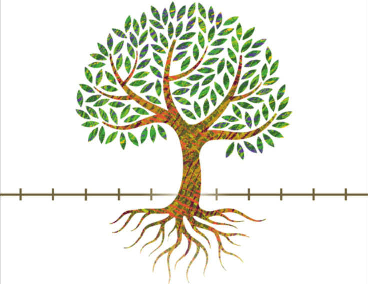
OXFORD



THE FOUNDATIONS OF MATHEMATICS

SECOND EDITION

IAN STEWART
AND DAVID TALL

THE FOUNDATIONS OF MATHEMATICS

Second Edition

IAN STEWART AND DAVID TALL

OXFORD
UNIVERSITY PRESS

OXFORD
UNIVERSITY PRESS

Great Clarendon Street, Oxford, OX2 6DP,
United Kingdom

Oxford University Press is a department of the University of Oxford.
It furthers the University's objective of excellence in research, scholarship,
and education by publishing worldwide. Oxford is a registered trade mark of
Oxford University Press in the UK and in certain other countries

First Edition published in 1977
Second Edition published in 2015

Impression: 1

Published in the United States of America by Oxford University Press
198 Madison Avenue, New York, NY 10016, United States of America

British Library Cataloguing in Publication Data
Data available

Library of Congress Control Number: 2014946122

ISBN 978–0–19–870644–1 (hbk.)
ISBN 978–0–19–870643–4 (pbk.)

Printed and bound by
CPI Group (UK) Ltd, Croydon, CR0 4YY

Links to third party websites are provided by Oxford in good faith and for information only. Oxford disclaims any responsibility for the materials contained in any third party website referenced in this work.

CONTENTS

Part I The Intuitive Background

Part II The Beginnings of Formalisation

Part III The Development of Axiomatic Systems

Part IV Using Axiomatic Systems

Part V Strengthening the Foundations

PART I
The Intuitive Background

The first part of the book reflects on the experiences that the reader will have encountered in school mathematics to use it as a basis for a more sophisticated logical approach that precisely captures the structure of mathematical systems.

Chapter 1 considers the learning process itself to encourage the reader to be prepared to think in new ways to make sense of a formal approach. As new concepts are encountered, familiar approaches may no longer be sufficient to deal with them and the pathway may have side-turnings and blind alleys that need to be addressed. It is essential for the reader to reflect on these new situations and to prepare a new overall approach.

Using a 'building' metaphor, we are surveying the territory to see how we can use our experience to build a firm new structure in mathematics that will make it strong enough to support higher levels of development. In a 'plant' metaphor, we are considering the landscape, the quality of the soil, and the climate to consider how we can operate to guarantee that the plants we grow have sound roots and predictable growth.

Chapter 2 focuses on the intuitive visual concept of a real number as a point on a number line and the corresponding symbolic representation as an infinite decimal, leading to the need to formulate a definition for the completeness property of the real numbers. This will lead in the long term to surprising new ways of seeing the number line as part of a wider programme to study the visual and symbolic representations of formal structures that bring together formal, visual, and symbolic mathematics into a coherent framework.

CHAPTER 1

Mathematical Thinking

Mathematics is not an activity performed by a computer in a vacuum. It is a human activity performed in the light of centuries of human experience, using the human brain, with all the strengths and deficiencies that this implies. You may consider this to be a source of inspiration and wonder, or a defect to be corrected as rapidly as possible, as you wish; the fact remains that we must come to terms with it.

It is not that the human mind cannot think logically. It is a question of different kinds of understanding. One kind of understanding is the logical, step-by-step way of understanding a formal mathematical proof. Each individual step can be checked but this may give no idea how they fit together, of the broad sweep of the proof, of the reasons that lead to it being thought of in the first place.

Another kind of understanding arises by developing a global viewpoint, from which we can comprehend the entire argument at a glance. This involves fitting the ideas concerned into the overall pattern of mathematics, and linking them to similar ideas from other areas. Such an overall grasp of ideas allows the individual to make better sense of mathematics as a whole and has a cumulative effect: what is understood well at one stage is more likely to form a sound basis for further development. On the other hand, simply learning how to 'do' mathematics, without having a wider grasp of its relationships, can limit the flexible ways in which mathematical knowledge can be used.

The need for overall understanding is not just aesthetic or educational. The human mind tends to make errors: errors of fact, errors of judgement, errors of interpretation. In the step-by-step method we might not notice that one line is not a logical consequence of preceding ones. Within the overall framework, however, if an error leads to a conclusion that does not fit into the total picture, the conflict will alert us to the possibility of a mistake.

For instance, given a column of a hundred ten-digit numbers to add up, where the correct answer is 137568304452, we might make an arithmetical error and get 137568804452 instead. When copying this answer we might make a second error and write 1337568804452. Both of these errors could escape detection. Spotting the first would almost certainly need a step-by-step check of the calculation. The second error, however, is easily detected because it does not fit into the overall pattern of arithmetic. A sum of 100 ten-digit numbers will be at most a twelve-digit number (since 9999999999 × 100 = 999999999900) and the final proposed answer has thirteen.

It is a combination of step-by-step and overall understanding that has the best chance of detecting mistakes; not just in numerical work, but in all areas of human understanding. The student must develop both kinds, in order to appreciate the subject fully and be an effective practitioner. Step-by-step understanding is fairly easy; just take one thing at a time and do lots of 'drill' exercises until the idea sinks in. Overall understanding is much harder; it involves taking a lot of individual pieces of information and making a coherent pattern out of them. What is worse is that having developed a particular pattern which suits the material at one stage, new information may arise which seems to conflict. The new information may be erroneous but it often happens that previous experiences that worked in one situation no longer operate in a new context. The more radical the new information is, the more likely that it does not fit, and that the existing overall viewpoint has to be modified. That is what this first chapter is about.

Concept Formation

When thinking about any area of mathematics, it helps to understand a little about how we learn new ideas. This is especially true of foundational issues, which involve revisiting ideas that we already think we know. When we discover that we do not—more precisely, that there are basic questions that we have not been exposed to—we may feel uncomfortable. If so, it's good to know that we are not alone: it happens to nearly everyone.

All mathematicians were very young when they were born. This platitude has a non-trivial implication: even the most sophisticated mathematician must have passed through the complex process of building up mathematical concepts. When first faced with a problem or a new concept, the mathematician turns it over in the mind, digging into personal experiences to see if it is like something that has been encountered before. This exploratory, creative phase of mathematics is anything but logical. It is only when the pieces begin to fit together and the mathematician gets a 'feel' for the concept, or the problem, that a semblance of order emerges. Definitions are formulated in ways that can be used for deduction, and there is a final polishing phase where the essential facts are marshalled into a neat and economical proof.

As a scientific analogy, consider the concept 'colour'. A dictionary definition of this concept looks something like 'the sensation produced in the eye by rays of decomposed light'. We do not try to teach the concept of colour to a child by presenting them with this definition. ('Now, Angela, tell me what sensation is produced in your eye by the decomposed light radiating from this lollipop . . . ') First you teach the concept 'blue'. To do this you show a blue ball, a blue door, a blue chair, and so on, accompanying each with the word 'blue'. You repeat this with 'red', 'yellow', and so on. After a while the child begins to get the idea; you point to an object they have not seen before and their response is 'blue'. It is relatively easy to refine this to 'dark blue', 'light blue', and so forth. After repeating this procedure many times, to establish the individual colours, you start again. 'The colour of that door is blue. The colour of this box is red. What colour is that buttercup?' If the response is 'yellow' then the concept 'colour' is beginning to develop.

As a child develops and learns scientific concepts they may eventually be shown a spectrum obtained by passing light through a prism. This may lead to learning about the wavelength of light, and, as a fully fledged scientist, being able to say with precision which wavelength corresponds to light of a particular colour. The understanding of the concept 'colour' is now highly refined, but it does not help the scientist to explain to a child what 'blue' is. The existence of a precise and unambiguous definition of 'blue' in terms of wavelength is of no use at the concept-forming stage.

It is the same with mathematical concepts. The reader already has a large number of mathematical concepts established in their mind: how to solve a quadratic equation, how to draw a graph, how to sum a geometric progression. They have great facility in arithmetical calculations. Our aim is to build on this wealth of mathematical understanding and to refine these concepts to a more sophisticated level. To do this we use examples, drawn from the reader's experience, to introduce new concepts. Once these concepts are established, they become part of a richer experience upon which we can again draw to aim even higher.

Although it is certainly possible to build up the whole of mathematics by axiomatic methods starting from the empty set, using no outside information whatsoever, it is also totally unintelligible to anyone who does not already understand the mathematics being built up. An expert can look at a logical construction in a book and say 'I guess that thing there is meant to be "zero", so that thing is "one", that's "two", . . . this load of junk must be the integers, . . . what's that? Oh, I think I see: it must be "addition". . . '.

The non-expert is faced with an indecipherable mass of symbols. It is never sufficient to define a new concept without giving enough examples to explain what it looks like and what can be done with it. Of course, an expert is often in a position to supply their own examples, and may not need much help.

Schemas

A mathematical concept, then, is an organised pattern of ideas that are somehow interrelated, drawing on the experience of concepts already established. Psychologists call such an organised pattern of ideas a 'schema'. For instance, a young child may learn to count ('one, two, three-four-five, once I caught a fish alive') progressing to ideas like 'two sweets', 'three dogs', . . . and eventually discovers that two sweets, two sheep, two cows have something in common, and that something is 'two'. He or she builds a schema for the concept 'two' and this schema involves the experience that everyone has two hands, two feet, last week we saw two sheep in a field, the fish-alive rhyme goes 'one, two, . . . ', and so on. It is really quite amazing how much information the brain has lumped together to form the concept, or the schema.

The child progresses to simple arithmetic ('If you have five apples and you give two away, how many will you have left?') and eventually builds up a schema to handle the problem 'What is five minus two?' Arithmetic has very precise properties. If 3 and 2 make 5, then 5 take away 2 leaves 3. The child discovers these properties by trying to make sense of arithmetic. It then becomes possible to use known facts to derive new facts. If the child knows that 8 plus 2 makes 10, then 8 plus 5 can be thought of as 8 plus 2 plus 3, so the sum is 10 plus 3, which is 13. Over time the child can build up a rich schema of whole number arithmetic.

At this point, if you ask 'What is five minus six?' the response is likely to be 'You can't do it', or perhaps just an embarrassed giggle that an adult should ask such a silly question. This is because the question does not fit the child's schema for subtraction: when thinking about 'five apples, take six away', this simply cannot be done. At a later stage, experiencing negative numbers will give the answer 'minus one'. What has happened? The child's original schema for 'subtraction' has been modified to accommodate new ideas—perhaps by thermometer scales, or the arithmetic of banking, or whatever—and the understanding of the concept changes. During the process of change, confusing problems will arise (what does minus one apple look like?) which may eventually be resolved satisfactorily (apples don't behave like thermometer readings).

A large part of the learning process involves making an existing schema more sophisticated, so that it can take account of new ideas. This process, as we have said, may be accompanied by a state of confusion. If it were possible to learn mathematics without becoming confused, life would be wonderful.

Unfortunately, the human mind does not seem to work that way. More than 2000 years ago, Euclid supposedly told King Ptolemy I that 'There is no royal road to geometry'. The next best thing is to recognise not just the confusion, but also its causes. At various stages in reading this book the reader will be confused. Sometimes, no doubt, the cause will be the authors' sloppiness, but often it will be the process of modifying personal knowledge to make sense of a more general situation. This type of confusion is creative, and it should be welcomed as a sign that progress is being made—unless it persists for too long. By the same token, once the confusion is resolved, a sudden clarity can appear with a feeling of great pleasure that the pieces fit together perfectly like a jigsaw. It is this feeling of perfect harmony that makes mathematics not only a challenge, but also an endeavour that leads to deep aesthetic satisfaction.

An Example

This way to develop new ideas is illustrated by the historical development of mathematical concepts—itself a learning process, but involving many minds instead of one. When negative numbers were first introduced, they met considerable opposition: 'You can't have less than nothing'. Yet nowadays, in this financial world of debits and credits, negative numbers are a part of everyday life.

The development of complex numbers is another example. Like all mathematicians, Gottfried Leibniz knew that the square of a positive number or of a negative number must always be positive. If i is the square root of minus one, then $i^2 = -1$, so i cannot be a positive or a negative number. Leibniz believed that it should therefore be endowed with great mystical significance: a non-zero number neither less than zero nor greater than zero. This led to enormous confusion and distrust concerning complex numbers; it persists to this day in some quarters.

Complex numbers do not fit readily into many people's schema for 'number', and students often reject the concept when it is first presented. Modern mathematicians look at the situation with the aid of an enlarged schema in which the facts make sense.

Imagine the real numbers marked on a line in the usual way:

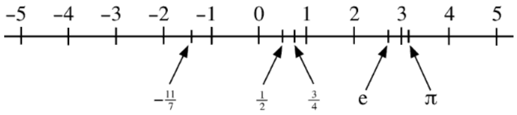


Fig. 1.1 The real numbers

Negative numbers are to the left of zero, positive to the right. Where does i go? It can't go to the left; it can't go to the right. The people whose schema does not allow complex numbers must argue thus: this means that it can't go anywhere. There is no place on the line where we can mark i, so it's not a number.

However, there's an alternative. We can visualise complex numbers as the points of a plane. (In 1758 François Daviet de Foncenex stated that it was pointless to think of imaginary numbers as forming a line at right angles to the real line. Fortunately others disagreed.) The real numbers lie along the 'x-axis', the number i lies one unit above the origin along the 'y-axis', and the number $x + \mathrm{i}y$ lies x units along the real line and then y units above it (change directions for negative x or y). The objection to i ('it can't lie anywhere on the line') is countered by the observation that it doesn't. It lies one unit above the line. The enlarged schema can accommodate the disturbing facts without any trouble.

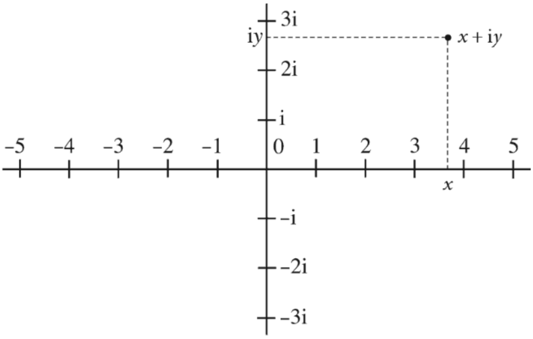


Fig. 1.2 Putting i in its place

This happens quite often in mathematics. When a particular situation is generalised to a new context, some properties operate in the same way as before, such as addition and multiplication both being commutative. But other properties (such as the order properties of real numbers) that work well in the original schema are no longer relevant in the extended schema (in this case the schema of complex numbers).

This is a very general phenomenon; it has happened not only to students, but to mathematicians throughout history, up to the present day. If you work in an established situation where the ideas have been fully sorted out, and the methods used are sufficient to solve all of the usual problems, it is not that difficult to teach an apprentice the trade. All you need is to grasp the current principles and develop fluency in the methods. But when there is a genuine change in the nature of the system, as happened when negative numbers were introduced in a world that only used natural counting numbers, or when complex numbers were encountered solving equations, then there is a genuine period of confusion for everyone. What are these newfangled things? They certainly don't work the way I expected them to!

This can cause deep confusion. Some conquer it by engaging with the ideas in a determined and innovative fashion; others suffer a growing feeling of anxiety, even revulsion and rejection.

One such major occasion began in the final years of the nineteenth century and transformed the mathematics of the twentieth and twenty-first centuries.

Natural and Formal Mathematics

Mathematics began historically with activities such as counting objects and measuring quantities, dealing with situations in the natural world. The Greeks realised that drawing figures and counting pebbles had more profound properties, and they built up the method of Euclidean proof in geometry and the theory of prime numbers in arithmetic. Even though they developed a Platonic form of mathematics that imagined perfect figures and perfect numbers, their ideas were still linked to nature. This attitude continued for millennia. When Isaac Newton studied the force of gravity and the movement of the heavenly bodies, science was known as 'natural philosophy'. He built his ideas about calculus on Greek geometry, and on algebra that generalised the natural operations of arithmetic.

The reliance on 'naturally occurring' mathematics continued until the late nineteenth century, when the focus changed from the properties of objects and operations to the development of formal mathematics based on set-theoretic definition and logical proof. This historical transition from natural to formal mathematics involved a radical change of viewpoint, leading to far more powerful insights into mathematical thinking. It plays an essential role in the shift from school geometry and algebra to formal mathematics at university.

Building Formal Ideas on Human Experience

As mathematics becomes more sophisticated, new concepts often involve some ideas that generalise, but others that operate in new ways. As the transition is made from school mathematics to formal mathematics, it may seem logical to start anew with formal definitions and learn how to make formal deductions from first principles. However, experience over the last half-century has shown that this is not a sensible idea. In the 1960s, schools tried a new approach to mathematics, based on set theory and abstract definitions. This 'new math' failed because, although experts might understand the abstract subtleties, learners need to build up a coherent schema of knowledge to make sense of the definitions and proofs. We now know more about how humans learn to think mathematically. This lets us give examples from practical research to show how students have interpreted ideas in ways that are subtly different from what is intended in the printed text. We mention this to encourage you to think carefully about the precise meanings involved, and to develop strong mathematical links between ideas.

It is helpful to read proofs carefully and to get into the habit of *explaining to yourself* why the definitions are phrased as they are and how each line of a proof follows from previous lines. (See the Appendix on Self-Explanation on page 377.) Recent research [3] has shown that students who make an effort to think through theorems for themselves benefit in the long run. Eye-tracking equipment has been used to study how students read pages from the first edition of this very book. There is a strong correlation between spending longer considering significant steps in a proof and obtaining higher marks on tests administered at a later stage. It's a no-brainer really. A stronger effort at making personal links gives you a more coherent personal schema of knowledge that will be of benefit in the long run.

You need to be sensible about how to proceed. In practice, it is not always possible to give a precise, dictionary definition for every concept encountered. We may talk about a set being 'a well-defined collection of objects', but we will be begging the question, since 'collection' and 'set' mean the same thing.

When studying the foundations of mathematics, we must be prepared to become acquainted with new ideas by degrees, rather than by starting from a watertight definition that can be assimilated at once. As we continue along that path, our understanding of an idea can become more sophisticated. We can sometimes reach a stage where the original vague definition can be reformulated in a rigorous context ('yellow is the colour of light with a wavelength of 5500 Å'). The new definition, seemingly so much better than the vague ideas that led to its formulation, has a seductive charm.

Wouldn't it be so much better to start from this nice, logical definition?

The short answer is 'no'.

In this book, we begin in Part I with ideas that you have met in school. We consider the visual number line, and how it is built up by marking various number systems, such as the whole numbers, 1, 2, 3, . . . ; then fractions between adjacent whole numbers; then signed numbers to the right and left of the origin, including signed whole numbers (the integers) and signed fractions (the rationals); then expanding to the real numbers including both rational and irrational numbers. In particular, we focus on natural ways to perform operations such as addition, multiplication, subtraction, and division, using whole numbers, fractions, decimals, and so on, to highlight properties that can be used as a basis for formal axioms for the various number systems.

Part II lays the foundations for set theory and logic, appropriate to the concept of proof used by mathematicians, with a sensible balance of logical precision and mathematical insight. In particular, the reader should note that it is essential to focus not only on what the definitions actually say, but also to be careful not to assume other properties that may arise not from the definition but from mental links set up by previous experience. For instance, students in school meet functions such as $y = x^2$ or $f(x) = \sin 3x$, which are always given by some kind of formula. However, the general notion of a function does not require a formula. All that is needed is that for each value of x (in a specified set) there is a single corresponding value of y. This broader definition applies to sets in general, not just to numbers. The properties that a defined concept must have are deduced from the definition by mathematical proof.

Part III develops the axiomatic structures appropriate for the succession of number systems, starting with axioms for natural numbers and proof by induction. The story continues by demonstrating how successive systems—integers, rationals, and real numbers—can be constructed from first principles using set-theoretic techniques. This process culminates in a list of axioms that defines the system of real numbers, with two operations (addition and multiplication) that satisfy specified properties of arithmetic and order, together with a 'completeness axiom' that states that any increasing sequence bounded above must tend to a limit. These axioms define a 'complete ordered field', and we prove that they specify the real numbers *uniquely*. Real numbers may be pictured as points on a line with the defined operations of addition, multiplication, and order, where the line is filled out to include irrational numbers such as $\sqrt{2}$ or π as infinite decimals that may be computed to any required accuracy as a finite decimal. For instance, $\sqrt{2}$ is 1·414 to 3 decimal places, π is approximately equal to the fraction 22/7, or may be calculated to any desired accuracy as a decimal, say 3·14 to two decimal places or 3·1415926536 to ten places.

Formal Systems and Structure Theorems

This sequence of development, building a formal system from a carefully chosen list of axioms, can be generalised to cover a wide range of new situations. It has a huge advantage compared to dealing with naturally occurring systems that are encountered in everyday life. The theorems that can be deduced from a given list of axioms using formal proof must hold in *any* system that satisfies the axioms—old or new. Formal theorems are *future-proofed*. The theorems apply not only to systems that are already familiar, but also to any new system that satisfies the given axioms. This releases us from the necessity of re-checking our beliefs in every new system we encounter. This is a major step forward in mathematical thinking.

Another more subtle development is that some theorems deduced within a formal system prove that the system has specific properties that allow it to be visualised in a certain way, and other properties that allow its operations to be carried out using symbolic methods. Such theorems are called *structure theorems*. For example, any complete ordered field has a unique structure that may be represented as points on a number line or as decimal expansions.

This shifts formal proof to a new level of power. Not only do we devote lengthy resources to develop a consistent approach to formal proof, ultimately we can develop new ways of thinking that blend together formal, visual, and symbolic ways of operation that combine human ingenuity and formal precision.

Using Formal Mathematics More Flexibly

In Part IV we show how these more flexible methods can be applied in various contexts, first by applying the ideas to group theory and then to two quite different extensions of finite ideas to infinite concepts. One is the extension of counting from finite sets to infinite sets, by saying that two sets have the same *cardinal number* if all their elements can be paired so that each element in one set corresponds to precisely one element in the other. Cardinal numbers have many properties in common with regular counting numbers, but they also have new and unfamiliar properties. For instance, we can take away an infinite subset (such as the even numbers) from an infinite set (such as the natural numbers) to leave an infinite subset (the odd numbers) with the same cardinal number of elements as the original set. By the same token, subtraction cannot be uniquely defined for infinite cardinal numbers, nor can division, so the reciprocal of an infinite cardinal number is not defined as a cardinal number.

The second extension places the real numbers, which form a complete ordered field, inside a larger (but not complete) ordered field. Here, an element k in the larger field may satisfy the order property '$k > r$ for every real number r'. In this sense, k is infinite: in the formally defined order, it is greater than *all* real numbers. Yet this k behaves quite differently from an infinite cardinal number, because it has a reciprocal $1/k$. Moreover, $1/k$ is smaller than any positive real number.

Upon reflection, we should not be surprised by these apparently contradictory possibilities, where an infinite number has a reciprocal in one system but not in another. The system of whole numbers that we use for counting does not provide reciprocals, but the systems of rational and real numbers do. If we select certain properties to generalise different systems, we should not be surprised if the generalisations are also different.

This brings us to an important conclusion. Mathematics is a living subject, in which seemingly impossible ideas may become possible in a new formal context, determined by stating appropriate axioms.

Writing over a century ago, when the new formal approach to mathematics was becoming widespread, Felix Klein [4] wrote:

> Our standpoint today with regard to the foundations is different from that of the investigators of a few decades ago; and what we today would state as ultimate principles, will certainly be outstripped after a time.

On the same page he noted:

> Many have thought that one could, or that one indeed must, teach all mathematics *deductively* throughout, by starting with a definite number of axioms and deducing everything from these by means of logic. This method, which some seek to maintain on the authority of Euclid, certainly does not correspond to the historical development of mathematics. In fact, mathematics has grown like a tree, which does not start from its tiniest roots and grow merely upward, but rather sends its roots deeper and deeper at the same time and rate that its branches and leaves are spreading upwards. Just so—if we may drop the figure of speech—mathematics began its development from a certain standpoint corresponding to normal human understanding and has progressed, from that point, according to the demands of science itself and of the then prevailing interests, now in one direction toward new knowledge, now in the other through the study of fundamental principles.

We follow this development throughout the book by starting from the experiences of students in school, digging deeper in Part II to find fundamental ideas that we use in Part III to build into formal structures for number systems, and expanding the techniques to wider formal structures in Part IV. In Part V, we close this introduction to the foundations of mathematics by reflecting on the deeper development of fundamental logical principles that become necessary to support more powerful mathematical growth in the future.

Exercises

The following examples are intended to stimulate you into considering your own thought processes and your present mathematical viewpoint. Many of them do not have a 'correct' answer, however it will be most illuminating for you to write out solutions and keep them in a safe place to see how your opinions may change as you read the text. Later in the book (at the end of chapters 6 and 12) you will be invited to reconsider your responses to these questions to see how your thinking has changed. Don't be afraid at this time to say that some of the ideas do not make sense to you at the moment. On the contrary, it is to your advantage to acknowledge any difficulties you may have. The intention of this book is that the ideas will become much clearer as you develop in sophistication.

1. Think how you think about mathematics. If you meet a new problem which fits into a pattern that you recognise, your solution may follow a time-honoured logical course, but if not, then your initial attack may be anything but logical. Try these three problems and do your best to keep track of the steps you take as you move towards a solution.
   (a) John's father is three times as old as John; in ten years he will only be twice John's age. How old is John now?
   (b) A flat disc and a sphere of the same diameter are viewed from the same distance, with the plane of the disc at right angles to the line of vision. Which looks larger?
   (c) Two hundred soldiers stand in a rectangular array, in ten rows of twenty columns. The tallest man in each row is selected and of these ten, *S* is the shortest. Likewise the shortest in each column is singled out and *T* is the tallest of these twenty. Are *S* and *T* one and the same? If not, what can be deduced about the relative size of *S* and *T*?

   Make a note of the way that you attempted these problems, as well as your final solution, if you find one.

2. Consider the two following problems:
   (a) Nine square metres of cloth are to be divided equally between five dressmakers; how much cloth does each one get?
   (b) Nine children are available for adoption and are to be divided equally between five couples; how many children are given to each couple?

   Both of these problems translate mathematically into:

   'Find x such that $5x = 9$'.

   Do they have the same solution? How can the mathematical formulation be qualified to distinguish between the two cases?

3. Suppose that you are trying to explain negative numbers to someone who has not met the concept and you are faced with the comment:

   > 'Negative numbers can't exist because you can't have less than nothing.'

   How would you reply?

4. What does it mean to say that a decimal expansion 'recurs'? What fraction is represented by the decimal $0{\cdot}333\ldots$? What about $0{\cdot}999\ldots$?

5. Mathematical use of language sometimes differs from colloquial usage. In each of the following statements, record whether you think that they are true or false. Keep them for comparison when you read chapter 6.
   (a) All of the numbers 2, 5, 17, 53, 97 are prime.
   (b) Each of the numbers 2, 5, 17, 53, 97 is prime.
   (c) Some of the numbers 2, 5, 17, 53, 97 are prime.
   (d) Some of the numbers 2, 5, 17, 53, 97 are even.
   (e) All of the numbers 2, 5, 17, 53, 97 are even.
   (f) Some of the numbers 2, 5, 17, 53, 97 are odd.

6. 'If pigs had wings, they'd fly.'
   Is this a logical deduction?

7. 'The set of natural numbers 1, 2, 3, 4, 5, . . . is infinite.' Give an explanation of what you think the word 'infinite' means in this context.

8. A formal definition of the number 4 might be given in the following terms.

   First note that a set is specified by writing its elements between curly brackets { } and that the set with no elements is denoted by ∅. Then we define

$$4 = \{\varnothing, \{\varnothing\}, \{\varnothing, \{\varnothing\}\}, \{\varnothing, \{\varnothing\}, \{\varnothing, \{\varnothing\}\}\}\}.$$

Can you understand this definition? Do you think that it is suitable for a beginner?

9. Which, in your opinion, is the most likely explanation for the equality

$$(-1) \times (-1) = +1?$$

   (a) A scientific truth discovered by experience.
   (b) A definition formulated by mathematicians as being the only sensible way to make arithmetic work.
   (c) A logical deduction from suitable axioms.
   (d) Some other explanation.

   Give reasons for your choice and retain your comments for later consideration.

10. In multiplying two numbers together, the order does not matter, $xy = yx$. Can you justify this result
   (a) when x, y are both whole numbers?
   (b) when x, y are any real numbers?
   (c) for any numbers whatever?

CHAPTER 2

Number Systems

The reader will have built up a coherent understanding of the arithmetic of the various number systems: counting numbers, negative numbers, and so on. But he or she may not have subjected the processes of arithmetic to close logical scrutiny. Later, we place these number systems in a precise axiomatic setting. In this chapter we give a brief review of how the reader may have developed their ideas about these systems. Although constant use of the ideas will have smoothed out many of the difficulties that were encountered when the concepts were being formed, these difficulties tend to reappear in the formal treatment and have to be dealt with again. It is therefore worth spending a little time to recall the development, before we plunge into the formalities.

The experienced reader may feel tempted to skip this chapter because of the very simple level of the discussion. Please don't. Every adult's ideas have been built up from simple beginnings as a child. When trying to understand the foundations of mathematics, it is important to be aware of the genesis of your own mathematical thought processes.

Natural Numbers

The natural numbers are the familiar counting numbers $1, 2, 3, 4, 5, \ldots$. Young children learn the names of these, and the order in which they come, by rote. Contact with adults leads the children to an awareness of the meaning that adults attach to phrases such as 'two sweets', 'four marbles'. Use of the word 'zero' and the concept 'no sweets' is more subtle and follows later.

To count a collection of objects, we point to them in turn while reciting 'one, two, three, . . .' until we have pointed to all of the objects, once each.

Next we learn the arithmetic of natural numbers, starting with addition. At this stage the basic 'laws' of addition (which we can express algebraically as the commutative law $a + b = b + a$, and the associative law

$a + (b + c) = (a + b) + c$) may or may not be 'obvious', depending on the approach used. If addition is introduced in terms of combining collections of real-world objects and then counting the result, then these two laws depend only on the tacit assumption that rearranging the collection does not alter the number of things in it. Similarly, one modern approach using coloured rods whose lengths represent the numbers (which are added by placing them end to end) makes commutativity and associativity so obvious that it is almost confusing to have them pointed out. However, if a child is taught addition by 'counting on', the story is quite different. To calculate 3 + 4, he or she starts at 3 and counts on four more places: 4, 5, 6, 7. The calculation 4 + 3 starts at 4 and counts on three places: 5, 6, 7. That the two processes yield the same answer is now much more mysterious. In fact children taught this way often have difficulty doing a calculation such as 1 + 17, but find 17 + 1 trivial!

Next we come to the concept of place-value. The number 33 involves two threes, but they don't mean the same thing. It must be emphasised that this is purely a matter of notation, and has nothing to do with the numbers themselves. But it is a highly useful and important notation. It can represent (in principle) arbitrarily large numbers, and is very well adapted to calculation. However, a precise mathematical description of the general processes of arithmetic in Hindu-Arabic place notation is quite complicated (which is why children take so long to learn them all) and not well adapted to, say, a proof of the commutative law. (This can be done, but it's harder than we might expect.) Sometimes a more primitive system has some advantages. For instance, the ancient Egyptians used the symbol | to represent 1, a hoop ∩ to represent 10, the end of a scroll 𓍢 for 100, with other symbols for 1000, etc. A number was written by repeating these symbols: thus 247 would have been written

𓍢𓍢∩∩∩∩|||||||

Adding in Egyptian is easy: all we do is to put the symbols together. Now the commutative and associative laws are obvious again. But the notation is less suited to computation. To recover place-notation from Egyptian we must supply some 'carrying rules', such as ||||||||||=∩ and insist that we never use any particular symbol more than nine times.

Before proceeding, we introduce a small amount of notation. We write **N** for the set of all natural numbers. The symbol ∈ will mean 'is an element of' or 'belongs to'. So the symbols

$$2 \in \mathbf{N}$$

are read as '2 belongs to the set of natural numbers', or in more usual language, '2 is a natural number'.

Fractions

Fractions are introduced into arithmetic to make division possible. It is easy to divide 12 into 3 parts: 12 = 4 + 4 + 4. It is not possible to divide, say, 11 into 3 equal parts if we insist that these parts are natural numbers. Hence we are led to define fractions as m/n where $m, n \in \mathbf{N}$ and $n \neq 0$. This introduces a new idea, that different fractions such as 2/4 and 3/6 can involve two different processes, where the first divides an object into 4 equal pieces and takes 2 of them to get 2 fourths while the second would divide the object into 6 equal pieces and take 3 to get 3 sixths. The processes are different, but the quantity produced is the same (a half). These fractions are said to be *equivalent*. Equivalent fractions, when marked on a number line, are marked at the same point.

This observation proves to be seminal throughout this book: equivalent concepts at one stage are often reconsidered as single entities later on. In this case equivalent fractions are considered as a single rational number.

Operations of addition and multiplication on the set **F** of fractions can be defined algebraically by the rules

$$\frac{m}{n} + \frac{p}{q} = \frac{mq + np}{nq},$$

$$\frac{m}{n} \times \frac{p}{q} = \frac{mp}{nq}.$$

It is straightforward (but somewhat tedious) to prove that if the fractions are replaced by equivalent fractions, these formulas for the operations yield equivalent results.

Integers

What fractions do for division, integers do for subtraction. A subtraction sum like 2 − 7 = ? cannot be answered in **N**. To do so, we introduce negative numbers. Children are often introduced to negative numbers in terms of a 'number line': a straight line with equally spaced points marked on it. One of them is called 0; then natural numbers 1, 2, 3, . . . are marked successively to the right, and negative numbers −1, −2, −3, . . . to the left.

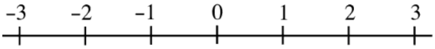


Fig. 2.1 The integers

This gives an extended number system called the 'integers'. An integer is either a natural number n, or a symbol $-n$ where n is a natural number, or 0. We use **Z** to denote the set of integers. (Z is the initial letter of 'Zahlen', the German for integers.)

In your own learning, you met counting numbers **N** before the integers **Z** were introduced. This step is usually motivated by thinking of a negative number as a 'debt'. Then we can see why we have the rule that 'minus times minus makes plus', because taking away a debt has the same result as giving the corresponding credit.

Sometimes in school mathematics, a distinction may initially be made between counting numbers, $1, 2, 3, \ldots$, and positive integers $+1, +2, +3, \ldots$ with their negative counterparts $-1, -2, -3, \ldots$. There are times when this distinction is useful or necessary. Indeed, later we start with counting numbers and show how to construct integers formally. In this process there *is* a difference between the two. However, if we carry on maintaining such distinctions, we will only be making unnecessary work for ourselves. For example, the symbolic statement $4 - (+2)$ (taking away $+2$ from 4) involves a different operation from $4 + (-2)$ (adding -2 to 4). However, it is clearly sensible to say that both equal $4 - 2$.

In the same way, later we start with counting numbers and use set theory to construct integers. This process leads to a different symbolism for counting numbers and positive integers; however, they clearly have the same properties, so it is sensible to think of them as being the same.

In set-theoretic notation, the symbol $\subseteq$ means 'is a subset of'. We then have

$$\mathbf{N} \subseteq \mathbf{Z},$$

where every natural number is also a (positive) integer. Similarly

$$\mathbf{N} \subseteq \mathbf{F}.$$

Rational Numbers

The system **Z** is designed to allow subtraction in all cases; the system **F** allows division (except by zero). However, in neither system are both operations always possible. To get both working at once we move into the system of rational numbers **Q** (for 'quotients'). This is obtained from **F** by introducing 'negative fractions' in much the same way that we obtained **Z** from **N**.

We can still represent **Q** by points on a number line, by marking fractions at suitably spaced intervals between the integers, with negative ones to the left of 0 and positive ones to the right. For example, 4/3 is marked one third of the way between 1 and 2, like this:

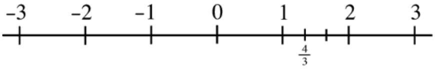


Fig. 2.2 Marking a rational number

The rules for adding and multiplying rational numbers are the same as for fractions, but now m, n, p, q are allowed to be integers rather than natural numbers.

Both **Z** and **F** are subsets of **Q**. We can summarise the relations between the four number systems so far encountered by the diagram:

$$\begin{tikzcd} & \mathbf{F} \arrow[dr, phantom, "\subsetneq" sloped] & \\ \mathbf{N} \arrow[ur, phantom, "\subsetneq" sloped] \arrow[dr, phantom, "\subsetneq" sloped] & & \mathbf{Q} \\ & \mathbf{Z} \arrow[ur, phantom, "\subsetneq" sloped] & \end{tikzcd}$$

Fig. 2.3 Four number systems

Real Numbers

Numbers can be used to measure lengths or other physical quantities. However, the Greeks discovered that there exist lines whose lengths, in theory, cannot be measured exactly by a *rational* number. They were magnificent geometers, and one of their simple but profound results was Pythagoras' theorem. Applied to a right-angled triangle whose two shorter sides have lengths 1, this implies that the hypotenuse has length x, where $x^2 = 1^2 + 1^2 = 2$.

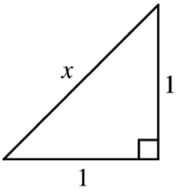


Fig. 2.4 Pythagoras and $\sqrt{2}$

However, x cannot be rational, because there is no rational number m/n such that $(m/n)^2 = 2$. To see why, we use the result that any natural number can be factorised uniquely into primes. For instance, we can write

$$360 = 2 \times 2 \times 2 \times 3 \times 3 \times 5$$

or

$$360 = 5 \times 2 \times 3 \times 2 \times 3 \times 2,$$

but however we write the factors we will always have one 5, two 3s, and three 2s. Using index notation we write

$$360 = 2^3 \times 3^2 \times 5.$$

We shall prove this unique factorisation theorem formally in chapter 8 but for the moment we assume it without further proof.

If we factorise any natural number into primes and then square, each prime will occur an even number of times. For instance,

$$360^2 = (2^3 \times 3^2 \times 5)^2 = 2^6 \times 3^4 \times 5^2,$$

and the indices 6, 4, 2 are all even. A general proof is not hard to find.

Now take any rational number m/n and square it. (Since m/n has the same square as $-m/n$, we may assume m and n positive.) Factorise m^2 and n^2 and cancel factors top and bottom if possible. Whenever a prime p cancels, then since all primes occur to even powers it follows that p^2 cancels. Hence, after cancellation, all primes still occur to even powers. But $(m/n)^2$ is supposed to equal 2, which has one prime (namely 2) which only occurs once (which is an odd power).

It follows that no rational number can have square 2, so the hypotenuse of the given triangle does not have rational length.

With a little more algebraic symbolism we can tidy up this proof and present it as a formal argument, but the above is all that we really need. The same argument shows that numbers like 3, 3/4, or 5/7 do not have rational square roots.

The implication is clear. If we want to talk of lengths like $\surd 2$, we must enlarge our number system further. Not only do we need rational numbers, we need 'irrational' ones as well.

Using Hindu-Arabic notation this can be done by introducing decimal expansions. We construct a right-angled triangle with sides of unit length, and using drawing instruments transfer the length of its hypotenuse to the number line. We then obtain a specific point on the number line that we call $\surd 2$. It lies between 1 and 2 and, on subdividing the unit length from 1 to 2 into ten equal parts, we find that $\surd 2$ lies between 1·4 and 1·5.

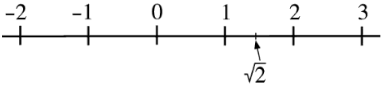


Fig. 2.5 Marking $\surd 2$

By further subdividing the distance between 1·4 and 1·5 into ten equal parts we might hope to obtain a better approximation to $\surd 2$.

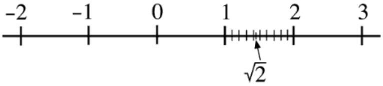


Fig. 2.6 Marking more accurately

Already in a practical situation we are reaching the limit of accuracy in drawing. We might imagine that in an accurate diagram we can look sufficiently close, or magnify the picture, to give the next decimal place. If we were to look at an actual picture under a magnifying glass, not only would the lengths be magnified, but so would the thickness of the lines in the drawing. This would not be a very satisfactory way to obtain a better estimate for $\sqrt{2}$.

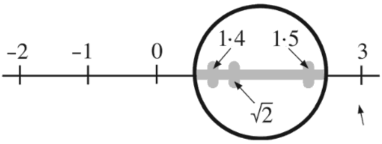


Fig. 2.7 Using a magnifying glass

Practical drawing is in fact extremely limited in accuracy. A fine drawing pen marks a line $0 \cdot 1$ millimetres thick. Even if we use a line 1 metre long as a unit length, since $0 \cdot 1$ mm $= 0 \cdot 0001$ metres, we could not hope to be accurate to more than four decimal places. Using much larger paper and more refined instruments gives surprisingly little increase in accuracy in terms of the number of decimal places we can find. A light year is approximately $9 \cdot 5 \times 10^{15}$ metres. As an extreme case, suppose we consider a unit length 10^{18} metres long. If a light ray started out at one end at the same time that a baby was born at the other, the baby would have to live to be over 100 years old before seeing the light ray. At the lower extreme of vision, the wavelength of red light is approximately 7×10^{-7} metres, so a length of 10^{-7} metres is smaller than the wavelength of visible light. Hence an ordinary optical microscope cannot distinguish points which are 10^{-7} metres apart. On a line where the unit length is 10^{18} metres we cannot distinguish numbers which are less than $10^{-7}/10^{18} = 10^{-25}$ apart. This means that we cannot achieve an accuracy of 25 decimal places by a drawing. Even this is a gross exaggeration in practice, where three or four decimal places is often the best we can really hope for.

Inaccurate Arithmetic in Practical Drawing

The inherent inaccuracy in practice leads to problems in arithmetic. If we add two inaccurate numbers, the errors also add. If we cannot distinguish errors less than some amount e, then we cannot tell the difference, in practice, between a and $a + \frac{3}{4}e$ and between b and $b + \frac{3}{4}e$. But adding, we can distinguish between $a + b$ and $a + b + \frac{3}{2}e$. When we come to multiplication, errors can increase even more dramatically. We cannot hope to get answers to the same degree of accuracy as the numbers used in the calculation.

If we use arithmetic to calculate all answers correct to a certain number of decimal places, the errors involved lead to some disturbing results. Suppose, for example, that we work to two decimal places ('rounding up' if the third place is 5 or more and down if it is less). Given two real numbers a and b, we denote their product correct to two decimal places by $a \otimes b$. For example, $3{\cdot}05 \otimes 4{\cdot}26 = 12{\cdot}99$ because $3{\cdot}05 \times 4{\cdot}26 = 12{\cdot}993$. Using this law of multiplication we find that

$$(1{\cdot}01 \otimes 0{\cdot}5) \otimes 10 \neq 1{\cdot}01 \otimes (0{\cdot}5 \otimes 10).$$

The left-hand side reduces to $0{\cdot}51 \otimes 10 = 5{\cdot}1$, whilst the right-hand side becomes $1{\cdot}01 \otimes 5 = 5{\cdot}05$. This is by no means an isolated example, and it shows that the associative law does not hold for $\otimes$.

If we further define $a \oplus b$ to be the sum correct to two decimal places, we will find other laws that do not hold, including the distributive law

$$a \otimes (b \oplus c) \overset{?}{=} (a \otimes b) \oplus (a \otimes c).$$

A Theoretical Model of the Real Line

We have just seen that if our measurement of numbers is not precise, then some of the laws of arithmetic break down. To avoid this we must make our notion of real number exact.

Suppose we are given a real number x on a theoretical real line, and we try to express it as a decimal expansion. As a starting point, we see that x lies between two integers.

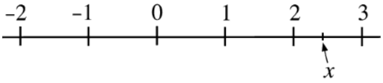


Fig. 2.8 Marking a real number

In the above example x is between 2 and 3, so x is 'two point something'. Next we divide the interval between 2 and 3 into ten equal parts.

Again, x lies in some sub-interval. In the picture, x lies between $2{\cdot}4$ and $2{\cdot}5$, so x is '$2{\cdot}4$ something'. To obtain a still better idea, we divide the interval between $2{\cdot}4$ and $2{\cdot}5$ into ten equal parts and repeat the process to find the next figure in the decimal expansion. Already, in a practical situation, we are reaching the limit of accuracy in drawing.

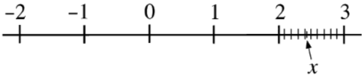


Fig. 2.9 Marking more accurately

For our theoretical picture we must imagine that we can look sufficiently closely, or magnify the picture, to read off the next decimal place. If we looked at an actual picture under a magnifying glass, not only would the lengths be magnified, but so would the thickness of the lines.

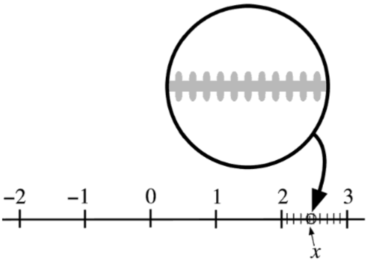


Fig. 2.10 Magnifying

This is not very satisfactory for getting a better estimate. We must, in the theoretical case, assume that the lines have no thickness, so that they are not made wider when the picture is magnified. We can represent this as a practical picture by drawing the magnified lines with the same drawing implements as before, and making them as fine as possible. In this case *x* lies between 2·43 and 2·44, so *x* is '2·44 something'.

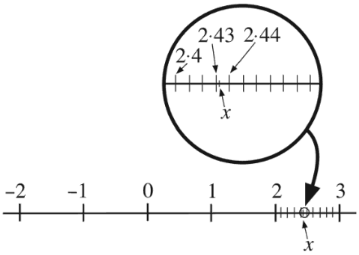


Fig. 2.11 Magnifying more accurately

Using this method we can, in theory, represent any real number as a decimal expansion to as many figures as we require. If we are careful to define this expansion to avoid ambiguity, two numbers will be different if, by calculating sufficiently many terms, we eventually obtain different answers for some decimal place.

We can express this theoretical method as follows in more mathematical terms.

(i) Given a real number x, find an integer a_0 such that

$$a_0 \leq x < a_0 + 1.$$

(ii) Find a whole number a_1 between 0 and 9 inclusive such that

$$a_0 + \frac{a_1}{10} \leq x < a_0 + \frac{a_1 + 1}{10}.$$

(iii) After finding $a_0, a_1, \ldots, a_{n-1}$, where $a_1, \ldots, a_{n-1}$ are integers between 0 and 9 inclusive, find the integer a_n between 0 and 9 inclusive for which

$$a_0 + \frac{a_1}{10} + \cdots + \frac{a_{\dot{n}}}{10^n} \leq x < a_0 + \frac{a_1}{10} + \cdots + \frac{a_n + 1}{10^n}.$$

This gives an inductive process which at the nth stage determines x to n decimal places:

$$a_0 \cdot a_1 a_2 \ldots a_n \leq x < a_0 \cdot a_1 a_2 \ldots a_n + 1/10^n.$$

The theoretically exact representation of the number x requires a decimal expansion

$$a_0 \cdot a_1 a_2 a_3 a_4 a_5 a_6 \ldots$$

that goes on forever. (Of course, if all a_n from some point on are zero, we omit them in normal notation; instead of 1066·31700000000 . . . we write 1066·317.) An infinite decimal is called a *real number*. The set of all real numbers is denoted by **R**.

In most practical situations we will need only a few decimal places. Earlier we saw that 25 decimal places are sufficient for all ratios of lengths within human visual capacity, and that two or three places are usually sufficient for many practical purposes.

Different Decimal Expansions for Different Numbers

If we expand a number x as above in an endless decimal, we say that $a_0 \cdot a_1 a_2 \ldots a_n$ is the expansion of x to the first n decimal places (without 'rounding up'). If two real numbers x and y have the same decimal expansion to n places then

$$a_0 \cdot a_1 \ldots a_n \leq x < a_0 \cdot a_1 \ldots a_n + 1/10^n,$$
$$a_0 \cdot a_1 \ldots a_n \leq y < a_0 \cdot a_1 \ldots a_n + 1/10^n.$$

The second line of inequalities can be rewritten as

$$-a_0 \cdot a_1 \ldots a_n - 1/10^n < -y \leq -a_0 \cdot a_1 \ldots a_n.$$

Adding this to the first line we obtain

$$-1/10^n < x - y < 1/10^n.$$

In other words, if two real numbers have the same decimal expansion to n places, then they differ by at most $1/10^n$.

If x and y are different numbers on the line and we wish to distinguish between them, all we need do is find n such that $1/10^n$ is less than their difference: then their expansion to n places will differ. This again exposes the deficiencies of practical drawing, where x and y might be too close to distinguish. In our theoretical concept of the real line, this distinction must always be possible. It is so important that it is worth giving it a name. The great Greek mathematician Archimedes stated a property that is equivalent to what we want, so we shall name our condition after him:

Archimedes' Condition: Given a positive real number ε, there exists a positive integer n such that $1/10^n < \varepsilon$.

Rationals and Irrationals

As we have seen, the real number $\sqrt{2}$ is irrational: so are many others. It is not always easy to prove a given number irrational. (It's moderately easy for e, less so for π, and there are many interesting numbers which mathematicians have been convinced for centuries are irrational, but have never proved them to be.) But just the fact that $\sqrt{2}$ is irrational implies that between any two rational numbers there exist irrational numbers. First we need:

Lemma 2.1: If m/n and r/s are rational, with $r/s \neq 0$, then $m/n + (r/s)\sqrt{2}$ is irrational.

Proof: Suppose that $m/n + (r/s)\sqrt{2}$ is rational, equal to p/q where p, q are integers. Solve for $\sqrt{2}$ to obtain

$$\sqrt{2} = (pn - mq)s/qnr$$

which is rational, contrary to the irrationality of $\sqrt{2}$. □

Proposition 2.2: Between any two distinct rational numbers there exists an irrational number.

Proof: Let the rational numbers be m/n and r/s, where $m/n < r/s$. Then

$$m/n < m/n + \frac{\sqrt{2}}{2}(r/s - m/n) < r/s$$

(because $\sqrt{2}/2 < 1$), and the number in the middle is irrational by the lemma. □

There is a corresponding result with 'rational' and 'irrational' interchanged:

Proposition 2.3: Between any two distinct irrational numbers there exists a rational number.

Proof: Let the irrational numbers be a, b with $a < b$. Consider their decimal expansions, and let the nth decimal place be the first in which they differ. Then

$$\begin{aligned} a &= a_0 \cdot a_1 \ldots a_{n-1} a_n \ldots, \\ b &= a_0 \cdot a_1 \ldots a_{n-1} b_n \ldots, \end{aligned}$$

where $a_n \neq b_n$. Let $x = a_0 \cdot a_1 \ldots a_{n-1} b_n$. Then x is rational and $a < x \leq b$. But since b is irrational, $x \neq b$, so we must have $a < x < b$. □

In fact, the exercises at the end of this chapter show that the rational and irrational numbers are mixed up in a very complicated way. One should not make the mistake of thinking that they 'alternate' along the real line.

The rational numbers may be characterised as those whose decimal expansions repeat at regular intervals (though we shall omit the proof). To be precise, say that a decimal is repeating if, from some point on, a fixed sequence of digits repeats indefinitely. For example, $1{\cdot}5432174174174174\ldots$ is a repeating decimal. We shall write it as $1{\cdot}5432\dot{1}7\dot{4}$, with dots over the end digits of the block that repeats.

The Need for Real Numbers

The Greeks' belief that all numbers are rational (enshrined in the mystic philosophy of the cult of Pythagoreans) led them to a logical impasse. Viewing the real numbers as infinite decimals helps to overcome this mental block, because it makes it clear that rational numbers, whose expansions repeat, do not exhaust the possibilities.

However, we have also seen that for practical purposes we do not need infinite decimals, nor even very long finite ones. Why go to all the trouble? One reason we have already noted: the arithmetic of decimals of limited length fails to obey the familiar laws which integers and rational numbers obey. A perhaps more serious reason arises in analysis.

Consider the function f given by

$$f(x) = x^2 - 2 \quad (x \in \mathbf{R}).$$

This is negative at $x = 1$, positive at $x = 2$. In between, it is zero at $x = \sqrt{2}$. However, if we restrict x to take only rational values, the function

$$f(x) = x^2 - 2 \quad (x \in \mathbf{Q})$$

is also negative at $x = 1$, positive at $x = 2$, but is not zero at any rational x in between, because $x^2 = 2$ has no rational solution.

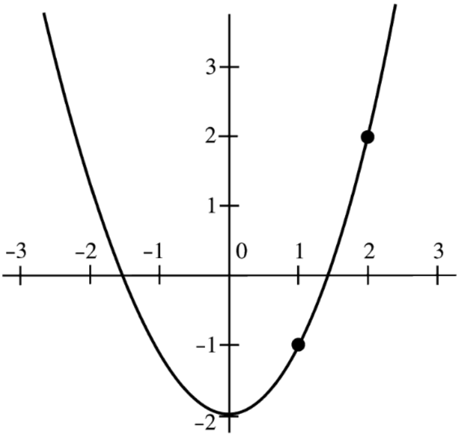


Fig. 2.12 No rational solution

This is a nuisance. A fundamental theorem in analysis asserts that if a continuous function is negative at one point and positive at another, then it must be zero in between. This is true for functions over the real numbers, but not for functions over the rationals. A civilisation such as that of the ancient Greeks, with no satisfactory method for handling irrational numbers, cannot build a theory of limits, or invent calculus.

Arithmetic of Decimals

The idea of infinite decimals representing real numbers is a useful one, but it is not well suited to numerical manipulations, nor to theoretical investigations beyond an elementary level. We add two finite decimals by starting at the right-hand end, but infinite decimals do not have right-hand ends, so there is nowhere to start.

We can instead start at the left-hand end, adding the first decimal places, then the first two, then the first three, and so on. To see what happens, try adding $2/3 = 0{\cdot}6$ and $2/7 = 0{\cdot}\dot{2}8571\dot{4}$ in this way.

$$\begin{array}{lll}
{\cdot}6 & +\,{\cdot}2 & = {\cdot}8 \\
{\cdot}66 & +\,{\cdot}28 & = {\cdot}94 \\
{\cdot}666 & +\,{\cdot}285 & = {\cdot}951 \\
{\cdot}6666 & +\,{\cdot}2857 & = {\cdot}9523 \\
{\cdot}66666 & +\,{\cdot}28571 & = {\cdot}95237 \\
{\cdot}666666 & +\,{\cdot}285714 & = {\cdot}952380.
\end{array}$$

The actual answer is $2/3 + 2/7 = 20/21 = {\cdot}\dot{9}5238\dot{0}$. Notice that adding the first decimal places does not give the answer to one decimal place, nor does adding the first two places give the first two places of the answer. This is precisely because of the possibility of 'carried' digits from later places affecting earlier ones.

However, in this example, successive terms increase and get closer and closer to the actual answer. The sequence of numbers ·8, ·94, ·951, ·9523, . . . is an increasing sequence of real numbers, and it 'tends to' 20/21 in the sense that the error can be made as small as we please by calculating enough decimal places.

In the next few sections we shall examine in detail the ideas required to make this concept precise. For theoretical purposes it is often easier to use increasing sequences (of approximations to a real number) rather than decimal expansions.

Sequences

A *sequence* of real numbers can be thought of as an endless list

$$a_1, a_2, a_3, a_4, \ldots, a_n, \ldots$$

where each term a_n is a real number. (Using set theory we shall give a more formal definition in chapter 5.)

Examples 2.4:

(1) The sequence of squares: $1, 4, 9, 16, \ldots$ where $a_n = n^2$.
(2) The sequence of decimal approximations to $\surd 2$ is 1·4, 1·41, 1·414, . . . where $a_n = \surd 2$ to n places.
(3) The sequence $1, 1\frac{1}{2}, 1\frac{5}{6}, \ldots$ where $a_n = 1 + \frac{1}{2} + \frac{1}{3} + \ldots + \frac{1}{n}$.
(4) The sequence $3, 1, 4, 1, 5, 9, \ldots$ where $a_n =$ the nth digit in the decimal expansion of π.

We often use the shorthand notation

$$(a_n)$$

for the sequence $a_1, a_2, \ldots$, where the nth term is placed in round brackets. Thus example (1) could be written (n^2).

Notice how general the concept of a sequence is. We can consider *any* endless list of numbers. It is not necessary that the nth term be defined by a 'nice formula', as long as we know what each a_n is supposed to be.

Sequences can be added, subtracted, or multiplied. It is necessary to define what we mean by this: the simplest way is to perform the operations on each pair of terms in corresponding positions. In other words, to add the sequences

$$a_1, a_2, \ldots$$

and

$$b_1, b_2, \ldots$$

means to form the sequence

$$a_1 + b_1, a_2 + b_2, \ldots.$$

For example, if $a_n = n^2$ and $b_n = 1 + \frac{1}{2} + \cdots + \frac{1}{n}$, then the nth term of $(a_n) + (b_n)$ is

$$n^2 + 1 + \tfrac{1}{2} + \cdots + \tfrac{1}{n}.$$

Since the nth term of the sequence $(a_n) + (b_n)$ is $a_n + b_n$, we can express the rule for addition as

$$(a_n) + (b_n) = (a_n + b_n).$$

Similarly the rules for subtraction and multiplication are

$$(a_n) - (b_n) = (a_n - b_n),$$
$$(a_n)(b_n) = (a_n b_n).$$

In the case of division we put

$$(a_n)/(b_n) = (a_n/b_n),$$

noting that this division can be carried out only when *all* terms b_n are non-zero.

Example 2.5: If $a_n = \sqrt{2}$ to n decimal places, and $b_n =$ the nth decimal place in π, then the first few terms of $(a_n)(b_n)$ are

$$1{\cdot}4 \times 3 = 4{\cdot}2$$

$$1{\cdot}41 \times 1 = 1{\cdot}41$$

$$1{\cdot}414 \times 4 = 5{\cdot}656$$

$$1{\cdot}4142 \times 1 = 1{\cdot}4142.$$

If you were given the sequence 4·2, 1·41, 5·656, 1·4142, could you have guessed the rule for the nth term? This drives home the point that in order to specify a sequence we must know in principle how to calculate *all* of its terms. In general, it is not enough to write down the first few terms and a few dots. The sequence $3, 1, 4, 1, 5, 9, \ldots$ certainly looks as if it consists of the digits of π. However, it might just as well be the sequence of digits of the number 355/113, which starts off the same way. This is why, in example (4), we specify the general rule for finding the nth term.

Nevertheless, you will often find mathematicians writing things like $2, 4, 8, 16, 32, \ldots$ and expecting you to infer that the nth term is 2^n. One aspect of learning mathematics is to understand how mathematicians actually work, and what their idiosyncrasies are: you should be prepared to accept slight differences in notation provided that the idea is clear from the context.

Order Properties and the Modulus

We digress to introduce an important concept. If x is a real number we define the modulus or absolute value of x to be

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

The graph of $|x|$ against x looks like:

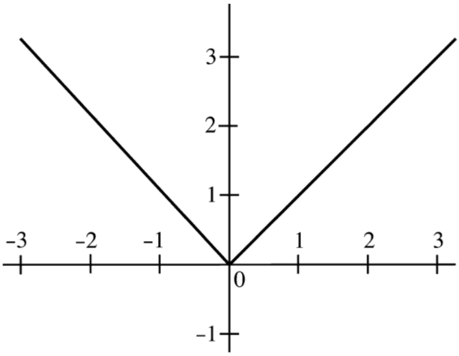


Fig. 2.13 The modulus function

The value of $|x|$ tells us how large or small x is, ignoring whether it is positive or negative. Perhaps the most useful fact about the modulus is the triangle inequality, so called because its generalisation to complex numbers expresses the fact that each side of a triangle is shorter than the other two put together. It is:

Proposition 2.6 (Triangle Inequality): If x and y are real numbers, then

$$|x + y| \leq |x| + |y|.$$

***Proof*:** The visual idea is that $|x+y|$ says how far from the origin $x+y$ is, and this is at most the sum of the distances $|x|$ and $|y|$ of x and y from the origin, being less if x and y have opposite sign. (Draw some pictures to check this.) The easiest way to prove it logically is to divide into cases, according to the signs and relative sizes of x and y.

(i) $x \geq 0, y \geq 0$. Then $x + y \geq 0$, so

$$|x + y| = x + y = |x| + |y|.$$

(ii) $x \geq 0, y < 0$. If $x + y \geq 0$ then

$$|x + y| = x + y < x - y = |x| + |y|.$$

On the other hand, if $x + y < 0$ then

$$|x + y| = -(x + y) = -x - y < |x| + |y|.$$

(iii) $x < 0, y \geq 0$ follows as in case (ii) with x and y interchanged.

(iv) $x < 0, y < 0$. Then $x + y < 0$, so

$$|x + y| = -x - y = |x| + |y|. \qquad \square$$

Be on the lookout for variations on this theme, such as

$$|x - y| + |y - z| \geq |x - z|,$$

which follows since $x - z = (x - y) + (y - z)$, so that $|x - y| + |y - z| \geq |x - z|$.

The modulus is most useful for expressing certain inequalities succinctly. For example,

$$a - \varepsilon < x < a + \varepsilon$$

can be written

$$-\varepsilon < x - a < \varepsilon,$$

which translates into

$$|x - a| < \varepsilon.$$

Convergence

Now we are ready to consider the general notion of representing a real number as a 'limit' of a sequence, rather than just being a particular decimal expansion. As an exercise, the reader should mark, to as large a scale and as accurately as possible, the numbers 1·4, 1·41, 1·414, 1·4142, $\sqrt{2}$, on the interval between 1 and 2.

The numbers 1·4, 1·41, 1·414, 1·4142, get closer and closer together, until they become indistinguishable from each other and from $\sqrt{2}$, up to the accuracy of the drawing. By drawing a more accurate picture we must go further along the sequence of decimal approximations to $\sqrt{2}$ before this happens. If we work to an accuracy of 10^{-8}, then from the eighth term onwards all points of the sequence are indistinguishable from $\sqrt{2}$.

This observation motivates the theoretical concept of convergence. Let ε be any positive real number (ε is the Greek letter epsilon, for 'e', and may be thought of as the initial letter of 'error'). For practical convergence of a sequence (a_n) to a limit l, if we are working to an accuracy ε, we require there to be some natural number N such that the difference between a_n and l has size less than ε when $n > N$. In other words, $|a_n - l| < \varepsilon$. In the following diagram we cannot distinguish points less than ε apart; in this case $N = 7$ and a_n is indistinguishable from l when $n > 7$.

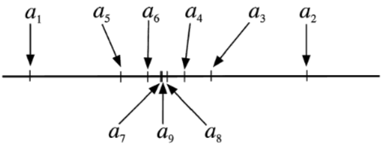


Fig. 2.14 Practical convergence

For theoretical convergence we ask that a similar phenomenon should occur for *all* positive ε. This is on the explicit understanding that smaller values of ε may require larger values of N. In this sense, N is allowed to depend on ε. Thus we reach:

Definition 2.7: A sequence (a_n) of real numbers *tends to a limit l* if, given any $\varepsilon > 0$, there is a natural number N such that

$$|a_n - l| < \varepsilon \text{ for all } n > N.$$

Mathematicians use various pieces of shorthand notation to express this concept. To say 'the sequence (a_n) tends to the limit l' we write

$$\lim_{n \to \infty} a_n = l,$$

or

$$a_n \to l \text{ as } n \to \infty.$$

The symbol '$n \to \infty$' is read as 'n tends to infinity' and is meant to remind us that we are interested in the behaviour of a_n as n becomes large (namely $n > N$ for an appropriately large number N).

The symbol ∞ has historical connotations that can have a variety of different meanings. We will return to these in chapters 14 and 15 to see that ideas that occurred in history and in the minds of growing students can be interpreted formally in very interesting ways. Until then, we will usually refrain from using the symbol and just write

$$\lim a_n = l.$$

Example 2.8: The sequence 1·1, 1·01, 1·001, 1·0001, . . . , for which $a_n = 1 + 10^{-n}$, tends to the limit 1. For, given $\varepsilon > 0$, we have to make

$$|1 + 10^{-n} - 1| < \varepsilon \text{ for } n > N$$

by finding a suitable N. But this follows from Archimedes' condition: if we find N to make $10^{-N} < \varepsilon$, then for all $n > N$ we have $10^{-n} < 10^{-N} < \varepsilon$. (If the theory of logarithms is available, we take $N > \log_{10}(l/\varepsilon)$.)

Definition 2.9: A sequence (a_n) which tends to a limit l is called *convergent.* If no limit exists, it is said to be *divergent.*

A convergent sequence can tend to only one limit. For suppose $a_n \to l$ and $a_n \to m$, where $l \neq m$. Take $\varepsilon = \frac{1}{2}|l - m|$. For large enough n,

$$|a_n - l| < \varepsilon, \ \ |a_n - m| < \varepsilon.$$

From the triangle inequality, $\left|l - m\right| < 2\varepsilon = \left|l - m\right|$, which is not the case.

In other words, if all the terms a_n must eventually be very close to l, they cannot also be very close to m, because this requires them to be in two different places at the same time.

Completeness

Definition 2.10: A sequence (a_n) is *increasing* if each $a_n \leq a_{n+1}$, so that

$$a_1 \leq a_2 \leq a_3 \leq \ \ldots .$$

Suppose that (a_n) is an increasing sequence. Either the terms a_n increase without limit, eventually becoming as large as we please, or else there must be some real number k such that $a_n \leq k$ for all n. An example of a sequence of the first type is 1, 4, 9, 16, 25, . . . ; one of the latter type is the sequence of decimal approximations to e: 2·7, 2·71, 2·718, 2·7182, . . . , every term of which is less than 3.

Definition 2.11: If there exists a real number k such that $a_n < k$ for all n we say that (a_n) is *bounded.*

If we draw the points of a bounded increasing sequence on a part of the real line we need only draw the interval between a_1 and k, since all the other points lie inside this. So a typical picture is:

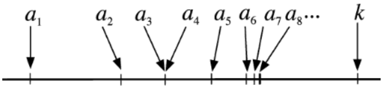


Fig. 2.15 A bounded increasing sequence

It seems visually evident that the terms become increasingly squashed together, and tend to some limit $l \leq k$. This intuition is correct if we consider sequences of real numbers and real limits, but it is wrong for sequences of rational numbers and *rational* limits. In fact the sequence of decimal approximations to $\sqrt{2}$ is an increasing sequence of rational numbers with no rational number as limit.

The fact that every bounded sequence of real numbers tends to a real number as limit as known as the *completeness* property of the real numbers. The origin of the name is that the rational numbers are 'incomplete' because numbers like $\sqrt{2}$ are 'missing'. As we consider a formal approach to the real numbers, we will see this idea in a new light.

We can make the completeness property of the reals very plausible in terms of our ideas about decimals. Let (a_n) be an increasing sequence of real numbers, with $a_n \leq k$ for all k.

The set of integers between $a_1 - 1$ and k is finite, so there is an integer b_0 that is the largest integer for which some term a_{n_0} of the sequence is $\geq b_0$. Now all terms a_n are less than $b_0 + 1$.

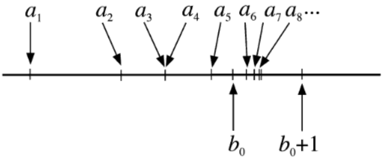


Fig. 2.16 Later terms between successive integers

We subdivide the interval from b_0 to $b_0 + 1$ into ten parts, and find b_1 so that some term $a_{n_1} \geq b_0 + b_1/10$, but no term $a_n \geq b_0 + (b_1+1)/10$. Continuing in this way we get a sequence of decimals

$$b_0, b_0 \cdot b_1, b_0 \cdot b_1 b_2, \ \ldots$$

such that for $n > n_r$ the term a_n lies between $b_0 \cdot b_1 b_2 \ldots b_r$ and $b_0 \cdot b_1 b_2 \ldots b_r + 1/10^r$. Then the real number

$$l = b_0 \cdot b_1 b_2 \ldots$$

has the property that $|a_n - l| < 1/10^r$ for all $n > n_r$. Hence $a_n \to l$ as $n \to \infty$.

It is easy to check that this l is less than or equal to k.

Decreasing Sequences

There is no need to be obsessed with increasing sequences.

Definition 2.12: A sequence (a_n) is *decreasing* if $a_n \geq a_{n+1}$ for all n. If it satisfies $a_n \geq k$ for all n then k is a *lower bound* and the sequence is *bounded below*. (To avoid ambiguity with increasing sequences we can now say 'bounded above' instead of 'bounded'.) There is a similar theorem concerning decreasing sequences, but instead of copying out the proof again and changing the inequalities we use a trick. If (a_n) is decreasing, then $(-a_n)$ is increasing. If $a_n \geq k$ for all n then $-a_n \leq -k$ for all n, so $(-a_n)$ is bounded above, hence tends to a limit l. It follows easily that $a_n \to -l$. Hence any decreasing sequence of real numbers bounded below by k tends to a limit $-l \geq k$.

Different Decimal Expansions for the Same Real Number

Previously we expanded a real number x as an infinite decimal, $x = a_0{\cdot}a_1a_2\ldots$, by using the inequalities

$$a_0 + \frac{a_1}{10} + \cdots + \frac{a_n}{10^n} \leq x < a_0 + \frac{a_1}{10} + \cdots + \frac{a_n + 1}{10^n},$$

where a_0 is an integer and a_n is an integer from 0 to 9 for $n \geq 1$. This condition can be written

$$a_0 \cdot a_1a_2 \ldots a_n \leq x < a_0 \cdot a_1a_2 \ldots a_n + 1/10^n. \tag{2.1}$$

This, used successively for $n = 1, 2, 3, \ldots$, gives a unique decimal expansion for any real number, and different real numbers have different decimal expansions. However, this is not quite the whole story since certain decimal expansions do not occur when we use condition (2.1). For example the expansion $0{\cdot}999999\ldots$, where $a_0 = 0$ and $a_n = 9$ for all $n \geq 1$, does not occur.

Why does this happen? Suppose there were a real number x with decimal expansion (according to (2.1)) $0{\cdot}999999\ldots$. Then

$$0 \cdot 999\ldots9 \leq x < 0 \cdot 999\ldots9 + 1/10^n,$$

where there are n 9s each time. Therefore

$$1 - (1/10^n) \leq x < 1,$$

or

$$0 < 1 - x \leq 1/10^n$$

for all $n \in \mathbf{N}$. But this is impossible by Archimedes' condition: since $1 - x > 0$ there must exist n with $1/10^n < 1 - x$.

The reason why this sequence of 9s cannot occur is our choice of inequalities in (2.1). If instead we use

$$a_0 \cdot a_1 a_2 \ldots a_n < x \leq a_0 \cdot a_1 a_2 \ldots a_n + 1/10^n \tag{2.2}$$

then we get an equally useful definition of the decimal expansion, and it is easy to see that the expansion of the number $x = 1$ now takes the form $0{\cdot}999999\ldots.$

However, the second rule (2.2) will now never give us the expansion $1{\cdot}000000\ldots.$

These are the only possibilities. For example, if a number x has two different decimal expansions, then, without loss in generality, we can take

$$x = a_0 \cdot a_1 \ldots a_{n-1} a_n \ldots = a_0 \cdot a_1 \ldots a_{n-1} b_n \ldots \text{ where } a_n < b_n.$$

Multiply through by 10^n to get

$$a_0 a_1 \ldots a_{n-1} a_n \cdot a_{n+1} \ldots = a_0 a_1 \ldots a_{n-1} b_n \cdot a_{n+1} \ldots \text{ where } a_n < b_n.$$

Subtracting the whole number $a_0 a_1 \ldots a_{n-1} a_n$ gives

$$0 \cdot a_{n+1} \ldots = k \cdot b_{n+1} \ldots \text{ where } k = b_{n+1} - a_{n+1} > 0 \text{ is a positive integer.}$$

But the first decimal is $0 \cdot a_{n+1} \ldots < 0 \cdot 999 \ldots \leq 1$ and the second exceeds the positive integer k. So they can be equal only if $k = 1$ and both decimals represent the same limiting value 1. In this case, $a_{n+1} = a_{n+2} = \cdots = 9$, $b_{n+1} = b_{n+2} = \cdots = 0$ and $b_n = a_n + 1$.

For example, $3{\cdot}14999\ldots$ equals $3{\cdot}15000\ldots.$

This proves that an infinite decimal expansion is unique, *except* when one representation is finite, given by (2.1), and the other ends in an infinite number of 9s, given by (2.2).

It is important not to think that $0{\cdot}99\ldots9\ldots$ is a number 'infinitely smaller' than 1. They are just two different ways of writing the same real number.

It is convenient to allow both notations because under certain circumstances a calculation may give rise to the infinite sequence of 9s. This will happen using the method given earlier to find the decimal expansion of the limit of a bounded increasing sequence.

Example 2.13: Suppose $a_1 = 1$ and in general $a_{n+1} = a_n + (\frac{1}{2})^{n-1}$, then trivially (a_n) is increasing and a calculation gives $a_n = 2 - \left(\frac{1}{2}\right)^{n-1}$, so the sequence is bounded above by 2. Using the same method to calculate the decimal expansion using definition (2.2) instead of (2.1), the limit of the sequence (a_n) is then found to be

$$b_0 \cdot b_1 b_2 \ldots b_n \ldots = 1 \cdot 99 \ldots 9 \ldots .$$

To cover all cases, we introduce the following:

Definition 2.14: The value of an infinite decimal $a_0 \cdot a_1 a_2 \ldots a_n \ldots$ is the limit l of the sequence (d_n) of decimals to n decimal places, where $d_n = a_0 \cdot a_1 a_2 \ldots a_n$.

Using this definition, 0·333 . . . *is* 1/3, and 0·999 . . . *is* 1.

Comment. Research has shown that most people initially believe that 0.999 . . . is 'just less than 1'. The psychological reason seems to be that we think of a sequence (a_n) not as a list of numbers but as a 'variable quantity' that varies as n varies. For example, if $a_n = 1/n$, then we tend to think of the nth term as varying with n and becoming dynamically smaller and smaller. The variable term in this case gets closer and closer to zero, but never equals zero. This dynamic intuition makes us believe that 0·999 . . . is 'just less than one' rather than equal to one. It can lead to resistance to accepting the definition of an infinite decimal being *defined* as the limiting value.

One of us taught an introductory course [5] on convergence using computers for students to investigate the numerical convergence of sequences to get the sense that if a sequence converged, then, to a given number of places, the sequence stabilised onto a fixed value. They were introduced to the idea that the limit was the *precise* value that the sequence stabilised on, leading to the formal definition of the limit l of a sequence (a_n), including the specific example that if $a_n = 1 - 1/10^n$ then the limit l equals 1. Before the course, as expected, 21 out of 23 stated that $0 \cdot \dot{9}$ was just less than one and only two said that it was equal to 1. After the course, the students remained of the same opinion. In a class discussion, the general opinion of the students was that they *knew* that the repeating decimal never reached 1, so trying to *define* it equal to one was not possible.

In order to make sense of formal mathematics, it is essential to get to know the definitions and to be aware of precisely what they say. Only then will it become possible to build up a coherent formal theory. In this case, the limit of a sequence (a_n) is defined to be *the fixed number l* that it approaches, as formulated in the definition.

Bounded Sets

By drawing the picture of a bounded increasing sequence we can actually see the limit process in action, as later terms in the sequence pack together