Preface to first edition

This book arose out of a one-semester course taught over a number of years both at the University of Notre Dame, Indiana, and at the University of Liverpool. The aim of the course is to introduce the concepts of algebra, especially group theory, by many examples and to relate them to some applications, particularly in computer science. The books which we considered for the course seemed to fall into two categories. Some were too elementary, proceeded at too slow a pace and had far from adequate coverage of the topics we wished to include. Others were aimed at a higher level and were more comprehensive, but had correspondingly skimpy presentation of the material. Since we could find no text which presented the material at the right level and in a way we felt appropriate, we prepared our own course notes: this book is the result. We have added some topics which are not always treated in order to increase the flexibility of the book as the basis for a course. The material in the book could be covered at an unhurried pace in about 48 lectures; alternatively, a 36-hour unit could be taught, covering Chapters 1, 2 (not Section 2.4), 4 and 5.

Preface to second edition

We have prepared this second edition bearing in mind the fact that students studying mathematics at university, at least in the UK, are less well prepared than in the past. We have taken more time to explain some points and, in particular, we have not assumed that students are comfortable reading formal statements of theorems and making sense of their proofs. Especially in the first chapter, we have added many comments designed to help the reader make sense of theorems and proofs. The more 'mathematically sophisticated' reader may, of course, read quickly through these comments. We have also added a few more straightforward exercises at the ends of some sections.

Two major changes in content have been made. In Chapter 3 we have removed material (some propositional logic, Boolean algebras and Karnaugh maps) around Boolean algebras. We have retained most of the material on propositional logic, added a section designed to help students deal with quantifiers and added a further section on proof strategies. The emphasis of this chapter is now on the use of logic within mathematics rather than on the Boolean structure behind propositional logic.

The second major change has been the inclusion of a new chapter, Chapter 6, on the algebra of polynomials. We emphasise the similarity with the arithmetic of integers, including the usefulness of the notion of congruence class, and we show how polynomials are used in constructing cyclic codes.

Introduction

'A group is a set endowed with a specified binary operation which is associative and for which there exist an identity element and inverses.' This, in effect, is how many books on group theory begin. Yet this tells us little about groups or why we should study them. In fact, the concept of a group evolved from examples in number theory, algebra and geometry and it has applications in many contexts. Our presentation of group theory in this book reflects to some extent the historical development of the subject. Indeed, the formal definition of an abstract group does not occur until the fourth chapter. We believe that, apart from being more 'honest' than the usual presentation, this approach has definite pedagogic advantages. In particular, the student is not presented with a seemingly unmotivated abstract definition but, rather, sees the sense of the definition in terms of the previously introduced special cases. Moreover, the student will realise that these concepts, which may be so glibly presented, actually evolved slowly over a period of time.

The choice of topics in the book is motivated by the wish to provide a sound, rigorous and historically based introduction to group theory. In the sense that complete proofs are given of the results, we do not depart from tradition. We have, however, tried to avoid the dryness frequently associated with a rigorous approach. We believe that by the overall organisation, the style of presentation and our frequent reference to less traditional topics we have been able to overcome this problem. In pursuit of this aim we have included many examples and have emphasised the historical development of the ideas, both to motivate and to illustrate. The choice of applications is directed more towards 'finite mathematics' and computer science than towards applications arising out of the natural sciences.

Group theory is the central topic of the book but the formal definition of a group does not appear until the fourth chapter, by which time the reader will

have had considerable practice in 'group theory'. Thus we are able to present the idea of a group as a concept that unifies many ideas and examples which the reader already will have met.

One of the objectives of the book is to enable the reader to relate disparate branches of mathematics through 'structure' (in this case group theory) and hence to recognise patterns in mathematical objects. Another objective of the book is to provide the reader with a large number of skills to acquire, such as solving linear congruences, calculating the sign of a permutation and correcting binary codes. The mastery of straightforward clearly defined tasks provides a motivation to understand theorems and also reveals patterns. The text has many worked examples and contains straightforward exercises (as well as more interesting ones) to help the student build this confidence and acquire these skills.

The first chapter of the book gives an account of elementary number theory, with emphasis on the additive and multiplicative properties of sets of congruence classes. In Chapter 2 we introduce the fundamental notions of sets, functions and relations, treating formally ideas that we have already used in an informal way. These fundamental concepts recur throughout the book. We also include a section on finite state machines. Chapter 3 is an introduction to the logic of mathematical reasoning, beginning with a detailed discussion of propositional logic. Then we discuss the use of quantifiers and we also give an overview of some proof strategies. The later chapters do not formally depend on this one. Chapter 4 is the central chapter of the book. We begin with a discussion of permutations as yet another motivation for group theory. The definition of a group is followed by many examples drawn from a variety of areas of mathematics. The elementary theory of groups is presented in Chapter 5, leading up to Lagrange's Theorem and the classification of groups of small order. At the end of Chapter 5, we describe applications to error-detecting and error-correcting codes. Chapter 6 introduces the arithmetic of polynomials, in particular the division algorithm and various results analogous to those in Chapter 1. These ideas are applied in the final section, which depends on Chapter 5, to the construction of cyclic codes.

Every section contains many worked examples and closes with a set of exercises. Some of these are routine, designed to allow the reader to test his or her understanding of the basic ideas and methods; others are more challenging and point the way to further developments.

The dependences between chapters are mostly in terms of examples drawn from earlier material and the development of certain ideas. The main dependences are that Chapter 5 requires Chapter 1 and the early part of Section 4.3

and, also, the examples in Section 4.3 draw on some of Chapter 1 (as well as Sections 4.1 and 4.2).

The material on group theory could be introduced at an early stage but this would not be in the spirit of the book, which emphasises the development of the concept. The formal material of the book could probably be presented in a book of considerably shorter length. We have adopted a more leisurely presentation in the interests of motivation and widening the potential readership.

We have tried to cater for a wide range in ability and degree of preparation in students. We hope that the less well prepared student will find that our exposition is sufficiently clear and detailed. A diligent reader will acquire a sound basic knowledge of a branch of mathematics which is fundamental to many later developments in mathematics. All students should find extra interest and motivation in our relatively historical approach. The better prepared student also should derive long-term benefit from the widening of the material, will discover many challenging exercises and will perhaps be tempted to develop a number of points that we just touch upon. To assist the student who wishes to learn more about a topic, we have made some recommendations for further reading.

Changes in teaching and examining mathematics in secondary schools in the UK have resulted in first-year students of mathematics having rather different skills than in the past. We believe that our approach is well suited to such students. We do not assume a great deal of background yet we do not expect the reader to be an uncritical and passive consumer of information.

One last word: in our examples and exercises we touch on a variety of further developments (for example, normal subgroups and homomorphisms) that could, with a little supplementary material, be introduced explicitly.

Advice to the reader

Mathematics cannot be learned well in a passive way. When you read this book, have paper and pen(cil) to hand: there are bound to be places where you cannot see all the details in your head, so be prepared to stop reading and start writing. Ideally, you should proceed as follows. When you come to the statement of a theorem, pause before reading the proof: do you find the statement of the result plausible? If not, why not? (try to disprove it). If so, then why is it true? How would you set about showing that it is true? Write down a sketch proof if you can: now try to turn that into a detailed proof. Then read the proof we give.

Exercises The exercises at the end of each section are not arranged in order of difficulty, but loosely follow the order of presentation of the topics. It is essential that you should attempt a good portion of these.

Understanding the proofs of the results in this book is very important but so also is doing the exercises. The second-best way to check that you understand a topic is to attempt the exercises. (The best way is to try to explain it to someone else.) It may be quite easy to convince yourself that you understand the material: but attempting the relevant exercises may well expose weak points in your comprehension. You should find that wrestling with the exercises, particularly the more difficult ones, helps you to develop your understanding. You should also find that exercises and proofs illuminate each other.

Proofs Although the emphasis of this text is on examples and applications, we have included proofs of almost all the results that we use. Since students often find difficulty with formal proofs, we will now discuss these at some length. Attitudes towards the need for proofs in mathematics have changed over the centuries.

The first mathematics was concerned with computations using particular numbers, and so the question of proof, as opposed to correctness of a

computation, never arose. Later, however, in arithmetic and geometry, people saw patterns and relationships that appeared to hold irrespective of the particular numbers or dimensions involved, so they began to make general assertions about numbers and geometrical figures. But then a problem arose: how may one be certain of the truth of a general assertion? One may make a general statement, say about numbers, and check that it is true for various particular cases, but this does not imply that it is always true.

To illustrate. You may already have been told that every positive integer greater than 1 is a product of primes, for instance $12 = 2 \cdot 2 \cdot 3$, $35 = 5 \cdot 7$, and so on. But since there are infinitely many positive integers it is impossible, by considering each number in turn, to check the truth of the assertion for every positive integer. So we have the assertion: 'every positive integer greater than 1 is a product of primes'. The evidence of particular examples backs up this assertion, but how can we be justifiably certain that it is true?

Well, we may give a proof of the assertion. A proof is a sequence of logically justified steps which takes us from what we already know to be true to what we suspect (and, after a proof has been found, know) to be true.

It is unreasonable to expect to conjure something from nothing, so we do need to make some assumptions to begin with (and we should also be clear about what we mean by a valid logical deduction). In the case of the assertion above, all we need to assume are the ordinary arithmetic properties of the integers, and the principle of induction (see Section 1.2 for the latter). It is also necessary to have defined precisely the terms that we use, so we need a clear definition of what is meant by 'prime'. We may then build on these foundations and construct a proof of the assertion. (We give one on p. 28.)

It should be understood that current mathematics employs a very rigorous standard of what constitutes a valid proof. Certainly what passed for a proof in earlier centuries would often not stand up to present-day criteria. There are many good reasons for employing such strict criteria but there are some drawbacks, particularly for the student.

A formal proof is something that is constructed 'after the event'. When a mathematician proves a result he or she will almost certainly have some 'picture' of what is going on. This 'picture' may have suggested the result in the first place and probably guided attempts to find a proof. In writing down a formal proof, however, it often is the case that the original insight is lost, or at least becomes embedded in an obscuring mass of detail.

Therefore one should not try to read proofs in a naive way. Some proofs are merely verifications in which one 'follows one's nose', but you will probably be able to recognise such a proof when you come across one and find no great

trouble with it – provided that you have the relevant definitions clear in your mind and have understood what is being assumed and what is to be proved. But there are other proofs where you may find that, even if you can follow the individual steps, you have no overview of the structure or direction of the proof. You may feel rather discouraged to find yourself in this situation, but the first thing to bear in mind is that you probably will understand the proof sometime, if not now, then later. You should also bear in mind that there is some insight or idea behind the proof, even if it is obscured. You should therefore try to gain an overview of the proof: first of all, be clear in your mind about what is being assumed and what is to be proved. Then try to identify the key points in the proof – there are no recipes for this, indeed even experienced mathematicians may find difficulty in sorting out proofs that are not well presented, but with practice you will find the process easier.

If you still find that you cannot see what is 'going on' in the proof, you may find it helpful to go through the proof for particular cases (say replacing letters with numbers if that is appropriate). It is often useful to ignore the given proof (or even not to read it in the first place) but to think how *you* would try to prove the result – you may well find that your idea is essentially the same as that behind the proof given (or is even better!).

In any event, do not allow yourself to become 'stuck' at a proof. If you have made a serious attempt to understand it, but to little avail, then *go on*: read through what comes next, try the examples, and maybe when you come back to the proof (and you should make a point of coming back to it) you will wonder why you found any difficulty. Remember that if you can do the 'routine' examples then you are getting something out of this text: understanding (the ideas behind) the proofs will deepen your understanding and allow you to tackle less routine and more interesting problems.

Background assumed We have tried to minimise the prerequisites for successfully using this book. In theory it would be enough to be familiar with just the basic arithmetic and order-related properties of the integers, but a reader with no more preparation than this would, no doubt, find the going rather tough to begin with. The reader that we had in mind when writing this book has also seen a bit about sets and functions, knows a little elementary algebra and geometry, and does know how to add and multiply matrices. A few examples and exercises refer to more advanced topics such as vectors, but these may safely be omitted.

1 Number theory

This chapter is concerned with the properties of the **set of integers** $\{\ldots, -2, -1, 0, 1, 2, \ldots\}$ under the arithmetic operations of addition and multiplication. We shall usually denote the set of integers by $\mathbb{Z}$. We shall assume that you are acquainted with the elementary arithmetical properties of the integers. By the end of this chapter you should be able to solve the following problems.

1. What are the last two digits of 3^{1000}?
2. Can every integer be written as an integral linear combination of 197 and 63?
3. Show that there are no integers x such that $x^5 - 3x^2 + 2x - 1 = 0$.
4. Find the smallest number which when divided by 3 leaves 2, by 5 leaves 3 and by 7 leaves 2. (This problem appears in *Sūn tzǐ suàn jīng* (*Master Sun's Arithmetical Manual*) which was written around the fourth century.)
5. How may a code be constructed which allows anyone to encode messages and send them over public channels, yet only the intended recipient is able to decode the messages?

1.1 The division algorithm and greatest common divisors

We will assume that the reader is acquainted with the elementary properties of the order relation '$\leq$' on the set $\mathbb{Z}$. This is the relation 'less than or equal to' which allows us to compare any two integers. Recall that, for example, $-100 \leq 2$ and $3 \leq 3$. The following property of the set $\mathbb{P} = \{1, 2, \ldots\}$ of **positive integers** is important enough to warrant a special name.

Well-ordering principle Any non-empty set, X, of positive integers has a smallest element (meaning an element which is less than or equal to every member of the set X).

You are no doubt already aware of this principle. Indeed you may wonder why we feel it necessary to state the principle at all, since it is so 'obvious'. It is, however, as you will see, a key ingredient in many proofs in this chapter. An equivalent statement is that one cannot have an unending, strictly decreasing, sequence of positive integers.

Note that the principle remains valid if we replace the set of positive integers by the set $\mathbb{N} = \{0, 1, 2, \ldots\}$ of **natural numbers**. But the principle fails if we replace $\mathbb{P}$ by the set, $\mathbb{Z}$, of all integers or, for a different kind of reason, if we replace $\mathbb{P}$ by the set of positive rational numbers (you should stop to think why). We use $\mathbb{Q}$ to denote the set of all **rational numbers** (fractions).

A typical use of the well-ordering principle has the following shape. We have a set X of positive integers which, for some reason, we know is non-empty (that is, contains at least one element). The principle allows us to say 'Let k be the least element of X'. You will see the well-ordering principle in action in this section.

The well-ordering principle is essentially equivalent to the method of proof by mathematical induction. That method of proof may take some time to get used to if it is unfamiliar to you, so we postpone mathematical induction until the next section.

The proof of the first result, Theorem 1.1.1, in this section is a good example of an application of the well-ordering principle. Look at the statement of the result now. It may or may not be obvious to you what the theorem is 'really saying'. Mathematical statements, such as the statement of 1.1.1, are typically both general and concise. That makes for efficient communication but a statement which is concise needs thought and time to draw out its meaning and, when faced with a statement which is general, one should always make the effort (in this context, by plugging in particular values) to see what it means in particular cases.

In this instance we will lead you through this process but it is something that you should learn to do for yourself (you will find many opportunities for practice as you work through the book).

The first sentence, 'Let a and b be natural numbers with $a > 0$', invites you to choose two natural numbers, call one of them a and the other b, but make sure that the first is strictly positive. We might choose $a = 175$, $b = 11$.

The second sentence says that there are natural numbers, which we will write as q and r, such that $0 \leq r < a$ and $b = aq + r$. The first statement, $0 \leq r < a$, says that r is strictly smaller than a (the '$0 \leq r$' is redundant since any natural number has to be greater than or equal to 0, it is just there for emphasis).

The second statement says that b is an integer multiple of a, plus r.

With our choice of numbers the second statement becomes: 'There are natural numbers q, r such that $0 \leq r < 175$ and $11 = 175q + r$'. In other words, we can write 11 as a non-negative multiple of 175, plus a non-negative number which is strictly smaller than 175. But that is obvious: take $q = 0$ and $r = 11$ to get $11 = 175 \cdot 0 + 11$.

You would be correct in thinking that there is more to 1.1.1 than is indicated by this example! You might notice that 1.1.1 says more if we take $b > a$. So let us try with the values reversed, $a = 11$, $b = 175$. Then 1.1.1 says that there are natural numbers q, r with $r < 11$ such that $175 = 11q + r$. How can we find such numbers q, r? Simply divide 175 by 11 to get a quotient (q) and remainder (r): $175 = 11 \cdot 15 + 10$, that is $q = 15$, $r = 10$.

So the statement of 1.1.1 is simply an expression of the fact that, given a pair of positive integers, one may divide the first into the second to get a quotient and a remainder (where we insist that the remainder is as small as possible, that is, strictly less than the first number).

Now you should read through the proof to see if it makes sense. As with the statement of the result we will discuss (after the proof) how you can approach such a proof in order to understand it: in order to see 'what is going on' in the proof.

Theorem 1.1.1 (Division Theorem) *Let a and b be natural numbers with $a > 0$. Then there are natural numbers q, r with $0 \leq r < a$ such that:*

$$b = aq + r$$

(*r is the* **remainder**, *q the* **quotient** *of b by a*).

Proof If $a > b$ then just take $q = 0$ and $r = b$. So we may as well suppose that $a \leq b$.

Consider the set of non-negative differences between b and integer multiples of a:

$D = \{b - ak : b - ak \geq 0 \text{ and } k \text{ is a natural number}\}$.

(If this set-theoretic notation is unfamilar to you then look at the beginning of Section 2.1.)

This set, D, is non-empty since it contains $b = b - a \cdot 0$. So, by the well-ordering principle, D contains a least element $r = b - aq$ (say). If r were not strictly less than a then we would have $r - a \geq 0$, and therefore

$$r - a = (b - aq) - a = b - a(q + 1).$$

So $r - a$ would be a member of D strictly less than r, contradicting the minimality of r.

Hence r does satisfy $0 \leq r < a$; and so r and q are as in the statement of the theorem. □

For example, if $a = 3$ and $b = 7$ we obtain $q = 2$ and $r = 1$: we have $7 = 3 \cdot 2 + 1$. If $a = 4$ and $b = 12$ we have $q = 3$ and $r = 0$: that is $12 = 4 \cdot 3 + 0$.

The symbol '□' above marks the end of a proof.

Comments on the proof Let us pull the above proof apart in order to see how it works.

You might recognise the content of the first sentence from the discussion before 1.1.1: it is saying that if $a > b$ then there is nothing (much) to do – we saw an example of that when we made the choice $a = 175$, $b = 11$. The next sentence says that we can concentrate on the main case where $a \leq b$.

The next stage, the introduction of the set D, certainly needs explanation. Before you read a proof of any statement you should (make sure you understand the statement! and) think how you might try to prove the statement yourself. In this case it is not so obvious how to proceed: you know how to divide any one number into another in order to get a quotient and a remainder, but trying to express this formally so that you can prove that it always works could be quite messy (though it is possible). The proof above is actually a very clever one: by focussing on a well chosen set it cuts through any messy complications and gives a short, elegant path to the end. So to understand the proof we need to understand what is in the set D.

Now, one way of finding q and r is to subtract integer multiples of a from b until we reach the smallest possible non-negative value. The definition of the set D is based on that idea. That definition says that the typical element of D is a number of the form $b - ak$, that is, b minus an integer multiple of a (well, in the definition k is supposed to be non-negative but that is not essential: we are after the *smallest* member of D and allowing k to be negative will not affect that). In other words, D is the set of non-negative integers which may be obtained by subtracting a non-negative multiple of a from b (so, in our example, D would contain numbers including 175 and $98 = 175 - 7 \cdot 11$).

What we then want to do is choose the least element of D, because that will be a number of the form $b - ak$ which is the smallest possible (without dropping to a negative number). The well-ordering principle guarantees that D, a set of natural numbers, has a smallest element, but only if we first check that D has at least one element. But that is obvious: b itself is in D.

So now we have our least element in D and, in anticipation of the last line of the proof, we write it as r. Of course, being a member of D it has the form $r = b - aq$ for some q (again, in anticipation of how the remainder of the

proof will go, we write q for this particular value of what we wrote as 'k' in the definition of D).

Rearranging the equation $r = b - aq$ we certainly have $b = aq + r$ so all that is left is to show that $0 \leq r < a$. We chose r to be in D and it is part of the definition of D that all its elements should be non-negative so we do have $0 \leq r$. All that remains is to show $r < a$.

The last part of the proof is an example of what is called 'proof by contradiction' (we discuss this technique below). We want to prove $r < a$ so we say, suppose not – then $r \geq a$ – but in that case we could subtract a at least once more from r and still have a number of the form $b - ak$ which is non-negative. Such a number would be an element of D but strictly smaller than r and that contradicts our choice of r as the smallest element of D. The conclusion is that we do, indeed, have $r < a$ and, with that, the proof is finished.

Proof by contradiction Suppose that we want to prove a statement. Either it is true or it is false. What we can do is suppose that it is false and then see where that leads us: if it leads us to something that is wrong then we must have started out by supposing something that is wrong. In other words, the supposition that the statement is false must be wrong. Therefore the original statement must be true.

For instance, suppose that we want to prove that there is no largest integer. Well, either that is correct or else there *is* a largest integer. So let us suppose for a moment that there is a largest integer n say. But then $n + 1$ is an integer which is larger than n, a contradiction (to n being the largest integer). So supposing that there is a largest integer leads to a contradiction and must, therefore, be false. In other words, there is no largest integer.

Definition Given two integers a and b, we say that a **divides** b (written '$a \mid b$') if there is an integer k such that $ak = b$.

For example, $7 \mid 42$ but 7 does not divide 40, we write $7 \nmid 40$ (it is true that $40/7$ makes sense as a rational number but here we are working in the integers and insist that k in the definition should be an integer: positive, negative or 0).

Thus a divides b exactly if, with notation as in Theorem 1.1.1, $r = 0$.

Note that this definition has the consequence (take $k = 0$) that every integer divides 0.

Another idea with which you are probably familiar is that of the greatest common divisor (also called highest common factor) of two integers a and b. Usually this is described as being 'the largest integer which divides both a and b'. In fact, it is not only 'the largest' in the sense that every other common divisor of a and b is less than it: it is even the case that every common divisor of a and b *divides* it.

This is essentially what the next theorem says. The proof should be surprising: it proves an important property of greatest common divisor that you may not have come across before, a property which we extract in Corollary 1.1.3.

Theorem 1.1.2 *Given positive integers a and b, there is a positive integer d such that*

(i) *d divides a and d divides b, and*
(ii) *if c is a positive integer which divides both a and b then c divides d* (*that is, any common divisor of a and b must divide d*).

Proof Let D be the set of all positive integers of the form $as + bt$ where s and t vary over the set of *all* integers:

$$D = \{as + bt : s \text{ and } t \text{ are integers and } as + bt > 0\}.$$

Since $a(a = a \cdot 1 + b \cdot 0)$ is in D, we know that D is not empty and so, by the well-ordering principle, D has a least element d, say. Since d is in D there are integers s and t such that

$$d = as + bt.$$

We have to show that any common divisor c of a and b is a divisor of d. So suppose that c divides a, say $a = cg$, and that c divides b, say $b = ch$. Then c divides the right-hand side $(cgs + cht)$ of the above equation and so c divides d. This checks condition (ii).

We also have to check that d does divide both a and b, that is we have to check condition (i). We will show that d divides a since the proof that d divides b is similar (a and b are interchangable throughout the statement and proof so 'by symmetry' it is enough to check this for one of them). Applying Theorem 1.1.1 to 'divide d into a', we may write

$$a = dq + r \text{ with } 0 \leq r < d.$$

We must show that $r = 0$. We have

$$\begin{aligned} r &= a - dq \\ &= a - (as + bt)q \\ &= a(1 - sq) + b(-tq). \end{aligned}$$

Therefore, if r were positive it would be in D. But d was chosen to be minimal in D and r is strictly less than d. Hence r cannot be in D, and so r cannot be positive. Therefore r is zero, and hence d does, indeed, divide a. □

Comment Note the structure of the last part of the proof above. We chose d to be minimal in the set D and then essentially said, 'The remainder r is an integer combination of a and b so, if it is not zero, it must be in the set D. But d was supposed to be the *least* member of D and $r < d$. So the only possibility is that $r = 0$.' There is a definite similarity to the end of the proof of 1.1.1.

Given any a and b as in 1.1.2, we claim that there is just one positive integer d which satisfies the conditions (i) and (ii) of the theorem. For, suppose that a positive integer e also satisfies these conditions. Applying condition (i) to e we have that e divides both a and b; so, by condition (ii) applied to d and with e in place of 'c' there, we deduce that e divides d. Similarly (the situation is symmetric in d and e) we may deduce that d divides e. So we have two integers, d and e, and each divides the other: that can only happen if each is $\pm$ the other. But both d and e are positive, so the only possibility is that $e = d$, as claimed.

Note the strategy of the argument in the paragraph above. We want to show that there is just one thing satisfying certain conditions. What we do is to take two such things (but allowing the *possibility* that they are equal) and then show (using the conditions they satisfy) that they *must* be equal.

Definition The integer d satisfying conditions (i) and (ii) of the theorem is called the **greatest common divisor** or **gcd** of a and b and is denoted (a, b) or $\gcd(a, b)$. Some prefer to call (a, b) the **highest common factor** or **hcf** of a and b. Note that, just from the definition, $(a, b) = (b, a)$.

For example, $(8, 12) = 4$, $(3, 21) = 3$, $(4, 15) = 1$, $(250, 486) = 2$.

Note It follows easily from the definition that if a divides b then the gcd of a and b is a. For instance $\gcd(6, 30) = 6$.

The proof of 1.1.2 actually showed the following very important property (you should go back and check this).

Corollary 1.1.3 *Let a and b be positive integers. Then the greatest common divisor, d, of a and b is the smallest positive integral linear combination of a and b.* (*By an* **integral linear combination** *of a and b we mean an integer of the form $as + bt$ where s and t are integers.*) *That is, $d = as + bt$ for some integers s and t.*

For instance, the gcd of 12 and 30 is 6: we have $6 = 30 \cdot 1 - 12 \cdot 2$. In Section 1.5 we give a method for calculating the gcd of any two positive integers.

We make some comment on what might be unfamiliar terminology. A 'Corollary' is supposed to be a statement that follows from another. So often, after a Theorem or a Proposition (a statement which, for whatever reason, is judged by the authors to be not quite as noteworthy as a Theorem) there might be one or more Corollaries. In the case above it was really a corollary of the proof, rather than the statement, of 1.1.2. The term 'Lemma', used below, indicates a result which we prove on the way to establishing something more notable (a Proposition or even a Theorem).

Before stating the next main theorem we give a preliminary result.

Lemma 1.1.4 *Let a and b be natural numbers and suppose that a is non-zero. Suppose that*

$b = aq + r$ *with q and r positive integers.*

Then the gcd of b and a is equal to the gcd of a and r.

Proof Let d be the gcd of a and b. Since d divides both a and b, d divides the (term on the) right-hand side of the equation $r = b - aq$: hence d divides the left-hand side, that is, d divides r. So d is a common divisor of a and r. Therefore, by definition of (a, r), d divides (a, r).

Similarly, since the gcd (a, r) divides a and r and since $b = aq + r$, (a, r) must divide b. So (a, r) is a common divisor of a and b and hence, by definition of $d = (a, b)$, it must be that (a, r) divides d.

It has been shown that d and (a, r) are positive integers which divide each other. Hence they are equal, as required. □

Discussion of proof of 1.1.4 Sometimes, if the structure of a proof is not clear to you, it can help to go through it with some or all 'x's and 'y's (or in this case, a and b) replaced by particular values. We illustrate this by going through the proof above with particular values for a and b.

Let us take $a = 30$, $b = 171$. In the statement of 1.1.4 we write b in the form $aq + r$, that is, we write 171 in the form $30q + r$. Let us take $q = 5$ so $r = 21$ and the equation in the statement of the lemma is $171 = 30 \cdot 5 + 21$ (but we do not have to take the form with smallest remainder r, we could have taken say $q = 3$ and $r = 81$, the conclusion of the lemma will still be true with those choices).

The proof begins by assigning d to be (30, 171). Then (says the proof) d divides both 30 and 171 so it divides the right-hand side of the rearranged equation $21 = 171 - 30 \cdot 5$ hence d divides the left-hand side, that is d divides 21. So d is a common divisor of 30 and 21. Therefore, by definition of the gcd (30, 21) it must be that d divides (30, 21).

Similarly, since (30, 21) divides both 30 and 21 and since $171 = 30 \cdot 5 + 21$ it must be that (30, 21) divides 171 and so is a common divisor of 30 and 171. Therefore, by definition of $d = (30, 171)$ we must have that (30, 21) divides d.

Therefore d and (30, 21) are positive integers which divide each other. The conclusion is that they must be equal: $(30, 171) = (30, 21)$. (Of course, you can compute the actual values of the gcd to check this but the point is that you do not need to do the computation to know that they are equal. In fact, the lemma that we have just proved is the basis of the practical method for computing greatest common divisors, so to say that we do not need this lemma because we can always compute the values completely misses the point!)

The next result appears in Euclid's *Elements* (Book VII Propositions 1 and 2) and so goes back as far as 300 BC. The proof here is essentially that given in Euclid (it also appears in the Chinese *Jiŭ zhāng sùan shù* (*Nine Chapters on the Mathematical Art*) which was written no later than the first century AD). Observe that the proof uses 1.1.1, and hence depends on the well-ordering principle (which was used in the proof of 1.1.1). Indeed it also uses the well-ordering principle directly. The (very useful) 1.1.3 is not explicit in Euclid.

Theorem 1.1.5 (Euclidean algorithm) *Let a and b be positive integers. If a divides b then a is the greatest common divisor of a and b. Otherwise, applying* 1.1.1 *repeatedly, define a sequence of positive integers $r_1, r_2, \ldots, r_n$ by*

$$\begin{aligned}
b &= aq_1 + r_1 \qquad (0 < r_1 < a),\\
a &= r_1q_2 + r_2 \qquad (0 < r_2 < r_1),\\
&\vdots\\
r_{n-2} &= r_{n-1}q_n + r_n \qquad (0 < r_n < r_{n-1}),\\
r_{n-1} &= r_nq_{n+1}.
\end{aligned}$$

Then r_n is the greatest common divisor of a and b.

Proof Apply Theorem 1.1.1, writing $r_1, r_2, \ldots, r_n$ for the successive *non-zero* remainders. Since $a, r_1, r_2, \ldots$ is a decreasing sequence of positive integers, it must eventually stop, terminating with an integer r_n which, because no non-zero remainder 'r_{n+1}' is produced must, therefore, divide r_{n-1}. Then, applying 1.1.4 to the second-to-last equation gives $(r_{n-2}, r_{n-1}) = (r_{n-1}, r_n)$ which, we have just observed, is r_n. Repeated application of Lemma 1.1.4, working back through the equations, shows that r_n is the greatest common divisor of a and b. □

Example Take $a = 30, b = 171$.

$$\begin{aligned} 171 &= 5 \cdot 30 + 21 \qquad \text{so } r_1 = 21 \qquad \text{and } (171, 30) = (30, 21); \\ 30 &= 21 + 9 \qquad \text{so } r_2 = 9 \qquad \text{and } (30, 21) = (21, 9); \\ 21 &= 2 \cdot 9 + 3 \qquad \text{so } r_3 = 3 \qquad \text{and } (21, 9) = (9, 3); \\ 9 &= 3 \cdot 3. \end{aligned}$$

Hence

$$(171, 30) = (30, 21) = (21, 9) = (9, 3) = 3.$$

If we wish to write the gcd in the form $171s + 30t$, we can use the above equations to 'solve' for the remainders as follows.

$$\begin{aligned} 3 &= 21 - 2 \cdot 9 \\ &= 21 - 2(30 - 21) \\ &= 3 \cdot 21 - 2 \cdot 30 \\ &= 3(171 - 5 \cdot 30) - 2 \cdot 30 \\ &= 3 \cdot 171 - 17 \cdot 30. \end{aligned}$$

The calculation may be conveniently arranged in a matrix format.

To find (a, b) as a linear combination of a and b, set up the partitioned matrix

$$\left(\begin{array}{cc|c} 1 & 0 & b \\ 0 & 1 & a \end{array}\right)$$

(this may be thought of as representing the equations: '$x = b$' and '$y = a$'). Set $b = aq_1 + r_1$ with $0 \leq r_1 < a$. If $r_1 = 0$ then we may stop since then $a = (a, b)$. If r_1 is non-zero, subtract q_1 times the bottom row from the top row to get (noting that $b - aq_1 = r_1$)

$$\left(\begin{array}{cc|c} 1 & -q_1 & r_1 \\ 0 & 1 & a \end{array}\right).$$

Now write $a = r_1q_2 + r_2$ with $0 \leq r_2 < r_1$. We may stop if $r_2 = 0$ since r_1 is then the gcd of a and r_1, and hence by 1.1.4 is the gcd of a and b. Furthermore, the row of the matrix which contains r_1 allows us to read off r_1 as a combination of a and b: namely $1 \cdot b + (-q_1) \cdot a = r_1$.

If r_2 is non-zero then we continue. Thus, if at some stage one of the rows is

$$n_i \quad m_i \mid r_i \quad (*)$$

representing the equation

$$bn_i + am_i = r_i,$$

and if the other row reads

$$n_{i+1} \quad m_{i+1} \mid r_{i+1} \quad (**)$$

then we set

$$r_i = r_{i+1}q_{i+2} + r_{i+2} \text{ with } 0 \leq r_{i+2} < r_{i+1}$$

and we subtract q_{i+2} times the second of these rows from the first and replace $(*)$ with the result.

Observe that these operations reduce the size of the (non-negative) numbers in the right-hand column, and so eventually the process will stop. When it stops we will have the gcd: moreover if the row containing the gcd reads

$$n \quad m \mid d$$

then we have the expression,

$$bn + am = d,$$

of d as an integral linear combination of a and b.

Example 1 We repeat the above example in matrix form: so $a = 30$ and $b = 171$.

$$\left(\begin{array}{cc|c} 1 & 0 & 171 \\ 0 & 1 & 30 \end{array}\right) \to \left(\begin{array}{cc|c} 1 & -5 & 21 \\ 0 & 1 & 30 \end{array}\right) \to \left(\begin{array}{cc|c} 1 & -5 & 21 \\ -1 & 6 & 9 \end{array}\right)$$
$$\to \left(\begin{array}{cc|c} 3 & -17 & 3 \\ -1 & 6 & 9 \end{array}\right) \to \left(\begin{array}{cc|c} 3 & -17 & 3 \\ -10 & 57 & 0 \end{array}\right).$$

So $(171, 30) = 3 = 3 \cdot 171 - 17 \cdot 30$.

Example 2 Take b to be 507 and a to be 391.

$$\left(\begin{array}{cc|c} 1 & 0 & 507 \\ 0 & 1 & 391 \end{array}\right) \to \left(\begin{array}{cc|c} 1 & -1 & 116 \\ 0 & 1 & 391 \end{array}\right) \to \left(\begin{array}{cc|c} 1 & -1 & 116 \\ -3 & 4 & 43 \end{array}\right)$$
$$\to \left(\begin{array}{cc|c} 7 & -9 & 30 \\ -3 & 4 & 43 \end{array}\right) \to \left(\begin{array}{cc|c} 7 & -9 & 30 \\ -10 & 13 & 13 \end{array}\right)$$
$$\to \left(\begin{array}{cc|c} 25 & -35 & 4 \\ -10 & 13 & 13 \end{array}\right) \to \left(\begin{array}{cc|c} 25 & -35 & 4 \\ -91 & 118 & 1 \end{array}\right)$$

$(507, 391) = 1 = -91 \cdot 507 + 118 \cdot 391$.

You may use whichever method you prefer for calculating gcds: the methods are essentially the same and it is only in the order of the calculations that they differ. The advantages of the matrix method are that there is less to write down and, at any stage, the calculation can be checked for correctness, since a row $u \quad v \mid w$ represents the equation $bu + av = w$. A disadvantage is that one has to put more reliance on mental arithmetic. Therefore it is especially important that, after finishing a calculation like those above, you should check the correctness of the final equation as a safeguard against errors in arithmetic.

A good exercise (if you have the necessary background) is to write a program (in pseudocode) which, given any two positive integers, finds their gcd as an integral linear combination. If you attempt this exercise you will find that any gaps in your understanding of the method will be highlighted.

The definition of greatest common divisor may be extended as follows.

Definition Let $a_1, \ldots, a_n$ be positive integers. Then their **greatest common divisor** $(a_1, \ldots, a_n)$, also written $\gcd(a_1, \ldots, a_n)$, is the positive integer m with the property that $m \mid a_i$ for each i and, whenever c is an integer with $c \mid a_i$ for each i, we have $c \mid m$.

This exists, and can be calculated, by using the case $n = 2$ 'and induction'. We discuss induction at length in the next section but here we will give a somewhat informal indication of how it is used.

We claim that $(a_1, \ldots, a_n) = ((a_1, \ldots, a_{n-1}), a_n)$. In other words, we can compute the gcd of n numbers, $a_1, \ldots, a_n$ by computing the gcd of the first $n - 1$ of them and then computing the gcd of *that* and the last number a_n. As for computing the gcd of $a_1, \ldots, a_{n-1}$ we compute *that* by computing the gcd of the first $n - 2$ numbers and then computing the gcd of that with a_{n-1}. Etc. So, in the end, all we need is to be able to compute the gcd of *two* numbers. Here is an example.

Suppose that we wish to compute $(24, 60, 30, 8)$. We claim that this is equal to $((24, 60, 30), 8)$ and that this is equal to $(((24, 60), 30), 8)$. Now we do some arithmetic and find that $(24, 60) = 12$, then $(12, 30) = 6$ and then $(6, 8) = 2$, so we conclude $(24, 60, 30, 8) = 2$.

After you have read the section on induction you can try, as an exercise, to give a formal proof that, for all n and integers $a_1, \ldots, a_n$, we have $(a_1, \ldots, a_n) = ((a_1, \ldots, a_{n-1}), a_n)$.

Definition Two positive integers a and b are said to be **relatively prime** (or **coprime**) if their greatest common divisor is 1: $(a, b) = 1$. Example 2 above shows that 507 and 391 are relatively prime.

We now give some properties of relatively prime integers. You are probably aware of these properties though you may not have seen them stated formally. A special case of (i) below is the deduction that since 15 and 8 are relatively prime and since 15 divides $8 \cdot 30 = 240$ it must be that 15 divides 30. A special case of (ii) is that since 15 and 8 are relatively prime and since 15 divides 360 and 8 divides 360 we must have that $15 \cdot 8 = 120$ divides 360. Perhaps by giving numerical values to a, b, and c in this way it all seems rather obvious but, beware: neither (i) nor (ii) is true without the assumption that a and b are relatively prime. If we were to replace 15 and 8 by, say 6 and 9, the statements (i) and (ii) would be false for some values of c. See Exercises 1.1.4 and 1.1.5 at the end of the section.

Theorem 1.1.6 *Let a, b, c be positive integers with a and b relatively prime. Then*

(i) *if a divides bc then a divides c,*
(ii) *if a divides c and b divides c then ab divides c.*

Proof (i) Since a and b are relatively prime there are, by 1.1.3, integers r and s such that

$$1 = ar + bs.$$

Multiply both sides of this equation by c to get

$$c = car + cbs. \qquad (*)$$

Since a divides bc, it divides the right-hand side of the equation and hence divides c.

(ii) With the above notation, consider equation $(*)$. Since a divides c, ab divides cbs and, since b divides c, ab divides car. Thus ab divides c as required. □

Comment Note how using 1.1.3 gives a beautifully simple argument – surely not the argument one would first think of trying.

The results of this section may be extended in fairly obvious ways to include negative integers. For example, to apply Theorem 1.1.1 with b negative and a positive it makes best sense to demand that the remainder 'r' still satisfy the inequality $0 \leq r < a$. This means that in order to divide the negative number b by a we do *not* simply divide the positive number $-b$ by a and then put a minus sign in front of everything.

Example To divide -9 by 4: find the multiple of 4 which is just below -9 (that is $-12 = 4(-3)$) and then write:

$$-9 = 4(-3) + 3,$$

noting that the remainder 3 satisfies $0 \leq 3 < 4$. (If we wrote $-9 = 4(-2) + -1$, then the remainder -1 would not satisfy the inequality $0 \leq r < 4$.)

So remember that the remainder should always be positive or zero.

A similar remark applies to Theorem 1.1.5: we require that the greatest common divisor always be positive.

Example The greatest common divisor of -24 and -102 equals the greatest common divisor of 24 and 102. To express it as a linear combination of -24 and -102, either we use the matrix method or we proceed as follows, remembering that remainders must always be non-negative:

$$\begin{aligned} -102 &= -24 \cdot 5 + 18 \\ -24 &= 18(-2) + 12 \\ 18 &= 12 \cdot 1 + 6 \\ 12 &= 6 \cdot 2. \end{aligned}$$

Hence the gcd of -24 and -102 is 6 and 6 is $-1(-102) + 4(-24)$.

To conclude this section, we note that there is the notion of **least common multiple** or **lcm** of integers a and b. This is defined to be the positive integer m such that both a and b divide m (so m is a common multiple of a and b), and such that m divides every common multiple of a and b. It is denoted by $\text{lcm}(a, b)$. The proof that such an integer m does exist, and is unique, is left as an exercise.

More generally, given non-zero integers $a_1, \ldots, a_n$, we define their **least common multiple**, $\text{lcm}(a_1, \ldots, a_n)$, to be the (unique) positive integer m which satisfies $a_i | m$ for all i and, whenever an integer c satisfies $a_i \mid c$ for all i, we have $m \mid c$.

For instance $\text{lcm}(6, 15, 4) = \text{lcm}(\text{lcm}(6, 15), 4) = \text{lcm}(30, 4) = 60$.

We shall see in Section 1.3 how to interpret both the greatest common divisor and the least common multiple of integers a and b in terms of the decomposition of a and b as products of primes.

All of the concepts and most of the results of this section are to be found in the *Elements* of Euclid (who flourished around 300 BC). Euclid's origins are unknown but he was one of the scholars called to the Museum of Alexandria. The Museum was a centre of scholarship and research established by Ptolemy, a general of Alexander the Great, who, after the latter's death in 323 BC, gained control of the Egyptian part of the empire.

The *Elements* probably was a textbook covering all the elementary mathematics of the time. It was not the first such 'elements' but its success was such that it drove its predecessors into oblivion. It is not known how much of the mathematics of the *Elements* originated with Euclid: perhaps he added no new results; but the organisation, the attention to rigour and, no doubt, some of the proofs, were his. It is generally thought that the algebra in Euclid originated considerably earlier.

No original manuscript of the *Elements* survives, and modern editions have been reconstructed from various recensions (revised editions) and commentaries by other authors.

Exercises 1.1

1. For each of the following pairs a, b of integers, find the greatest common divisor d of a and b and express d in the form $ar + bs$:
   (i) $a = 7$ and $b = 11$; (ii) $a = -28$ and $b = -63$;
   (iii) $a = 91$ and $b = 126$; (iv) $a = 630$ and $b = 132$;
   (v) $a = 7245$ and $b = 4784$; (vi) $a = 6499$ and $b = 4288$.
2. Find the gcd of 6, 14 and 21 and express it in the form $6r + 14s + 21t$ for some integers r, s and t.
   [Hint: compute the gcd of two numbers at a time.]
3. Let a and b be relatively prime integers and let k be any integer. Show that b and $a + bk$ are relatively prime.
4. Give an example of integers a, b and c such that a divides bc but a divides neither b nor c.
5. Give an example of integers a, b and c such that a divides c and b divides c but ab does not divide c.
6. Show that if $(a, c) = 1 = (b, c)$ then $(ab, c) = 1$ (this is Proposition 24 of Book VII in Euclid's *Elements*).
7. Explain how to measure 8 units of water using only two jugs, one of which holds precisely 12 units, the other holding precisely 17 units of water.

1.2 Mathematical induction

We can regard the positive integers as having been constructed in the following way. Start with the number 1. Then add 1 to get 2. Then add 1 to get 3. Add 1 again to get 4. And so on. That is, we start with a certain base, 1, and then, again and again without ending, we add 1. In this way we generate the positive integers. A construction of this sort (begin with a base case then apply a process again and again) is described as an **inductive construction**. Here is another example.

Define a sequence of integers as follows. Set $a_1 = 2$, define a_{n+1} inductively by the formula $a_{n+1} = 2a_n + 1$. So $a_2 = 2a_1 + 1 = 2 \cdot 2 + 1 = 5$, $a_3 = 2a_2 + 1 = 2 \cdot 5 + 1 = 11$, $a_4 = 2a_3 + 1 = 2 \cdot 11 + 1 = 23$, and so on. You might notice that if we add 1 to any of the numbers that we have generated so far we obtain a multiple of 3. You might check a few more values and see that this still seems to be a property of the numbers generated in this sequence. But how can we *prove* that this holds for every number in the sequence? Obviously we cannot check each one, because the sequence never ends. What we can do is use a proof by induction. Essentially this is a proof that uses the way that the sequence is generated by a base number together with a rule ($a_{n+1} = 2a_n + 1$) which is applied again and again. At the base case we can just compute: adding 1 to a_1 gives $1 + 2 = 3$ which is certainly divisible by 3. At the 'inductive step', where we go from a_n to a_{n+1}, we argue as follows. Suppose that we know that $a_n + 1$ is a multiple of 3, say $a_n + 1 = 3k$. Then $a_{n+1} + 1 = (2a_n + 1) + 1 = 2a_n + 2 = 2(a_n + 1) = 2(3k)$: which is certainly a multiple of 3. It follows that for every n it is true that $a_n + 1$ is a multiple of 3.

Use of the induction principle can take very complicated forms but, at base, is the fact that the positive integers are constructed by starting somewhere and then applying a 'rule' again and again.

Here is a very abstract statement of the induction principle. In the statement, '$P(n)$' is any mathematical assertion involving the positive integer n (think of 'n' as standing for an integer variable, as in the assertion '$\frac{1}{2} + \frac{1}{3} + \cdots + \frac{1}{n}$ is not an integer').

Induction principle Let $P(n)$ be an assertion involving the positive integer variable n. If

(a) $P(1)$ holds and
(b) whenever $P(k)$ holds so also does $P(k + 1)$

then $P(n)$ holds for every positive integer n.

That is, if we can prove the 'base case' at $n = 1$ and then, if we have an argument that proves the $k + 1$ case from the k case, then we have the result for all positive integers.

The typical structure of a proof by induction is as follows.

Base case – show $P(1)$;

Induction step – assume that $P(k)$ holds (this assumption is the **induction hypothesis**) and *deduce* that $P(k + 1)$ follows.

Then the **conclusion** (by the induction principle) is that $P(n)$ holds for all positive integers n.

Example Show that the sum $1 + 2 + \cdots + n$ of the first n positive integers is $n(n + 1)/2$.

This may be proved using the induction principle: for the assertion $P(n)$ we take '$1 + 2 + \cdots + n = n(n + 1)/2$'.

First, the *base case* holds because when $n = 1$ the left-hand side and right-hand side of the formula are both equal to 1: so the formula is valid for $n = 1$.

For the *induction step*, the induction hypothesis, $P(k)$, is that

$$1 + 2 + \cdots + k = k(k + 1)/2$$

(so we *assume* this, and have to *prove* $P(k + 1)$).

The statement $P(k + 1)$ concerns the sum of the first $k + 1$ positive integers, so let us try writing down this sum and then using the above equation to replace the sum of the first k terms. We get

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

Simplifying the right-hand side gives

$$k(k + 1)/2 + (k + 1) = (k + 1)(k/2 + 1) = (k + 1)(k + 2)/2.$$

Thus we have deduced

$$1 + 2 + \cdots + k + (k + 1) = (k + 1)(k + 2)/2 \qquad (= (k + 1)\{(k + 1) + 1\}/2)$$

which is the required assertion, $P(k + 1)$.

It follows by induction that the formula is valid for every $n \geq 1$.

Example For each positive integer n let

$$a_n = 4^{2n-1} + 3^{n+1}.$$

We will show that, for all positive integers n, a_n is divisible by 13.

For the proposition $P(n)$ we take: 'a_n is divisible by 13.'

The *base case* is $n = 1$. In that case a_n equals $4 + 9 = 13$, which certainly is divisible by 13.

For the *induction step* we assume the induction hypothesis, that $4^{2k-1} + 3^{k+1}$ is divisible by 13: so $4^{2k-1} + 3^{k+1} = 13r$ for some integer r. We must deduce that 13 divides

$$4^{2(k+1)-1} + 3^{(k+1)+1} = 4^{2k+1} + 3^{k+2}.$$

To see this we note that

$$\begin{aligned} 4^{2k+1} + 3^{k+2} &= 4^2(4^{2k-1}) + 3(3^{k+1}) \\ &= 16(4^{2k-1} + 3^{k+1}) - 16(3^{k+1}) + 3(3^{k+1}) \\ &= 16(13r) - 13(3^{k+1}). \end{aligned}$$

It is clear that 13 divides the right-hand side of this expression and so 13 divides $4^{2k+1} + 3^{k+2}$, as required.

It follows by induction that $4^{2n-1} + 3^{n+1}$ is divisible by 13 for every positive integer n.

We should be clear about the following point. Induction is a *form* of argument: if we want to use it then we have to assume $P(k)$ and try to prove $P(k+1)$, but induction does not tell us *how* to do that. In the examples we have given above, we just had to rearrange equations a bit: but it is not always so easy!

We say a bit more about **definition by induction** (sometimes termed **definition by recursion**). This is even used, for example, in defining the positive powers of an integer a. Informally one says: '$a^1 = a, a^2 = a \cdot a, a^3 = a \cdot a \cdot a$, and so on'. More formally, one proceeds by setting $a^1 = a$ (the 'base case') and then inductively defining $a^{k+1} = a^k \cdot a$ (think of a^k as being already defined). Another example of this occurs in defining the factorial symbol $n!$ Here $0!$ is defined to be 1 and, inductively, $(n+1)!$ is defined to be $(n+1) \times n!$ (Thus $4!$ is $4 \cdot 3! = 4 \cdot 3 \cdot 2! = 4 \cdot 3 \cdot 2 \cdot 1! = 4 \cdot 3 \cdot 2 \cdot 1 \cdot 0! = 4 \cdot 3 \cdot 2 \cdot 1 \cdot 1 = 24$.) An informally presented definition by induction is usually signalled by use of '...' or a phrase such as 'and so on'. For other examples of definition by induction see Exercises 1.2.3 and 1.2.9.

As another example of proof by induction, we establish the binomial theorem, 1.2.1 below. Supposing that you have not seen this before, stated in this generality, how can you make sense of the statement of the theorem which, at first sight, might look rather complicated?

Try substituting in values: in this case giving values to all of n, x and y would probably obscure what is being said (remember that one reason for using letters to stand for numbers is that it allows clearer statements!). We will leave x and y as variables but try out giving values to n.

For $n = 1$ you should check that the statement becomes $(x+y)^1 = 1 \cdot x^1 + 1 \cdot y^1$, not very exciting!

For $n = 2$ we obtain $(x+y)^2 = 1 \cdot x^2 + 2 \cdot x^1 y^1 + 1 \cdot y^2$ and for $n = 3$ the statement becomes $(x+y)^3 = 1 \cdot x^3 + 3 \cdot x^2 y^1 + 3 \cdot x^1 y^2 + 1 \cdot y^3$. You should write out the corresponding statements for $n = 4$ and $n = 5$. This

exhibits the theorem as a very general statement covering some familiar special cases. You might also notice that the coefficients occurring are those seen in Pascal's Triangle, the first few rows of which are shown below.

$$\begin{array}{ccccccc} & & & 1 & & & \\ & & 1 & & 1 & & \\ & 1 & & 2 & & 1 & \\ 1 & & 3 & & 3 & & 1 \\ 1 & & 4 & & 6 & & 4 & & 1 \end{array}$$

Pascal's Triangle is formed by adding pairs of adjacent numbers in one row to give the numbers in the next row: you should look out for where that rule occurs in the proof.

Theorem 1.2.1 *Let n be a positive integer and let x, y be any numbers. Then*

$$(x+y)^n = \binom{n}{0} x^n + \cdots + \binom{n}{i} x^{n-i} y^i + \cdots + \binom{n}{n} y^n$$

where for $0 \leq k \leq n$, $\binom{n}{k}$ is defined to be $\frac{n!}{k!(n-k)!}$ (and is known as a binomial coefficient).

Proof Observe that, for any $n \geq 1$, $\binom{n}{n} = \frac{n!}{n!0!} = 1$ and $\binom{n}{0} = \frac{n!}{0!n!} = 1$. For the base case, $n = 1$, the theorem asserts that $(x+y)^1 = \binom{1}{0}x^1 + \binom{1}{1}y^1$ which, by the observation just made, is true. Now suppose that the result holds for $n = k$ (induction hypothesis). Then, using the induction hypothesis, we have

$$\begin{aligned} (x+y)^{k+1} &= (x+y)(x+y)^k \\ &= (x+y)\left(\binom{k}{0} x^k + \binom{k}{1} x^{k-1}y^1 + \cdots + \binom{k}{k-1} x^1 y^{k-1} + \binom{k}{k} y^k\right). \end{aligned}$$

When we multiply this out, the term involving x^{k+1} is

$$\binom{k}{0} x^{k+1} = x^{k+1} = \binom{k+1}{0} x^{k+1},$$

and that involving y^{k+1} is

$$\binom{k}{k} y^{k+1} = y^{k+1} = \binom{k+1}{k+1} y^{k+1}.$$

The term involving $x^{k+1-i}y^i$ $(1 \leq i \leq k)$ is obtained as the sum of two terms,

namely

$$x\binom{k}{i}x^{k-i}y^i + y\binom{k}{i-1}x^{k-(i-1)}y^{i-1}.$$

This simplifies to

$$\left(\binom{k}{i} + \binom{k}{i-1}\right)x^{k+1-1}y^i.$$

We must show that the coefficient, $\binom{k}{i} + \binom{k}{i-1}$, of $x^{k+1-i}y^i$ is $\binom{k+1}{i}$. We have

$$\begin{aligned}
\binom{k}{i} + \binom{k}{i-1} &= \frac{k!}{i!(k-1)!} + \frac{k!}{(i-1)!(k-(i-1))!} \\
&= \frac{k!}{i!(k-i)!} + \frac{k!}{(i-1)!(k-i+1)!} \\
&= \frac{k!}{i\cdot(i-1)!(k-i)!} + \frac{k!}{(i-1)!(k-i+1)\cdot(k-i)!} \\
&= \frac{k!}{(i-1)!(k-i)!}\left(\frac{1}{i} + \frac{1}{k-i+1}\right) \\
&= \frac{k!}{(i-1)!(k-i)!}\cdot\frac{k+1}{i\cdot(k-i+1)} \\
&= \frac{(k+1)\cdot k!}{i\cdot(i-1)!\cdot(k-i+1)\cdot(k-i)!} = \frac{(k+1)!}{i!(k+1-i)!} \\
&= \binom{k+1}{i}
\end{aligned}$$

as required. □

(The rule for forming Pascal's Triangle was the last part of the proof, where we showed that $\binom{k}{i} + \binom{k}{i-1} = \binom{k+1}{i}$.)

Next we show that the principle of mathematical induction may be deduced from the well-ordering principle. There is no harm in skipping this proof in your first reading.

Theorem 1.2.2 *The well-ordering principle implies the principle of mathematical induction.*

Proof Suppose that the assertion $P(n)$ satisfies the conditions for the induction principle: so $P(1)$ holds and whenever $P(k)$ holds, so also does $P(k+1)$. Let S be the set of positive integers m for which $P(m)$ is false. There are two cases:

either S is the empty set (that is the set with no elements) or else S is non-empty. We will see that the second case leads to a contradiction.

If S is not empty then we can apply the well-ordering principle, to deduce that S has a least element, which we call t. Since $P(1)$ holds we know that 1 is not in S and so t must be greater than 1. Hence $t - 1$ is positive. The definition of t as the least element of S implies that $P(t - 1)$ does hold. Since $t = (t - 1) + 1$ it follows, by our assumption on P (take $k = t - 1$), that $P(t)$ holds. This is a contradiction to the fact that t is in S.

Thus the hypothesis that S is non-empty allows us to derive a contradiction, and so it must be the case that S is the empty set. In other words, $P(n)$ is true for every positive integer n. □

Comment Note that this is essentially a proof by contradiction: we set S to be the set of positive integers where $P(n)$ is false; we showed that, if S is non-empty, then one can derive a contradiction; so we concluded that S must be empty, in other words, we concluded that $P(n)$ is true for all $n > 0$.

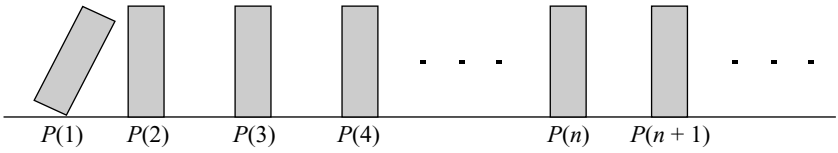
In fact, the converse of the above result is also true: the well-ordering principle can be deduced from the principle of mathematical induction. So the two principles are logically equivalent. We do not need this fact but we indicate its proof in Exercise 1.2.10.

There are some useful variations of the induction principle: let $P(n)$ be an assertion as before.

(a) If P holds for an integer n_0 and if, for every integer $k \geq n_0$, $P(k)$ implies $P(k + 1)$, then P holds for all integers $k \geq n_0$.
(b) If $P(0)$ holds and if, for each $k \geq 0$, from the hypothesis that P holds for all non-negative $m \leq k$ one may deduce that $P(k + 1)$ holds, then $P(n)$ holds for every natural number n.

The first variation simply says that the induction need not start at $n = 1$: for example, it may be appropriate to start with the base case being at $n = 0$.

The second of these variations is known variously as **strong induction**, **complete induction** or **course of values induction** and is a very commonly used form of the induction principle (of course the '0' in its statement could as well be replaced by any integer 'n_0' as in (a) and the conclusion would be modified accordingly). Several examples of its use will occur later (in the proof of 1.3.3, for example). This variation takes note of the fact that, if in an induction we have reached the stage where we have that $P(k)$ is true then, in getting there, we also showed that $P(k - 1)$, $P(k - 2)$, ... (down to the base case) are true, so it is legitimate to use all this information (not just the fact that $P(k)$ is true) in trying to prove that $P(k + 1)$ holds.



Base Case
$n = 1$
(first domino falls over)

Induction Step
(The dominoes are close enough that if the nth falls over then it knocks the $(n + 1)$st over also)

Fig. 1.1

Fig. 1.1 sometimes is used to illustrate the idea of proof by induction.

Imagine a straight line of dominoes all standing on end: these correspond to the integers. They are sufficiently close together that if any one of them falls, then it will knock over the domino next to it: that corresponds to the induction step (from $P(k)$ we get $P(k + 1)$). One of the dominoes is pushed over: that corresponds to the base case. Now imagine what happens.

In these terms, the principle of strong induction says that to knock over the $(k + 1)$st domino we are not restricted to using just the force of the kth domino: we can also, if we can, use the fact that *all* the previous dominoes have fallen over.

The well-ordering principle was explicitly recognised long before the principle of induction.

Since the well-ordering principle expresses an 'obvious' property of the positive integers, one would not expect to see it stated until there was some recognition of the need for mathematical assertions to be backed up by proofs from more or less clearly stated axioms. There is a perfectly explicit statement of it in Euclid's *Elements*. It is not, however, stated as one of his axioms but, rather, is presented as an obvious fact in the course of one of the proofs (Book VII, proof of Proposition 31, our Theorem 1.3.3), which is by no means the first proof in the *Elements* where it is used (e.g. it is implicit in the proofs of Propositions 1 and 2 in Book VII: our 1.1.4 and 1.1.5). In Euclid, the principle is of course applied to the set of positive numbers rather than to the set of natural numbers, for it was to be many centuries before zero would be recognised as a number (especially in Europe).

There are instances in Euclid's *Elements* of something approaching proof by induction, though not in a form that would be recognised today as correct.

It is not unusual for a student new to the idea of proof by induction to 'prove' that $P(n)$ is true, by just checking it for the first few values of n and

then claiming that it necessarily holds also for all greater values of n. In fact, up into the seventeenth century this was not an uncommon method of 'proof'. For example, Wallis in his *Arithmetica Infinitorum* of 1655–6 made much use of such procedures, and he was heavily criticised (in 1657) by Fermat for doing so. By 1636 Fermat had used the principle of induction in a way we would now regard as valid, and Blaise Pascal in his *Triangle Arithmétique* of 1653 spells out the details of a proof by induction. As Fermat points out in his criticism of Wallis' methods, one may manufacture an assertion ($P(n)$) which is true for small values (of n) but which fails at some large value (also see Exercise 1.2.11 below). Actually, many (but not all!) of these early 'proofs' by induction are easily modified to give rigorous proofs because, although their authors used particular numbers, their arguments often apply equally well to an arbitrary positive integer.

Exercises 1.2

1. Define a sequence a_n $(n \geq 1)$ of integers by $a_1 = 1$, $a_{n+1} = 2a_n + 1$ for $n \geq 2$. Compute the values of a_i for $i = 1, \ldots, 5$. Prove by induction that for all $n \geq 1$, $a_n + 1$ is a power of 2.
2. Prove that for all positive integers n,

$$1 + 4 + \cdots + n^2 = n(n+1)(2n+1)/6 = \frac{1}{3}n^3 + \frac{1}{2}n^2 + \frac{1}{6}n.$$

3. The **Fibonacci sequence** is the sequence 1, 1, 2, 3, 5, 8, 13,... where each term is the sum of the two preceding terms. Show that every two successive terms of the Fibonacci sequence are relatively prime.
   [Hint: write down an explicit definition (by induction) of this sequence.]
4. We saw in the text that the sum of the first n positive integers is given by a quadratic (degree 2) polynomial in n. From Exercise 1.2.2 above you see that the sum of the squares of the first n positive integers is given by a polynomial in n of degree 3. Given the information that the formula for the sum of the cubes of the first n positive integers is given by a polynomial in n of degree 4, find this polynomial. [Hint: suppose that the polynomial is of the form $an^4 + bn^3 + cn^2 + dn + e$ for certain constants $a, \ldots, e$, then express the sum of the first $n + 1$ cubes in two different ways.]
5. Prove that for all positive integers n,

$$\frac{1}{3} + \frac{1}{15} + \cdots + \frac{1}{(2n-1)(2n+1)} = \frac{n}{2n+1}.$$

6. Find a formula for the sum of the first n odd positive integers.

7. Prove that if x is not equal to 1 and n is any positive integer then
$$1 + x + x^2 + \cdots + x^n = \frac{1 - x^{n+1}}{1 - x}.$$
8. (i) Show that, for every positive integer n, $n^5 - n$ is divisible by 5.
(ii) Show that, for every positive integer n, $3^{2n} - 1$ is divisible by 8.
9. Given that $x_0 = 2$, $x_1 = 5$ and
$$x_{n+2} = 5x_{n+1} - 3x_n$$
for n greater than or equal to 0, prove that
$$2^n x_n = (5 + \sqrt{13})^n + (5 - \sqrt{13})^n$$
for every natural number n.
10. Show that the principle of induction implies the well-ordering principle. [Hint: let X be a set of positive integers which contains no least element; we must show that X is empty. Define L to be the set of all positive integers, n, such that n is not greater than or equal to any element in X. Show by induction that L is the set of all positive integers, and hence that X is indeed empty.]
11. Consider the assertion: $(*)$ 'for every prime number n, $2^n - 1$ is a prime number' (a positive integer is prime if it cannot be written as a product of two strictly smaller positive integers). Taking n to be 2, 3, 5 in turn, the corresponding values of $2^n - 1$ are 3, 7, 31 and these certainly are prime. Is $(*)$ true? (Also see Exercise 1.3.6.)
12. The following arguments purport to be proofs by induction: are they valid?
(a) This argument shows that all people have the same height. More formally, it is shown that if X is any set of people then each person in X has the same height as every person in X. The proof is by induction on n the number of people in X.
Base case $n = 1$. This is clear, since if the set X contains just one person then that person certainly has the same height as him / herself.
Induction step. We assume that the result is true for every set of $k - 1$ people (the induction hypothesis), and deduce that it is true for any set X containing exactly k people.
Choose any person a in X, and let Y be the set X with a removed. Then Y contains exactly $k - 1$ people who, by the induction hypothesis, must all be of the same height: h metres, say.
Choose any other person b (say) in X, and let Z be the original set X with b removed. Since Z has just $k - 1$ people, the induction hypothesis applies, to give that the people in Z all have the same height, let us say k metres.
Now let c be any person in X other than a or b. Since b and c both are in Y,

each is h metres tall. Since a and c both are in Z, each is k metres tall. So (consider the height of c) $h = k$. But that means that a and b are of the same height.

Therefore, since a and b were arbitrary members of X, it follows that the people in X all have the same height. Thus the induction step is complete, and so the initial assertion follows by induction.

(b) To establish the formula $1 + 2 + \cdots + n = \frac{n^2}{2} + \frac{n}{2} + 1$.

Assume inductively that the formula holds for $n = k$; thus

$$1 + 2 + \cdots + k = \frac{k^2}{2} + \frac{k}{2} + 1.$$

Add $k + 1$ to each side to obtain

$$1 + 2 + \cdots + k + (k + 1) = \frac{k^2}{2} + \frac{k}{2} + 1 + (k + 1).$$

The term on the left-hand side is $1 + 2 + \cdots + (k + 1)$, and the term on the right-hand side is easily seen to be equal to

$$\frac{(k + 1)^2}{2} + \frac{k + 1}{2} + 1.$$

Thus the induction step has been established and so the formula is correct for all values of n.

1.3 Primes and the Unique Factorisation Theorem

Definition A positive integer p is **prime** if p has exactly two positive divisors, namely 1 and p.

Thus, for example, 5 is prime since its only positive divisors are itself and 1, whereas 4 is not prime since it is divisible by 1, 4 and 2.

Notes (i) The definition implies that 1 is not prime since it does not have two distinct positive divisors.

(ii) The smallest prime number is therefore 2 and this is the only even prime number, since any other even positive integer n has at least three distinct divisors (namely 1, 2 and n).

(iii) We may begin listing the primes in ascending order:

2, 3, 5, 7, 11, 13, 17, 19, 23, 29, ...

If one wishes to continue this list beyond the first few primes, then it is not very efficient to check each number in turn for **primality** (the property of being prime). A fairly efficient, and very old, method for generating the list of primes is the Sieve of Eratosthenes, described below.

Eratosthenes of Cyrene (c.280–c.194 BC) is probably more widely known for his estimate of the size of the earth: he obtained a circumference of 250 000 stades (believed to be about 25 000 miles); the actual value varies between 24 860 and 24 902 miles.

The Sieve of Eratosthenes To find the primes less than some number n, prepare an array of the integers from 2 to n. Save 2 and then delete all multiples of 2. Now look for the next undeleted integer (which will be 3), save it and delete all its multiples. The smallest undeleted number will be the next prime, 5. Continue in this way to find all the primes up to n. In fact, it will turn out that you can stop this process once you have reached the greatest integer which is less than or equal to the square root of n, in the sense that any integers left undeleted at this stage will be prime. (You will be asked to think about this in Exercise 1.3.2 at the end of the section.)

As an exercise, you might like to write a computer program which, given a positive integer n, will use the sieve of Eratosthenes to find all the prime numbers up to n. In fact, such a program is one of the standard benchmark tests which is used to evaluate the speed of a computer. Also you should use the sieve to find all prime numbers less than or equal to $n = 50$ (you will be asked to do this for a larger value of n in the exercises).

The first result of this section describes a very useful property of primes: the property is characteristic of these numbers and is sometimes used as the definition of prime. The theorem occurs as Proposition 30 in Book VII of Euclid's *Elements*.

Theorem 1.3.1 *Let p be a prime integer and suppose that a and b are integers such that p divides ab. Then p divides either a or b.*

Proof Since the only positive divisors of p are 1 and p, it follows that the greatest common divisor of p and a is either p, in which case p divides a, or 1. So, if p does not divide a, the greatest common divisor of a and p must be 1. The result now follows by applying Theorem 1.1.6(i) (with p, a, b in place of a, b, c). □

Comment This is short but quite subtle: there is not an immediately obvious connection between the property of being prime and having the property expressed in 1.3.1, but probably you were already aware that primes have that property (just through experience with numbers). But how to prove it? The concept of greatest common divisor is the key to the short and simple proof above.

It is not difficult to see that any integer p which has the property expressed in Theorem 1.3.1 must be prime.

To see how the statement of Theorem 1.3.1 can fail if p is not prime consider: 4 divides $6 \cdot 2 = 12$ yet 4 divides neither 6 nor 2.

Notice that, as is usual in mathematics, the term 'or' is used in the inclusive sense: so the conclusion of Theorem 1.3.1 is more fully expressed as 'p divides a or b or both'.

The next result is an extension of Theorem 1.3.1. It provides an illustration of one kind of use of the principle of mathematical induction.

Lemma 1.3.2 *Let p be a prime and suppose that p divides the product*

$$a_1 a_2 \dots a_r.$$

Then p divides at least one of $a_1, a_2, \dots, a_r$.

Proof The proof is by induction on the number, r, of factors a_i. The base case is trivial since the 'product' is then just a_1. We therefore suppose inductively that if p divides a product of the form

$$b_1 b_2 \dots b_{r-1}$$

then p divides at least one of $b_1, b_2, \dots, b_{r-1}$.

Suppose then that p divides the product $a_1 \dots a_r$. We want to write this as a product, $b_1 \dots b_{r-1}$, of $r-1$ integers. All we have to do is multiply the last two together. So define b_i to be equal to a_i for $i \leq r-2$ and let b_{r-1} be the product $a_{r-1}a_r$: thus we think of bracketing the product of the a_i in the following way:

$$a_1 a_2 \dots a_{r-2}(a_{r-1}a_r)$$

as a product of $r-1$ integers. It follows by induction that either p divides one of $a_1, a_2, \dots a_{r-2}$ or p divides $a_{r-1}a_r$ and, in the latter case, Theorem 1.3.1 implies that p divides a_{r-1} or a_r. So, either way, we conclude that p divides one of the original integers $a_1, \dots, a_r$. □

Discussion The key to this, which is the 'obvious' (but nevertheless has to be proved) extension of 1.3.1 is temporarily to bracket together and multiply two of the numbers so that the product of r integers is 'reduced' to a product of $r-1$ integers. That allowed us to apply the induction hypothesis. Then, we were able to get back to the original list of r integers because p was prime so, by 1.3.1, if it divided the product of the last two numbers, it must have been a divisor of at least one of those numbers.

The following result is sometimes referred to as the Fundamental Theorem of Arithmetic. It says that, in some sense, the primes are the multiplicative building blocks from which every (positive) integer may be produced in a unique way.

Therefore positive integers, other than 1, which are not prime are referred to as **composite**. The distinction between prime and composite numbers, and the importance of this distinction, was recognised at least as early as the time of Philolaus (who died around 390 BC).

Theorem 1.3.3 (The Unique Factorisation Theorem for Integers) *Every positive integer n greater than or equal to 2 may be written in the form*

$$n = p_1 p_2 \dots p_r$$

where the integers $p_1, p_2, \dots, p_r$ are prime numbers (which need not be distinct) and $r \geq 1$. This factorisation is unique in the sense that if also

$$n = q_1 q_2 \dots q_s$$

where $q_1, q_2, \dots, q_s$ are primes, then $r = s$ and we can renumber the q_i so that $q_i = p_i$ for $i = 1, 2, \dots, r$. In other words, up to rearrangement, there is just one way of writing a positive integer as a product of primes.

Proof The proof is in two parts. We show in this first part, using strong induction, that every positive integer greater than or equal to 2 has a factorisation as a product of primes.

The base case holds because 2 is prime. If n is greater than 2, then either n is prime, in which case n has a factorisation (with just one factor) of the required form, or n can be written as a product ab where $1 < a < n$ and $1 < b < n$. In this latter case, apply the inductive hypothesis to deduce that both a and b have factorisations into primes: so juxtaposing the factorisations of a and b (i.e. putting them next to each other), we obtain a factorisation of n as a product of primes.

For the second part of the proof, we use the standard form of mathematical induction, this time on r, the number of prime factors, to show that any positive integer which has a factorisation into a product of r primes has a *unique* factorisation.

To establish the base case ($r = 1$ so n is prime) let us suppose that n is a prime which also may be expressed as:

$$n = q_1 q_2 \dots q_s.$$

If we had $s \geq 2$ then n would have distinct divisors $1, q_1, q_1 q_2$, contradicting that it is prime. So $s = 1$, and the base case is proved.

Now take as induction hypothesis the statement 'any positive integer greater than 2 which has a factorisation into $r - 1$ primes has a unique factorisation (in the above sense)'. Suppose that

$$n = p_1 p_2 \cdots p_r = q_1 q_2 \cdots q_s$$

are two prime factorisations of n. We show that, up to rearrangement, they are the same.

Since p_1 divides n it divides one of $q_1, q_2, \ldots, q_s$ by Lemma 1.3.2. It is harmless to renumber the q_i so that it is q_1 which p_1 divides. Since q_1 is prime it must be that p_1 and q_1 are equal. We may therefore cancel $p_1 = q_1$ from each side to get

$$p_2 p_3 \cdots p_r = q_2 q_3 \cdots q_s.$$

Since the integer on the left-hand side is a product of $r - 1$ primes, the induction hypothesis allows us to conclude that $r - 1$ is equal to $s - 1$, and hence that r is equal to s, and also that, after renumbering, $p_i = q_i$ for $i = 2, \ldots, r$ and hence, since we already have $p_1 = q_1$, for $i = 1, \ldots, r$. □

The first part of Theorem 1.3.3 (existence of the decomposition) occurs, stated in a some what weaker form, as Proposition 31 in Book VII of the *Elements*. It is in the proof of this that Euclid clearly asserts the well-ordering principle. Euclid's argument is in essence the same as that given above.

Comment The first part of the proof was done efficiently but that, to some extent, obscures the simplicity of the idea, which is as follows.

If n is not prime, then factor it, as ab. If a is not prime then factor it, similarly for b. Continue: that is, keep splitting any factors which are not prime. This cannot go on forever – the integers we produce are decreasing and each factorisation gives strictly smaller numbers. So eventually the process stops, with a product of primes.

For instance $588 = 4 \times 147 = (2 \times 2) \times (7 \times 21) = (2 \times 2) \times (7 \times (3 \times 7))$ (see Fig. 1.2).

The second part of the proof is based on the observation that if a prime divides one side of the equation then it divides the other, so we can cancel it from each side, then we continue this process, which can only halt with the equation $1 = 1$. Therefore there must have been the same number of primes, and indeed the same primes, on each side of the original equation. In giving a more formal proof, we rearranged this idea and used induction.

The following result, that there are infinitely many primes, and its elegantly simple proof appear in Euclid's *Elements* (Book IX, Proposition 20).

Corollary 1.3.4 *There are infinitely many prime integers.*

Proof Choose any positive integer n and suppose that $p_1, p_2, \ldots, p_n$ are the first n prime numbers. We will show that there is a prime number different from