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.

Cambridge University Press
978-0-521-54050-6 - Numbers, Groups and Codes: Second Edition
J . F. Humphreys and M. Y. Prest
Excerpt
More information

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:

$D = \{b - ak : b - ak \geq 0$ and k 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.

Cambridge University Press
978-0-521-54050-6 - Numbers, Groups and Codes: Second Edition
J . F. Humphreys and M. Y. Prest
Excerpt
More information

Hence r does satisfy $0 \le r < a$; 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 $a \le b$.

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.

Cambridge University Press
978-0-521-54050-6 - Numbers, Groups and Codes: Second Edition
J . F. Humphreys and M. Y. Prest
Excerpt
More information

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

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

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 ± 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.

Cambridge University Press
978-0-521-54050-6 - Numbers, Groups and Codes: Second Edition
J . F. Humphreys and M. Y. Prest
Excerpt
More information

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).

CAMBRIDGE

Cambridge University Press
978-0-521-54050-6 - Numbers, Groups and Codes: Second Edition
J . F. Humphreys and M. Y. Prest
Excerpt
More information

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*

$$\begin{aligned}
b &= aq_1 + r_1 \qquad (0 < r_1 < a),\\
a &= r_1 q_2 + r_2 \qquad (0 < r_2 < r_1),\\
&\vdots\\
r_{n-2} &= r_{n-1} q_n + r_n \qquad (0 < r_n < r_{n-1}),\\
r_{n-1} &= r_n q_{n+1}.
\end{aligned}$$

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. □

Cambridge University Press
978-0-521-54050-6 - Numbers, Groups and Codes: Second Edition
J . F. Humphreys and M. Y. Prest
Excerpt
More information

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

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

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 $0 \le r_1 < a$. 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 $a = r_1 q_2 + r_2$ with $0 \le r_2 < r_1$. 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

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

representing the equation

$$bn_i + am_i = r_i,$$