\}$ we have $c_{m+i} = a_i \, \forall i \in \mathbb{N}$. It follows that (c_i) converges to l.

Proof

(a) Suppose that we are given some $\varepsilon > 0$. We must show that there is $N \in \mathbb{N}$ with the property that $|b_n - l| < \varepsilon$ when $n \geq N$. Now, we know that we can choose $N \in \mathbb{N}$ with the property that if $n \geq N$, then $|a_n - l| < \varepsilon$. This very N will do the job, because if $n \geq N$, then $b_n = a_{n+m}$ and $n + m \geq n \geq N$. Thus

$$|b_n - l| = |a_{n+m} - l| < \varepsilon.$$

(b) Given any $\varepsilon > 0$, choose $N_1 \in \mathbb{N}$ so that if $n \geq N_1$, then $|a_n - l| < \varepsilon$. Let $N = N_1 + m$. Now if $n \geq N$ we have $|c_n - l| = |a_{n-m} - l|$, but also since $n \geq N$ it follows that $n - m \geq N - m = N_1$. It now follows that $|a_{n-m} - l| < \varepsilon$ and we are done.

□

Corollary 6.2

Given a sequence (a_i) converging to l you can insert or delete finitely many terms in the sequence and the resulting sequence will still converge to l. This is because any changes that you make can be excised by pruning an initial segment of the altered sequence, and this same shortened sequence could also have been obtained by deleting an initial fragment of the original sequence. Now apply both parts of Proposition 6.4.

Thus the convergence or otherwise of a sequence is determined solely by its tail, i.e. the behaviour of the terms a_n when n is large.

Proposition 6.5

Suppose that (a_n) is a convergent sequence. It follows that (a_i) is bounded.

Proof

Let the limit of the sequence be l. Taking ε to be 1, we know that there is a natural number N with the property that if $i \geq N$, then $|a_i - l| < 1$, and so $l - 1 < a_i < l + 1 \leq |l| + 1$. For these values of i we have $-a_i < -l + 1 \leq |l| + 1$. Thus both a_i and $-a_i$ are bounded above by $|l| + 1$ when $i \geq N$. We conclude that $|a_i| < |l| + 1$ when $i \geq N$.

Let $M = \max\{|l| + 1, |a_i| \mid 1 \leq i < N\}$. Thus M is the maximum of a non-empty finite set of real numbers, so it is properly defined, and by the choice of M we have $|a_i| \leq M$ for every $i \in \mathbb{N}$.

□

EXERCISES

6.5 (a) Give an example of a sequence (a_n) which is not convergent, but with the property that the sequence $(|a_i|)$ (i.e. the sequence whose i-th term is $|a_i|$) is convergent.

(b) Give an example of two unbounded sequences (a_i) and (b_i) with the property that the sequence $(a_i b_i)$ (the sequence whose i-th term is $a_i b_i$) is bounded.

(c) Give an example of two unbounded sequences (a_i) and (b_i) with the property that the sequence $(a_i b_i)$ (the sequence whose i-th term is $a_i b_i$) is convergent.

6.6 Suppose that (a_n) is a sequence converging to a limit l.

(a) Prove that the sequence $(|a_i|)$ converges to $|l|$.

(b) If (a_n) confusedly converges to l, then (a_n) converges to l. However, if (a_n) converges to l, it does not follow that (a_n) confusedly converges to l. Justify these remarks.

You can go on to prove a variety of theorems about convergent sequences. We have already remarked that sequences can be added, subtracted and multiplied term by term. You have to be a little cautious about division, because of the possibility that some entries may be 0 even though the whole sequence is not the sequence (0) consisting solely of zeros. From the point of view of convergence, we can always discard finitely many terms of a sequence without changing the limit; this gives us a route to division in favourable circumstances.

Proposition 6.6

Suppose that (a_n) and (b_n) are sequences, and that they converge to l and m respectively. For each $n \in \mathbb{N}$ let $c_n = a_n + b_n$, and $d_n = a_n b_n$. It follows that (c_n) converges to $l + m$ and d_n converges to lm.

Remark 6.2

As you would expect!

Proof

First we deal with the sum of (a_n) and (b_n). Suppose that $\varepsilon > 0$. Let $\varepsilon_1 = \varepsilon/2$. Choose $N_1 \in \mathbb{N}$ such that if $n \geq N_1$, then $|a_n - l| < \varepsilon_1$. Choose $N_2 \in \mathbb{N}$ such that if $n \geq N_2$, then $|b_n - m| < \varepsilon_1$. We are allowed to select N_1 and N_2 since we are given the convergence of (a_n) and (b_n) to their respective limits.

Let $N = \max\{N_1, N_2\}$, so if $n \geq N$, then $n \geq N_1$ and $n \geq N_2$. Thus for $n \geq N$ we have

$$\begin{aligned}|c_n - (l+m)| = |(a_n + b_n) - (l+m)| = |(a_n - l) + (b_n - m)| \\ \leq |a_n - l| + |b_n - m| < \varepsilon_1 + \varepsilon_1 = \varepsilon\end{aligned}$$

and this part of the result is proved.

Now for (d_n). This will require a little more cunning, so we prepare the ground. Notice the crucial use of the triangle inequality in what follows.

$$\begin{aligned}|d_n - lm| &= |a_n b_n - lm| = |a_n b_n - l b_n + l b_n - lm| \\ &\leq |a_n b_n - l b_n| + |l b_n - lm| = |b_n||a_n - l| + |l||b_n - m|. \quad (6.3)\end{aligned}$$

Aside: Now we break off from the official proof, to discuss how we are going to tackle this. We want to argue that for sufficiently large n, the expression $|b_n||a_n - l| + |l||b_n - m|$ can be made as small as you like. This will put the squeeze on $|d_n - lm|$. By looking sufficiently far along the two sequences (a_n) and (b_n) we will be able to make $|a_n - l|$ and $|b_n - m|$ as small as we please. Now $|l|$ is a constant, so we will be able to make $|l||b_n - m|$ small. The problem is with the other term $|b_n||a_n - l|$. Far along the sequence, $|a_n - l|$ will become small. However, there is an irritating multiplier $|b_n|$ to worry about. Might it be that $|b_n|$ gets big and so neutralizes the attempts of $|a_n - l|$ to force the product to be small? Well no, that cannot happen, because (b_n) converges to a limit, so that $|b_n|$ will also converge to a limit – and will not be able to wander off becoming bigger and bigger (see Proposition 6.5). Enough of this chatter! Let us put on formal attire, and write out the proof properly. **End of aside.**

Choose $N_0 \in \mathbb{N}$ such that if $n \geq N_0$, then $|b_n - m| < 1$. For these values of n we have

$$||b_n| - |m|| \leq |b_n - m| < 1,$$

and so $-1 + |m| < |b_n| < |m| + 1$. We isolate the important fact that if $n \geq N_0$, then $|b_n| < |m| + 1$, and moreover $|m| + 1$ is at least 1 and so is not 0.

Now suppose that we have an arbitrary $\varepsilon > 0$. Let

$$\varepsilon_1 = \frac{\varepsilon}{2(|m|+1)} \text{ and } \varepsilon_2 = \frac{\varepsilon}{2(|l|+1)}. \tag{6.4}$$

Note that $|l| + 1$ is non-zero for the same reason that $|m| + 1$ is non-zero, and that $\varepsilon_1, \varepsilon_2 > 0$. Choose $N_1 \in \mathbb{N}$ such that if $n \geq N_1$, then $|a_n - l| < \varepsilon_1$. Choose $N_2 \in \mathbb{N}$ such that if $n \geq N_2$, then $|b_n - m| < \varepsilon_2$. Now put $N = \max\{N_0, N_1, N_2\}$. Now if $n \geq N$, then by the inequality (6.3) we have

$$\begin{aligned} |d_n - lm| &\leq |b_n||a_n - l| + |l||b_n - m| \\ &< (|m|+1)\frac{\varepsilon}{2(|m|+1)} + |l|\frac{\varepsilon}{2(|l|+1)} \leq \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon. \end{aligned}$$

This is exactly what was required.

□

The fact that multiplying convergent sequences leads to a convergent sequence has various consequences. For example, a constant sequence (c), where every term is c, is certainly convergent, and its limit is c. Thus if (a_i) is convergent and $c \in \mathbb{R}$, then (ca_i) is convergent. Moreover, if the limit of the first sequence is l, then the limit of the second sequence is cl.

In fact an energetic reader might verify that the collection of all convergent sequences of real numbers forms an abstract vector space over the field of real numbers. Scalar multiplication is effected via the constant sequences. In particular, the constant sequence whose entries are all -1 is convergent, so if (a_i) converges, then so does $(-a_i)$. Recall from Remark 4.1 what you have to check in order to verify that a set is a vector space.

Definition 6.5 (non-standard)

We say that a sequence (a_i) *bonverges* to a limit l exactly when for all $\varepsilon > 0$ we have $|a_i - l| < \varepsilon$ for all except for finitely many values of i.

The phrase "for all except for finitely many values of i" deserves some amplification. In this context, it means that the set $\{i \mid i \in \mathbb{N}, |a_i - l| \geq \varepsilon\}$ is finite. Notice that the empty set is finite, so it may be that $\{i \mid i \in \mathbb{N}, |a_i - l| \geq \varepsilon\}$ is empty, and so $|a_i - l| < \varepsilon$ for all $i \in \mathbb{N}$.

Instead of developing the concept of bonvergence, we immediately show that it is not a new idea at all, so that no-one will bother to talk about bonvergence.

Proposition 6.7

The sequence (a_i) converges to the limit l if and only if the sequence (a_i) bonverges to the limit l.

Proof

This is an "if and only if" result, so we have to prove the implication both ways. First suppose that (a_i) converges to the limit l. Given any $\varepsilon > 0$ there is a natural number N with $|a_i - l| < \varepsilon$ whenever $n \geq N$ – we know this because (a_i) converges to l. Thus, except (possibly) for the finite number of terms a_i where $i < N$ we have $|a_i - l| < \varepsilon$. Thus (a_i) bonverges to l.

Now for the other half of the proof. Suppose that (a_i) bonverges to the limit l. Given any $\varepsilon > 0$ there are only finitely many subscripts r such that $|a_r - l| \geq \varepsilon$. Call the largest of these troublesome subscripts N (if there are no troublesome subscripts, let $N = 0$). Now if $n \geq N + 1$ we have $|a_n - l| < \varepsilon$. Thus (a_i) converges to l.

□

Now, you could react by thinking that you have just wasted your time by thinking about a concept which is not new. However, this is not the point. Look at the definition of bonvergence again. It does not depend on the order in which the terms of the sequence occur. You can rearrange a sequence how you please; it does not alter whether or not it bonverges to l. However, thanks to Proposition 6.7 we can deduce that the same is true of convergence to l. Thus we have discovered an important consequence.

Corollary 6.3

A sequence (a_i) converges to l if and only if any rearrangement of the sequence converges to l.

By the way, I invented the terms bonvergence and confused convergence to make a point. This terminology is not standard, and if you use it in mathematical circles, please do not expect to be understood.

6.5 The Completeness Axiom

Examples of sets which are bounded above include all finite subsets of $\mathbb{R}$, the set of negative integers, and $\{x \mid x \in \mathbb{R}, x^2 < 2\}$.

Throughout this chapter we have apparently been talking about $\mathbb{R}$. However, looking back, you will see that every single thing we have done works just as well in $\mathbb{Q}$, or indeed in any field intermediate between $\mathbb{Q}$ and $\mathbb{R}$. We now introduce an important special property of $\mathbb{R}$ called *completeness*. The rationals do not have this property. Completeness can be expressed in many equivalent ways, and we choose one which is easy to understand.

The Completeness Axiom

Suppose that S is a non-empty subset of the real numbers, and that S is bounded above. Let

$$U(S) = \{u \mid u \in \mathbb{R},\ \ s \leq u\ \forall s \in S\}.$$

Thus $U(S)$ is the set of all upper bounds for S. The completeness axiom asserts that $U(S)$ contains a least element u_{min}.

Remark 6.3

In the notation of the completeness axiom, we see that $u_{\text{min}} \in U(S)$ and $u_{\text{min}} \leq u\,\forall u \in U(S)$. The real number u_{min} is called (depending on your whim) either the ***supremum of** S* or the ***least upper bound*** of S. The choice is a matter of taste. Those who prefer the descriptive phrase tend to abbreviate it to l.u.b.; perhaps the "supremum" terminology is slightly better, because it relates directly to S, rather than to $U(S)$, and we have the neat notation $\sup S$ for the supremum of S.

Example 6.1

(a) $\sup \emptyset$ does not exist since $\emptyset$ is not non-empty.

(b) $\sup \mathbb{N}$ does not exist, since $\mathbb{N}$ is not bounded above.

(c) $\sup\{1,2,3\} = 3$.

(d) $\sup\{x \mid x \in \mathbb{R}, x^2 < 2\} = \sqrt{2}$.

(e) $\sup\{x \mid x \in \mathbb{R}, x^2 \leq 2\} = \sqrt{2}$.

These examples are extremely instructive. To start with, when a set has a supremum, the supremum may or may not be in the set. Moreover, it is now plain that $\mathbb{Q}$ does not satisfy the completeness axiom by Example 6.1(d) with $\mathbb{R}$ replaced by $\mathbb{Q}$. Recall that $\sqrt{2} \notin \mathbb{Q}$ by Proposition 2.6.

The real numbers also satisfy a lower bound analogue of the completeness axiom. This is easy to see; multiply by -1, apply the completeness axiom, then multiply by -1 again. Thus if S is a non-empty subset of the real numbers, and S is bounded below, then there is a *greatest lower bound* for S. This is also called the *infimum* of S, and is written $\inf S$.

6.6 Limits of Sequences Revisited

The joy of the completeness axiom is that it has the consequence, put informally, that if a sequence looks like it wants to converge, then it does converge. Let us see what this means, and build towards formalizing this glib summary in Theorem 6.2.

Definition 6.6

A sequence (a_i) is said to be *monotone increasing* if whenever $i < j$, then $a_i \leq a_j$.

Thus as you look along a monotone increasing sequence the terms never get smaller. The sequence (b_i) where $b_i = i \; \forall i \in \mathbb{N}$ is an example of a monotone increasing sequence, as are constant sequences. In a perfect world, we might have called our notion *monotone non-decreasing*, since we are allowing $i < j$ and $a_i = a_j$, but life is too short. We use the terminology that (a_i) is *strictly monotone increasing* to mean that if $j < k$, then $a_j < a_k$.

Proposition 6.8

Let (a_i) be a monotone increasing sequence and suppose that (a_i) is bounded above (see Definition 6.2). It follows that (a_i) is a convergent sequence.

Proof

Let $a = \sup\{a_i \mid i \in \mathbb{N}\}$, which exists by the completeness axiom. We shall show that (a_i) converges to a. Suppose that $\varepsilon > 0$ is given. There must exist $N \in \mathbb{N}$ such that $a - \varepsilon < a_N$, otherwise $a - \varepsilon$ would be an upper bound for $\{a_i \mid i \in \mathbb{N}\}$, and yet be smaller that the least such upper bound a. Now if $n \geq N$ we have $a - \varepsilon < a_N \leq a_n \leq a$ so $|a_n - a| < \varepsilon$ and we are done.

□

Observe that there is a corresponding notion of a *(strictly) monotone decreasing* sequence, and a decreasing analogue of Proposition 6.8.

Definition 6.7

Suppose that (a_i) is a sequence of real numbers, and that α is an *order-preserving* injection from $\mathbb{N}$ to $\mathbb{N}$. A sequence (b_i) whose i-th term is $a_{\alpha(i)}$ is called a *subsequence* of (a_i).

To say that α is order preserving means that when $j, k \in \mathbb{N}$ and $j < k$ then $\alpha(j) < \alpha(k)$. This will probably be a bit abstract for some tastes. Informally then, a subsequence of a sequence (a_i) is obtained by omitting some entries, but still leaving infinitely many terms (in the same order that they appeared in the original sequence) to form the subsequence. You might omit the first billion terms, or omit alternate terms, or omit all but the terms a_p where p is a prime number. A good notation for a subsequence is (a_{n_i}), where the i-th term is the n_i-th term of (a_j), and it is understood that if $i_1 < i_2$ then $n_{i_1} < n_{i_2}$.

We gather together some useful observations in the next result.

Lemma 6.1

Suppose that (a_i) is a sequence with subsequence (a_{n_i}).

(a) If (a_i) is bounded, then (a_{n_i}) is bounded.

(b) If (a_i) converges to l, then (a_{n_i}) converges to l.

(c) If (a_i) converges to l, and each $a_i \in I$, where $I = [c, d]$ is a closed interval, then $l \in I$.

Proof

The proofs of parts (a) and (b) are immediate from the definitions. As for part (c), we suppose (for contradiction) that $l > d$ (the case $l < c$ is similar). Let

$\varepsilon = l - d > 0$. Thus $|a_i - l| \geq \varepsilon$ for all $i \in \mathbb{N}$, so (a_i) does not converge to l which is absurd.

□

The following result is a gem. The result itself is interesting, the proof is sweet, and the proposition has important consequences.

Proposition 6.9

Any sequence (a_i) has a monotone subsequence.

Remark 6.4

The strategy for showing this is as follows. We set up a dichotomy between two situations, exactly one of which must occur. In one case we exhibit a strictly monotone decreasing subsequence of (a_i), and in the other case we exhibit a (non-strictly) increasing subsequence of (a_i).

Proof

Let $J = \{j \mid a_j > a_i \forall i > j\} \subseteq \mathbb{N}$. The set J is either finite or infinite.

If J is infinite, let the k-th element of J in ascending order be n_k. Now if $u, v \in \mathbb{N}$ and $u < v$, then $n_u < n_v$ so $a_{n_u} > a_{n_v}$. Thus (a_{n_k}) is a strictly monotone decreasing subsequence of (a_i).

On the other hand, if J is finite (this includes the case that J is empty), then we put $m = \text{Max}(J \cup \{0\})$. Let $m_1 = m + 1$. Suppose that $m_i \in \mathbb{N}$ is defined, then we may choose a natural number m_{i+1} such that $m_{i+1} > m_i$, and $a_{m_i} \leq a_{m_{i+1}}$ since $m_i > m$. It follows that (a_{m_i}) is a (non-strictly) monotonic increasing sequence. Thus either we construct a strictly monotone decreasing subsequence, or we fail but discover a (not necessarily strictly) increasing subsequence instead.

□

As a corollary we obtain a famous result. Note that a bounded sequence is one which is bounded both above and below.

Theorem 6.1 (Bolzano–Weierstrass)

Any bounded sequence has a convergent subsequence.

Proof

This follows directly from Propositions 6.8 and 6.9 and Lemma 6.1(a).

□

This leads us to the highlight of this section. There is a nasty weakness in the definition of a convergent sequence. If you want to show that a sequence is convergent, you have to know the number l to which it converges, otherwise you can't apply Definition 6.3. Now, when you are doing mathematics, and you stumble across a sequence which you hope converges, it is highly unlikely that your invisible friend will whisper the limit in your ear. What you need is an *intrinsic* criterion of convergence; something which depends only upon the sequence itself, and not on the knowledge of l.

We make the key definition.

Definition 6.8

A sequence (a_n) of real numbers is called a *Cauchy sequence* exactly when the following condition is satisfied: for any $\varepsilon > 0$ there is $N \in \mathbb{N}$ such that if $n, m \in \mathbb{N}$ and both $n, m \geq N$, then $|a_n - a_m| < \varepsilon$.

Informally, you don't know that the terms are approaching a limit, but you do know that the terms are approaching one another.

Theorem 6.2

The sequence (a_n) of real numbers is convergent if and only if it is a Cauchy sequence.

Proof

$\Rightarrow$) If (a_n) converges to l, and $\varepsilon > 0$, choose $N \in \mathbb{N}$ such that if $n \geq N$, then $|a_n - l| < \varepsilon/2$. Now if $n, m \geq N$, then $|a_n - a_m| \leq |a_n - l| + |l - a_m| < \varepsilon$.
$\Leftarrow$) Set $\varepsilon = 1$. There is $N \in \mathbb{N}$ such that for all $n, m \geq N$ we have $|a_n - a_m| < 1$. In particular $||a_n| - |a_N|| \leq |a_n - a_N| < 1$. Thus for every $i \in \mathbb{N}$ we have $|a_i| \leq \max\{|a_j| + 1 \mid 1 \leq j \leq N\}$.

The sequence (a_i) also contains a monotone subsequence (a_{n_i}) by Proposition 6.9, which is also bounded by Lemma 6.1. The monotone subsequence converges to a limit l by Proposition 6.8. Now we show that l is the limit of the Cauchy sequence (a_n). Suppose that $\varepsilon > 0$ is given. Choose $M_1 \in \mathbb{N}$ so that if $n, m \geq M_1$, then $|a_n - a_m| < \varepsilon/2$. We can do this by the Cauchy condition.

Now choose M_2 so that if $i \geq M_2$ then $|a_{n_i} - l| < \varepsilon/2$. Let $M = \max\{M_1, M_2\}$. Select $k \geq M$ so that $n_k \geq M$. Now we deploy the triangle inequality using a standard ruse. For $n \geq M$ we have

$$|a_n - l| = |a_n - a_{n_k} + a_{n_k} - l| \leq |a_n - a_{n_k}| + |a_{n_k} - l| < \varepsilon/2 + \varepsilon/2 = \varepsilon$$

and we are done.

□

EXERCISES

6.7 Let a_k be the truncation of the decimal representation of π after the k-th decimal place. Show that (a_k) is a Cauchy sequence, and conclude that this sequence converges to a real number.

6.8 (a) Suppose that (a_n) is a bounded sequence. For each $j \in \mathbb{N}$ let $b_j = \sup\{a_m \mid m \geq j\}$. Show that (b_n) is a convergence sequence, the limit of which is called the limit supremum of (a_n), written as $\limsup(a_n)$.

(b) Define an analogous sequence using infima instead of suprema. Show that $\liminf(a_n) \leq \limsup(a_n)$ in the obvious notation.

(c) Show that $\liminf(a_n) = \limsup(a_n)$ if and only if (a_n) is a convergent sequence, and in that event, both limits are $\lim(a_n)$.

6.7 Series

One of the main reasons for studying sequences is to study infinite sums. For example,

$$1 + 1/2 + 1/4 + 1/8 + \ldots$$

or put more neatly $\sum_{i=0}^{\infty} 2^{-i}$, or even $\sum 2^{-i}$ provided that the range of summation is clear from the context. No amount of staring into space will tell you what this infinite sum is. Infinite processes must be tamed by definitions. We use our knowledge of sequences to assign a meaning to the sum. You simply form the *sequence of partial sums.* This sequence is

$$1, \left(1 + \frac{1}{2}\right), \left(1 + \frac{1}{2} + \frac{1}{4}\right), \left(1 + \frac{1}{2} + \frac{1}{4} + \frac{1}{8}\right), \ldots$$

or rather

$$1,\ 3/2,\ 7/4,\ 15/8, \ldots.$$

You define the infinite sum to be the limit of this sequence – if it converges. In this case it does, the limit is 2 and, once we have made a proper definition (and not before), we will be able to write $\sum_{i=0}^{\infty} 2^{-i} = 2$.

We make the definition formally.

Definition 6.9

Suppose that $\sum_{i=1}^{\infty} a_i$ is an infinite series. Form a sequence (s_j) by letting $s_j = \sum_{i=1}^{j} a_i$ for every $j \in \mathbb{N}$. We say the series $\sum_{i=1}^{\infty} a_i$ converges to the sum l exactly when the sequence (s_j) converges to the limit l. If the sequence (s_j) does not converge, then we say that the series $\sum_{i=1}^{\infty} a_i$ is not convergent.

In this definition, the quantities s_j are called ***partial sums***, because they are obtained by adding up a finite part of the series. In fact s_j is the sum of the first j terms of the series. The sequence (s_j) is called the ***sequence of partial sums*** of $\sum_{i=1}^{\infty} a_i$.

Whenever you are staring at an infinite series, remember that it is a sequence in disguise. In particular, the routine results about sequences carry over, and you can add and subtract convergent infinite series term by term and get predictable results. Multiplication is a more delicate business, though there is a sensible way to define the product of two infinite series, but it is **not** term by term.

We have an easy necessary condition for the convergence of a series.

Proposition 6.10

If the series $\sum_i a_i$ converges, then the sequence (a_i) converges to 0.

Proof

The sequence (s_i) of partial sums of the series must be a Cauchy sequence. Thus for any $\varepsilon > 0$, there is $M \in \mathbb{N}$ such that if $n, m \geq M$, then $|s_n - s_m| < \varepsilon$. Let $N = M + 1$, then if $n \geq N$ we have $|a_n| = |s_n - s_{n-1}| < \varepsilon$. Thus (a_i) converges to 0.

□

One of the most striking results about infinite sequences is Corollary 6.3 which ensures that you can rearrange the terms of a convergent sequence and be confident that the new sequence will converge to the same limit. You might expect that if $a_1 + a_2 + a_3 + \ldots$ is a convergent infinite sum, then you should be able to rearrange the summands as you please and get the same answer. After all, addition is commutative, so this looks very likely. Astonishingly, it is false. For example, consider the sequence $1 - 1/2 + 1/3 - 1/4 + 1/5 - \ldots$ which of course means $1 + (-1/2) + (1/3) + (-1/4) + (1/5) + \ldots$. One can show that this series sums to $\log_e 2$. However, you can rearrange the order of the terms so that the series sums to π, or -13, or indeed any real number which you nominate.

We explore this series in some detail, since it is highly instructive. Consider the infinite series

$$\sum_{n=1}^{\infty} 1/n = 1 + 1/2 + 1/3 + 1/4 + \ldots$$

The terms being added are getting smaller and smaller, so it looks as though it might be possible to add up this series – at least Proposition 6.10 is not an obstruction to convergence. Let

$$t_r = \sum_{n=2^{2r-1}+1}^{2^{2r+1}} 1/n.$$

This looks a little intimidating. t_r is the sum of the reciprocals of consecutive natural numbers, the smallest being $2^{2r-1}+1$ and the largest being 2^{2r+1}. The first couple of terms are

$$t_1 = (1/3 + 1/4) + (1/5 + 1/6 + 1/7 + 1/8) \geq 1/2 + 1/2 = 1$$

and

$$t_2 = (1/9 + \ldots + 1/16) + (1/17 + \ldots + 1/32) \geq 1/2 + 1/2 = 1.$$

An induction argument shows that $t_k \geq 1$ for all $k \in \mathbb{N}$. Thus

$$\sum_{n=1}^{8} 1/n = 1 + 1/2 + t_1 > 2,$$

$$\sum_{n=1}^{32} 1/n = 1 + 1/2 + t_1 + t_2 > 3$$

and an induction on k shows that $\sum_{n=1}^{2^{2k-1}} 1/n \geq k$ for every natural number k. Thus the partial sums of $\sum 1/n$ are not bounded above and so the sum does not converge, since, thanks to Proposition 6.5, a convergent sequence is bounded.

Note that it follows that starting with any given $1/k$, one can add reciprocals of consecutive natural numbers so that $(1/k + 1/(k+1) + \ldots 1/m)$ is larger than any natural number you please.

Suppose that we look at alternate terms

$$\begin{aligned} 2\left(1+\frac{1}{3}+\cdots+\frac{1}{2m-1}\right) &= \left(1+\frac{1}{3}+\cdots+\frac{1}{2m-1}\right) \\ &\quad + \left(1+\frac{1}{3}+\cdots+\frac{1}{2m-1}\right) \\ &\geq \left(1+\frac{1}{3}+\cdots+\frac{1}{2m-1}\right)+\left(\frac{1}{2}+\frac{1}{4}+\cdots+\frac{1}{2m}\right) \\ &= \left(1+\frac{1}{2}+\frac{1}{3}+\frac{1}{4}+\cdots+\frac{1}{2m}\right). \end{aligned}$$

It follows that the partial sums of $\sum_{k=0}^{\infty} \frac{1}{2k+1}$ are not bounded above and so also do not converge. Similarly

$$2\left(\frac{1}{2}+\frac{1}{4}+\cdots+\frac{1}{2m}\right) = \left(\frac{1}{2}+\frac{1}{4}+\cdots+\frac{1}{2m}\right)+\left(\frac{1}{2}+\frac{1}{4}+\cdots+\frac{1}{2m}\right)$$

$$\geq \left(\frac{1}{2}+\frac{1}{4}+\cdots+\frac{1}{2m}\right)+\left(\frac{1}{3}+\frac{1}{5}+\cdots+\frac{1}{2m+1}\right) = \frac{1}{2}+\frac{1}{3}+\frac{1}{4}+\cdots+\frac{1}{2m+1}$$

By the same reasoning as before we see that the sum $1/k + 1/(k+2) + \ldots + 1/(k+2t)$ of alternate reciprocals, starting with any natural number k, can be made larger than any natural number you please by suitable choice of t.

We now illustrate how to rearrange the sum $\sum(-1)^n/n$ to add to 2. The same method would work, when modified, to make the sum converge to the real number of your choice. Start by adding the reciprocals of the odd numbers until you overshoot the target, which is 2.

$$1 + 1/3 + 1/5 + \ldots + 1/13 \leq 2$$

is not quite far enough but

$$1 + 1/3 + 1/5 + \ldots + 1/13 + 1/15 > 2$$

takes us home. Now start adding in reciprocals of negative even numbers until you first undershoot 2. You do this immediately because

$$1 + 1/3 + 1/5 + \ldots + 1/13 + 1/15 - 1/2 < 2.$$

Now go back to the original plan, and add on reciprocals of odd numbers until you overshoot again. This happens with

$$(1+1/3+\ldots+1/15)+(-1/2)+(1/17+1/19+\ldots+1/41) = \frac{27438157533868993}{13691261858724450}$$

which is a touch bigger than 2.

Repeat this procedure for ever. You will overshoot and undershoot 2 infinitely many times. However, when you overshoot, you do so by at most the last number to be added. Similarly, when you undershoot, you undershoot by at most the number being subtracted. These "errors" shrink towards 0 as the process continues, and so the sequence converges to 2.

The series

$$1 + 1/2 + 1/3 + 1/4 + \ldots$$

is called the *harmonic series*. Its partial sums are intimately related to the logarithm function. In fact the sequence whose n-th term is $(\sum_{k=1}^{n} 1/k) / \log n$ is a convergent sequence whose limit is 1. Informally then, $\log n$ is a very good approximation to $\sum_{k=1}^{n} 1/k$.

Definition 6.10

A series $\sum a_n$ is *absolutely convergent* if and only if $\sum |a_n|$ is convergent.

Notice that a convergent series of non-negative terms is automatically absolutely convergent. It turns out that you can rearrange the order of the terms in an absolutely convergent series, and it will still converge to the same limit. We do not prove that in this book, though you might like to think about why this is true.

We now show that an absolutely convergent series is convergent.

Proposition 6.11

Let $\sum_{i=1}^{\infty} a_i$ be an infinite series of real numbers. Suppose that the series $\sum_{i=1}^{\infty} |a_i|$ converges. It follows that the series $\sum_{i=1}^{\infty} a_i$ converges.

Proof

Since $\sum_{i=1}^{\infty} |a_i|$ converges, it follows that its sequence of partial sums must be a Cauchy sequence by Theorem 6.2. Our strategy is to show that the partial sums of $\sum_{i=1}^{\infty} a_i$ also form a Cauchy sequence, and then another application of Theorem 6.2 takes us home.

Suppose that $\varepsilon > 0$. There is $N \in \mathbb{N}$ such that if $n, m \geq N$, then

$$\left| \sum_{i=1}^{n} |a_i| - \sum_{i=1}^{m} |a_i| \right| < \varepsilon.$$

We seek to show that if $n, m \geq N$, then

$$\left| \sum_{i=1}^{n} a_i - \sum_{i=1}^{m} a_i \right| < \varepsilon.$$

If $n = m$, this is clear. We suppose, without loss of generality, that $n < m$, then for $N \leq n < m$ we have

$$\left| \sum_{i=1}^{n} a_i - \sum_{i=1}^{m} a_i \right| = \left| \sum_{i=n+1}^{m} a_i \right| \leq \sum_{i=n+1}^{m} |a_i|.$$

The final inequality is by an induction argument based on the triangle inequality. Now for $N \leq n < m$ we have

$$\sum_{i=n+1}^{m} |a_i| = \left| \sum_{i=1}^{n} |a_i| - \sum_{i=1}^{m} |a_i| \right| < \varepsilon.$$

Our strategy is successful and the proof is complete.

□

EXERCISES

6.9 Suppose that $0 \leq a_n \leq b_n \forall n \in \mathbb{N}$. Suppose furthermore that $\sum_{i=1}^{\infty} b_n$ is convergent. Prove that $\sum_{i=1}^{\infty} a_n$ is convergent.

6.10 (a) Suppose $\alpha > 1$ is fixed. By drawing a graph or otherwise, prove that

$$\sum_{i=2}^{n} \frac{1}{i^\alpha} \leq \int_1^n \frac{1}{x^\alpha} dx.$$

Show that there is M such that $\int_1^n \frac{1}{x^\alpha} dx < M$ for all $n \in \mathbb{N}$ and conclude that $\sum_{i=1}^{\infty} n^{-\alpha}$ is convergent.

(b) Suppose now that $\alpha = 1$. Use a similar argument to show that $\sum_{i=1}^{\infty} n^{-1}$ is not convergent.

6.11 A piece of elastic is 1 metre long. It is fixed to a point, and held horizontally. A very lazy spider tries to walk from the fixed end along the elastic at a constant speed of 1 cm per day. At the end of 24 hours, and at the end of every subsequent 24-hour period, an arachnophobe hand stretches the far end of the elastic away from the fixed point by a distance of 1 metre. You should either estimate how long it will take the spider to reach the far end of the elastic, or prove that it will never do so.

7 Mathematical Analysis

Mathematical analysis – or just *analysis* – is calculus with attitude. It is the subject which picks up the calculus, and shines a blazing light on every tiny detail of how it works, justifying many familiar methods of differentiation and integration in the process. After that, once we have a way of understanding what happens, we can move into new areas – calculus in many dimensions, or the study of solutions of differential equations. All the time you carry with you your analysis skills. These keep you honest, and force you to check that your mathematical activities are legitimate.

Do you remember the definition of a function? It is a rule $f : A \to B$ which assigns to each element $x \in A$ a specific element of the set B. We focus on the case that $A \subseteq B = \mathbb{R}$. Everyday functions include those which send x to x^2, y to $\sin y$, and z to $|z|$. If you draw graphs of these functions you will see that the first two are very nice – you can draw the graphs without taking your pen from the paper and without making any sharp turns. What you actually draw is a bit of the graph near the origin – since your paper is of finite extent, you have no choice.

7.1 Continuity

The graph of the function which sends x to $|x|$ is an infinite V-shape with the sharp bit at the origin. You can draw it in one go, but there is one particular place where things take a dramatic turn.

G. Smith, *Introductory Mathematics: Algebra and Analysis*

Now think about the *Heaviside step function* $H(x)$. This is a function of which electrical engineers are very fond. We define it by $H(x) = 1$ if $x > 0$ and $H(x) = 0$ if $x \leq 0$. Please draw its graph now. It consists of two horizontal lines, or rather half-lines. This is very dull function once you are away from 0 since there is no difficulty in drawing the graph without taking the pen from the paper. There are no kinks, nor anything but unremitting dullness. Where does the leap happen? Well, not until after 0, but before any positive number, as you pass from left to right.

Let us isolate two features concerned with the smoothness of the graph of a function. The idea of drawing the graph without removing the pen will be captured by the mathematical notion of *continuity*. Of course it may be that the graph can be drawn without removing the pen in one region, but not in another. Thus continuity is a local idea, and we should first get control of the idea of a function being continuous at a particular value a in its domain. Once that is done, we can then consider what happens if the function is continuous at all points of its domain. The idea that the graph of a function is free of leaps and sharp kinks sufficiently near $a \in \operatorname{dom}(f)$ corresponds to the function being *differentiable* at a. The function defined by the formula $|x|$ is continuous for all real values of x. It is also differentiable except when $x = 0$. Geometrically there is no reasonable definition of the tangent to the graph of $|x|$ at the origin.

Let us make some proper definitions.

Definition 7.1

A function $f(x)$ is *continuous* at the real number a exactly when the following condition holds: given any $\varepsilon > 0$ there is $\delta > 0$ such that $|f(x) - f(a)| < \varepsilon$ when $|x - a| < \delta$. Here we implicitly insist that not only is $|x - a| < \delta$, but also $x \in \operatorname{dom}(f)$, a condition which we may omit as "understood" in future.

You can also phrase this in other ways.

Definition 7.2 (different form of words)

A function $f(x)$ is *continuous* at the real number a exactly when the following condition holds: given any $\varepsilon > 0$ there is $\delta > 0$ such that if $|x - a| < \delta$, then $|f(x) - f(a)| < \varepsilon$.

Note that you are allowed to choose δ in response to ε, just as when dealing with sequences when you are allowed to choose $N \in \mathbb{N}$ in response to ε. This is very important. This definition may look a little scary at first, but perhaps Figure 7.1 will help. We will try to tame the definition. First of all, there are a

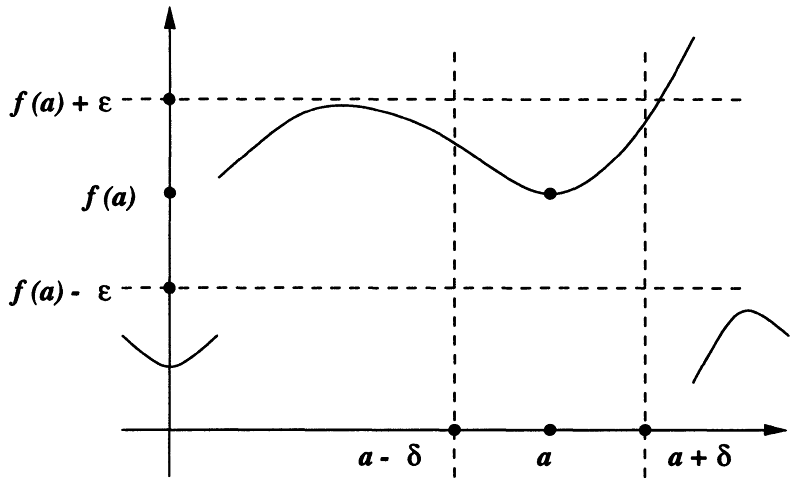

Fig. 7.1. A snapshot of δ doing the job for a particular ε

couple of expressions involving the modulus symbol. Remember that if x and y are real numbers, then $|x-y| = |y-x|$ is just the distance between x and y when you think of x and y as being points on the real line. Geometrically then, $|x-a| < \delta$ means that the distance from x to a is less than δ. We know that there is an open interval centred on a of length 2δ inside which x lives.

Now read the definition again: given any $\varepsilon > 0$ (*think of ε as a bound on the allowable εrror*), there is $\delta > 0$ (*think of $\delta > 0$ as a δeviation from a*) such that if $|x-a| < \delta$, then $|f(x) - f(a)| < \varepsilon$.

Now we put it conceptually: given any $\varepsilon > 0$, a bound on allowable error, there is a permitted deviation δ so that if the distance of x from a is less than the permitted deviation δ, then the distance of $f(x)$ from $f(a)$ is less than the allowable error ε.

Note, by the way, that small margins of error are more difficult to arrange than big ones. What matters is what happens when ε shrinks towards 0. A δ which works for a particular value of ε will also work for any larger value of ε.

We look at some reassuring examples.

Example 7.1

The most boring function of all: a constant function. Consider the function h defined by $h(x) = c$ for some fixed $c \in \mathbb{R}$ and for all $x \in \mathbb{R}$. Choose and fix a real number a. We will show that h is continuous at a. Given any $\varepsilon > 0$, let $\delta = 1$ (or any positive real number which takes your fancy). Now if $|x-a| < \delta$, then $|h(x) - h(a)| = |0| = 0 < \varepsilon$.

That was quite simple; we didn't learn much because constant functions are so easy to handle. However, it is reassuring that constant functions are continuous at all $a \in \mathbb{R}$, according to our definition.

Example 7.2

The Heaviside function $H(x)$ takes only two distinct values, one when $x > 0$, and another when $x \leq 0$. In this sense it is almost, but not exactly, constant. Arguments similar to those used in the previous example show that if $a \in \mathbb{R}$ and $a \neq 0$, then $H(x)$ is continuous at a. However, we expect things to go wrong at $a = 0$.

Let $\varepsilon = 1/2$ and suppose that $\delta > 0$. Let $x = \delta/2$, so $|x - 0| = \delta/2 < \delta$. However $|H(x) - H(0)| = 1 > 1/2 = \varepsilon$. Thus no $\delta > 0$ will do the job required in the definition of continuity when $\varepsilon = 1/2$, so $H(x)$ is not continuous at 0. Note that the same problem would arise for any $\varepsilon \in (0, 1]$. If however, we look at $\varepsilon > 1$ then any positive δ will do the job. For continuity, you need to able to find a δ in response to any ε. Also, you expect it to be more difficult to find δ when ε is small, as in this example.

Example 7.3

Consider the identity function $i : \mathbb{R} \to \mathbb{R}$ defined by $i(x) = x$ for every $x \in \mathbb{R}$. The graph of this function is (geometrically) a straight line through the origin. This time there is a little more work to do. Choose and fix a real number a. We will show that i is continuous at a. Given any $\varepsilon > 0$, let $\delta = \varepsilon$. *Note that this time we are choosing δ which actually depends on ε – this is the usual state of affairs.* Now suppose that $|x - a| < \delta = \varepsilon$, then $|i(x) - i(a)| = |x - a| < \varepsilon$. Thus i is continuous at a. However, a was arbitrary so i is continuous everywhere.

In that example, we need not have chosen δ to be the same as ε. The only property that the positive real number δ needs to have is that $\delta \leq \varepsilon$. Thus we might have chosen δ to be $\varepsilon/2$, $34\varepsilon/55$, $\min\{\sqrt{\varepsilon}, \varepsilon^2\}$ or in any one of myriad ways. Of course, you have a duty not to generate irrelevant complications. If you have a choice, choose naturally or elegantly wherever possible.

Example 7.4

Now be brave, and consider the squaring function defined by the formula $s(x) = x^2$ for every $x \in \mathbb{R}$. You should be very familiar with the graph of this function.

Its graph is a nose-down parabola. and the lowest point of the graph is at the origin. Continuity at each $a \in \mathbb{R}$ is clear. However, clear is not good enough; we have to use our definition. Choose and fix a real number a. We will show that s is continuous at a.

If $a = 0$, then put $\delta = \sqrt{\varepsilon}$. Now if $|x - a| = |x - 0| = |x| < \delta = \sqrt{\varepsilon}$, then $|x^2 - a^2| = |x^2| = |x|^2 < (\sqrt{\varepsilon})^2 = \varepsilon$ and we are done.

Now for the main case. We assume that $a \neq 0$. Given any $\varepsilon > 0$, ***abracadabra*** let $\delta = \min\{\varepsilon/3|a|, |a|\} > 0$. Where did that come from? Wait and see. Notice we have used the fact that $a \neq 0$.

Suppose that $|x - a| < \delta$. We need to show that it follows that $|x^2 - a^2| < \varepsilon$. First a little algebra on the side. Observe that

$$|x^2 - a^2| = |(x - a)(x + a)| = |x - a||x + a|.$$

It seems that in order to force $|x^2 - a^2|$ to be small, you have to worry about the size of $|x + a|$. The triangle inequality will rush to our aid.

$$|x + a| = |x - a + 2a| \leq |x - a| + |2a| = |x - a| + 2|a|.$$

Under the assumption that $|x - a| < \delta$ we conclude that $|x + a| \leq \delta + 2|a| \leq 3|a|$. Here we have just used the fact that $\delta \leq |a|$ which is true by definition of δ. It also follows that

$$|x^2 - a^2| = |x - a||x + a| < \delta \cdot 3|a| \leq \frac{\varepsilon}{3|a|} \cdot 3|a| = \varepsilon.$$

Think about what you have to do; you have to select δ so that at a later stage, everything will work out well. Sometimes the later argument may be very complicated, and you many want to choose δ to satisfy a variety of conditions. What you do is to select δ to be the minimum of a finite number of expressions, each one carefully chosen so that a fragment of the argument will work out nicely. This technique is illustrated by the next example.

Example 7.5

Consider the function $g(x) = x^n$ for some fixed natural number n and all $x \in \mathbb{R}$. We will show that this function is continuous (irrespective of the value of n), using induction on n. The base case has been done already ($n = 1$), and so has the next case ($n = 2$). That is logically irrelevant, but was good exercise. Assume that $n \geq 2$, then

$$|x^n - a^n| = |x^n - xa^{n-1} + xa^{n-1} - a^n| \leq |x^n - xa^{n-1}| + |xa^{n-1} - a^n|$$

using the triangle inequality. This simplifies to say that

$$|x^n - a^n| \leq |x||x^{n-1} - a^{n-1}| + |a|^{n-1}|x - a|.$$

The occurrence of $|x|$ is a bit awkward, so we use the triangle inequality again via $|x| \leq |x-a| + |a|$ to get rid of it. Thus

$$|x^n - a^n| \leq |x-a||x^{n-1} - a^{n-1}| + |a||x^{n-1} - a^{n-1}| + |a|^{n-1}|x-a|.$$

Choose $\delta_1 > 0$ such that if $|x-a| < \delta_1$, then $|x^{n-1} - a^{n-1}| < \varepsilon/3$. You can do this by induction on n.

Choose $\delta_2 > 0$ such that if $|x-a| < \delta_2$, then $|a||x^{n-1} - a^{n-1}| < \varepsilon/3$. If $a = 0$ you choose any value of δ_2 which you please. If $a \neq 0$ you just choose δ_2 so that $|x^{n-1} - a^{n-1}| < \varepsilon/3|a|$ by induction on n.

Choose $\delta_3 > 0$ such that if $|x-a| < \delta_3$, then $|a|^{n-1}|x-a| < \varepsilon/3$. If $a = 0$ you can select δ_3 arbitrarily. If $a \neq 0$ then let $\delta_3 = \varepsilon/3|a|^{n-1}$.

Now put $\delta = \min\{\delta_1, \delta_2, \delta_3, 1\}$. If $|x-a| < \delta$, then

$$|x^n - a^n| < 1 \cdot \varepsilon/3 + \varepsilon/3 + \varepsilon/3 = \varepsilon$$

and we are done.

After these examples, some theory is refreshing.

Proposition 7.1

Suppose that $f : I \to \mathbb{R}$ where I is an open interval. Suppose that $a \in I$ and that f is continuous at a. It follows that there is $\eta > 0$ such that f is a bounded function when you restrict its domain to $I \cap (a - \eta, a + \eta)$.

Proof

Let $M = |f(a)| + 1$. Let $\varepsilon = 1$ and choose $\eta > 0$ so that if $x \in I$ and $|x - a| < \eta$, then $|f(x) - f(a)| < 1$. Thus, for such x we have

$$f(a) - 1 < f(x) < f(a) + 1 < |f(a)| + 1 = M.$$

Also $-f(a) - 1 < -f(x) < -f(a) + 1$ so

$$-f(x) < -f(a) + 1 \leq |f(a)| + 1 = M.$$

In any event, $|f(x)| < M$ for x in the specified range and we are done.

□

Example 7.6

Consider the function defined by $f(x) = 1/x$. We must worry about $f(0)$. Well, you have two choices: patch or discard. You can make a special case out of 0, and put $f(0) = 42$ for example. Alternatively, you can throw away 0 (and possibly other stuff) from the domain. You might want to decide that the domain of the function is the positive reals $\mathbb{R}_+$. We (arbitrarily) adopt the second strategy. However, in the light of Proposition 7.1, we know that no definition of $f(0)$ will render our function continuous at the origin, since it is not bounded in any interval surrounding 0, even though 0 is removed from the interval.

We shall show that this function is continuous at each positive real number a. Suppose that $\varepsilon > 0$. We want to show that there is $\delta > 0$ with the property that if $x \in \mathbb{R}_+$ and $|x - a| < \delta$, then $|1/x - 1/a| < \varepsilon$.

We investigate the expression $|1/x - 1/a|$ which we need to bound by ε. Now

$$|1/x - 1/a| = \left|\frac{a-x}{xa}\right|$$

so adopt the strategy of letting $\delta = \min\{\frac{a}{2}, \frac{a^2\varepsilon}{2}\}$. This may look a little like magic, so we explain the idea. By selecting $\delta \leq \frac{a}{2}$ the condition $|x - a| < \delta$ will ensure that $x > a/2$. This in turn ensures that x cannot be close to 0 where the trouble lives.

We have

$$|1/x - 1/a| = \left|\frac{a-x}{xa}\right| = \frac{1}{x} \cdot \frac{|a-x|}{a}$$

Now by virtue of $x > a/2$ we know that $1/x < 2/a$, and the condition that $\delta \leq \frac{a^2\varepsilon}{2}$ can come into play via

$$|1/x - 1/a| < \frac{2}{a} \cdot \frac{a^2\varepsilon}{2a} = \varepsilon.$$

That was slightly cunning.

We now have a list of many functions which ought to be continuous at sensibly chosen points of their domains, and now we know they are. Here is a more disturbing example.

Consider the function $f : \mathbb{R} \to \mathbb{R}$ defined by $f(r) = 0$ if $r \in \mathbb{R} \setminus \mathbb{Q}$, and if $q = a/b$ with $a \in \mathbb{Z}$ and $b \in \mathbb{N}$ in lowest terms (the greatest common divisor of a and b is 1), then $f(q) = 1/b$.

This function will give our definition of continuity a thorough test. First of all, let us imagine the graph of f. Loosely speaking, "most" numbers are irrational. If we throw out all the rational numbers from our "x-axis", the function is constant, and its graph coincides with the "x-axis". However, putting the rationals back in, there are points marked in the upper half plane above each non-zero rational number. Above 1 is $(1, 1)$, above 2 is $(2, 1)$, above $14/91$ is $(14/91, 1/13) = (2/13, 1/13)$ and above $-3/4$ is $(-3/4, 1/4)$. This function is very jumpy, but the following result holds.

Proposition 7.2

This function f is continuous at all irrational points a, and at 0. It is not continuous at any non-zero rational point.

Proof

Suppose that a is an irrational number, and that we are given $\varepsilon > 0$. Choose $M \in \mathbb{N}$ such that $1/M < \varepsilon$. For each natural number y let $S_y = \{z/y \mid z \in \mathbb{Z}\}$. Thus each S_y is an infinite set but its intersection with any bounded interval is finite (since its elements are spaced out regularly, at distances $1/y$ apart). Now consider the interval $(a - \varepsilon, a + \varepsilon)$. Let $T_M = \bigcup_{i=1}^{M} S_i$ and define $\overline{T}_M = (a - \varepsilon, a + \varepsilon) \cap T_M = \bigcup_{i=1}^{M} (a - \varepsilon, a + \varepsilon) \cap S_i$ which is the union of finitely many finite sets and so is finite. Now consider $D_M = \{|t - a| \mid t \in \overline{T}_M\}$. Now elements of $\overline{T}_M$ are rational and a is not rational so $0 \notin D_M$. Thus, D_M is a finite set of positive numbers. It therefore has a positive minimum element δ.

Now if $x \in \mathbb{R}$ and $|x - a| < \delta$ then $x \notin \overline{T}_M$ (otherwise we would violate the minimality of δ). Thus x is either irrational, or it is rational in lowest terms u/v with $u \in \mathbb{Z}$ and $v \in \mathbb{N}$ with $v > M$. Thus either $f(x) = 0 < \varepsilon$ or $f(x) = 1/v < 1/M < \varepsilon$, and this part of the proof is complete. We point out that a similar but simpler analysis will yield the result that the function is continuous at 0. We leave the details as an exercise.

Now suppose instead that a is a non-zero rational number. Thus $a = p/q$, the latter being a ratio of integers in lowest terms with $q \in \mathbb{N}$ and $p \neq 0$. Now $f(a) = 1/q > 0$. Let $\varepsilon = 1/q$. Now suppose (for contradiction) that there is $\delta > 0$ such that if $|x - a| < \delta$, then $|f(x) - 1/q| < 1/q$. Choose $N \in \mathbb{N}$ sufficiently large so that $\sqrt{2}/N < \delta$. Put $t = 1/q + \sqrt{2}/N$ so t is irrational (if t were rational then $N(t - 1/q) = \sqrt{2}$ would be rational which is absurd). Thus $|t - 1/q| < \delta$ and t is irrational so $f(t) = 0$ and so $|f(t) - 1/q| = 1/q = \varepsilon$ and so it is not the case that $|f(t) - 1/q| < \varepsilon$. Thus no δ will do the job, and f is not continuous at irrational a.

□

Proposition 7.2 is quite remarkable in view of Proposition 1.1. The rational and irrational points are completely intermingled, and the function switches between continuity and discontinuity infinitely many times as a, the point in question, moves through any interval of positive length.

There are ways to make new functions from old. For example you can add, subtract, multiply and compose functions from $\mathbb{R}$ to $\mathbb{R}$. Thus from functions defined by the formulas x^2 and x^3 you can build functions defined by formulas x^2+x^3, x^2-x^3, x^5 and x^6 corresponding, respectively, to these four procedures.

As you might expect, continuity at a point of the domain is preserved by these operations. The next theorem says it precisely.

Theorem 7.1

Suppose that $f, g : \mathbb{R} \to \mathbb{R}$ are functions which are continuous at $a \in \mathbb{R}$. It follows that the sum $f + g$ and the product fg are both continuous at a. Moreover the composition $f \circ g$ is continuous at a provided f is continuous at $g(a)$ (and for this part the continuity of f at a is irrelevant – unless $g(a) = a$).

Proof

We begin, as usual, with an application of the triangle inequality. We have

$$|(f+g)(x) - (f+g)(a)| = |f(x) + g(x) - f(a) - g(a)|$$

$$= |f(x) - f(a) + g(x) - g(a)| \leq |f(x) - f(a)| + |g(x) - g(a)|.$$

Thus you can make $(f+g)(x)$ close to $(f+g)(a)$ by forcing $f(x)$ close to $f(a)$ and $g(x)$ close to $g(a)$. Given any $\varepsilon > 0$, choose $\delta_1, \delta_2 > 0$ so that if $|x-a| < \delta_1$, then $|f(x) - f(a)| < \varepsilon/2$ while if $|x - a| < \delta_2$, then $|g(x) - g(a)| < \varepsilon/2$. Let $\delta = \min\{\delta_1, \delta_2\}$. Now if $|x - a| < \delta$, then $|(f+g)(x) - (f+g)(a)| \leq |f(x) - f(a)| + |g(x) - g(a)| < \varepsilon/2 + \varepsilon/2 = \varepsilon$.

Now for the product. Notice that

$$|f(x)g(x) - f(a)g(a)| = |f(x)g(x) - f(x)g(a) + f(x)g(a) - f(a)g(a)|$$

$$\leq |f(x)g(x) - f(x)g(a)| + |f(x)g(a) - f(a)g(a)|$$

$$= |f(x)||g(x) - g(a)| + |g(a)||f(x) - f(a)|.$$

By continuity we know how to force $g(x)$ to be close to $g(a)$ and $f(x)$ to be close to $f(a)$. The trouble is caused by the term $|f(x)|$. We need a ruse.

Choose $\delta_1 > 0$ so that if $|x - a| < \delta_1$, then $|f(x) - f(a)| < 1$. Now by the variation on the triangle inequality $||f(x)| - |f(a)|| \leq |f(x) - f(a)| < 1$. Thus $|f(a)| - 1 < |f(x)| < |f(a)| + 1$. The final inequality is what we want. Notice that $|f(a)| + 1$ is not 0 so we can take its reciprocal.

Now we are in business. Choose δ_2 so that if $|x-a| < \delta_2$, then $|g(x)-g(a)| < \varepsilon/2(|f(a)|+1)$. Choose δ_3 so that if $|x-a| < \delta_3$, then $|g(a)||f(x)-f(a)| < \varepsilon/2$. We have seen something like this before. If $g(a) = 0$, then you choose $\delta_3 = 1$ or indeed how you please. If $g(a) \neq 0$, then you choose δ_3 so that if $|x - a| < \delta_3$, then $|f(x) - f(a)| < \varepsilon/2|g(a)|$.

Now putting $\delta = \min\{\delta_1, \delta_2, \delta_3\}$, we finish in the same way we did in Example 7.5.

Composition of functions turns out to be relatively easy. Suppose that you wish to force $|f(g(x)) - f(g(a))| < \varepsilon$. You choose δ_1 so that if $|z - g(a)| < \delta_1$, then $|f(z) - f(g(a))| < \varepsilon$. This works because f is continuous at $g(a)$. Now let $\varepsilon_1 = \delta_1$, and select δ so that if $|x - a| < \delta$, then $|g(x) - g(a)| < \varepsilon_1 = \delta_1$. It then follows of course that $|f(g(x)) - f(g(a))| < \varepsilon$ and we are done.

□

In hindsight, much of our earlier work is rendered pointless. Constant functions are continuous, and if f is continuous so is the function $2f$, or indeed cf where c is any real number. The function $x \mapsto x$ (the identity function) is continuous, so by the product result, the function defined by the formula x^2, and by induction the functions defined by the formula x^n, are all continuous at all points $a \in \mathbb{R}$.

In fact both sine and cosine are continuous functions at all $a \in \mathbb{R}$, as is the exponential function ($\sin x$, $\cos x$ and e^x). Using our results we now know that functions we build from these using addition, multiplication and composition will all be continuous. Thus the function defined by the formula $\sin(\cos(x^{13} + e^x))$ is continuous at all $a \in \mathbb{R}$.

The expressions $\sec x$, $\operatorname{cosec} x$, $\tan x$ and $\cot x$ do not define maps from $\mathbb{R}$ to $\mathbb{R}$ (remember these are respectively $1/\cos x$, $1/\sin x$, and the quotients $\sin x/\cos x$, and $\cos x/\sin x$. The problem in each case is the same; at certain places the denominator is zero. For example, $\sin x$ vanishes at all integer multiples of π (and nowhere else). The easy way to define the cosecant (cosec) function is to jettison these multiples of π from the domain. Let $A = \mathbb{R}\backslash\{k\pi \mid k \in \mathbb{Z}\}$ so $\operatorname{cosec} : A \to \mathbb{R}$ is the map defined by $\operatorname{cosec}(x) = 1/\sin x$.

Since continuity is all about what happens near to a point of the domain, everything works just as well for cosecant. This is a function (with a slightly peculiar domain) which is continuous at every point of its domain.

EXERCISES

7.1 Prove that the function defined by the formula $|x|$ is continuous at all $x \in \mathbb{R}$.

7.2 Using the definition of the sine and cosine functions given at the beginning of Section 3.4, prove that sine and cosine are continuous functions at all points $a \in \mathbb{R}$.

7.3 (a) Suppose that $a, b \in \mathbb{R}$. Show that

$$\max\{a, b\} = \frac{a+b}{2} + \frac{|a-b|}{2}.$$

(b) Suppose that $f, g : \mathbb{R} \to \mathbb{R}$ are continuous at the point $a \in \mathbb{R}$. Define a new function by $h : \mathbb{R} \to \mathbb{R}$ by $h(x) = \max\{f(x), g(x)\}$. Prove that h is continuous at a.

Definition 7.3

Suppose that I is an interval. A function $f : I \to \mathbb{R}$ is *continuous on I* if it is continuous at each point $a \in I$.

We now introduce a way of taking discussion about continuity, and turning it into an equivalent discussion about sequences. This is an attractive move, because all those excellent notions and theorems of Chapter 6 (Cauchy sequences, bounded sequences have convergent subsequences, and so on) can come into play when we are reasoning about continuity. It may also be that you find the following sequential characterization of continuity at $a \in \mathbb{R}$ more appealing than the classical "ε, δ" Definition 7.1. Read on, and make a judgement.

Proposition 7.3

Suppose that I is an interval. A function $f : I \to \mathbb{R}$ is continuous at $a \in I$ if and only if the following condition is satisfied: whenever (a_i) is a sequence of elements of I converging to a, it follows that the sequence $(f(a_i))$ converges to $f(a)$.

Proof

Suppose that f is continuous at a and that (a_i) is a sequence in I converging to a. We need to show that $(f(a_i))$ converges to $f(a)$. Given any $\varepsilon > 0$ we must find $N \in \mathbb{N}$ with the property that if $n \geq N$, then $|f(a_i) - f(a)| < \varepsilon$. By the continuity of f at a we may find $\delta > 0$ such that $|f(x) - f(a)| < \varepsilon$ when $x \in I$ and $|x - a| < \delta$. Let this δ play the rôle of ε in the the definition of (a_i) converging to a. Thus there is $N \in \mathbb{N}$ with the property that if $i \geq N$, then $|a_i - a| < \delta$, and so $|f(a_i) - f(a)| < \varepsilon$ and we are exactly half way home.

Start again, and suppose that whenever (a_i) is a sequence of elements of I converging to a, it follows that the sequence $(f(a_i))$ converges to $f(a)$. We need to show that f is continuous at $a \in I$. Given any $\varepsilon > 0$, we must show there is $\delta > 0$ such that if $|x - a| < \delta$, then $|f(x) - f(a)| < \varepsilon$. Suppose, for contradiction, that this is not the case. Now our remarks about quantifiers in Section 1.11 come into their own. We seek the negation of a rather complex statement. The statement itself is

$$\forall a \in I \ \forall \varepsilon > 0 \ \exists \delta > 0 \ \forall x \in (a-\delta, a+\delta) \cap I \ \ (|f(x) - f(a)| < \varepsilon). \tag{7.1}$$

The negation of this statement is

$$\exists a \in I \ \exists \varepsilon > 0 \ \forall \delta > 0 \ \exists x \in (a-\delta, a+\delta) \cap I \ \ (|f(x) - f(a)| \geq \varepsilon). \tag{7.2}$$

Now our assumption (for contradiction) is that condition (7.2) holds. Focus upon the a and ε mentioned in condition (7.2). Construct a sequence (a_i) as follows. For each i, let $\delta = 1/i$, and let a_i be an x having the property described in condition (7.2). Now $|a - a_i| < 1/i$ for every $i \in \mathbb{N}$, so it follows that (a_i) converges to a. On the other hand $|f(a_i) - f(a)| > \varepsilon \, \forall i \in \mathbb{N}$. Thus it cannot be that $f(a_i)$ converges to $f(a)$. This is absurd by hypothesis. This contradiction brings the argument to a close.

□

Theorem 7.2

Let I be a closed interval, and $f : I \to \mathbb{R}$ be continuous on I. It follows that the function is bounded above, and there is $a \in I$ such that $f(a) = \sup\{f(x) \mid x \in I\}$.

Proof

We first show that $f(x)$ is bounded above. Suppose (for contradiction) that it is not. Choose $x_1 \in I$ such that $f(x_1) > 1$. Suppose that $x_k \in I$ has been defined. Choose $x_{k+1} \in I$ such that $f(x_{k+1}) > \max\{f(x_k), k+1\}$. We have constructed a sequence (x_i) of points of I. It also follows that the sequence $(f(x_i))$ is strictly monotone increasing.

Now (x_k) is a bounded sequence and so has a convergent subsequence (x_{k_i}) thanks to Theorem 6.1. Let the limit be b. Now $b \in I$ since I is a closed interval and Lemma 6.1 applies. Using Proposition 7.3 we know that $(f(x_{k_i}))$ converges to a limit, so it is a bounded sequence by Proposition 6.5. However $k_i \geq i$ for all $i \in \mathbb{N}$ (that being the nature of subsequences), so $f(x_{k_i}) \geq f(x_i) > i$ since $(f(x_i))$ is monotone increasing. Thus $(f(x_{k_i}))$ is not a bounded sequence. Earlier we proved it was. This contradiction establishes that f is bounded above.

Let $M = \sup\{f(x) \mid x \in I\}$. Thus for every $\eta > 0$ there is $y \in I$ such that $f(y) \in (M - \eta, M)$. Choose a a sequence (a_i) of points of I where for each k we insist that $a_k = y$ in the condition when $\eta = 1/k$.

We have constructed a sequence (a_i) of elements of I with the property that $(f(a_i))$ is a sequence converging to M. Choose a convergent subsequence (a_{m_i}) of the bounded sequence (a_i), and let (a_{m_i}) have limit $a \in I$ using Lemma 6.1. The sequence $(f(a_{m_i}))$ is a subsequence of a sequence converging to M, and so converges to M itself by Lemma 6.1. Now by Proposition 7.3 we conclude that $f(a) = M$ and we are done.

□

We are now in a position to prove one of the major theorems of basic analysis. First, a non-mathematical perspective. Suppose that you note the temperature outside every day at noon. Suppose that yesterday it registered 15^o and today it registered 28^o. Perhaps you live in Alaska and use the Fahrenheit scale, or somewhere more temperate and use the Celsius scale. Notice that temperature as registered by a mercury thermometer varies in a continuous way – and here we are using the word continuous in a non-technical sense. As time goes by, the temperature may rise or fall. You can (informally) deduce that at at least one time between those two midday measurements the temperature outside was exactly 17.29^o. You know this because the mercury level rises and falls smoothly, so if it is going to change from 15^o to 28^o it is going to have to pass through 17.29^o on the way.

We have a highly technical mathematical definition of continuity. If we have captured the intuitive idea of continuity well, then we would expect to be able to turn this example about temperature into a mathematical theorem. Evidence that we have made a fine definition of continuity is provided by the next result. Figure 7.2 may be suggestive.

Theorem 7.3 (Intermediate Value Theorem)

Suppose that $f : [a, b] \to \mathbb{R}$ is a continuous function on $I = [a, b]$. Suppose that $c \in (f(a), f(b))$. It follows that there is $x \in I$ such that $f(x) = c$.

Proof

Since $c \in (f(a), f(b))$ it follows that $f(a) < f(b)$. Let $T = \{y \mid y \in I, f(y) < c\}$. Notice that T is non-empty because $a \in T$. Also T is bounded above by b and so $x = \sup T$ exists. Moreover, since I is a closed interval, we have $x \in I$. Construct a sequence in T inductively. Choose $x_i \in T$ such that $x_i \in (x - 1/i, x]$. This can be done because $x - 1/i$ is not an upper bound for T. By design, (x_i) converges to x. Using Proposition 7.3 we conclude that $(f(x_i))$ converges to $f(x) = w$ and $f(x_i) < c$ for all $i \in \mathbb{N}$. It cannot be that $w > c$ or else $|w - f(x_i)| \geq w - c \, \forall i \in \mathbb{N}$ which is absurd. Thus $w \leq c$.

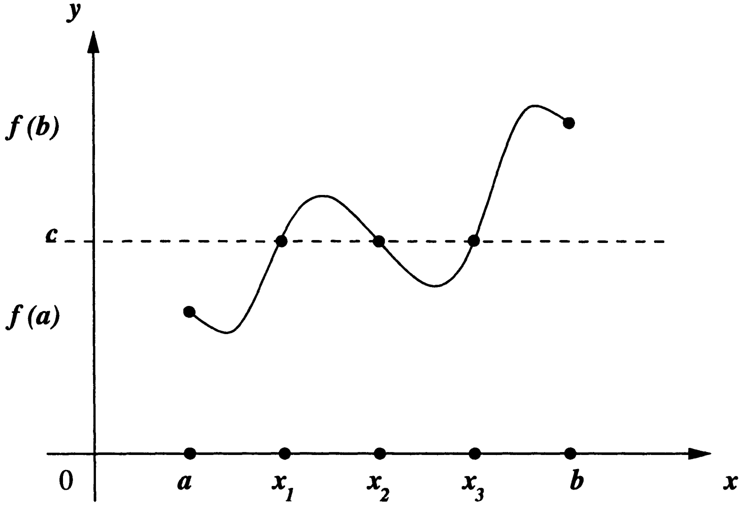

Fig. 7.2. A picture of the Intermediate Value Theorem

Notice that $w \leq c < f(b)$ so $x \neq b$. Now define another sequence inductively. Choose $z_i \in (x, b]$ such that $|x - z_i| < 1/i$. Now (z_i) converges to x and so $(f(z_i))$ converges to $f(x) = w$. Now $f(z_i) \geq c \forall i \in \mathbb{N}$ so by an argument similar to that in the previous paragraph, $w = f(x) = \lim f(z_i) \geq c$.

We have now proved that $c \leq w \leq c$ so $f(x) = w = c$ as required.

□

7.2 Limits

Sometimes you want to investigate "functions" which are not defined at a particular point. For example, the formula $(\sin x)/x$ does not define a map from $\mathbb{R}$ to $\mathbb{R}$. You can't evaluate the formula at $x = 0$ because $0/0$ is not defined. Alternatively, you might be interested in what happens to $(x^3 - 6x + 2)/(2x^3 - 1)$ as x gets arbitrarily large or arbitrarily negative, and you can not simply evaluate at $\pm\infty$.

We need proper definitions, otherwise we do not know what we are talking about. In this next definition, we need to talk about a *punctured* open interval. This means an open interval with a point removed. Similarly one can talk about a punctured plane, which is $\mathbb{R}^2$ with a point removed.

Definition 7.4

Suppose that $a < b < c$ are real numbers, and $f : (a,c) \setminus \{b\} \to \mathbb{R}$. Thus f is a function from a punctured open interval to $\mathbb{R}$. The puncture is at b, and we do not care whether f is not defined at b, or f was defined at b but we are ignoring the fact. Even if $f(b)$ exists, it is irrelevant to this definition. We say that $f(x)$ tends to the limit l as x tends to b if the following condition holds.

$$\forall \varepsilon > 0 \, \exists \delta > 0 \text{ such that if } x \in (a,c) \text{ and } 0 < |x-b| < \delta, \text{ then } |f(x) - l| < \varepsilon.$$

In words then, for every $\varepsilon > 0$ (no matter how small, and the smaller ε is, the tougher things will be), we can always find $\delta > 0$ (and we are allowed to know what ε is before we choose δ), so that if x is in the domain but is not b, and the distance between b and x is smaller than δ, then the distance between $f(x)$ and l is less than ε.

Remark 7.1

We write $f(x) \to l$ as $x \to b$ or equivalently $\lim_{x\to b} f(x) = l$.

You can then develop a theory of limits along the lines of the theory of limits of sequences as we did in Chapter 6.

EXERCISES

7.4 In the notation we have just established, show that if both $f(x) \to l$ as $x \to a$ and $f(x) \to m$ as $x \to a$, then $l = m$.

7.5 Suppose that we have two functions f_1, f_2, both from the punctured (a,c) to $\mathbb{R}$, the puncture being at b in both cases. Suppose that as $x \to b$ we have $f_1(x) \to l_1$ and $f_2(x) \to l_2$.

(a) Let g be the sum of these functions, so $g(x) = f_1(x) + f_2(x)$. Show that $g(x) \to l_1 + l_2$ as $x \to b$.

(b) Let h be the product of these functions, so $h(x) = f_1(x)f_2(x)$. Show that $h(x) \to l_1 l_2$ as $x \to b$.

Now for the very beginning of the theory of differentiation.

Let $f : \mathbb{R} \to \mathbb{R}$ be a function, and suppose that $a \in \mathbb{R}$. Suppose that f is continuous at a. Thus as x gets close to a then $f(x)$ gets close to $f(a)$. Differentiation can be viewed as analysing the relationship between these two approaches. Does $f(x)$ approach $f(a)$ at exactly the same rate that x approaches

a? Does it matter if x approaches a from above or below (right or left in the usual picture)? It might be that as x approaches a at a steady rate, $f(x)$ enjoys a wild and jerky path approaching $f(a)$.

We measure the approach by looking at the ratio

$$\frac{f(x) - f(a)}{x - a}$$

as x gets closer and closer to a. This quantity has geometric meaning. It is the slope of the chord joing $(a, f(a))$ to $(x, f(x))$ in the graph of f. There is a picture of this in Figure 7.3.

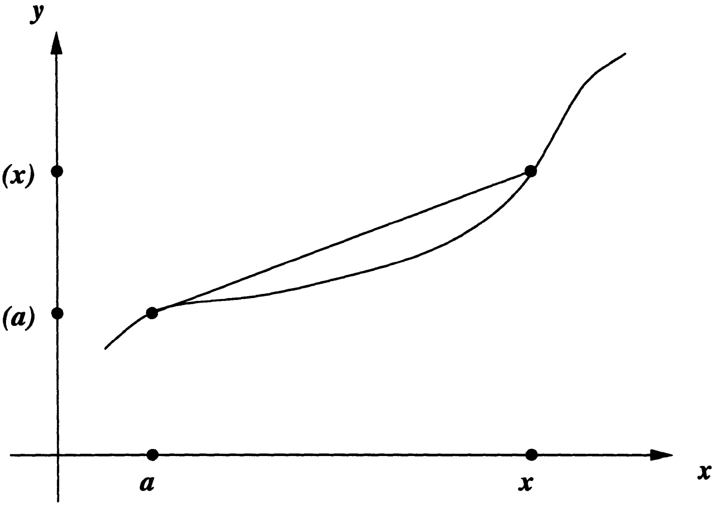

Fig. 7.3. Genesis of differentiation

There is no point in trying to calculate this ratio when $x = a$ because you are faced with $0/0$ which is not defined. However, we have introduced exactly the right tool in the form of a limit.

Definition 7.5

Suppose that $f(x)$ is defined in an open interval surrounding a, and that

$$\lim_{x \to a} \frac{f(x) - f(a)}{x - a} = l,$$

then we say that f is differentiable at a, and l is called the derivative of f at a.

8
Creating the Real Numbers

In this final chapter we give, in kit form only, two sound constructions of the real numbers $\mathbb{R}$. This takes us back to the spirit of Chapter 1. Recall that we can build $\mathbb{N}$ from $\emptyset$, and that it is an easy step to build $\mathbb{Z}$ from $\mathbb{N}$. If you have done all the exercises, you will know how to build $\mathbb{Q}$ from $\mathbb{Z}$. We also showed how to construct $\mathbb{C}$ from $\mathbb{R}$ in Chapter 3. The glaring omission is the construction of $\mathbb{R}$. So far we have relied on our intuition about $\mathbb{R}$. If you want to be sure about the existence and nature of $\mathbb{R}$, read on.

The first construction, due to Dedekind, is both quick and easy to understand, and uses the usual ordering of $\mathbb{Q}$ as its main ingredient. The second method, which uses devices called Cauchy sequences already introduced in Chapter 6, is a little more difficult conceptually. However, it wins in the long run because the construction generalizes to many more situations. We give an example of this in Section 8.3.

8.1 Dedekind's Construction

Definition 8.1

A *Dedekind cut* or *cut* C of $\mathbb{Q}$ consists of a partition of $\mathbb{Q}$ into two non-empty subsets C_- and C_+. These sets are required to have two properties.

G. Smith, *Introductory Mathematics: Algebra and Analysis*

(i) If $x \in C_-$ and $y \in C_+$, then $x < y$.

(ii) C_- has no maximum.

What we would like to do is to let each real number correspond to one of these cuts. The square root of 2 would correspond to (or be!) the cut A_-, A_+ consisting of

$$A_- = \{q \mid q \in \mathbb{Q}, q \leq 0\} \cup \{q \mid q \in \mathbb{Q}, q > 0, q^2 < 2\}.$$

and

$$A_+ = \{q \mid q \in \mathbb{Q}, q > 0, q^2 \geq 2\}.$$

The rational number t corresponds to a cut B_-, B_+ where B_+ has a minimal element. Specifically we have

$$B_- = \{q \mid q \in \mathbb{Q}, q < t\} \text{ and } B_+ = \{q \mid q \in \mathbb{Q}, q \geq t\}.$$

We need to be able to do everything with cuts that we want to do with real numbers. Addition and multiplication must be available, just to start with. Then we need to check that all field axioms are satisfied, and that the order $<$ which we want to use in the real numbers can be captured in some way by these cuts. We have to show that the ordering interacts with the algebraic laws in exactly the way we want. This is a great pile of routine and uninspiring (but important) work. If you want to be *really sure* that the real numbers exist, you have to run away and do all that. Of course, some very careful people have done it already, but if you have the appropriate psychology, you may have no choice but to do it yourself.

Addition is easy. If C_-, C_+ and D_-, D_+ are two cuts, define their sum to be E_-, E_+ where

$$E_- = \{x + y \mid x \in C_-,\ y \in D_-\} \text{ and } E_+ = \{x + y \mid x \in C_+,\ y \in D_+\}$$

Additive inverses are a little delicate. Suppose that A_-, A_+ is a cut. You can consider $-A_+, -A_-$ where $-X = \{-x \mid x \in X\}$. This may not quite be a cut, because it is possible that $-A_+$ has a maximal element. If so, let it migrate over to $-A_-$. When this is done, you have the additive inverse of A_-, A_+.

Multiplication of C_-, C_+ and D_-, D_+ is a little more delicate. You must not simply multiply the sets together, element by element, to form (as it were) C_-D_- and C_+D_+, because these sets will not be disjoint (why?). In this notation $XY = \{xy \mid x \in X \text{ and } y \in Y\}$.

If either C_+ or D_+ contains no negative rationals, then define the product of C_-, C_+ and D_-, D_+ to be E_-, E_+ where $E_+ = C_+D_+$ and $E_- = \mathbb{Q} \setminus E_+$. In the event that both C_+ or D_+ contain negative rationals, define the product of C_-, C_+ and D_-, D_+ to be the product of their additive inverses.

We have an ordering via $A_-, A_+ \leq B_-, B_+$ if and only if $A_- \subseteq B_-$. It is a purely mechanical procedure to check that the field axioms work for the cuts, and that $<$ interacts correctly with the algebra.

The completeness axiom (what this game is all about) is fine. Suppose that $\{A(i)_-, A(i)_+ \mid i \in I\}$ is a non-empty collection of cuts. To say that this collection is bounded above means that there is a cut X_-, X_+ with $A(i)_- \subseteq X_- \, \forall i \in I$. Now the supremum S_-, S_+ of our collection will be defined by $S_- = \cup_{i \in I} A(i)_{\subseteq} X_-$, and $S_+ = \mathbb{Q} \setminus S_-$. Notice that S_- is a union of a non-empty collection of non-empty sets and so is non-empty, and $S_- \subseteq X_-$ so $S_- \neq \mathbb{Q}$ since $X_+ \neq \emptyset$ and $S_- \cap X_+ = \emptyset$.

8.2 Construction via Cauchy Sequences

Now, throw away Dedekind's construction, and start again. We can't talk about sequences of rational numbers converging to $\sqrt{2}$ because $\sqrt{2}$ doesn't exist yet! However, Pythagoras's Theorem requires $\sqrt{2}$ to exist in any number system capable of measuring length in the plane. Now, measuring lengths in the plane is a fundamental activity, and if we can't even do that, we might as well give up mathematics. This highlights the inadequacy of $\mathbb{Q}$, and makes it imperative that we construct a number system which is up to the task.

So, we are temporarily back in the mathematical dark ages before anyone has constructed the real numbers. All we have is $\mathbb{Q}$. From this point of view, a sequence of rationals either converges to a rational number, or it doesn't converge at all. Our problem is to somehow get round the problem that we want sequences to converge to limits which don't yet exist, and then define these limits to be the real numbers, which will then exist. Well, that looks very doubtful, and is not quite what we do, but it is close. We remind the reader of a definition given in Chapter 6.

Definition 8.2

A sequence (a_n) of rational numbers is called a *Cauchy sequence* exactly when the following condition is satisfied: for any $\varepsilon > 0$ there is $N \in \mathbb{N}$ such that if $n, m \in \mathbb{N}$ and both $n, m \geq N$, then $|a_n - a_m| < \varepsilon$.

EXERCISES

8.1 Show that the sum and product of Cauchy sequences are Cauchy sequences.

Notice that any convergent sequence of rationals is a Cauchy sequence. We have a (rational) limit l, and given any ε we can choose N such that if $n, m \geq N$, then $|a_n - l| < \varepsilon/2$. In turn, this ensures that

$$|a_n - a_m| = |a_n - l + l - a_m| \leq |a_n - l| + |a_m - l| < \varepsilon/2 + \varepsilon/2 = \varepsilon.$$

The plan is to define an equivalence relation on our Cauchy sequences, and these equivalence classes will be our newly minted real numbers. We first need a definition on the way to defining the equivalence relation.

Definition 8.3

A *null sequence* is a sequence of rational numbers which converges to 0.

Now for the equivalence relation. Suppose that (a_n) and (b_n) are Cauchy sequences of rationals. We write $(a_n) \sim (b_n)$ if and only if $(a_n - b_n)$ is a null sequence. Thus $(a_n) \sim (b_n)$ means that the terms of the two sequences are ultimately close together. Now, there is checking to be done here. Reflexivity and symmetry are trivial, and transitivity comes easily enough with a little help from the triangle inequality. Details are omitted.

Let $\mathbb{R}$ be the set consisting of all the equivalence classes. We now need to define various operations. This is routine and dull, much as in the method that uses Dedekind cuts. However, let us sketch what must be done.

First the algebra; given two real numbers (equivalence classes) x and y, choose any rational Cauchy sequence $(a_i) \in x$ and any rational Cauchy sequence $(b_i) \in y$. Form the Cauchy sequence $(a_i + b_i)$, and then define $x + y$ to be the equivalence class of $(a_i + b_i)$. Define products in a similar manner. There is one important detail that must be verified. We have, as it were, intruded on the privacy of x and y by making selections of $(a_i) \in x$ and $(b_i) \in y$. We have to worry that if we were to make a different selection, and follow our procedure, then we might end up with a different value of $x+y$. Well, by careful bookkeeping, you can show that the intrusion has not caused any trouble, and $x + y$ is independent of the choices of (a_i) and (b_i).

There is also a technical difficulty with multiplicative inverses. Suppose that (a_i) is a Cauchy sequence in the equivalence class $x \neq 0$. We want to let x^{-1} be the equivalence class that contains the Cauchy sequence (a_i^{-1}). There are two problems: if some number a_j is 0, then a_j^{-1} does not exist. Even if this problem is solved, there is still the matter of showing that the sequence (a_i^{-1}) is a Cauchy sequence. The first problem is solved by observing that as $x \neq 0$, we do not have $x \sim (0)$ where (0) is the zero sequence. After a little argument it follows that in the sequence (a_i), the entry 0 can occur only finitely many times. The trick is to throw away finitely many (but enough) terms from the start of the sequence to get rid of all zero entries. You have to check that the resulting pruned sequence is in the same equivalence class of course, but that is routine, and so is the verification that the sequence of the inverses of its terms is a Cauchy sequence.

We need to define an ordering on our newborn reals. Suppose then that x, y are equivalence classes of Cauchy sequences. Choose $(a_i) \in x$ and $(b_i) \in y$. Write $x < y$ if and only if $a_i < b_i$ for all except finitely many values of i.

One should now go in for deep accountancy, and verify all the axioms and interactions.

EXERCISES

8.2 Show that every real number has a unique decimal representation, provided we stipulate that representations ending in infinitely many 9s are not permitted (else $0.\overline{9} = 1.0$ and so on).

8.3 Show that every real number has a unique "binary equivalent of a decimal" expansion. Thus every real number can be uniquely represented as

$$\pm \sum_{i=k}^{\infty} a_i 2^{-i}$$

for some integer k, with each a_i being 0 or 1, and (a_i) is not ultimately the constant sequence (1).

8.4 (J. F. Toland) By mixing up the ideas in the previous two questions, and those in the discussion following Proposition 6.10 (or otherwise), construct a function $f : \mathbb{R} \to \mathbb{R}$ which has the property that it assumes every real value infinitely often on each interval of positive length in the domain. Conclude that a function can satisfy the (intermediate value) Theorem 7.3 without being continuous.

8.3 A Sting in the Tail: p-adic numbers

By now you have done a lot of reading, and many exercises, and perhaps it is time we had a conversation. "Did you ever want to write

$$1 + 2 + 2^2 + 2^3 + \ldots = -1?$$

After all

$$\frac{1}{1-x} = \sum_{i=0}^{\infty} x^i \tag{8.1}$$

so putting $x = 2$ in Equation (8.1) yields the result."

"Hang on", I hear you cry, "you can't do that because the sum in Equation (8.1) only converges when $|x| < 1$."

"That's all right then", I reply, "because $|2| < 1$."

At this point I imagine that an uneasy silence fills the room, and you are probably wondering if you should have been reading this book after all. After a little time, I helpfully add "But what I mean by $|2|$ is $1/2$."

We make a new definition of $|q|$ for $q \in \mathbb{Q} \setminus \{0\}$. Now $q = 2^i(a/b)$ where a and b are odd integers and $i \in BbbZ$. Since what follows is not the same as the old definition of modulus, we should not really use exactly the same name notation. We will use a subscript 2. Define $|2^i(a/b)|_2$ to be 2^{-i}. In addition, we define $|0|_2$ to be 0. Now we isolate some of the features of this new function from $\mathbb{Q}$ to $\mathbb{Q}$.

Proposition 8.1

The new modulus function satisfies the following conditions for all $x, y \in \mathbb{Q}$.

(a) $|x|_2 \geq 0$, and $|x|_2 = 0$ exactly when $x = 0$.

(b) $|xy|_2 = |x|_2|y|_2$.

(c) $|x + y|_2 \leq \max\{|x|_2, |y|_2\}$

(d) $|x + y|_2 \leq |x|_2 + |y|_2$.

Proof

Parts (a) and (b) are easy exercises, left to the reader, and part (d) follows immediately from parts (a) and (c). It remains to discuss part (c).

Suppose that $x = 2^i a/b$ and $y = 2^j c/d$ with i, j integers, a, b, c, d all odd integers and $b, d > 0$. Suppose, without loss of generality, that $|x|_2 \leq |y|_2$, so $i \geq j$. Now $|x + y|_2 = |2^j(2^{i-j}ad + bc)/bd)|_2$. If $i > j$ then $2^{i-j}ad + bc$ is odd so $|x+y|_2 = 2^{-j} = \max\{|x|_2, |y|_2\}$. On the other hand, if $i = j$ then $ad + bc$ is even and so then $|x+y|_2 < 2^{-j} = \max\{|x|_2, |y|_2\}$. □

We now give $\mathbb{Q}$ a new "geometry". Well, you will have to be a little generous to call it a geometry, but what we are going to do is to define the distance between each pair of rational numbers x and y not to be $|x-y|$, but instead to be $|x-y|_2$. This is an interesting "geometry" to work in. The triangle inequality holds because

$$|x+y|_2 \leq \max\{|x|_2, |y|_2\} \leq |x|_2 + |y|_2,$$

as do many of the algebraic properties of the ordinary modulus function. In this strange 'geometry', which is not easy to visualize, all triangles are isosceles, thanks to the proof of part (c). To be explicit, if $a, b, c \in \mathbb{Q}$ let $a - b = x$, $b - c = y$. The proof of Proposition 8.1 shows that either $|x|_2 = |y|_2$ (and Δabc is isosceles), or $|a-c|_2 = \max\{|x|_2, |y|_2\}$ (and Δabc is isosceles).

In this geometry 1024 is very close to 0, because $1024 = 2^{10}$ so $|1024 - 0|_2 = 2^{-10}$, but 1023 is distance 1 from 0. The distances between points are completely unrelated to the usual distance when you think of $\mathbb{Q}$ as embedded in $\mathbb{R}$, and identify $\mathbb{R}$ with the real line.

It gets better. The function $|\cdot|_2$ has sufficiently many properties in common with the ordinary modulus function $|\cdot|$ that one can build equivalence classes of Cauchy sequences, and construct a number system $\mathbb{Q}_2$ (the 2-adic numbers) in just the same way that we constructed $\mathbb{R}$. In fact there is nothing special about 2; given any prime number p you can build the p-adic numbers $\mathbb{Q}_p$. All you have to do is to begin by writing each non-zero rational as $p^i a/b$ where $i, a \in \mathbb{Z}$, $b \in \mathbb{N}$ and p divides neither a nor b. Define $|p^i a/b|_p = p^{-i}$ and follow your nose. You can then go on to study sequences and series in $\mathbb{Q}_p$. In $\mathbb{Q}_p$ we have $\sum_{i=0}^{\infty} p^i = (1-p)^{-1}$. In fact $\mathbb{Q}_p$ is easier to work in than $\mathbb{R}$. For example, any sequence in $\mathbb{Q}_p$ which converges to 0 can be turned into a convergent series by insinuating plus signs. Compare this with $\mathbb{R}$ where the sequence $(1/n)$ converges to 0, but $\sum 1/n$ does not converge.

These fields $\mathbb{Q}_p$ are not just accidents caused by a lucky definition of $|x|_p$. They play an important rôle in the theory of numbers.

Further Reading

In all cases, the editions mentioned are the cheapest currently available (1997) according to their list price. Many of these books are therefore the paperback editions of books also available in hard cover. Moreover, some books may come back into print in the future. Watch the book web site for updated information.

First, we mention one non-mathematical book in this text. That is Bertrand Russell's *History of Western Philosophy.* This book is a lot funnier than its title might suggest. A recent edition was published by Routledge in 1993 (ISBN 0415078547). This large volume contains information on various mathematician-philosophers, including the school of Pythagoras.

Another non-specialist book which definitely merits attention is *What is mathematics?* by R. Courant and H. Robbins. Oxford Universitry Press published an edition revised by I. Stewart in 1996 (ISBN 0195105192). This is a wide-ranging exploration of mathematical ideas.

An inspiring but demanding undergraduate algebra text is I. N. Herstein's *Topics in Algebra*, Wiley, 1975 (ISBN 0471010901). This marvellous book does assume that you are very interested. A more accessible text is J. B. Fraleigh's *First course in abstract algebra*, the 5th revised edition was published by Addison-Wesley in 1994 (ISBN 0201592916).

There are many linear algebra texts available. The classic text for the good student who enjoys pure mathematics is *Finite Dimensional Vector Spaces* by P. R. Halmos published by Springer UTM (ISBN 0387 90093 4). This is not a book for the faint-hearted, and the gothic lettering of the vector spaces can be a little intimidating at first. There are many other linear algebra texts suitable for students with a wide range of abilities. For example, a very accessible text for those whose tastes are not so abstract is T. S. Blyth and E. F. Robertson's *Basic Linear Algebra* published by Springer in 1998 (ISBN 3540761225).

For analysis, an excellent starting book is J. A. Green's *Sequences and Series* currently out of print. It was published by Routledge and Kegan Paul. It was once very common, so libraries may have it. You may wish to pursue the topic through one of the many modern texts. For example W. Y. Hsiang's *A Concise Introduction to Calculus* was published by World Scientific 1995 (ISBN 9810219016). M. Spivak's *Calculus* has been a modern standard, but is (amazingly) out of print. However, the world is full of Spivak's Calculus, and if you look in a second hand bookshop near a university, you should be able to find a copy (with luck). A strong student may wish to look at T. Apostol's *Mathematical Analysis*, Addison-Wesley, 1974 (ISBN 0201002884). The truly exceptional student may wish to read *Principles of Mathematical Analysis* by W. Rudin, published by McGraw in 1976 (ISBN 0070856133). This book takes no prisoners.

For those students interested in the set theoretic foundations of mathematics, an obvious starting point is *Naive Set Theory* by P. Halmos, published by Springer-Verlag. The most recent edition came out in 1994 (ISBN 0387900926). However, from a completely different perspective, in order to understand how to construct a proof, one cannot do better than D. L. Johnson's *Elements of Logic via Numbers and Sets* which is a Springer SUMS book published in 1998 (ISBN 3540761233).

Solutions

Chapter 1

1.1 (a) B, (b) A, (c) E, (d) B, (e) G, (f) H, (g) G, (h) H, (j) A, (k) H, (l) A, (m) C.

1.2 (a) (iii), (b) (iii), (c) (ii), (d) (i), (e) (ii), (f) (i), (g) (i), (h) (iii), (j) (iii), (k) (i), (l) (i), (m) (ii).

1.3 (a) $A \cap B \cap C$, $A' \cap B \cap C$, $A \cap B' \cap C$, $A \cap B \cap C'$, $A' \cap B' \cap C$, $A \cap B' \cap C'$, $A' \cap B \cap C'$, $A' \cap B' \cap C'$.
(b) Use De Morgan's laws (generalized to apply to three sets, and even to arbitrary collections of sets) on the answers to part (a). For example, $A \cap B \cap C = (A \cap B \cap C)'' = (A' \cup B' \cup C')'$. The others are entirely similar.

1.4 (a) This is not a proof, but an instruction on how to perform the proof (to save paper). Take each element of $\{1, 2, 3\}$ and check that it is an element of $\{1, 1, 2, 3\}$. Next take each element of $\{1, 1, 2, 3\}$ and check that it is an element of $\{1, 2, 3\}$. When you have done this, you have verified that $\{1, 2, 3\} = \{1, 1, 2, 3\}$. (b) $2^3 = 8$. (c) 1 (since $P(\emptyset) = \{\emptyset\}$). (d) 2 (since $P(\{\emptyset\}) = \{\emptyset, \{\emptyset\}\}$). (e) 2^{65536} (Don't try to write it out!).

1.5 (a) Pictures omitted. (b) Claim: For all X, Y subsets of a universe U we have that $(X \cup Y)' = X' \cap Y'$. First we prove $(X \cup Y)' \subseteq X' \cap Y'$. Choose $a \in (X \cup Y)'$ so $a \notin X$ and $a \notin Y$. Thus $a \in X' \cap Y'$. The reverse inclusion is similar, as is the other law of De Morgan. (c) Let $A = \{2x \mid x \in \mathbb{N}\}$ and $B = \mathbb{N} \setminus A$. There are many other examples.

1.6 (a) $n2^{n-1}$ since there are n choices for a, and 2^{n-1} choices for S, since we can choose the set $S \setminus \{a\}$ in 2^{n-1} ways. (b) Same as part (a). (c) 3^n because you wish to count the ways in which you can build S, T and $A \setminus (S \cup T)$ with S and T disjoint. You have three choices as to where each of the n elements are placed, and so 3^n possibilities altogether. Try small n to convince yourself. (d) 3^n. This follows from part (c) by looking at S' and T' in the universe A.

1.7 (a) $(f \circ g)(x) = x^2 + 2x + 1 \,\forall x \in \mathbb{Z}$. $(g \circ f)(x) = x^2 + 1 \ \forall x \in \mathbb{Z}$. (b) $f^n(x) = x^{2^n} \ \forall x \in \mathbb{Z}$. Note that this means $x^{(2^n)}$ and not $(x^2)^n$. $g^n(x) = x + n \,\forall x \in \mathbb{Z}$. (c) g and g^2 are bijections.

1.8 There are many possible answers. (a) $f(x) = 2x \,\forall x \in \mathbb{N}$. (b) $f(1) = 1$ and $f(x) = x - 1$ if $x \in \mathbb{N}$ and $x > 1$. (c) We define maps f and g as follows.

$f(1) = 2$, $f(2) = 1$, and $f(i) = i$ if $i \geq 3$. $g(1) = 1$, $g(2) = 3$, $g(3) = 2$ and $g(i) = i$ if $t \geq 4$. Now $(f \circ g)(1) = 2 \neq 3 = (g \circ f)(1)$, so $f \circ g \neq g \circ f$.

1.9 There are many answers. (a) Define f by $f(1) = 0$, $f(2n) = n \forall n \in \mathbb{N}$ and $f(2n+1) = -n \forall n \in \mathbb{N}$. (b) Define g by $g(x) = x - 1$ if $x \in \mathbb{N}$, and $g(x) = x$ if $x \in \mathbb{Z} \setminus \mathbb{N}$.

1.10 (a) Yes it does follow. For if there is $x \in A$ with $\beta(x) \neq \gamma(x)$, then $\alpha(\beta(x)) \neq \alpha(\gamma(x))$ since α is injective. This is absurd because $\alpha \circ \beta = \alpha \circ \gamma$. Therefore $\beta(x) = \gamma(x)$ for every $x \in A$, so $\beta = \gamma$. This is an argument by contradiction (see Chapter 2). (b) No. Let $A = \mathbb{Z}$, let f be the identity map and g be defined by $g(x) = -x \forall x \in \mathbb{Z}$. (c) Yes. We know $f^2 = g^2$ and $f^3 = g^3$. Now $f^2 \circ f = g^2 \circ g = f^2 \circ g$. If $x \in A$, then $f^2(f(x)) = f^2(g(x))$. However, f^2 is the composition of two injections so is injective. Thus $f(x) = g(x)$. Now x was arbitrary so $f = g$. Note that we only needed f^2 injective. Surjectivity is irrelevant.

1.11 (a) n^n, (b) $n!$, (c) $n!$, (d) $n!$.

1.12 (a) T, (b) T, (c) T, (d) F, (e) T, (f) F, (g) T, (h) T, (j) F, (k) F, (l) F.

1.13 (a) Use associativity three times. (b) Let $U = \{1, 2, 3\}$. Let f be the identity map, let g swap 1 and 2 and fix 3. Let h swap 2 and 3 and fix 1. (c) Assume, for contradiction, that $\alpha \neq \mathrm{Id}_C$. Thus there is $c \in C$ with $\alpha(c) \neq c$. Define $\beta : C \to C$ by $\beta(x) = c \forall x \in C$. Now $(\alpha \circ \beta)(c) = \alpha(c)$ but $(\beta \circ \alpha)(c) = c$. Thus $\alpha \circ \beta \neq \beta \circ \alpha$. This is absurd, by hypothesis, so $\alpha = \mathrm{Id}_C$. (d) $|D| = 2$. You need $|D| \geq 2$ else you can't have a non-identity bijection from D to D. The situation does happen when $D = 2$ and there only two bijections to consider. When $D \geq 3$ no non-identity bijection from D to D commutes with all the others. In fact, for such D and non-identity bijection f, consider a pair of elements of D not swapped by f. Now consider the bijection which swaps this pair, and leaves everything else fixed. When you have finished considering, you are done.

1.14 Yes, you can do it. Make 26 adjacent copies of the word by applying f_1 5 times (to create 32 copies of the word) and then erasing 6 copies of the word using f_2 repeatedly. Next erase the first 25 letters until you have a z at the front. Move that z to the rear by duplicating the whole current word and then deleting from the rear until you stop just before the z which is a copy of the one at the front. Now delete from the front until you reach the first y and so on. At the end you delete from the front to leave the alphabet reversed. Note that this method will enable you to generate any finite sequence of letters (including repetitions) from the original alphabet in the correct order.

1.15 (a) Suppose $b \in B$. Now $f(g(b)) = b$ so b is in the image of f. However, b was arbitrary so f is surjective. (b) Suppose that $a_1, a_2 \in A$ and $f(a_1) = f(a_2)$. Now $h(f(a_1)) = h(f(a_2))$ so $\mathrm{Id}_A(a_1) = \mathrm{Id}_A(a_2)$. Thus $a_1 = a_2$ and so f is injective. (c) Bijectivity is immediate.

$$h = h \circ \mathrm{Id}_B = h \circ (f \circ g) = (h \circ f) \circ g = \mathrm{Id}_A \circ g = g.$$

1.16 (a) $A \times A$ has cardinality n^2. A relation is a subset of $A \times A$, so is a subset of a set of size n^2. The power set of a set of size n^2 has cardinality 2^{n^2}. (b) 2^{n^2-n}. (c) $2^{(n^2+n)/2}$. (d) $2^{(n^2-n)/2}$.

1.17 (a) and (b) are routine. (c) is more interesting. The equivalence classes are the rational numbers. You have just built $\mathbb{Q}$ from $\mathbb{N}$ and $\mathbb{Z}$. Congratulations.

1.18 Many answers. RST use $=$; RS say $x \sim x \forall x$, and $0 \sim 1$, $1 \sim 0$, $1 \sim 2$, $2 \sim 1$ and nothing else; RT use $\leq$; ST say $a \sim b$ if and only if $ab \neq 0$ (notice that $0 \not\sim 0$); R say $x \sim x \forall x \in \mathbb{Z}$, $0 \sim 1, 1 \sim 2$ and nothing else; S say $a \sim b$ if and only if $a \neq b$; T say $<$; an example of a relation which is neither reflexive, symmetric nor transitive is $a \sim b$ if and only if $a - b$ is a prime number.

1.19 The "proof" breaks down at the stage "Choose any $b \in A$ such that $a \sim b$". You have no reason to suppose that any such b exists.

1.20 (a) R is reflexive if and only if R includes the line "$y = x$". (b) R is symmetric if and only if R is unchanged if you reflect it in the line "$y = x$". (c) Both of (a) and (b) hold simultaneously.

1.21 (a) There are five equivalence relations because there are 5 partitions of a set with 3 objects. The partitions can be the two trivial ones: three singleton sets, or the whole set. The non-trivial partitions consist of the singleton and a disjoint set containing two elements. There are 3 such partitions. (b) There are 15 equivalence classes on a set with 4 elements. The equivalence classes must have one of the following shapes: four singletons (1), one doubleton and two singletons (6), two doubletons (3), one triple and one singleton (4), and the whole set (1).

1.22 (a) Yes it is. Routine check. (b) A natural transversal is $I \times I$ where $I = \{r \mid r \in \mathbb{R},\ 0 \leq r < 1\}$. Join the top and bottom of I to form a tube, and bend the tube round to join the two circles as well. You have the surface of a doughnut (a torus in mathematical language). As a point moves around $\mathbb{R}^2$ in a continuous way, keep track of where its representative in the transversal goes. If you don't do this joining process, the dot may go off the top of the square $I \times I$ and reappear instantly on the bottom, or disappear from the left edge and reappear magically on the right. If you make the identifications as suggested, the representative in the transversal will move continuously on the torus without any disturbing jumps. (c) Any sphere centred at the origin (or even with the origin in its interior) will do nicely.

1.23 (a) $\emptyset = (2,1)$. (b) $\{1\} = [1,1]$. (c) If $(a,b) \neq \emptyset$ then $a < b$. Thus (a,b) contains at least two distinct points by Proposition 1.1. However $\{1\}$ contains exactly one point, so is neither empty nor infinite. Thus $\{1\}$ is not an open interval.

1.24 (a) Let I, J be intervals. Suppose that $a, b \in I \cap J$ and $a < x < b$. It follows that $x \in I$ since I is an interval, and similarly $x \in J$. Thus $x \in I \cap J$ so $I \cap J$ is an interval. (b) Let $I_1 = (a,b)$ and $I_2 = (c,d)$ then $I_1 \cap I_2 = (\max\{a,c\}, \min\{b,d\})$ is an open interval. (c) Take the answer to part (b) and use square brackets instead of round ones.

1.25 (a)

$$[0,1] = \cap_{i=1}^{\infty}(0 - 1/n, 1 + 1/n).$$

(b) No. If $x \in \cup_{\lambda \in \Lambda} I_\lambda$ where each I_λ is an open interval, then there is $\mu \in \Lambda$ such that $x \in I_\mu$. Now there is $\eta > 0$ so that $(x - \eta, x + \eta) \subseteq I_\mu$ and so $(x - \eta, x + \eta) \subseteq \cup_{\lambda \in \Lambda} I_\lambda$. However, if we apply this (for contradiction) to $[0,1] = \cup_{\lambda \in \Lambda} I_\lambda$ and put $x = 1$ we obtain an absurdity.

1.26 The child was right if the original ruler was either open $(0,1)$ or closed $[0,1]$. However, if the ruler was half-open, the teacher (unknowingly) had a point: $(0,1] = (0,1/2] \cup (1/2,1]$.

Chapter 3

3.1 (a) $-i$, (b) $1 + i$, (c) $2i$, (d) 16, (e) 16, (f) $-5 - 9i$, (g) $(-1 - i\sqrt{3})/2$, (h) 1, (j) 1, (k) $(1 - i)/2$, (l) 1, (m) $(5 + i)/13$.

3.2 (a) $i, -i$ (b) $\pm i\sqrt{2}$ (c) $\pm(1+i)/\sqrt{2}$ (d) $(-1 \pm i\sqrt{3})/2$ (e) $1, (-1 \pm i\sqrt{3})/2$ (f) $i, -5i$ (g) $-2i \pm 1$.

3.3 (a) By definition, $1 + (-1) = 0$. Multiply by (-1) and use Proposition 3.2. Thus $(-1) + (-1)(-1) = 0$. Add 1 to each side on the left, and use additive associativity and the fact that 0 is an additive identity. (b) From Proposition 3.2 we have $0 = 0 \cdot f = (1 + (-1)) \cdot f = f + (-1)f$. Add $-f$ to each side on the left, and use additive associativity, and the fact that 0 is an additive identity, so $-f = (-1)f$. (c) Use part (b) twice, and multiplicative associativity. (d) Use part (b), and both associativity and commutativity of multiplication to obtain $(-f)(-g) = (-1)^2(fg)$. Using part (a) $(-1)^2 = 1$, and 1 is a multiplicative identity so $(-f)(-g) = fg$. (e) Add $-g$ to each side. Use part (b) so $(f + (-1))g =$

0. Either $g = 0$ or g has a multiplicative inverse g^{-1}. In the latter event, multiply by g^{-1} on the right, use associativity of multiplication so $(f + (-1)) \cdot 1 = f + (-1) = 0$. Add 1 to each side on the right, and use associativity of addition, and finally the fact that 0 is an additive identity to yield that $f = 1$. (f) Very similar to part (e).

3.4 (a) Suppose f is an isometry (a distance-preserving map). Let $b = f(0)$ and $a = f(1) - f(0)$. $|1 - 0| = 1$ so $f(1) - f(0) = \pm 1$. Define a map h by $h(x) = ax + b$. Check that h preserves distances. Also h agrees with f at 0 and 1, so $h = f$. Conversely, we have already checked that a map h given by such a linear formula preserves distances. (b) Same as part (a).

3.5 There are reflections in straight lines through the origins, and rotations about the origin. Use geometrical arguments.

3.6 (a) Expand $(a + b)\overline{(a + b)}$ in two ways. (b) Replace b by $-b$ in part (a), and add this equation to the old part (a) equation. (c) The sum of the squares of the diagonals of a parallelogram is equal to the sum of the squares of the four sides of the parallelogram.

3.7 (a) $m = a^2 + b^2$ and $n = c^2 + d^2$ for integers a, b, c, d. Thus $m = |a + ib|^2$ and $n = |c + id|^2$. Now $mn = |(a + ib)(c + id)|^2$ is the sum of the squares of the real and imaginary parts of $(a + ib)(c + id)(= ac - bd + i(ad + bc))$, and so is the sum of two perfect squares. (b) $97 = 9^2 + 4^2$ and $1000001 = 1000^2 + 1^2$. Use part (a) to discover that $97,000,097 = 8996^2 + 4009^2$. (c) If z is a complex number, then $\mathrm{Re}(z) \leq |z|$. Now let $z = (a + ib)(c - id)$.

3.8 (a) $\cos 0 = 1$, $\cos \pi/2 = 0$, $\cos \pi/3 = 1/2$, $\cos \pi = -1$. Also $\cos 4 \cdot 0 = 1$, $\cos 4 \cdot \pi/2 = \cos 2\pi = 1$, $\cos 4 \cdot \pi/3 = \cos 4\pi/3 = -1/2$, and $\cos 4\pi = 1$. Plug in the numbers. (b) Take the real parts. Thus $\cos 4\theta = \cos^4\theta - 6\cos^2\theta \sin^2\theta + \sin^4\theta$ $= \cos^4\theta - 6\cos^2\theta + 6\cos^4\theta + (1 - \cos^2\theta)^2 = 8\cos^4\theta - 8\cos^2\theta + 1$.

3.9 (all parts) It is a saw tooth; 0 at integer values x, then climbing to the value μ at $x + \mu$ for μ in the range $0 \leq \mu < 1$. This function has period p for all $p \in \mathbb{N}$. The fundamental period is 1.

3.10 (all parts) Ξ has all positive rationals as periods, and no others. This is easy to check. There is no shortest period so no fundamental period.

3.11 (a) $\forall x \in \mathbb{R}$ we have $f(x) = f(-x)$ and $f(x) = -f(-x)$. Thus $f(x) = 0 \, \forall x \in \mathbb{R}$. (b) $g(-x) = g(x) \, \forall x \in \mathbb{R}$ so $g(x)$ is even. (c) $\mathrm{even}(x) = (f(x) + f(-x))/2$. $\mathrm{odd}(x) = (f(x) - f(-x))/2$. By design, these are even and odd functions (respectively), and $f(x) = \mathrm{even}(x) + \mathrm{odd}(x)$. (d) Suppose $f = e_1 + o_1 = e_2 + o_2$ are two rival expressions of f as sums of even and an odd functions in the obvious notation. Now $e_1 - e_2 = o_2 - o_1$ is both an even and odd function, so by part (a) is 0. Thus $e_1 = e_2$ and $o_1 = o_2$. (e) If f is even, then its even part is f and its odd part is 0. (f) The even part of $p(x)$ is $-x^4 + 3x^2 + \sin(-x^2) + \cos(x^3)$. The odd part is $2x^5$.

3.12 $\cos\theta = (e^{i\theta} + e^{-i\theta})/2$. Similarly for $\cos 4\theta$. Substitute in.

3.13 Let $w = e^{iz}$, then w satisfies $w^2 - 4iw - 1 = 0$. Thus $w = i(2 \pm \sqrt{3})$. Let $z = a + ib$ for real a and b so $e^{-b}e^{ia} = (2 + \sqrt{3})e^{i\pi/2}$ or $|2 - \sqrt{3}|e^{i\pi/2}$. Thus $b = -\log(2 \pm \sqrt{3})$ and $a = \pi/2 + 2k\pi$.

3.14 $2\sinh x \cosh x = 2(e^x + e^{-x})(e^x - e^{-x})/4 = (e^{2x} - e^{-2x})/2 = \sinh 2x$. Next we can differentiate this equation, or verify by direct calculation, that $\cosh 2x = \cosh^2 x + \sinh^2 x$. Finally $\tanh 2x = \sinh 2x/\cosh 2x = 2\tanh x/(1 + \tanh^2 x)$.

3.15 For all $x \in \mathbb{R}$ we have $-\cosh x = (-e^x - e^{-x})/2 < (e^x - e^{-x})/2 = \sinh x < (e^x + e^{-x})/2 = \cosh x$. Divide through by the positive quantity $\cosh x$ and you are done.

3.16 We seek all $z \in \mathbb{C}$ such that $(w_1 w_2)^z = w_1^z w_2^z$ for all $w_1, w_2 \in \mathbb{C}$. There is only a problem when $w_1 \neq 0 \neq w_2$. Let $w_j = e^{u_j} e^{i\theta_j}$ for $j = 1, 2$ in standard form, so $\theta_j \in (-\pi, \pi]$ for $j = 1, 2$. When $\theta_1 + \theta_2 \notin (-\pi, \pi]$ we will have that $\theta_1 + \theta_2$ differs from an element of $(-\pi, \pi]$ by 2π. For our equation to always be valid, we need that $e^{2\pi i z} = 1$. This happens if and only if $z \in \mathbb{Z}$.

Chapter 4

4.1 (4.4) $||\mathbf{x}||^2 = \sum_i x_i^2 = \langle \mathbf{x}, \mathbf{x} \rangle$. (4.5) $||\lambda \mathbf{x}|| = \sqrt{\sum_i \lambda_i^2 x_i^2} = |\lambda| \cdot ||\mathbf{x}||$. (4.6) $\langle \mathbf{x}, \mathbf{y} \rangle = \sum_i x_i y_i = \sum_i y_i x_i = \langle \mathbf{y}, \mathbf{x} \rangle$. (4.7) $\langle \mathbf{x}+\mathbf{y}, \mathbf{z} \rangle = \sum_i (x_i + y_i) z_i = \sum_i (x_i z_i + y_i z_i) = \sum_i x_i z_i + \sum_i y_i z_i = \langle \mathbf{x}, \mathbf{z} \rangle + \langle \mathbf{y}, \mathbf{z} \rangle$. (4.8) $\langle \mathbf{x}, \mathbf{y} + \mathbf{z} \rangle = \langle \mathbf{x}, \mathbf{y} \rangle + \langle \mathbf{x}, \mathbf{z} \rangle$ by an argument similar to the previous answer. (4.9) $\lambda \langle \mathbf{x}, \mathbf{y} \rangle = \sum_i \lambda x_i y_i = \langle \lambda \mathbf{x}, \mathbf{y} \rangle$. (4.10) $\lambda \langle \mathbf{x}, \mathbf{y} \rangle = \langle \mathbf{x}, \lambda \mathbf{y} \rangle$ by an argument similar to the previous answer.

4.2 It is enough to do the problem when $\mathbf{u}, \mathbf{v}$ are unit vectors (i.e. have length 1). The reader should construct the geometrical picture from the instructions which follow. Let the angle between $\mathbf{u}$ and $\mathbf{v}$ be θ. Let $\mathbf{i}, \mathbf{j}$ and $\mathbf{k}$ be unit vectors in the direction of the co-ordinate axes. Observe that $\mathbf{u} = \cos \alpha \mathbf{i} + \cos \beta \mathbf{j} + \cos \gamma \mathbf{k}$ where α, β and γ are the angles between $\mathbf{u}$ and the axes. This is just a matter of dropping perpendiculars, and reading the result off from the right-angled triangles. Drop a perpendicular from the tip of the position vector $\mathbf{u}$ to the line through the position vector $\mathbf{v}$. The side of this triangle in the direction of $\mathbf{v}$ will have length $\cos \theta$. Let $\mathbf{v}$ make angles α', β' and γ' with the axes. If perpendiculars are dropped to the line through $\mathbf{v}$ from the tips of the position vectors $\cos \alpha \mathbf{i}, \cos \beta \mathbf{j}$ and $\cos \gamma \mathbf{k}$, the resulting triangles will have sides in the direction of $\mathbf{v}$ of lengths $\cos \alpha \cos \alpha'$, $\cos \beta \cos \beta'$ and $\cos \gamma \cos \gamma'$. From the picture, which you should draw, it follows that $\cos \alpha \cos \alpha' + \cos \beta \cos \beta' + \cos \gamma \cos \gamma' = \cos \theta$. Thus $\langle \mathbf{u}, \mathbf{v} \rangle = \cos \theta$.

4.3 The answer is -2 no matter how you do it.

4.4 Subtract the first and then the second rows from the third and get a row of zeros. Expand using this row to see that the determinant is 0.

4.5 (a) $m_X(\mathbf{a} + \mathbf{b}) = X(\mathbf{a} + \mathbf{b})^T = X(\mathbf{a}^T + \mathbf{b}^T) = X\mathbf{a}^T + X\mathbf{b}^T = m_X(\mathbf{a}) + m_X(\mathbf{b})$. (b) $m_X(\lambda \mathbf{a}) = X((\lambda \mathbf{a})^T) = \lambda X(\mathbf{a}^T) = \lambda m_X(\mathbf{a})$.

4.6 Let the i-th row be $\mathbf{e}_i$. If $\mathbf{x} \in \mathbb{R}^n$, then $\mathbf{x} = (x_1, x_2, \ldots, x_n) = \sum_i x_i \mathbf{e}_i$ so the rows of the identity matrix span $\mathbb{R}^n$. As for linear independence, suppose that $\sum_i \lambda_i \mathbf{e}_i = 0$, then $(\lambda_1, \lambda_2, \ldots, \lambda_n) = \mathbf{0}$ so $\lambda_i = 0 \ \forall i \in \{1, 2, \ldots, n\}$.

4.7 A non-trivial linear relation among the terms of the subsequence would be a non-trivial linear relation among the terms of the original linearly independent sequence.

Chapter 5

5.1 (a) $(1,2,3,4,5)$, (b) $(1,4,2,5)$, (c) $(1,4,5,3,2)$, (d) $(3,4)$.

5.2 $(1,2,3,4)$.

5.3 $(1,5,2,3)$.

5.4 $(1,2,3)$ or $(1,3,2)$ or Id or $(23,45,666)$ for example.

5.5 (a) $(1,5,2,3)$, (b) $(1,5,2,3)$, (c) $(1,4,3,2)$ (d) id (e) $(1,3,4,2,5)$.

5.6 $(2,3)$.

5.7 (a) β, (b) β, (c) id, (d) α, (e) ε, (f) id, (g) id, (h) id.

5.8 Permutations which are products of disjoint cycles of the same length.

5.9 (a) 2 (b) 3 (c) 3 (d) 1 (e) 2 (f) 5.

5.10 (a) False (b) $(1,2)(1,3)(1,4)\ldots(1,n) = (1,2,3,4,\ldots,n)$. (c) Any permutation is a product of cycles, and now use part (b).

5.11 The group has order (= size) 8. Let the corners of the paper be labelled 1,2,3,4 consecutively. The elements of the group are id, $(1,2,3,4)$, $(1,4,3,2)$, $(1,3)(2,4)$, $(1,3)$, $(2,4)$, $(1,4)(2,3)$, $(1,2)(3,4)$. This is the *dihedral group* of order 8.

5.12 The group has order (= size) 10. Let the corners of the paper be labelled 1, 2, 3, 4, 5 consecutively. The group elements are id, $(1,2,3,4,5)$, $(1,3,5,2,4)$, $(1,4,2,5,3)$, $(1,5,4,3,2)$, $(2,5)(3,4)$, $(1,3)(4,5)$, $(1,5)(2,4)$, $(1,2)(3,5)$, $(1,4)(2,3)$. This is the *dihedral group* of order 10.

5.13 (a) 6, elements id, $(1,2)$, $(1,3)$, $(2,3)$, $(1,2,3)$, and $(1,3,2)$.(b) 2, elements id, $(1,2)$. (c) 1. element id.

5.14 Routine

5.15 This is really the associative law of multiplication. Notice that associativity and commutativity are therefore intimately related.

5.16 (a) $c = ah$ and $d = k^{-1}b$ for unique h and k. Thus $h = a^{-1}c$ and $k = bd^{-1}$. Now $ab = cd$ iff $a^{-1}c = bd^{-1}$. This happens iff $k = h^{-1}$. (b) Using part (a), we see that we will have such an inequality iff $h \in X$ and $h^{-1} \in Y$, i.e. iff $h \in X \cap Y$. (b) Routine. Apply Proposition 5.6 Count $X \times Y$ in two ways. It is both $|X| \cdot |Y|$ (obviously) and $|XY| \cdot |X \cap Y|$ (since, using part (b), every product xy is repeated $|X \cap Y|$ times). (d) By part (c) we have that $2^{10}/|X \cap Y| \leq 2^8$, so $|X \cap Y| \geq 4$. Also $|X \cap Y|$ divides 32 by Lagrange's Theorem so $|X \cap Y| = 4, 8, 16$ or 32.

5.17 (a) Choose $x \in U$ so $U = \{hx \mid h \in H\}$. Now $\{u^{-1} \mid u \in U\} = \{x^{-1}h^{-1} \mid h \in H\}$. Thus $\{u^{-1} \mid u \in U\} = x^{-1}H$ is a left coset. (b) Inversion (as in part (a)) induces the appropriate bijection.

5.18 Assume $H \neq 1$. Choose $m \in \mathbb{Z} \setminus \{0\}$ with $c^m \in H$. Inverting if necessary, we may assume $m \in \mathbb{N}$. Now revise m to be the minimal natural number such that $c^m \in H$ (since there is one, there must be a minimal one!). Now suppose $c^t \in H$ for $t \in \mathbb{Z}$. Divide t by m with remainder r in the range $0 \leq r < m$. Thus for some $q \in \mathbb{Z}$ we have $t = qm + r$. Now $c^t(c^m)^{-q} = c^r \in H$ since both $c^t, c^m \in H$. Thus $c^r \in H$. Now $0 \leq r < m$ contradicts the minimality of m unless $r = 0$. Now m divides t and c^t is a power of c^m. Thus $H = \langle c^m \rangle$.

5.19 Suppose $G = \langle x \rangle$ is cyclic. For any integers n, m we have $x^n x^m = x^{n+m} = x^m x^n$ so G is abelian.

5.20 Let the distinct prime divisors of n be $p_1, p_2, \ldots, p_k$. The number we want is

$$n(1 - \frac{1}{p_1})(1 - \frac{1}{p_2}) \ldots (1 - \frac{1}{p_k}) = n \prod_{i=1}^{k} (1 - \frac{1}{p_i}).$$

5.21 ζ is a bijection, and so too therefore is ζ^{-1}. Suppose $x, y \in G$. Thus $x = \zeta(a)$ and $y = \zeta(b)$ for some $a, b \in G$. Now

$$\zeta^{-1}(xy) = \zeta^{-1}(\zeta(a)\zeta(b)) = \zeta^{-1}(\zeta(ab)) = ab = \zeta^{-1}(x)\zeta^{-1}(y).$$

Thus ζ^{-1} preserves structure as required.

5.22 $\tau^2 = \mathrm{id}_{\mathbb{Z}}$ so τ is a bijection by Proposition 1.3.

$$\tau(x + y) = -(x + y) = (-x) + (-y) = \tau(x) + \tau(y).$$

Thus τ is an isomorphism of groups.

5.23 $\sigma : P \to P$ is an isomorphism (check). If P is replaced by $\mathbb{Q}$ it is not a surjection, so cannot be an isomorphism. Recall that 2 is not the square of a rational number by Theorem 2.6.

Chapter 6

6.1 (a) The constant sequence (0). (b) $((-1)^i)$. (c) See the answer to part (b). (d) No, since if (a_i) were such a sequence, then for each i we would have $a_i = a_{i+3}$ but a_{i+3} would have the opposite sign to a_i. This is absurd.

6.2 (a) $|\sin(i)| \leq 1 \, \forall i \in \mathbb{N}$ since $|\sin x| \leq 1$ for all real x. (b) $(i(-1)^i)$. (c) If (a_i) is bounded, then there is $M \in \mathbb{R}$ such that $|a_i| \leq M \, \forall i \in \mathbb{N}$. Thus $-M \leq a_i \leq M \, \forall i \in \mathbb{N}$ so (a_i) is bounded above and below. Conversely, if (a_i) is bounded above and below, there are $L, K \in \mathbb{R}$ such that $L \leq a_i \leq K \, \forall i \in \mathbb{N}$. Thus $-a_i \leq -L$ and $|a_i| \leq \max\{-L, K\}$ for all $i \in \mathbb{N}$.

6.3 (a) $\{x \mid x \in \mathbb{R},\ |x| < 1\}$. (b) Yes. This is because $|x| < -17\,\forall x \in \emptyset$. This is an example of vacuous reasoning. The elements of the empty set have any properties you like, because there are no such elements, so you cannot find one (there does not exist one) with the property that it does not have the property! Thus they all have the property! (c) No. We have just shown that the empty set is bounded. It is therefore not unbounded. If you were under the illusion that you can use vacuous reasoning to deduce whatever you like about the empty set, read on, for that is wrong. The point is that if you wanted to prove that the empty set was not bounded by 3, you would have to exhibit $y \in \emptyset$ such that $|y| \geq 3$. You can't do this, since there are no elements in $\emptyset$. The same argument works with any real number instead of 3. Another property the empty set doesn't enjoy is being non-empty. (d) Yes. Let the subsets of $\mathbb{R}$ be A and B, and be bounded by a and b respectively. Now $|x| \leq \max\{a, b\}\ \ \forall x \in A \cup B$ so $A \cup B$ is bounded. (e) Yes. Since $A \cap B \subseteq A \cup B$ we know $y \leq \max\{h, b\}\,\forall y \in A \cap B$.

6.4 (a) $(1/i)$. (b) The bound $\max\{|b_i| \mid i \in I\}$ will do. Note that $\{|b_i| + 1 \mid i \in I\}$ is a finite set of real numbers so we can take its maximum. (c) Let $a_i = i$ and $b_i = -i$ for every $i \in \{\mathbb{N}\}$. Now (a_i) and (b_i) are unbounded but $(a_i) + (b_i) = (0)$ is bounded since $|0| \leq 0$. (d) Suppose (a_i) and (b_i) are sequences and that $|a_i| < L\,\forall i \in \mathbb{N}$, and that $|b_i| < M\,\forall i \in \mathbb{N}$. Now $|a_i + b_i| \leq |a_i| + |b_i| < L + M$ for every $i \in \mathbb{N}$. Note the use of the triangle inequality.

6.5 (a) $((-1)^i)$. (b) Let $a_{2n} = n\forall n \in \mathbb{N}$, and $a_{2n-1} = 0\,\forall n \in \mathbb{N}$. Thus (a_i) is unbounded. Let $b_{2n} = 0\,\forall n \in \mathbb{N}$, and $ba_{2n-1} = n\,\forall n \in \mathbb{N}$. Thus (b_n) is unbounded. However $(a_n b_n) = (0)$ is bounded. (c) See part (b).

6.6 (a) Given $\varepsilon > 0 \exists N \in \mathbb{N}$ such that if $i \geq N$, then $|a_i - l| < \varepsilon$, so $||a_i| - |l|| \leq |a_i - l| < \varepsilon$. We have used Proposition 6.2. (b) If (a_n) confusedly converges to l, then there exists $N \in \mathbb{N}$ such that if $n \geq N$, then $a_n = l$, so $|a_n - l| = 0 < \varepsilon$ irrespective of the value of the positive quantity ε. Thus (a_n) converges to l. Conversely, $(1/n)$ converges to 0, but does not confusedly converge to 0.

6.7 (a_k) is a Cauchy sequence since terms beyond the k-th differ only in the decimal places $k + 1$ onwards; differences are therefore bounded by 10^{-k}. The sequence therefore converges (by which result?).

6.8 (a) (b_n) is monotone decreasing by design, and bounded below by $(\inf\{a_i \mid i \in \mathbb{N}\})$. Thus (b_n) converges (by which result?). (b) Similar, using a monotone increasing sequence (c_n). (c) Notice that $c_n \leq b_n$ for all $n \in \mathbb{N}$. Argue that $\lim(c_n) \leq \lim(b_n)$ by assuming the opposite (for contradiction). For the rest of the argument, and some discussion, see the web site.

6.9 Let $b = \sum b_i$ then b is the supremum of the partial sums $\sum_{i=1}^{j} b_i$ by the proof of Proposition 6.8. Now $\sum_{i=1}^{j} a_i < \sum_{i=1}^{j} b_i\,\forall j \in \mathbb{N}$. Thus the sequence of partial sums of the series $\sum a_i$ is monotone increasing and bounded above, and so convergent by Proposition 6.8.

6.10 (a) The relevant diagram consists of the portion of the graph of $1/x^\alpha$ in the first quadrant. You can fit a box of height 1/2 and width 1 between 1 and 2 and under the graph. To the right you fit a box of height 1/3 and width 1 and so on. Compare the area in the first $n - 1$ boxes with the area under the graph to obtain the desired inequality (note that $1/x^\alpha$ is a decreasing function in this quadrant). Now $\int_1^n x^{-\alpha}dx = [x^{1-\alpha}/(1-\alpha)]_1^n = 1/(\alpha - 1)(1 - n^{1-\alpha})$. Thus $\int_1^n x^{-\alpha}dx \leq M = 1/(\alpha - 1)\,\forall n \in \mathbb{N}$. The partial sums of $\sum n^{-\alpha}$ form a monotone increasing sequence bounded above by M, and so converge by Proposition 6.8. (b) This time you insert a box of height 1/2 and width 1 between 2 and 3, and a box of height 1/3 and width 1 between 3 and 4 etc. Now show that the partial sums of $\sum 1/n$ can be made arbitrarily large, by comparing them with the area under the graph. The point is that $\log n$ assumes arbitrarily large values for sufficiently large n.

6.11 A little algebra shows that the spider will travel along a proportion of $0.01(1 + (1/2) + (1/3) + \ldots + (1/k))$ of the elastic after k days. We need to estimate when the sum of the first k terms of the harmonic series will exceed 100. Using the (excellent) approximation of $\log k$ we conclude that the spider will reach the end of the elastic after e^{100} days. We crudely estimate this as follows (this is not sound, but gives the general idea). e^2 is about $8 = 2^3$, and 2^{10} is about 1000. Thus our (very loose) estimate for e^{100} is 2^{150} which is about 10^{45}. Say there are 1000/3 days in a year; our estimate is 3×10^{42} years. Current estimates of the age of our universe are about 13×10^9 years, so the spider has a way to go yet.

Chapter 7

7.1 For any $\varepsilon > 0$ let $\delta = \varepsilon$. If $|x - a| < \delta$, then $||x| - |a|| \leq |x - a| < \delta = \varepsilon$.

7.2 Draw a picture. $(\cos x - \cos a)^2 + (\sin x - \sin a)^2$ is the square of the distance between $(\cos a, \sin a)$ and $(\cos x, \sin x)$. The straight line distance between these points is bounded by the length of the arc of the unit circle joining these points. Thus

$$(\cos x - \cos a)^2 + (\sin x - \sin a)^2 \leq |x - a|^2.$$

Thus both $|\cos x - \cos a| < |x - a|$ and $|\sin x - \sin a| < |x - a|$. Given any ε, let $\delta = \varepsilon$, and check that everything works.

7.3 (a) Test the two cases $a \leq b$ and $b < a$. (b) Use part (a), and the answer to Exercise 7.1, and the results in the chapter about how to build new continuous functions from old ones. You therefore need not get involved with ε, δ arguments.

7.4 Routine.

7.5 Routine.

Chapter 8

8.1 For the sum use $|a_n + b_n - a_m - b_m| \leq |a_n - a_m| + |b_n - b_m|$ by the triangle inequality. For the product use $|a_n b_n - a_m b_m| = |a_n b_n - a_n b_m + a_n b_m - a_m b_m| \leq |a_n||b_n - b_m| + |b_m||a_n - a_m|$. The differences are good news, the multipliers $|a_n|$ and $|b_m|$ are not. However, use Proposition 6.5 to tame the problem.

8.2 Tedious. See the web site mentioned in the preface.

8.3 Similar to the previous question.

8.4 For each x, let the binary (binary–decimal) expansion of the fractional part of x ($x - \lfloor x \rfloor$) be $\sum_{i=1}^{\infty} a_i 2^{-i}$. Map x to $\sum_{i=1} (-1)^{a_i}/i$ if the sum converges, and 0 if it doesn't. This function assumes all real values on any interval which contains all real numbers represented by an initial starting string $b_1 b_2 \ldots b_k$ followed by a tail of any legal form. This stems from the ideas following Proposition 6.10. Any interval will contain such a collection of real numbers, provided k is chosen sufficiently large. Any non-trivial interval may be partitioned into an arbitrary number of non-trivial subintervals, so every real number is assumed as a value infinitely often.

Index

alternating, 123
Argand diagram, 7, 63, 71
argument, 21
associative law, 9
Atanasov, 49

basis, 122
bijection, 30
binary operation, 44
Bobker, David, 45
Bolzano-Weierstrass theorem, 172
bonvergence, 167
bounded sets and sequences, 159

cardinality, 3
Cartesian product, 16
Cauchy sequence, 199
Cauchy-Schwarz inequality, 101
closed operation, 4
codomain, 21
cofactor, 111
commutative law, 8
complement, 10
completeness, 169
complex conjugate, 72
complex number, 7, 56, 63
- imaginary part, 64
- modulus, 71
- polar form, 77
- real part, 64
conjecture, 50
constant sequence, 158
continuity at a point, 182
continuity on an interval, 191
coset, 144
cosh (hyperbolic cosine), 85
cosine, 76
counter-example, 47

De Moivre's theorem, 77
De Morgan's laws, 10
Dedekind cut, 197
degree of a permutation group, 132
determinant, 109
difference of sets, 10
differentiation, 195
dimension, 122
disjoint, 8
disjoint cycles, 128
distributive law, 10
domain, 21
double implication, 54
double implication $\Leftrightarrow$, 54

empty set, 7
equation
- quadratic, 58
- root, 58
equivalence class, 34
equivalence relation, 32
even functions, 79

Fermat, 50
Fermat's last theorem, 50
Fermat's two squares theorem, 51
field, 4, 65
for all ($\forall$), 19
Frege, 14
function, *see* map
fundamental period, 78

fundamental theorem of algebra, 68

graph, 29
greatest lower bound, 170
group
- cyclic, 146
- index, 145
- order, 138
- permutation group, 132
- subgroup, 142
- symmetric, 132
group axioms, 137

Halmos, 15
Heaviside step function, 182
homomorphism, 150, 151

identity map, 22
identity permutation, 130
if and only if, 54
iff, 54
image of a map, 22
imaginary part, 64
implication, 53
implication symbol $\Rightarrow$, 54
inclusive or, 8
index, 145
induction, 39
- complete, 43
- simple, 40, 43
inductive hypothesis, 40
infimum, 170
integer part function, 79
integers, 3
intermediate value theorem, 193
intersection, 8
interval
- closed, 37
- open, 37
inverse map, 30
irrational number, 5
isomorphism, 148

join, 10

kernel, 151
Kronecker delta, 108

Lagrange's theorem, 145
least upper bound, 169
limit of a function, 195
limit of a sequence, 162
linear algebra, 57
linear dependence, 120
linear independence, 120
linear map, 114
logarithm, 81, 90

map, 20
- bijective, 23
- codomain of a map, 21
- composition, 23
- domain of a map, 21
- image, 22
- injective, 23
- inverse, 30
- surjective, 23
mapping, *see* map
matrix, 106
- addition, 106
- adjugate, 117
- cofactor, 111
- determinant, 109, 110
- identity, 108
- inverse, 109
- minor, 111
- multiplication, 106
- orthogonal, 124
- transpose, 106, 123
meet, 10
method of descent, 50
minor, 111
modulus
- complex number, 71
modulus of a real number, 156
monotone sequences, 170
multilinear map, 114

natural logarithms, 5
natural numbers, 3
negation, 19
negation of symbols, 4
null set, 7

odd functions, 79
order of a group, 138
order of a group element, 138
order of a permutation, 133
order of a permutation group, 132
ordered pair, 16

p-adic numbers, 202
paradox, 13
partition, 33
Peano, 11
periodic, 78
permutation, 125
- cycle, 127

– even and odd, 112
– group, 132
– identity, 132
– inverse permutations, 130
– transposition, 134
– two-sided inverse, 132
polar form, 77
position vector, 103
power set, 15
prime numbers, 44
purely imaginary, 64
Pythagoras, 5, 51

quadratic equations, 58
quantifiers, 19

rational numbers, 4
real numbers, 5
real part, 64
relation, 33
– antisymmetric, 32
– equivalence, 32
– reflexive, 32
– symmetric, 32
– transitive, 2, 32
relations, 31
root, 58
Russell, 14

scalar product, 99
sequence
– absolutely convergent, 178
– alternating, 159
– Cauchy, 199
– Cauchy sequence, 173
– convergence to a limit, 162
– convergent, 175
– null, 200
– subsequence, 171
set
– cardinality, 3
– complement, 10
– De Morgan's laws, 10
– difference, 10
– element, 1
– equality, 2
– finite, 3
– intersection, 8
– join, 10
– meet, 10
– notation, 4
– power set, 15
– subset, 2, 16
– union, 8
– universal, 10
set difference, 17
sets, 1
– of sets, 10
sine, 76
sinh (hyperbolic sine), 85
span of vectors, 121
subgroup, 142
supremum, 169
symmetric, 123
symmetric group, 132

theorem
– fundamental – of algebra, 68
– Bolzano-Weierstrass, 172
– De Moivre's, 77
– Fermat's last, 50
– Fermat's two squares, 51
– intermediate value, 193
– Lagrange, 145
there exist (∃), 19
transposition, 112, 134
transversal, 35

union, 8

vector
– dot product, 99
– modulus, 98
– scalar product, 99
– span, 121
vector space, 121
– basis, 122
– dimension, 122
vector space
– abstract , 98
Venn diagram, 17
von Neumann, 11

Wiles, 50