To apply the decoding table we replace each of these received words by the entry on the top row which lies above it in the table:

00000 01011 01011 00000 11101 11101 11101 11101 00000.

Then we recover what, we hope, is the original message by extracting the first two digits of each of these words: 00 01 01 00 11 11 11 11 00.

We see then that we have corrected all the single errors which have occurred (but not the double or triple errors). In practice the probability of even a single error occurring should be small, and the probability of two or more errors correspondingly much smaller.

The entries in the first column of our coset decoding table are called **coset leaders**. The maximum likelihood decoding assumption corresponds to the choice of coset leaders to be of minimum weight. The received word is decoded as the codeword to which it is closest. This is a reasonable way to proceed but, as in the last example, if the number of errors is too high we may be led to an incorrect decoding. In any given coset there may be more than one word of the minimum weight for that coset (as in the last two rows in the example). The choice of which of these is to be coset leader corresponds to the fact that words which contain a comparatively large number of errors may be of equal distance from more than one codeword. Thus in the example above, choosing 10001 as coset leader means that we decode 10001 as 00000 and 01100 as 11101. But if we had chosen (as we could have) 01100 as coset leader then we would decode 01100 as 00000 and 10001 as 11101. So there can be a certain arbitrariness when dealing with words which contain a large number of errors.

If one found in practice that one was having to use the last two rows of the above decoding table, one would conclude that the rate of errors was too high for the code to deal with effectively.

Let us construct the decoding table for Example 3 before considering how one may avoid having to construct (and store) the whole table in the case where the code is given by a generator matrix.

Example Let $f : \mathbf{B}^3 \longrightarrow \mathbf{B}^6$ and let the generating matrix be as shown: we also list the codewords.

$$G = \begin{pmatrix} 1 & 0 & 0 & 1 & 1 & 1 \\ 0 & 1 & 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 & 1 & 1 \end{pmatrix}$$

000	000000	100	100111
001	001111	101	101000
010	010101	110	110010
011	011010	111	111101

We array the codewords along the top row and then look for words of minimum length not already included in the table, and array their cosets as described.

000000	001110	010101	011011	100111	101001	110010	111100
000001	001111	010100	011010	100110	101000	110011	111101
000010	001100	010111	011001	100101	101011	110000	111110
000100	001010	010001	011111	100011	101101	110110	111000
001000	000110	011101	010011	101111	100001	111010	110100
010000	011110	000101	001001	110111	111001	100010	101100
100000	101110	110101	111011	000111	001001	010010	011100
000011	001101	010110	011000	100100	101010	110001	111111

Then the message:

$$010111 \quad 111111 \quad 010000 \quad 101110 \quad 101110 \quad 011011$$

would be corrected as

$$010101 \quad 111100 \quad 000000 \quad 001110 \quad 001110 \quad 011011.$$

Definition Given the $m \times n$ generator matrix $G = (I_m A)$ we define the corresponding **parity-check matrix** H to be the $n \times (n - m)$ matrix

$$\begin{pmatrix} A \\ I_{n-m} \end{pmatrix}.$$

For example, if G is the matrix in Example 2 above then the corresponding parity-check matrix H is

$$\begin{pmatrix} 1 & 1 & 0 \\ 0 & 1 & 1 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

Given a word w in $\mathbf{B}^n$ we define the **syndrome** of w to be the matrix product wH in $\mathbf{B}^{n-m}$.

Theorem 5.4.5 *Let H be the parity-check matrix associated with a given code. Then w is a codeword if and only if its syndrome wH is the zero element in $\mathbf{B}^{n-m}$.*

Proof To see this, note that w is a codeword if and only if w has the form uv where $v = uA$. This equation may be rewritten as

$$\mathbf{0} = uA - vI_{n-m} = uA + vI_{n-m} = (uv)H = wH$$

where '$\mathbf{0}$' is the zero $(n-m)$-tuple: that is, $wH = \mathbf{0}$. Thus we see that w is a codeword if and only if wH is $\mathbf{0}$ □

Corollary 5.4.6 *Two words are in the same row of the coset decoding table if and only if they have the same syndrome.*

Proof Two words u and v are in the same row of the decoding table if and only if they differ by a codeword w, say $u = v + w$, that is, if and only if $u - v = w \in W$. Since $(u - v)H = uH - vH$ and since $wH = 0$ exactly if $w \in W$ (by Theorem 5.4.5), we have that u and v are in the same row exactly if $uH = vH$, as required. □

We construct a decoding table with syndromes by adding an extra column at the left which records the syndrome of each row (and is obtained by computing the syndrome of any element on that row). This makes it easier to locate the position of any given word in the table (compute its syndrome to find its row). It is also quite useful in the later stages of constructing the table: when we want to check whether or not a given word already is in the table, we can compute its syndrome and see if it is a new syndrome or not. Also, it is not necessary to construct or record the whole table: it is enough to record just the column of syndromes and the column of coset leaders. Then, given a word w to decode, compute its syndrome, add to (subtract from, really) w the coset leader u which has the same syndrome – the word $w + u$ will then be the corrected version of w – finally read off the first m digits to reconstruct the original word.

This means that it is sufficient to construct a two-column decoding table, one which contains just the column of coset leaders and the column of syndromes. The advantage of using a table showing only coset leaders and syndromes is well illustrated by the next example.

Example Let $f : \mathbf{B}^8 \longrightarrow \mathbf{B}^{12}$ be defined using the generator matrix

$$G = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 0 & 1 \end{pmatrix}$$

There are $2^8 = 256$ codewords and the minimum distance between codewords is 3. To see this, note that there is a codeword of weight 3, for example that given by the second row of G (it is $(01000000) \cdot G$). Any other codeword is obtained by adding together rows of G and so must have weight at least 2 in the first 8 entries. Since the entries of different rows in positions 9 to 12 are different, there can be no two codewords which are distance 2 from each other.

Thus the code detects two errors and corrects one error. This code is as effective as our examples from $\mathbf{B}^3$ to $\mathbf{B}^6$ but is considerably more efficient (we do not have to double the number of digits sent: rather just send half as many again). The parity check matrix H is the 12×4 matrix

$$\begin{pmatrix} 1 & 1 & 1 & 0 \\ 0 & 1 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 0 & 0 & 1 & 1 \\ 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 \\ 1 & 0 & 0 & 1 \\ 1 & 1 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

There will be 16 cosets in our table showing the syndrome then coset leader.

Syndrome	Coset leader
0000	000000000000
0001	000000000001
0010	000000000010
0100	000000000100
1000	000000001000
1101	000000010000
1001	000000100000
0101	000001000000
1100	000010000000
0011	000100000000
1010	001000000000
0110	010000000000
1110	100000000000
1011	000100001000
0111	000100000100
1111	100000000001

This table was produced by writing the 12 'unit error' vectors for the coset leaders and listing the appropriate syndromes (the rows of the parity check matrix). This gives 13 rows, counting the first. The final three rows are obtained by listing the three elements of $\mathbf{B}^4$ which do not occur as syndromes in the first 13 rows and seeing how to express these as combinations of the known syndromes. For example, we see by inspection that the syndrome 1011 does not occur in the first 13 rows. It may be expressed in several ways as combinations of the syndromes in the first 13 rows, for example

$$1011 = 1000 + 0011 = 1010 + 0001.$$

The first corresponds to the choice of 000100001000 as coset leader, the second to 001000000001. As we have seen, neither of these is 'correct', rather each is just a choice of how to correct words with more than one error.

Now to correct the message

$$000010000100 \quad 110110010000 \quad 001100101111$$

we compute the syndrome of each word. They are

$$1000 \quad 1010 \quad 1111$$

(thus none of the these is a codeword). The corresponding coset leaders are

$$000000100001 \quad 000000000001 \quad 100000000000.$$

Each of these is added to the corresponding received word so as to obtain a

codeword, and thus we correct to get

$$000010100101 \quad 110110010001 \quad 101100101110.$$

Now we see how much more efficient it may be to compute and store only coset leaders with syndromes: the table above contains $16 \cdot (12 + 4) = 256$ digits; how many digits would the full decoding table contain? It would have 16 rows, each containing 256 twelve-digit words, that is $16 \cdot 256 \cdot 12 = 49152$ digits!

The key point which makes the above example so efficient is that the rows of the parity-check matrix are non-zero and distinct. Such a code clearly corrects one error (by the argument used at the beginning of the above example). The extreme example of such a code occurs when the rows of the parity-check matrix contain all the $2^n - 1$ non-zero vectors in $\mathbf{B}^n$. Whenever this is the case, every vector v is a single error away from being a codeword, say $v = w + e_i$, and its syndrome is that of e_i which is the ith row of H. The code associated with such a parity check matrix is known as a **Hamming code**. In such a code the spheres of radius 1 centred on the codewords partition the entire 'space' of words ('radius' here is measured with respect to the distance function $d(u, v)$ between words).

Hamming codes are examples of 'perfect codes': codes in which the codewords (of length n say) are evenly distributed throughout the words of length n.

Example When $n = 3$ one possible parity-check matrix associated with a Hamming code is

$$\begin{pmatrix} 1 & 1 & 1 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 0 & 1 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

The corresponding generator matrix is

$$\begin{pmatrix} 1 & 0 & 0 & 0 & 1 & 1 & 1 \\ 0 & 1 & 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 & 1 & 1 \end{pmatrix}$$

and the syndrome plus coset leader decoding table is

Syndrome	Coset leader
000	0000000
111	1000000
110	0100000
101	0010000
011	0001000
100	0000100
010	0000010
001	0000001

Example Let us give one more example of constructing the two-column decoding table. Consider the coding function $f : \mathbf{B}^3 \longrightarrow \mathbf{B}^6$ with generating matrix (and codewords) as shown.

$$G = \begin{pmatrix} 1 & 0 & 0 & 1 & 0 & 1 \\ 0 & 1 & 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & 1 & 1 & 0 \end{pmatrix}$$

000	000000
001	001110
010	010011
011	011101
100	100101
101	101011
110	110110
111	111000

Since the minimum weight of a non-zero codeword is 3, the code detects two errors and corrects one error.

The parity-check matrix is

$$\begin{pmatrix} 1 & 0 & 1 \\ 0 & 1 & 1 \\ 1 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

There are $2^{6-3} = 8$ cosets of the group of 8 codewords in the group of 64 words of length 2^6, so there will be 8 rows in the decoding table, which is as shown below:

Syndrome	Coset leader
000	000000
001	000001
010	000010
100	000100
110	001000
011	010000
101	100000
111	100010

Now suppose that a message is encrypted using the number-to-letter equivalents

000	A	001	C	010	E	011	N
100	O	101	R	110	S	111	T

and then is sent after applying the coding function f. Suppose that the message received is

$$101110 \quad 100001 \quad 101011 \quad 111011 \quad 010011 \quad 011110 \quad 111000.$$

If we were not to attempt to correct the message and simply read off the first three digits of each word, we would obtain

$$101 \quad 100 \quad 101 \quad 111 \quad 010 \quad 011 \quad 111$$

which, converted to alphabetical characters, gives us the nonsensical

RORTENT.

But we note that errors have occurred in transmission, since some of the received words are not codewords, so we apply the correction process. The syndromes of the words received are obtained by forming the products wH where w ranges over the (seven) received words and H is the parity-check matrix above. They are

$$101 \quad 100 \quad 000 \quad 011 \quad 000 \quad 011 \quad 000.$$

The corresponding coset leaders are

$$100000 \quad 000100 \quad 000000 \quad 010000 \quad 000000 \quad 010000 \quad 000000.$$

The corrected message, obtained by adding the coset leaders to the corresponding received words, is therefore

$$001110 \quad 100101 \quad 101011 \quad 101011 \quad 010011 \quad 001110 \quad 111000.$$

Extracting the initial three digits of each word gives

$$001 \quad 100 \quad 101 \quad 101 \quad 010 \quad 001 \quad 111$$

and this, converted to alphabetical characters, yields the original message

CORRECT.

Error-correcting codes were introduced in the late 1940s in order to protect the transmission of messages. Although motivated by this engineering problem, the mathematics involved has become increasingly sophisticated. A context where they are of great importance is the sending of information to and from space probes, where retransmission is often impossible. Examples are the pictures sent back from planets and comets.

An application to group theory occurs in the idea of the group of a code. Given a code with codewords of length n, the group of the code consists of the permutations in $S(n)$ that send codewords to codewords. (A permutation π acts on a codeword by permuting its letters according to π.) There is a very important example of a code with words of length 24, known as the Golay code, whose group is the Mathieu group M_{24} which is one of the sporadic simple groups. This is just one example of the interaction between group theory and codes.

Exercises 5.4

1. Refer back to the example at the beginning of this section on ISBN numbers. One of the most commonly made errors in transcribing numbers is the interchange of two adjacent digits: thus 3 540 90346 could become 3 540 93046. Show that the check digit at the end of the ISBN will also detect this kind of error.
2. For each of the following generator matrices, say how many errors the corresponding code detects and how many errors it corrects:

$$\begin{pmatrix} 1 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 & 1 \\ 0 & 0 & 1 & 0 & 1 \end{pmatrix}; \quad \begin{pmatrix} 1 & 0 & 0 & 1 & 0 & 1 \\ 0 & 1 & 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 1 & 0 & 1 \end{pmatrix};$$

$$\begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 1 & 0 & 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 1 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 & 1 & 1 & 1 \end{pmatrix}; \quad \begin{pmatrix} 1 & 0 & 0 & 0 & 1 & 0 & 1 & 1 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 1 & 0 & 0 & 1 & 1 \end{pmatrix}.$$

3. Let $f : \mathbf{B}^3 \longrightarrow \mathbf{B}^9$ be the coding function given by

$$f(abc) = abcabc\bar{a}\bar{b}\bar{c}$$

where $\bar{x}$ is 1 if x is 0 and $\bar{x}$ is 0 if x is 1. List the eight codewords of f. Show that f does not give a group code. How many errors does f detect and how many errors does it correct?

4. Give the complete coset decoding table for the code given by the generator matrix

$$\begin{pmatrix} 1 & 0 & 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 & 1 & 1 \end{pmatrix}.$$

5. For the code given by the 8×12 generator matrix on p. 247, correct the following message:

101010101010 111111111100 000001000000 001000100010 001010101000.

6. Write down the two-column decoding table for the code given by the generator matrix

$$\begin{pmatrix} 1 & 0 & 0 & 1 & 1 & 0 & 1 \\ 0 & 1 & 0 & 1 & 1 & 1 & 0 \\ 0 & 0 & 1 & 1 & 0 & 1 & 1 \end{pmatrix}.$$

   Use this table to correct the message

1100011 1011000 0101110 0110001 1010110.

7. Let $f : \mathbf{B}^3 \longrightarrow \mathbf{B}^6$ be given by the generator matrix

$$\begin{pmatrix} 1 & 0 & 0 & 1 & 0 & 1 \\ 0 & 1 & 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 0 & 1 & 1 \end{pmatrix}.$$

   Write down the two-column decoding table for f. A message is encoded using the letter equivalents

000	blank	100	A	010	E	001	T
110	N	101	R	011	D	111	H

011011 110000 010110 100000 110110 110111 011111.

   Decode the received message.

Summary of Chapter 5

This chapter was an introduction to basic group theory. Although groups themselves were defined in Chapter 4, we did little beyond offering definitions and examples there. In the present chapter, we first investigated the power of the four group axioms and discussed equation-solving in groups as well as some simple facts about orders of elements in groups. An idea of great importance is

that of a subgroup (a non-empty set which itself satisfies the four group axioms using the same law of composition as that of the group). Associated with each subgroup, there is a partition of the group into distinct (left or right) cosets. The fundamental result (Lagrange's Theorem) is that the number of distinct left cosets of a subgroup H in a group G is equal to $o(G)/o(H)$. This result has many elementary consequences and we saw some of the power of this result in the process of classifying (producing a list of) groups with a small number of elements (this was done in Section 5.3).

In the final section of the chapter, we considered the application of the decomposition of a group into cosets of a subgroup to the elementary theory of error-correcting codes. These codes provide a systematic way to send messages, with some extra information (check digits) in such a way that an error occurring in the original messages will not just be noticed (detected) by the receiver but, in many cases, may even be corrected.

6 Polynomials

6.1 Introduction

We have mentioned polynomials on a couple of occasions already, but now is the time to take a closer look at them.

A (**real**) **polynomial function** f is a map from the set $\mathbb{R}$ to itself, where the value, $f(x)$, of the function f at every (real) number x is given by a formula which is a (real) linear combination of non-negative-integral powers of x (the same formula for all values of x).

An example of a polynomial function is the function which cubes any number x and adds 1 to the result: we write $f(x) = 1 \cdot x^3 + 1$ or, more usually, $f(x) = x^3 + 1$ since a coefficient of 1 before a power of x is normally omitted.

An expression, such as $x^3 + 1$ or $x^6 - 3x^2 + \frac{1}{2}$, which is a (real) linear combination of non-negative-integral powers of x (and which, therefore, defines a polynomial function) is usually referred to as a **polynomial with coefficients in** $\mathbb{R}$ (or **with real coefficients**). It is also, of course, possible to consider polynomials with other kinds of coefficients: for example we might wish to allow coefficients which are complex numbers; or we might wish only to consider polynomials with rational coefficients, etc. In such cases we refer to polynomials with coefficients from $\mathbb{C}$, or $\mathbb{Q}$, etc.

Notice that the following polynomial expressions all define the same function: $x^3 + 2x - 1$, $x^3 + 0x^2 + 2x - 1$, $-1 + 2x + x^3 + 0x^5$. That is, adding a term with 0 coefficient makes no difference (to the function defined) nor, because addition of real numbers is commutative, does rearranging terms. We wish to regard these three expressions (and all others we can get from them by adding terms with 0 coefficient and by rearranging terms) as being 'the same' polynomial. In other words, we regard two polynomial expressions as being equivalent if we can get from one to the other by rearranging terms and

adding or deleting terms with 0 coefficient. It is normal to say and write that such expressions are 'equal' rather than 'equivalent': so we write, for example, $-1 + 2x + x^3 = x^3 + 2x - 1$.

A typical polynomial can, therefore, be written in the form

$$a_0 + a_1 x + \cdots + a_i x^i + \cdots$$

If we want to make this look more uniform we may write

$$a_0 x^0 + a_1 x^1 + \cdots + a_i x^i + \cdots$$

We say that $a_i x^i$ is a **term** of the polynomial and that a_i is the **coefficient** of x^i. We do require that a polynomial should only have finitely many non-zero terms, that is, $a_i = 0$ for all but a finite number of values of i. We say that the power x^i **appears** in the polynomial if $a_i \neq 0$. So x^3 appears in $x^3 + 0x^2 + 2x - 1$ but x^2 does not.

We use notation such as $f(x)$, $g(x)$, $r(x)$, etc. for polynomials but sometimes we drop the '(x)', writing f, g, r etc. We also use the same notation for the functions defined by polynomials.

Summation notation gives a compact way of writing polynomials: in this notation a typical polynomial $f(x)$ has the form $f(x) = \sum_{i=0}^{n} a_i x^i$; the other way of writing this is $a_0 + a_1 x + \cdots + a_n x^n$ (where we have replaced $a_0 x^0$ which equals $a_0 \cdot 1$ by a_0, and $a_1 x^1$ is written more simply as $a_1 x$). If $a_n \neq 0$, in other words if x^n is the highest power of x which appears in the polynomial, then we call $a_n x^n$ and a_n the **leading term** and **leading coefficient** respectively and we say that that the **degree** of $f(x)$ is n and write $\deg(f(x)) = n$. For example the degree of $x^3 + 2x - 1$ is 3.

The zero polynomial is a special case since it has no non-zero coefficients. We shall use the convention that the degree of the zero polynomial is -1 (since doing so makes some things easier to state), although some authors prefer to say that it has degree $-\infty$ or simply say that its degree is undefined.

It is clear from our definition of degree that polynomials of degree one (also known as **linear** polynomials) are those of the form $f(x) = ax + b$ (where a, b are real numbers and $a \neq 0$). Polynomials of degree 2 (**quadratic** polynomials) are those of the form $f(x) = ax^2 + bx + c$ (where a, b, c are in $\mathbb{R}$ and $a \neq 0$). Polynomials of degrees 3, 4 and 5 are also referred to as **cubic, quartic** and **quintic** polynomials respectively. Notice that a polynomial of degree 0 is one of the form $f(x) = a$ (with $a \neq 0$); these, and the zero polynomial, are also referred to as **constant** polynomials. The function defined by a constant polynomial is, of course, a constant function (its value does not depend on x).

Example Let $f(x) = -5x^7 - 6x^4 + 2$, $g(x) = 1 - 4x$ and $h(x) = 5$. Then $\deg(f) = 7$, $\deg(g) = 1$, $\deg(h) = 0$. The leading terms of f, g and h are, respectively, $-5x^7$, $-4x$, 5 and their leading coefficients are -5, -4 and 5.

We have indicated already that the coefficients of a polynomial may be drawn from a range of different mathematical structures, but for our purposes we shall initially regard the coefficients as coming from the set, $\mathbb{R}$, of real numbers. We denote the set of all polynomials whose coefficients are real numbers by $\mathbb{R}[x]$. The 'x' which occurs in the expression of a polynomial should be thought of as a variable which may be substituted by an arbitrary (real) number α. We describe this as **evaluating** the polynomial **at** $x = \alpha$ and we write $f(\alpha)$ for the (real) number which is obtained by replacing every occurrence of x in the expression of $f(x)$ by α. For example, if $f(x) = -5x^7 - 6x^4 + 2$ and $\alpha = -2$ then $f(\alpha) = -5(-2)^7 - 6(-2)^4 + 2 = 546$.

The reader is probably accustomed to drawing graphs to illustrate polynomial functions. Thus a linear function has, as its graph, a straight line. The slope, and any point of intersection of the line with the x-axis and with the y-axis, are easily determined from the coefficients a, b appearing in the formula, $f(x) = ax + b$, defining the values of $f(x)$. We will be concerned with the general question of where the graph of a polynomial function intersects the x-axis: we say that the real number α is a **zero** (or **root**) of the polynomial f if $f(\alpha) = 0$ (that is, if the graph of $y = f(x)$ crosses the x-axis where $x = \alpha$).

For a quadratic polynomial function, the corresponding graph may cross or touch the x-axis in two, one or no points depending on the values of the constants a, b and c in the formula $f(x) = ax^2 + bx + c$ defining the function. We have the well known formula for determining, in terms of the coefficients a, b and c, these points. Namely, the zeros of f are given by the formula

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

It follows from this that f will have a repeated zero when $b^2 = 4ac$, no real zero when $b^2 - 4ac$ is negative and two distinct real zeros when $b^2 - 4ac$ is positive.

There are similar, though more complicated, formulae which give the zeros of general polynomials of degree 3 and 4. It will not be necessary here to know, or use, these formulae. As we mentioned in the historical remarks at the end of Section 4.3, the search for a formula for the zeros of a general polynomial of degree 5 was one of the ideas which lay at the origins of group theory. Recall that Abel proved that there is no formula of this general sort (i.e. involving arithmetic operations and taking roots) which gives the zeros of a general quintic polynomial. Nevertheless, for *some* types of quintics (and higher-degree polynomials) there is a formula. Galois gave the exact conditions on a polynomial for there to be a formula for its zeros. He did this by associating a group to each polynomial and then he was able to interpret the existence of such a formula for the zeros of the polynomial in terms of the structure of this group.

Note that a constant polynomial cannot have a zero unless it is actually the zero polynomial (and then every number is a zero!) so, when we make general statements about zeros of polynomials we sometimes have to insert a clause excluding constant polynomials.

We noted above that not every polynomial has a (real) zero. For example, if x is a real number, x^2 is never negative (and is only zero if x is zero), so $x^2 + 1$ is never zero. Thus the polynomial $x^2 + 1$ has no real roots. (The reader may be aware that, in order to find a zero of this equation, we need to use complex numbers. It is a very general result, beyond the scope of this book, that every non-constant real polynomial has a zero, which may be a complex number. Another fact that we will need to use later is that if the real polynomial f has a complex zero α, then the complex conjugate of α is also a zero. A general introduction to and summary of some basic facts about complex numbers is given in the Appendix.)

In this chapter we will see that there are some remarkable similarities between the set, $\mathbb{Z}$, of integers with its operations of addition and multiplication and the set, $\mathbb{R}[x]$, of polynomials with its algebraic operations – these we now define.

Just as for the set of integers, we have two algebraic operations on the set $\mathbb{R}[x]$. You have almost certainly met these before, at least in the context of specific examples.

First we have addition of polynomials. Given f and g in $\mathbb{R}[x]$, say

$$f(x) = a_0 + a_1x + \cdots + a_nx^n + \cdots$$

and

$$g(x) = b_0 + b_1x + \cdots + b_nx^n + \cdots,$$

we define their **sum** by

$$f(x) + g(x) = (a_0 + b_0) + (a_1 + b_1)x + \cdots + (a_n + b_n)x^n + \cdots$$

Of course all but a finite number of the coefficients $(a_i + b_i)$ will be zero (because this is so for $f(x)$ and $g(x)$). In fact, if f has degree n, so

$$f(x) = a_0 + a_1x + a_2x^2 + \cdots + a_nx^n,$$

with $a_n \neq 0$ and g has degree m, so

$$g(x) = b_0 + b_1x + \cdots + b_mx^m,$$

with $b_m \neq 0$ then we have

$$(f + g)(x) = (a_0 + b_0) + (a_1 + b_1)x + (a_2 + b_2)x^2 + \cdots + (a_k + b_k)x^k$$

where k is the larger of n and m.

There are a couple of points to make about this last formula. First, if, say $n > m$, so $k = n$, then all the coefficients, $b_{m+1} \cdots b_n$ of g beyond b_m, are 0. But it is useful to have them there so that we can write down a formula for the sum $f + g$ in a uniform way. We will see something similar when we define multiplication of polynomials.

The other point to make is that it is clear from the last formula that the degree of $f + g$ is less than or equal to k, the larger of n and m. It is possible for the degree of $f + g$ to be strictly smaller than this (for example, if $f = x^2 + x$ and $g = -x^2 - 1$) because the leading terms might cancel; this can only happen, though, if $\deg(f) = \deg(g)$.

The reader should be able to see that this definition is just a formal way of stating what is probably obvious: we obtain $f + g$ by adding together corresponding powers of x in f and g.

(We also remark that we have found it convenient to write general polynomials with 'lowest powers first' whereas, in numerical examples we usually write the highest powers of x first.)

Example The sum of the quadratic polynomial $f(x) = 2x^2 - 5x + 3$ and the linear polynomial $g(x) = 5x - 2$ is the polynomial $(f + g)(x) = 2x^2 + 1$ (that is, $(2 + 0)x^2 + (-5 + 5)x + (3 - 2)$).

As we stated in Section 4.4, the set, $\mathbb{R}[x]$, of polynomials forms an Abelian group under addition, with the polynomial 0 as its identity and the inverse of the polynomial $f(x) = a_0 + a_1x + a_2x^2 + \cdots + a_nx^n$ being the polynomial $-f$ given by $-f(x) = -a_0 - a_1x - a_2x^2 - \cdots - a_nx^n$.

The second basic operation on the set of polynomials is multiplication. If

$$f(x) = a_0 + a_1x + \cdots + a_nx^n$$

and

$$g(x) = b_0 + b_1x + \cdots + b_mx^m$$

then

$$(fg)(x) = c_0 + c_1x + \cdots + c_{n+m}x^{n+m}$$

where,

$$\begin{aligned}
c_0 &= a_0b_0 \\
c_1 &= a_0b_1 + a_1b_0 \\
c_2 &= a_0b_2 + a_1b_1 + a_2b_0 \\
&\vdots \\
c_i &= a_0b_i + a_1b_{i-1} + a_2b_{i-2} + \cdots + a_ib_0 \\
&\vdots
\end{aligned}$$

(Bear in mind our convention that any undefined coefficients should be taken to be zero: refer back to the comments after the definition of addition of polynomials.)

The definition might look quite complicated but it is only saying the obvious thing: to obtain the formula for fg, take the formulae for f and g and multiply them together, gathering together all the coefficients of the same power of x. If this is not obvious to you now then try multiplying together two polynomials with general coefficients, say $f(x) = a_0 + a_1x + a_2x^2 + a_3x^3$ and $g(x) = b_0 + b_1x + b_2x^2$, and then gathering together terms with the same power of x to see where the above expressions for c_0, c_1, etc. come from. The formulae above for the coefficients, c_i, of a product are useful when dealing with polynomials of large degree but, for 'small' polynomials, in practice we use the procedure of multiplying out and gathering terms together.

One consequence of this definition is that, provided f and g are non-zero, the degree of fg is the sum of the degree of f and the degree of g because if, with notation as above, $\deg(f) = n$, so $a_n \neq 0$ and $\deg(g) = m$, so $b_m \neq 0$, then (you should check), every coefficient c_l with $l > n + m$ is 0 and the coefficient, c_{n+m}, of x^{n+m} is a_nb_m, which is non-zero.

Example The product of the quadratic polynomial $f(x) = 2x^2 - 5x + 3$ and the linear polynomial $g(x) = 5x - 2$ is the polynomial of degree three $(fg)(x) = (2x^2 - 5x + 3)(5x - 2)$. Multiplying out and rearranging, we obtain

$$(fg)(x) = 10x^3 - 4x^2 - 25x^2 + 10x + 15x - 6 = 10x^3 - 29x^2 + 25x - 6.$$

The set, $\mathbb{R}[x]$, of polynomials equipped with these two operations satisfies most of the familiar laws of algebra. For example one can check that the distributive law, namely

$$f(x)h(x) + g(x)h(x) = (f(x) + g(x))h(x)$$

holds for polynomials. One may also check that we have the commutative law for multiplication of polynomials: $f(x)g(x) = g(x)f(x)$. These, and all other such elementary properties, will be used without comment from now on. We do not include their (straightforward) proofs but we do comment that they depend ultimately on the fact that the corresponding properties (such as commutativity and distributivity) hold for real numbers. These properties are summarised in the statement that $\mathbb{R}[x]$ is a commutative ring.

We define subtraction in the obvious way: the **difference** $f - g$ of the polynomials f and g, where $f(x) = a_0 + a_1x + a_2x^2 + \cdots + a_nx^n$ and

$g(x) = b_0 + b_1x + \cdots + b_mx^m$ with, say, $m \leq n$ is given by
$(f - g)(x) = (a_0 - b_0) + (a_1 - b_1)x + (a_2 - b_2)x^2 + \cdots + (a_n - b_n)x^n$.
Note that $f - g = f + (-g)$.

The situation for division is much more complicated (and interesting), and will be discussed in the next section.

We emphasise that the facts we have needed to use in our discussion are the basic properties (such as commutativity and distributivity) of the algebraic operations on $\mathbb{R}$: it is these which underpin the corresponding properties of $\mathbb{R}[x]$. For this reason, we could equally have considered $\mathbb{C}[x]$, the set of polynomials with complex coefficients of a complex variable x, in place of $\mathbb{R}[x]$. The definition of addition and multiplication of polynomials would be as above and we would have exactly the same basic algebraic properties as for $\mathbb{R}[x]$. However, one major difference is that every non-constant polynomial, f, in $\mathbb{C}[x]$ has a zero: that is, there is a complex number α such that $f(\alpha) = 0$.

This process, changing coefficients, does not end here.

Given any prime number p, we can consider (as in Chapter 1) the set, $\mathbb{Z}_p$, of congruence classes modulo p. Again, this set is a ring and so we can consider $\mathbb{Z}_p[x]$, the set of polynomials **over** $\mathbb{Z}_p$ (that is, with coefficients in $\mathbb{Z}_p$). Everything (definitions and basic algebraic properties) is as before. We do, of course, when adding and multiplying such polynomials, have to calculate the coefficients of the sum and the product using arithmetic modulo our prime p. Thus if $p = 2$, $(x^2 + x + 1) + (x^2 + 1) = x$ and when $p = 3$, $(x + 2)^2 = x^2 + x + 1$. Similarly, when evaluating such polynomials, we have to calculate modulo p. You might wonder why we do this only for prime numbers p: surely in almost everything we have said so far we could replace the prime p by any integer $n \geq 2$. That is correct and it is really only when we come to division (in the next section) that we do need p to be prime. For remember that if n is not prime then $\mathbb{Z}_n$ is not a field and this means that we cannot always divide by non-zero elements. For the next section we certainly need to be in a context where we can always divide by non-zero coefficients.

There is an important feature of polynomials with coefficients in $\mathbb{Z}_p$. We can, in principle, find all the zeros of any polynomial. Because there are only a finite number of elements in $\mathbb{Z}_p$, we can simply substitute each element of $\mathbb{Z}_p$ in turn. For small p this process will be simple to apply and will give us the zeros of any given polynomial. For example, as a polynomial in $\mathbb{Z}_2$, the quadratic $x^2 + 1$ takes the value 1 when $x = 0$ and when $x = 1$ takes the value $1 + 1$, which is zero modulo 2, so 1 is a zero. In fact it is clear that, as a polynomial in $\mathbb{Z}_2$, $x^2 + 1$ is equal to $(x + 1)^2$, so this polynomial actually has a repeated zero. However the polynomial $f(x) = x^2 + x + 1$ has no zeros in $\mathbb{Z}_2$ because $f(0) = 1 = f(1)$.

Exercises 6.1

1. Add the following pairs of polynomials:

   (i) the real polynomials $x^2 + 7x + 3$ and $x^2 - 5x - 3$;
   (ii) the real polynomials $x^3 - 2x^2 + x - 1$ and $-x^3 - x^2 + x + 1$;
   (iii) the complex polynomials $x^2 + 7x + 3$ and $x^2 - \mathrm{i}5x - 3\mathrm{i}$;
   (iv) the complex polynomials $x^3 - 2\mathrm{i}x^2 + \mathrm{i}x - \mathrm{i}$ and $-x^3 - \mathrm{i}x^2 + \mathrm{i}x + \mathrm{i}$;
   (v) the polynomials over $\mathbb{Z}_3$, $x^2 + 2x + 1$ and $x^2 + 2x + 2$;
   (vi) the polynomials over $\mathbb{Z}_3$, $x^3 + 2x^2 + x - 1$ and $2x^3 - x^2 + x + 1$.

2. Multiply the following pairs of polynomials:

   (i) the real polynomials $x^2 + 7x + 3$ and $x + 1$;
   (ii) the real polynomials $x^3 - 2x^2 + x - 1$ and $x^2 + x + 1$;
   (iii) the complex polynomials $x^2 + 7x + 3$ and $\mathrm{i}x + 3$;
   (iv) the complex polynomials $\mathrm{i}x^3 - 2x^2 + x - \mathrm{i}$ and $\mathrm{i}x^2 + \mathrm{i}x - \mathrm{i}$;
   (v) the polynomials over $\mathbb{Z}_2$, $x^2 + x + 1$ and $x^2 + x + 1$;
   (vi) the polynomials over $\mathbb{Z}_2$, $x^3 + x^2 + x + 1$ and $x^2 + x + 1$.

3. Find a zero of each of the given polynomials:

   (i) the real polynomial $x^3 - 3x^2 + 4x - 2$;
   (ii) the complex polynomial $x^3 - 7x^2 + x - 7$;
   (iii) the polynomial $x^3 + 4x^2 + 2x + 4$ over $\mathbb{Z}_5$.

6.2 The division algorithm for polynomials

Definition We say that the polynomial g **divides** the polynomial f if there is a polynomial q such that $f = qg$, that is, $f(x) = q(x)g(x)$. We start with a result which gives us a partial answer to the question of when one polynomial divides another.

Proposition 6.2.1 *Let s and t be polynomials of degree n and m respectively, say*

$$s(x) = a_0 + a_1x + \cdots + a_nx^n \qquad \text{and} \qquad t(x) = b_0 + b_1x + \cdots + b_mx^m$$

with $a_n \neq 0$ and $b_m \neq 0$. Assume that $n \geq m$. Then the degree of the polynomial

$$u(x) = s(x) - (a_n/b_m)x^{n-m}t(x)$$

is strictly less than the degree of s.

Proof Since $s(x)$ and $x^{n-m}t(x)$ are both of degree n (note that the leading term of $x^{n-m}t(x)$ is $b_m x^m \cdot x^{n-m} = b_m x^n$), the degree of $u(x)$ can be at most n. However the coefficient of x^n in $s(x)$ is a_n and the coefficient of this power of x in the polynomial $(a_n/b_m)x^{n-m}t(x)$ is $(a_n/b_m)b_m = a_n$. Thus the coefficient of x^n in $u(x)$ is zero, and the degree of u is indeed less than n. □

We shall refer to the term $(a_n/b_m)x^{n-m}$ as a **partial term** in the division of s by t. Using the above result repeatedly is the key to dividing one polynomial, f say, by another, g. (Here we mean 'dividing' to obtain a quotient and a remainder: only if the remainder is zero do we say that 'g divides f'.)

Example Divide the polynomial $f(x) = x^4 - 3x^2 + 2x - 4$ by the polynomial $g(x) = x^2 - 3x + 2$.

We first apply Proposition 6.2.1 with $s = f$ and $t = g$, so $a_0 = -4$, $a_1 = 2$, $a_2 = -3$, $a_3 = 0$ and $a_4 = 1$. Also $b_0 = 2$, $b_1 = -3$ and $b_2 = 1$. According to Proposition 6.2.1, the polynomial $u_1(x) = f(x) - (1/1)x^2 g(x)$ will have degree less than 4. Indeed, we can calculate to see that $u_1(x) = 3x^3 - 5x^2 + 2x - 4$. Next apply Proposition 6.2.1 again, with u_1 in place of s, and g in place of t, to obtain a polynomial

$$\begin{aligned} u_2(x) &= u_1(x) - (3/1)xg(x) \\ &= \cdots \\ &= 4x^2 - 4x - 4. \end{aligned}$$

We can repeat this process once more. We obtain

$$u_3(x) = u_2(x) - 4g(x) = 4x^2 - 4x - 4 - 4(x^2 - 3x + 2) = 8x - 12.$$

We summarise this calculation by rearranging the equations above to obtain

$$\begin{aligned} f(x) &= x^2 g(x) + u_1(x) \\ u_1(x) &= 3xg(x) + u_2(x) \\ u_2(x) &= 4g(x) + u_3(x) \end{aligned}$$

and so

$$\begin{aligned} f(x) &= x^4 - 3x^2 + 2x - 4 \\ &= x^2 g(x) + u_1(x) \\ &= x^2 g(x) + (3xg(x) + u_2(x)) \\ &= (x^2 + 3x)g(x) + (4g(x) + u_3(x)) \\ &= (x^2 + 3x)g(x) + 4g(x) + (8x - 12) \\ &= (x^2 + 3x + 4)g(x) + (8x - 12) \\ &= (x^2 + 3x + 4)(x^2 - 3x + 2) + (8x - 12). \end{aligned}$$

This type of calculation is often presented in the following 'long division' format.

$$\begin{array}{rl|l} & x^2 + 3x + 4 & \\ \hline x^2 - 3x + 2 \mid & x^4 - 3x^2 + 2x - 4 & \\ & x^4 - 3x^3 + 2x^2 & \\ \hline & 3x^3 - 5x^2 + 2x - 4 & \\ & 3x^3 - 9x^2 + 6x & \\ \hline & 4x^2 - 4x - 4 & \\ & 4x^2 - 12x + 8 & \\ \hline & 8x - 12 & \end{array}$$

We give some words of explanation about the way this calculation has been set out. The polynomial g is written on the second line to the left of the '|' sign with the polynomial f to its right. The top line records our partial terms. The line below f is obtained by multiplying g by the first partial term. The line below that is then obtained by subtracting polynomials. The next line arises from multiplying g by the next partial term and the following line is again obtained by subtraction. Continue in this way, alternating the products of g by the partial terms with the results of subtracting two polynomials, until we arrive at a polynomial of degree less than that of g (in this case the linear polynomial $8x - 12$). This is the remainder when f is divided by g.

The reader may need practice to be able to carry out this process with confidence, and several examples are provided in the end-of-section exercises. It is clear that we can repeat this process for any given polynomials f and g and hence we can write f in the form $qg + r$, where q is the polynomial obtained by adding together the partial terms and where r will either be the zero polynomial or have degree strictly less than the degree of g. Thus, in the above example $q(x) = x^2 + 3x + 4$ and $r(x) = 8x - 12$. The process we have just described for computing these polynomials q and r from f and g can be used as the basis of a proof of the next result. However, we prefer to give a different proof of this basic fact about division of polynomials in order to bring out more clearly the connection between polynomials and integers.

Theorem 6.2.2 (The Division Theorem for Polynomials) *Let f and g be polynomials* (*with real coefficients*) *with the degree of g being greater than zero* (*that is, g is not a constant polynomial*). *Then there are polynomials q and r, such that $f = qg + r$, where the degree of r is strictly less than that of g.*

Proof If the degree of g is greater than that of f, then we may take $q = 0$ and $r = f$. We may, therefore suppose that m, the degree of g is less than or equal to n,

the degree of f. For definiteness, suppose that $f(x) = a_0 + a_1x + a_2x^2 + \cdots + a_nx^n$ and $g(x) = b_0 + b_1x + \cdots + b_mx^m$. Consider the set S of those polynomials which are obtained by subtracting a polynomial multiple of g from f:

$$S = \{f - gt : t \text{ is in } \mathbb{R}[x]\}.$$

This set is non-empty since it contains $f (= f - 0 \cdot g)$. Now denote by D the set of degrees of those polynomials in S. Since f is in S, we see that D is also non-empty. Applying the well-ordering principle to D, a non-empty set of integers ≥ -1, we see that D has a least element k, say, so we can select a polynomial $r \in S$ of degree k. Since r is in S we have $r = f - qg$ for some q. Suppose that $r(x) = c_0 + c_1x + \cdots + c_kx^k$.

Next we wish to show that the degree of r is less than m. To show this, we assume that this is not so and obtain a contradiction. To do this, apply Proposition 6.2.1 with the polynomial r here playing the role of the polynomial s in the lemma and the polynomial g here playing the role of t in the lemma. Then the polynomial

$$r - (c_k/b_m)x^{k-m}g = f - qg - (c_k/b_m)x^{k-m}g = f - (q + (c_k/b_m)x^{k-m})g$$

has degree less than k (the degree of r). Since this polynomial is clearly in S, its degree is in D, contrary to the definition of k. Thus, the degree of r must be strictly less than m. □

This proof is an almost word-for-word generalisation of the division theorem for integers (Theorem 1.1.1). We have a simple, but important consequence of this result.

Corollary 6.2.3 *Given a polynomial f, the value $x = \alpha$ is a zero of f if and only if f is divisible by $x - \alpha$.*

Proof Suppose first that α is a zero of f. We apply the division theorem with $g(x) = x - \alpha$, so $f = qg + r$ where r has degree less than 1 (the degree of g). Therefore r is a constant c, say, and so $f(x) = (x - \alpha)q(x) + c$. Evaluating at $x = \alpha$, we obtain $f(\alpha) = (\alpha - \alpha)q(\alpha) + c$. However, we know that $f(\alpha) = 0$ (α is a zero of f) so we obtain $c = 0$, and conclude that $x - \alpha$ divides f.

For the converse, if f is divisible by $x - \alpha$, we have $f(x) = (x - \alpha)q(x)$, for some polynomial $q(x)$. Put $x = \alpha$ to get $f(\alpha) = (\alpha - \alpha)q(\alpha) = 0 \cdot q(\alpha) = 0$, so α is a zero of f. □

Note that nothing in our proof of Theorem 6.2.2 or its corollary requires us to work over $\mathbb{R}$ and so both results hold over $\mathbb{C}$ and over $\mathbb{Z}_p$. The corollary is very useful (in conjunction with the quadratic formula) in factorising (writing as a product of polynomials of smaller degree) polynomials.

Example 1 Factorise the real polynomial $f(x) = x^3 + 6x^2 + 11x + 6$. A reasonable place to start is to compute the value of f at small values (positive, negative and zero) of x. In this case, $f(0) = 6$, $f(1) = 1 + 6 + 11 + 6 = 24$, $f(-1) = -1 + 6 - 11 + 6 = 0$. Therefore -1 is a zero of f and so by the corollary $x - (-1) = x + 1$ divides f. Carry out the division to see that $f(x) = (x + 1)(x^2 + 5x + 6)$. Next, we can factorise the quadratic $x^2 + 5x + 6$, either by using the formula to find the zeros, or by noticing that $x^2 + 5x + 6 = (x + 2)(x + 3)$. Thus $f(x) = (x + 1)(x + 2)(x + 3)$.

Of course, not all polynomials with integer coefficients will have small integer roots. It is difficult, in general, to find roots of real polynomials, which is why people looked for something like the quadratic formula (that is, a formula which gives an exact expression) which would work for polynomials of higher degree.

Example 2 Factorise the polynomial $f(x) = x^3 + 2x^2 + x + 2$ over $\mathbb{Z}_3$.

Since $\mathbb{Z}_3$ is conveniently small, it is easy to find all zeros of $f(x)$. We substitute the three possible values of x into f to find: $f(0) = 2$, $f(1) = 1 + 2 + 1 + 2 = 0$ and $f(2) = 2^3 + 2 \cdot 2^2 + 2 + 2 = 2$. Thus the only zero is $x = 1$ and so $(x - 1) = (x + 2)$ divides f. We can therefore write $f(x)$ in the form $(x + 2)(ax^2 + bx + c)$. This equals $ax^3 + (2a + b)x^2 + (2b + c)x + 2c$. Setting this equal to $x^3 + 2x^2 + x + 2$ and equating coefficients gives $a = 1$, $2a + b = 2$, so $b = 0$, and $2c = 2$, so $c = 1$. With these values we do also have $2b + c = 1$. Thus $f(x) = (x + 2)(x^2 + 1)$. The quadratic $x^2 + 1$ has value 1 when $x = 0$, value 2 when $x = 1$ and value 2 when $x = 2$, so this quadratic has no zeros over $\mathbb{Z}_p$. It follows that we cannot factorise this quadratic into a product of two linear factors, since such a factorisation would, by Corollary 6.2.3 lead to zeros of $x^2 + 1$. Since the degree of a product of two polynomials is the sum of their degrees, there is no further possible factorisation of $x^2 + 1$.

For our next result, we return to the order of development followed in Chapter 1 and, continuing along this road, introduce the greatest common divisor of two polynomials.

Proposition 6.2.4 *Let f and g be non-zero polynomials. Then there is a polynomial $d(x)$ such that*

(i) *$d(x)$ divides both $f(x)$ and $g(x)$, and*
(ii) *if $c(x)$ is any polynomial which divides both $f(x)$ and $g(x)$, then $c(x)$ divides $d(x)$.*

Proof Let S be the set of polynomials of degree ≥ 0 of the form $f(x)r(x) + g(x)s(x)$, as $r(x)$ and $s(x)$ vary over the set of all polynomials. Consider the

set D of integers which are the degrees of the polynomials in S. Since $f(x) = 1 \cdot f(x) + 0$, the degree of $f(x)$ is in D, so D is a non-empty set of non-negative integers. Now apply the well-ordering principle to the set D to select a polynomial $d(x)$ in S such that the degree of $d(x)$ is the smallest integer in D. Since $d(x)$ is in S we have $d(x) = f(x)r(x) + g(x)s(x)$ for some polynomials r and s. We show that d has properties (i) and (ii).

If the polynomial $c(x)$ divides $f(x)$, say $f(x) = c(x)u(x)$ and $c(x)$ divides $g(x)$, say $g(x) = c(x)v(x)$ then

$$\begin{aligned} d(x) &= f(x)r(x) + g(x)s(x) \\ &= c(x)u(x)r(x) + c(x)v(x)s(x) \\ &= c(x)(u(x)r(x) + v(x)s(x)) \end{aligned}$$

so $c(x)$ divides $d(x)$. (Therefore condition (ii) is satisfied.)

To show that $d(x)$ divides $f(x)$, we use Theorem 6.2.2 to write $f(x) = d(x)q(x) + r(x)$ where r has degree strictly less than the degree of $d(x)$. Then

$$\begin{aligned} r(x) &= f(x) - q(x)d(x) \\ &= f(x) - q(x)(f(x)r(x) + g(x)s(x)) \\ &= f(x)(1 - q(x)r(x)) - g(x)s(x)q(x) \end{aligned}$$

is in S or is 0. If $r(x)$ were non-zero then its degree would be in D. But the degree of $d(x)$ was the smallest element of D. Therefore $r(x)$ must be zero and hence $d(x)$ divides $f(x)$. A similar proof shows that $d(x)$ divides $g(x)$. □

Remark By now it should be clear, if you check back to Chapter 1, what our strategy in this chapter is. We are repeating much of that earlier chapter, with essentially the same definitions and very much the same proofs. The details of the proofs need some care and there is the odd change of emphasis in order to take account of the fact that we are dealing with polynomials rather than integers. The general principle to follow in making this change from integers to polynomials is to replace the size of a positive integer by the degree of a polynomial.

This situation, where strong similarities between different situations are seen, is a familiar one in mathematics and it leads to the search for the common content in what might initially seem like very different contexts. Typically this common content is abstracted into ideas and definitions which apply in various contexts, and usually not just in those which motivated the definitions. We have already seen this in basic group theory, where common features of permutations, number systems and other mathematical objects were extracted and developed in a general, more abstract, context.

In presenting mathematics one has a choice. It is possible to present the abstract mathematics first, develop theorems in that context, and then apply them to various special cases. That is an efficient approach (one does not prove the 'same' result over and over again in different contexts) but introducing the abstract ideas right at the beginning presents the student with a steep 'learning curve'. The alternative approach, which we have adopted in this book, is to develop various motivating examples and then extract their common content, so providing the reader with a somewhat longer, but less steep, path.

In the case of the material we are discussing now and the strongly similar material in Chapter 1 there is, indeed, a common generalisation: to what are called 'Euclidean rings'. To treat these would take us beyond the introductory character of this book but the interested reader should look at, for example, [Allenby] or [Fraleigh] listed in the references towards the end of the book.

We can now define a **greatest common divisor** of two polynomials f and g as a polynomial d which satisfies the two conditions of Proposition 6.2.4. Such a polynomial is not unique (which is why we say 'a' greatest common divisor rather than 'the' greatest common divisor). Nevertheless, the degree of any greatest common divisor of f and g is uniquely determined, since if c and d are both greatest common divisors, then c divides d (so the degree of c is less than or equal to that of d) and d divides c (so the degree of d is less than or equal to that of c). It follows that there is a non-zero polynomial of degree 0 (in other words a non-zero constant), λ, such that $c = \lambda d$. This shows that the only difference between any two greatest common divisors of f and g is that each is a multiple of the other by a non-zero constant.

As with integers, the process of finding a greatest common divisor of polynomials uses repeated application of the division theorem and is illustrated by the following example.

Example Find a greatest common divisor $d(x)$ of the polynomials $f(x) = x^4 - 5x^3 + 7x^2 - 5x + 6$ and $g(x) = x^3 - 6x^2 + 11x - 6$.

First we use the division algorithm to write

$$x^4 - 5x^3 + 7x^2 - 5x + 6 = (x + 1)(x^3 - 6x^2 + 11x - 6) + 2x^2 - 10x + 12$$

(Although, we have chosen a reasonably easy first example, it is in general best not to try such calculations in your head: write them down on a sheet of paper. You can, and should, check by multiplying out the right-hand side.)

We now use the division algorithm again, this time applied to our original polynomial g in place of f and with the remainder, $2x^2 - 10x + 12$, in place of g, giving

$$x^3 - 6x^2 + 11x - 6 = (x/2 - 1/2)(2x^2 - 10x + 12) + 0.$$

Therefore $2x^2 - 10x + 12$ is one of the greatest common divisors of our given polynomials f and g ($x^2 - 5x + 6$ is another). The result which explains why this process yields a greatest common divisor for the initial polynomials is given next and its proof is, again, essentially the same as that of the corresponding result in Chapter 1.

Proposition 6.2.5 *Let f, g be polynomials with g non-zero. Suppose that $f = gq + r$. Then a greatest common divisor of f and g is equal to a constant multiple of any greatest common divisor of g and r.*

Proof First suppose that d is a greatest common divisor of f and g. Since d divides both f and g, d divides $f - gq = r$. It follows that d is a common divisor of g and r. Therefore if e is the greatest common divisor of g and r, we deduce that d must divide e. As a consequence of this, the degree of d is less than or equal to the degree of e.

Next, e is a common divisor of g and r, so e divides $f = gq + r$ and hence is a common divisor of f and g. It then follows from the definition of d that e divides d, so the degree of e is less than or equal to the degree of d. We conclude that e and d have equal degrees and, since each divides the other, one is a constant multiple of the other, as required. □

We now have the main result of this section.

Theorem 6.2.6 (The Euclidean algorithm for polynomials) *Let f, g be polynomials. If g divides f then g is a greatest common divisor for f and g. Otherwise apply Theorem* 6.2.2 *to obtain a sequence of non-zero polynomials $r_1, \ldots, r_n$ satisfying*

$$\begin{aligned} f &= gq_1 + r_1 \\ g &= r_1q_2 + r_2 \\ r_1 &= r_2q_3 + r_3 \\ &\vdots \\ r_{n-2} &= r_{n-1}q_n + r_n \\ r_{n-1} &= r_nq_{n+1} \end{aligned}$$

Then r_n is a greatest common divisor for f and g.

Proof Apply Proposition 6.2.1 repeatedly, denoting by $r_1, \ldots, r_n$ the non-zero remainders. Since the degrees of the polynomials $g, r_1, \ldots, r_n$ form a strictly decreasing sequence of non-negative integers, this process must terminate. This

implies that there exists an integer n such that r_n divides r_{n-1}. Then r_n is a greatest common divisor of r_n and r_{n-1}. Proposition 6.2.5 then implies that r_n is a greatest common divisor for r_{n-1} and r_{n-2}. Repeated application of Proposition 6.2.5 implies that r_n is a greatest common divisor for f and g. □

In Chapter 1, we explained a matrix method to record the calculations made in obtaining the greatest common divisor of two integers. Although this could be done in a similar way for polynomials, each step in the calculation would involve a long division of polynomials of the type already discussed. As it is much more difficult to do this calculation in one's head, it is just as easy actually to carry out the step-by-step process explained in Theorem 6.2.6 when dealing with polynomials.

Also, just as with integers, it is possible, after having computed the greatest common divisor, to work back through the equations and to obtain an expression for any greatest common divisor d of f and g in the form $d(x) = f(x)s(x) + g(x)t(x)$ for some polynomials s, t. This reverse process is illustrated in the following examples.

Example 1 Find a greatest common divisor $d(x)$ for the polynomials

$$f(x) = x^5 + 3x^4 - x^2 + 3x - 6 \text{ and } g(x) = x^4 + x^3 - x^2 + x - 2$$

and express $d(x)$ as a polynomial combination of f and g.

We start by writing f as a multiple of g plus a remainder. Only the result of this calculation is given, but it should be made clear that this was done by long division of polynomials, in a calculation which the reader should check. The result of this first step is to obtain

$$f(x) = (x + 2)g(x) - x^3 + 3x - 2.$$

Thus, in the notation of Theorem 6.2.6, $q_1 = x + 2$ and $r_1 = -x^3 + 3x - 2$. Then

$$g(x) = (-x - 1)(-x^3 + 3x - 2) + 2x^2 + 2x - 4$$

so $q_2 = -(x + 1)$ and $r_2 = 2x^2 + 2x - 4 = 2(x^2 + x - 2)$. Applying the process once more, we find that

$$-x^3 + 3x - 2 = 2(x^2 + x - 2)((-x + 1)/2).$$

Thus the calculation of Theorem 6.2.6 would appear in this case as

$$\begin{aligned} f(x) &= (x + 2)g(x) - x^3 + 3x - 2 \\ g(x) &= (-x - 1)(-x^3 + 3x - 2) + 2x^2 + 2x - 4 \\ -x^3 + 3x - 2 &= (2x^2 + 2x - 4)((-x + 1)/2). \end{aligned}$$

Thus a greatest common divisor for f and g is $d(x) = 2x^2 + 2x - 4$ (or we could, instead, take $\frac{1}{2}$ times this, that is, $x^2 + x - 2$). Make this the subject of the second equation to obtain $d(x) = g(x) + (x + 1)(-x^3 + 3x - 2)$. Now make the cubic, $-x^3 + 3x - 2$, the subject of the first equation to obtain $f(x) - (x + 2)g(x) = -x^3 + 3x - 2$. Finally substitute this expression for the cubic into the equation for $d(x)$ to obtain

$$\begin{aligned} d(x) &= g(x) + (x + 1)(f(x) - (x + 2)g(x)) \\ &= (x + 1)f(x) + (1 - (x + 1)(x + 2))g(x) \end{aligned}$$

which we can simplify to $d(x) = (x + 1)f(x) - (x^2 + 3x + 1)g(x)$.

At this stage, one should multiply out, as a check.

Example 2 As a second example, we take $f(x)$ to be the quartic equation $x^4 + x^3 - x - 1$ and $g(x)$ to be the cubic $x^3 - 2x^2 + 2x - 1$. Then, going through the steps of Theorem 6.2.6 as in the previous example we obtain

$$\begin{aligned} f(x) &= (x + 3)g(x) + 4x^2 - 6x + 2 \\ g(x) &= \frac{1}{8}(2x - 1)(4x^2 - 6x + 2) + \frac{3}{4}x - \frac{3}{4} \\ 4x^2 - 6x + 2 &= \frac{3}{4}(x - 1)\left(\frac{4}{3}(4x - 2)\right). \end{aligned}$$

(Once again the reader needs to take an active part in checking these calculations.) Then $d(x) = \frac{3}{4}(x - 1)$ (or, if you prefer, $x - 1$). Working back through the equations we have

$$\begin{aligned} d(x) &= \frac{3}{4}(x - 1) \\ &= g(x) - \frac{1}{8}(2x - 1)(4x^2 - 6x + 2) \\ &= g(x) - \frac{1}{8}(2x - 1)(f(x) - (x + 3)g(x)) \\ &= g(x)\left(1 + \frac{1}{8}(2x - 1)(x + 3)\right) - \frac{1}{8}(2x - 1)f(x) \\ &= \frac{1}{8}(2x^2 + 5x + 5)g(x) - \frac{1}{8}(2x - 1)f(x) \end{aligned}$$

(alternatively, $x - 1 = \frac{1}{6}(2x^2 + 5x + 5)g(x) - \frac{1}{6}(2x - 1)f(x)$).

We make a couple of remarks. First, it is quite possible for two polynomials to have greatest common divisor 1 (equivalently, any non-zero constant): for

instance the greatest common divisor of the real polynomials $x - 1$ and $x + 2$ is 1. Second, when you express the greatest common divisor of f and g as a combination of f and g the only fractions which appear should be numerical fractions (like $\frac{3}{4}$) not polynomial fractions (nothing like $\frac{1}{x-1}$ for instance)!

Example 3 As a final example in this section, we consider a case where our polynomials are over $\mathbb{Z}_2$.

Find a greatest common divisor for the polynomials $g(x) = x^4 + x^3 + x + 1$ and $f(x) = x^3 + x + 1$.

The long division process proceeds exactly as over $\mathbb{R}$, giving

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

At the next step we obtain

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

Thus 1 is a greatest common divisor for $f(x)$ and $g(x)$. Also $1 = f(x) - (x + 1)(x^2 + x)$, so we obtain

$$\begin{aligned} 1 &= f(x) - (x + 1)(g(x) - (x + 1)f(x)) \\ &= f(x)(1 + (x + 1)^2) - (x + 1)g(x) \\ &= x^2 f(x) + (x + 1)g(x) \end{aligned}$$

where, in the last line, we have used the fact that we are working over $\mathbb{Z}_2$ (so $-1 = 1$, $-x = x$ etc.).

Exercises 6.2

1. For each of the following pairs of polynomials, f, g, write f in the form $qg + r$ with either $r = 0$ or the degree of r less than that of g:
   (i) the real polynomials $f(x) = x^4 + x^3 + x^2 + x + 1$ and $g(x) = x^2 - 2x + 1$;
   (ii) the real polynomials $f(x) = x^3 + x^2 + 1$ and $g(x) = x^2 - 5x + 6$;
   (iii) the polynomials $f(x) = x^3 + x^2 + 1$ and $g(x) = x^2 - 5x + 6$ over $\mathbb{Z}_5$.
2. Factorise, as far as possible, the given polynomials:
   (i) $x^3 - x^2 - 4x + 4$ over $\mathbb{R}$;
   (ii) $x^3 - 3x^2 + 3x - 2$ over $\mathbb{R}$;
   (iii) $x^3 - 3x^2 + 3x - 2$ over $\mathbb{C}$;
   (iv) $x^3 - 3x^2 + 3x - 2$ over $\mathbb{Z}_7$;
   (v) $x^3 + x^2 + x + 1$ over $\mathbb{Z}_2$.

3. Find a greatest common divisor, $d(x)$, for each of the following pairs of polynomials and express $d(x)$ as a polynomial combination of the given pair:
   (i) the polynomials $x^3 + 1$ and $x^2 + x - 1$ over $\mathbb{R}$;
   (ii) the polynomials $x^4 + x + 1$ and $x^3 + x + 1$ over $\mathbb{Z}_2$;
   (iii) the polynomials $x^3 - \mathrm{i}x^2 + 2x - 2\mathrm{i}$ and $x^2 + 1$ over $\mathbb{C}$.
4. Prove that if f is any polynomial and α any number and if we write $f(x) = (x - \alpha)g(x) + r(x)$ then $r(x)$ is the constant $f(\alpha)$.

6.3 Factorisation

We start with an important definition.

Definition A non-constant polynomial f is said to be **irreducible** if the only way to write f as a product of two polynomials, $f = gh$, is for one of g and h to be a (non-zero) constant polynomial.

Example A polynomial of degree 1, that is, one of the form $x - \alpha$, must be irreducible.

Remark The reason why a new word 'irreducible' is used here is that it will be convenient to use the word 'prime' to describe a rather different concept. Thus we say that a non-constant polynomial f is **prime** if, whenever f divides a product, rs, of polynomials, then either f divides r or f divides s. In fact we shall show that these two ideas, prime and irreducible, coincide for polynomials (as they do for integers, see Theorem 1.3.1). In some rings these concepts differ.

Proposition 6.3.1 *Let f be an irreducible polynomial and suppose that r and s are polynomials such that f divides rs. Then f divides either r or s. That is, every irreducible polynomial is prime.*

Proof Since f is irreducible, a greatest common divisor for f and r is either a constant polynomial c, or is (a scalar multiple of) f. In the latter case f divides r, so the only case we need consider is when f does not divide r and hence when this greatest common divisor is a constant. Then, by the division algorithm, there are polynomials u, v such that $c = fu + rv$. Multiply this equation by s to obtain

$$c \cdot s = f \cdot u \cdot s + r \cdot s \cdot v$$

Since f divides rs by hypothesis, f divides the right-hand side, so f divides cs which, since c is a constant, implies that f divides s, as required. □

Again, it may be helpful to compare this proof with the corresponding proof 1.1.6(i) from Chapter 1.

As with integers, we may easily extend this result using induction. Its proof will be one of the end-of-section exercises.

Corollary 6.3.2 *Let f be an irreducible polynomial and suppose that f divides the product $f_1 \dots f_r$. Then f divides at least one of $f_1, f_2, \dots, f_r$.*

In Proposition 6.3.1 we see that an irreducible polynomial is prime. The converse is easy.

Proposition 6.3.3 *Let f be a prime polynomial, then f is irreducible.*

Proof Suppose that $f(x)$ is a prime polynomial and that $f(x)$ has a factorisation as $g(x)h(x)$. Then $f(x)$ divides $g(x)h(x)$, so f must divide one of g or h. If, say, f has degree n, g has degree m and h has degree k, then we have that $n = m + k$ (since $f = gh$), but also n is less than or equal to either m or k (since f divides either g or h). We conclude that one of m or k is zero, so either g or h is a non-zero constant polynomial. □

Remark Since primes and irreducibles coincide we can go on to consider the issue of unique factorisation. In doing this, we mimic the result, Theorem 1.3.3 in Chapter 1, and its proof. It may be an instructive exercise to re-read the proof of that earlier theorem before reading the proof below.

Theorem 6.3.4 *Every non-constant polynomial f can be written in the form*

$$f = f_1 \cdots f_r$$

where $f_1, \dots, f_r$ are irreducible polynomials. Furthermore this decomposition is unique in the sense that if also $f = g_1 \cdots g_r$, then $r = s$, and we may renumber the polynomials g_i so that each g_i is a constant multiple of the corresponding f_i (for $1 \leq i \leq r$).

Proof The proof is in two parts, the first to show that such a decomposition into irreducibles exists and the second to show that this decomposition is unique in the sense explained in the statement of the theorem.

We first show that f has a decomposition into irreducibles, using strong induction on the degree of f. The base case is for polynomials of degree 1 (linear polynomials). Since these are clearly irreducible, this case holds. Now suppose, by strong induction, that if g is any polynomial of degree less than or equal to k, then g has a decomposition of the required form. Let f be a polynomial of degree $k+1$. Then either f is irreducible (in which case the result holds with $r = 1$), or f has a factorisation as gh, with neither g nor h a constant polynomial. In that case, since $k+1$ is the sum of the degrees of g and h, our inductive hypothesis implies that each of g and h has degree less than or equal to k and hence has a decomposition into irreducibles. Writing these two decompositions next to each other gives the required decomposition of f.

For the second part of the proof, we use standard mathematical induction, this time on r, to show that any non-constant polynomial which has a factorisation into r irreducibles has a unique factorisation.

To establish the base case, $r = 1$, we suppose that f is an irreducible polynomial which may also be expressed as a product of non-constant irreducible polynomials $f = g_1 \cdots g_s$. If $s \geq 2$, we would contradict the fact that f is irreducible, so $s = 1$. Thus f can only be 'factorised' as a constant multiple of an irreducible g, which would therefore itself be a constant multiple of f.

Now suppose that $r > 1$ and take as our inductive hypothesis the fact that any non-constant polynomial which has a decomposition into $r-1$ irreducibles has a unique decomposition (in the above sense). Suppose that

$$f = f_1 \cdots f_r = g_1 \cdots g_s$$

are two decompositions of f into irreducible polynomials. Since f_1 divides $g_1 \cdots g_s$ and f_1 is irreducible, hence prime, f_1 divides g_i for some i. After renumbering, we may suppose that g_1 is divisible by f_1. Since f_1 and g_1 are both irreducible, this means that g_1 is a constant multiple $c_1 f_1$ say. Now divide throughout by f_1 to obtain

$$f_2 \cdots f_r = c_1 g_2 \cdots g_s.$$

Since the polynomial on the left-hand side is a product of $r-1$ irreducibles (and, on the right, the non-zero constant term can be absorbed into g_2), the inductive hypothesis allows us to conclude that $r - 1 = s - 1$ (so $r = s$) and, after renumbering, each g_i is a constant multiple of f_i for $i = 2, \ldots, r$ and hence for $i = 1, \ldots r$. □

Example If $f(x) = x^3 + 2x^2 - x - 2$ then, noting that $f(1) = 0 = f(-1)$, we easily obtain the factorisation $f(x) = (x-1)(x+1)(x+2)$ into irreducible polynomials.

Example The question of whether a polynomial is irreducible depends on where the coefficients come from. Thus $f(x) = x^2 + 1$ is irreducible as a real polynomial, since the only way it could be reducible would be if it were a product of two linear polynomials. Then it would have two real zeros, which is impossible, as we have seen. However, over the complex numbers $x^2 + 1 = (x + \mathrm{i})(x - \mathrm{i})$. Again as we have seen, $f(x) = (x + 1)^2$ over $\mathbb{Z}_2$. However, in $\mathbb{Z}_3$, $f(0) = 1$, $f(1) = 2$ and $f(2) = 2$, so in this case, $f(x)$ has no linear factor (by Corollary 6.2.3) so f is irreducible over $\mathbb{Z}_3$.

A very deep result (the Fundamental Theorem of Algebra), well beyond the scope of this book, says that every non-constant polynomial f over $\mathbb{C}$ has a complex zero, α_1, say. Then by Proposition 6.2.4, we may write $f(x)$ as $(x - \alpha_1) f_1(x)$ with $f_1(x)$ a polynomial of degree one less than the degree of f. Clearly, by repeating this procedure, we may write each non-constant polynomial in $\mathbb{C}$ as a product of linear polynomials. Thus if f is an irreducible complex polynomial we deduce that f must be a linear polynomial.

Corollary 6.3.5 *Every irreducible polynomial in $\mathbb{C}[x]$ is linear. Thus, if f is a complex polynomial of degree n, with a_n being the coefficient of x^n, then there are complex numbers $\alpha_1, \alpha_2, \ldots, \alpha_n$ such that*

$$f(x) = a_n(x - \alpha_1)(x - \alpha_2)\ldots(x - \alpha_n).$$

The situation is different for real polynomials. As we have mentioned (and prove in our Appendix), it is easy to show that if $f(x)$ is a real polynomial and α is a complex zero for $f(x)$, then the complex conjugate $\bar{\alpha}$ will be a root of $f(x)$. By unique factorisation this means that

$$(x - \alpha)(x - \bar{\alpha}) = x^2 - (\alpha + \bar{\alpha})x + \alpha\bar{\alpha}$$

divides $f(x)$. Since $\alpha + \bar{\alpha}$ and $\alpha\bar{\alpha}$ are both real, this quadratic is over the reals. So, given a polynomial f in $\mathbb{R}$, we first regard it as a complex polynomial in disguise (so each real coefficient is now regarded as a complex number which happens to be real). Then there is a factorisation of f as a polynomial in $\mathbb{C}[x]$, expressing f as a product of linear (complex) factors. Since each complex root will occur together with its conjugate 'partner', we may group each such pair together, as above, to produce a real quadratic polynomial (with no real roots). In particular, this implies that the irreducible real polynomials are either linear, or certain quadratics (those with no real zeros).

Corollary 6.3.6 *Irreducible real polynomials are either linear or quadratic. Thus, if f is a real polynomial of degree n, there are integers r and m with*

$n = m + 2r$ such that, if a_n is the coefficient of x^n, then there are real numbers $\alpha_1, \alpha_2, \ldots, \alpha_m$ and irreducible quadratic polynomials $x^2 + b_i x + c_i$ $(1 \leq i \leq r)$ with b_i, c_i real numbers such that

$$f(x) = a_n(x - \alpha_1)(x - \alpha_2)\ldots(x - \alpha_n)(x^2 + b_1 x + c_1)\ldots(x^2 + b_r x + c_r).$$

The situation is nothing like as straightforward as this for polynomials over $\mathbb{Z}_p$. There are irreducible polynomials in this case of arbitrarily large degrees. Since there will only be a finite number of polynomials of a given degree over $\mathbb{Z}_p$, we can work (inductively) up to a given polynomial f of degree n by considering all the polynomials over $\mathbb{Z}_p$ of degree less than n and then checking which of these are irreducible (that is, not themselves divisible by any polynomials of smaller degree in our list) and checking which actually divide f. However this is by no means a short calculation, especially since it can be shown that for any integer n there is an irreducible polynomial of degree n over $\mathbb{Z}_p$. A few examples over $\mathbb{Z}_p$ will illustrate this.

Example 1 Consider $f(x) = x^4 + x^3 + x^2 + x + 1$ over $\mathbb{Z}_2$. Since $f(0) = 1 = f(1)$, f has no zeros over $\mathbb{Z}_2$. This means that $f(x)$ has no linear factor, so if $f(x)$ does factorise as $g(x)h(x)$, the only possibility is that g and h are both quadratic. Then these factors would have the forms $g(x) = ax^2 + bx + c$ and $h(x) = dx^2 + ex + f$, where each of a, b, c, d, e and f is 0 or 1. Then

$$x^4 + x^3 + x^2 + x + 1 = (ax^2 + bx + c)(dx^2 + ex + f).$$

Expanding gives

$$(ax^2 + bx + c)(dx^2 + ex + f) = adx^4 + (ae + bd)x^3 + (be + af + cd)x^2 + (ce + bf)x + cf.$$

Equating coefficients of x^4 we see that $ad = 1$. Since each of a, d is 0 or 1 this implies that $a = d = 1$ Similarly, looking at the constant term shows that $cf = 1$, so $c = f = 1$. So now we have that

$$x^4 + x^3 + x^2 + x + 1 = (x^2 + bx + 1)(x^2 + ex + 1).$$

Now equate coefficients of x^3 to see that $(e + b) = 1$ (so one of e and b is 0, the other is 1), and of x^2 to obtain $be + 1 + 1 = 1$ (so $be = 1$). Since these equations, $e + b = 1$ and $be = 1$, have no common solution, there cannot be any such factorisation and we conclude that the polynomial f is irreducible.

Example 2 Consider $f(x) = x^4 + 1$ over $\mathbb{Z}_3$. Again we start by looking for zeros of f. Now, $f(0) = 1$, $f(1) = 2$ and $f(2) = 1 + 1 = 2$, so there are

no linear factors. Suppose that $f(x)$ were a product of quadratic factors, so

$$x^4 + 1 = (ax^2 + bx + c)(dx^2 + ex + f).$$

As in our previous example, we see that $ad = 1$ (so neither a nor d can be zero, and in fact $a = d$). Also $cf = 1$ (so $c = f$). Without loss of generality (check that you see why), we may suppose that $a = d = 1$. From the coefficients of x^3, we note that $(e + b) = 0$, so we have $e = -b$. Comparing coefficients of x^2, we have $0 = af + be + cd = f - b^2 + c = 2c - b^2$ so $2c = b^2$. Trying $c = 1$ gives the equation $2 = b^2$, which has no solution. Trying $c = 2$ we have $1 = b^2$, which does have a solution, try $b = 1$, so $e = -b = -1 = 2$. Finally, the coefficient of x gives $0 = ce + bf = -cb + bf = (-c + f)b$ which is consistent and gives no new information. So these equations do have a solution: $a = 1, b = 1, c = 2$ and $d = 1, e = 2, f = 2$ and we have the factorisation

$$x^4 + 1 = (x^2 + x + 2)(x^2 + 2x + 2) \qquad (= (x^2 + x - 1)(x^2 - x - 1))$$

(which you should check). Since neither of these quadratics has a zero (otherwise f would have a zero), our given quartic polynomial is a product of two irreducible quadratics.

The cases of polynomials over $\mathbb{R}$ and over $\mathbb{C}$ are unusual in that we can give an explicit description of the irreducibles. The situation for polynomials over $\mathbb{Z}_p$ is much more like that for integers. Just as we do not have a way to describe uniformly all prime numbers, so we cannot describe uniformly all irreducible polynomials over $\mathbb{Z}_p$.

Exercises 6.3

1. Prove, by induction on r, that if f is an irreducible polynomial and f divides the product $f_1 \cdots f_r$, then f divides one of $f_1, f_2, \ldots, f_r$.
2. Use Theorem 1.6.3 to factorise $x^{p-1} - 1$ over $\mathbb{Z}_p$.
3. Find all irreducible quadratic polynomials, with leading coefficient 1, over $\mathbb{Z}_p$ when p is 2, 3.
4. Find real numbers a, b such that the quartic polynomial $x^4 + 1$ has a decomposition as a product of two quadratics $(x^2 + ax + 1)(x^2 + bx + 1)$. Noting that $x^2 - y^2 = (x - y)(x + y)$, factorise $x^8 - 1$ over $\mathbb{R}$. Hence find a factorisation of $x^8 - 1$ over $\mathbb{C}$. Finally, factorise the polynomial $x^8 - 1$ as a product of irreducibles over $\mathbb{Z}_3$.
5. Find all irreducible cubic polynomials over $\mathbb{Z}_2$.
6. Give examples of polynomials f, g and h such that f divides gh, but f divides neither g nor h.

6.4 Polynomial congruence classes

To generalise the idea of congruence classes from Chapter 1, we take a fixed non-constant polynomial f and say that two polynomials r and s are **congruent modulo** f if $r - s$ is divisible by f. We then write

$$r \equiv s \mod f.$$

Example In $\mathbb{R}[x]$, we may take $f(x) = x^2 - 1$. Congruence classes modulo this polynomial have somewhat undesirable properties.

Let $r(x)$ be a polynomial of degree greater than 1. Then we can use the division algorithm for polynomials to write r in the form $qf + t$, with t being of degree strictly less than the degree of f (which is 2). That is, $t = ax + b$ for some, possibly zero, $a, b \in \mathbb{R}$. Thus every polynomial is congruent modulo $x^2 - 1$ to either a constant or a linear polynomial.

Consider the linear polynomials $x - 1$ and $x + 1$. Neither of these is congruent to the zero polynomial (because f does not divide either $x - 1$ or $x + 1$). However $(x + 1)(x - 1) = x^2 - 1$, so their product is congruent to the zero polynomial. (As in the case of integers modulo n we say that we have 'zero-divisors'.) The same complication will clearly arise whenever we take $f(x)$ to be a reducible polynomial, since if $f(x) = g(x)h(x)$ then the product of $g(x)$ and $h(x)$ will be congruent to zero. For this reason, we will sometimes restrict ourselves to the case when the polynomial $f(x)$ is irreducible. The situation is precisely analogous to that which arose when we considered $\mathbb{Z}_n$ when n is not a prime and so, for many purposes, we concentrated on $\mathbb{Z}_p$ only for p a prime.

Notation and terminology In line with the notation of Chapter 1, we will write $[r]_f$ to denote the set of all those polynomials s which are congruent to r modulo f. This set is referred to as the **polynomial congruence class** of r modulo f.

The argument used in the first part of the example above shows that each polynomial congruence class modulo f has a **standard representative** – the unique polynomial in the class which is of degree less than the degree of f. It can be found by taking any polynomial, r, in the class, applying the division algorithm to write $r = qf + t$ with $\deg(t) < \deg(f)$: then t is the standard representative of the class of r. To see that t is unique, suppose that $s \equiv r \mod f$: then $s = pf + r$ for some polynomial p. Since $r = qf + t$ this gives $s = (p + q)f + t$ and we see that r and s have the same remainder when divided by f.

We now consider operations on polynomial congruence classes.

Definition Fix a non-constant polynomial f, and let r and s be any other polynomials. Then we define the **sum** and **product** of the polynomial congruence classes of r and s as follows

$$[r]_f + [s]_f = [r+s]_f \qquad \text{and} \qquad [r]_f[s]_f = [rs]_f.$$

As for the integers, there is a potential problem with this definition. We have defined the sum and the product of two congruence classes by reference to particular examples of polynomials in the classes. However, we need to be sure that if we chose to represent $[r]_f$ (or $[s]_f$) by some other polynomial in the class, then we would get the same congruence class for the sum (and for the product). We check this, in another proof which precisely generalises that of the corresponding result in Chapter 1 (Theorem 1.4.1)

Theorem 6.4.1 *Let f be a non-constant polynomial, and let r, s, t be any polynomials. Suppose that $[r]_f = [t]_f$. Then*

(i) $[r+s]_f = [t+s]_f$, *and*
(ii) $[rs]_f = [ts]_f$.

Proof (i) Since $[r]_f = [t]_f$, f divides $r - t$, so we can write $r = t + kf$, for some polynomial k. Therefore

$$\begin{aligned}[r+s]_f &= [t+kf+s]_f \\ &= [s+t+kf]_f \\ &= [s+t]_f \text{ (by definition of congruence classes)}\end{aligned}$$

as required.

(ii) With the above notation,

$$\begin{aligned}[rs]_f &= [(t+kf)s]_f \\ &= [ts+kfs]_f \\ &= [ts]_f\end{aligned}$$

as required. □

By applying this result twice, we obtain the following easy consequence.

Corollary 6.4.2 *If $[r]_f = [t]_f$ and $[s]_f = [u]_f$ then*

(i) $[r+s]_f = [t+u]_f$
(ii) $[rs]_f = [tu]_f$.

Note that nowhere in the above discussion did we need to specify whether the polynomials in question had coefficients which were real, complex or over $\mathbb{Z}_p$. We therefore see that these two results are independent of where the coefficients come from. We now consider some examples of polynomial congruence classes.

Example 1 In $\mathbb{R}[x]$, we take $f(x)$ to be the irreducible polynomial $x^2 + 1$. As in the first example of this section, because this is a polynomial of degree 2 each polynomial congruence class has, for its standard representative, a constant or a linear polynomial. If r and s are standard representatives of their polynomial classes we will have that $r + s$ is a standard representative of its class but this need not be so for the product rs. The formula for the product of two polynomial congruence classes is different from that in $\mathbb{R}[x]$ itself because we replace every polynomial by a standard representative. For instance, take $r(x) = x + 1$ and $s(x) = x + 2$. Then $[rs]_f = [x^2 + 3x + 2]_f = [(x^2 + 1) + 3x + 1]_f = [3x + 1]_f$. The general formula can be computed as follows:

$$\begin{aligned}[ax + b]_f[cx + d]_f &= [acx^2 + (bc + ad)x + bd]_f \\ &= [ac(x^2 + 1) + (bc + ad)x + bd - ac]_f \\ &= [(bc + ad)x + bd - ac]_f.\end{aligned}$$

Even though this is a somewhat suprising outcome, it may be familar to those who know the formula for the product of two complex numbers (see the Appendix for a discussion of multiplication of complex numbers). To see how this connection arises, identify the variable x (or, rather, its polynomial congruence class) with the complex number i (not an entirely unreasonable thing to do, since we are putting $x^2 + 1$ congruent to 0 so the class of x will satisfy the equation $[x]_f^2 = -[1]_f$). Then we can regard $[ax + b]_f$ as $a\mathrm{i} + b$, $[cx + d]_f$ as $c\mathrm{i} + d$, and we have 'rediscovered' the formula for determining the real and imaginary parts of the product of two complex numbers.

Example 2 Now we work over $\mathbb{Z}_2$, and take $f(x)$ to be the cubic equation $x^3 + x + 1$. Since $f(0) = 1$ and $f(1) = 1 + 1 + 1 = 1$, $f(x)$ has no linear factors and so, since it has degree 3, f must be irreducible. Every congruence class will have, as standard representative, a polynomial of degree less than or equal to 2. So in this example our polynomial congruence classes are represented by polynomials of degree less than 3. Since all coefficients are 0 or 1, there are only eight such polynomials:

$$0, 1, x, x + 1, x^2, x^2 + 1, x^2 + x, x^2 + x + 1.$$

Now, computing successive powers of x, we get $x, x^2, x^3 = (x^3 + x + 1) + (x + 1)$. Thus $[x^3]_f = [x + 1]_f$. Similarly $[x^4]_f = [x^3 x]_f = [x^2 + x]_f$. Then $[x^5]_f = [x^2 + x + 1]_f$, $[x^6]_f = [x^2 + 1]_f$ and $[x^7]_f = [1]_f$. Thus each of the seven non-zero polynomial congruence classes occurs as a power of $[x]_f$. This will not be the case in all examples but, in a case like this, where the set of non-zero polynomial congruence classes may be represented by powers of $[x]_f$, we say that $[x]_f$ is a **primitive** polynomial congruence class (meaning that every other polynomial congruence class is a power of this one).

It is clearly the case that the set of polynomial congruence classes is closed under addition. Also, the set of non-zero polynomial congruence classes is closed under multiplication. Furthermore, each non-zero polynomial congruence class has an inverse in the sense of the following definition.

Definition A polynomial congruence class $[r]_f$ has an **inverse** if there is a polynomial congruence class $[s]_f$ such that $[r]_f[s]_f = [1]_f$ (then we write $[r]_f^{-1} = [s]_f$). Note that this equation means that $rs = tf + 1$ for some polynomial t. Clearly the congruence class of the zero polynomial cannot have an inverse.

Example In the example above, where we work over $\mathbb{Z}_2$ and take $f(x)$ to be the cubic equation $x^3 + x + 1$, the inverses are as follows:

element	1	x	$x + 1$	x^2	$x^2 + 1$	$x^2 + x$	$x^2 + x + 1$
inverse	1	$x^2 + 1$	$x^2 + x$	$x^2 + x + 1$	x	$x + 1$	x^2

The set of non-zero congruence classes of polynomials modulo $f = x^3 + x + 1 \in \mathbb{Z}_2[x]$ is an example of a **field** in that the set is closed under addition, the non-zero elements are closed under multiplication and all have inverses (for the exact definition see Section 4.4). This is a field with a finite number of elements (namely 8). Fields with a finite number of elements were first discussed by our young French genius Galois. He proved that any finite field has a prime power number of elements, and that this number uniquely determines the structure of the field. For this reason the notation $GF(p^n)$ is often used for a finite (or 'Galois') field with p^n elements: so we would write this one as $GF(2^3)$ or just as $GF(8)$.

Next we see that we always have inverses of non-zero congruence classes provided we take f to be an irreducible polynomial.

Proposition 6.4.3 *Let f be an irreducible polynomial. Then every non-zero polynomial congruence class modulo f has an inverse.*

Proof Let r be a polynomial not divisible by f (so $[r]_f$ is not equal to $[0]_f$). Consider a greatest common divisor of r and f. Such a polynomial is not a multiple of f (since r is not divisible by f), but must divide f. Since f is irreducible, this greatest common divisor must, therefore, be a non-zero constant polynomial c, say.

Then there are polynomials u and v, which we can find as in the previous section, such that

$$c = ur + vf.$$

Since c is a non-zero constant, it is valid to divide through by c to obtain

$$1 = u_1 r + v_1 f$$

where $u_1 = u/c$ and $v_1 = v/c$. This means that

$$[1]_f = [u_1 r]_f = [u_1]_f [r]_f,$$

so $[u_1]_f$ is an inverse for $[r]_f$. □

This result is a generalisation of two results from Chapter 1, namely Corollary 1.4.5 and Corollary 1.4.6.

Remark A general method for constructing a finite field with p^n elements now becomes clear. We search among the polynomials of degree n over $\mathbb{Z}_p$ for an irreducible one, f, say (we have stated, but not proved, that there will always be such a polynomial). Then we form the set of polynomial congruence classes modulo f. This carries the operations of addition and multiplication and we know that, since f is irreducible, each non-zero polynomial congruence class has an inverse. It is then not difficult to see that we have constructed our desired field with p^n elements. It can also be shown that, in this general situation, there will always be a primitive polynomial congruence class (not usually $[x]_f$, however).

Before finishing this section, we look at a further example where the polynomial $f(x)$ is not irreducible. This example will occur again in the next section.

Example Let n be any integer and take $f(x) = x^n - 1$ over $\mathbb{Z}_p$. This is always reducible, since $f(1) = 0$. As we have seen, within the set of polynomial congruence classes, there will, therefore, be zero-divisors and such a class will not have an inverse. Nevertheless, we can still consider the set of polynomial congruence classes modulo f (even though they will not form a field). As before, standard representatives for these will be all polynomials of degree less than n.

Now, if we multiply the polynomial

$$g(x) = a_0 + a_1x + a_2x^2 + \cdots + a_{n-1}x^{n-1}$$

by the linear polynomial x, we see that

$$\begin{aligned}[xg(x)]_f &= [a_0x + a_1x^2 + \cdots + a_{n-2}x^{n-2} + a_{n-1}x^n]_f \\ &= [a_0x + a_1x^2 + \cdots + a_{n-2}x^{n-2} + a_{n-1} \cdot 1]_f \\ &= [a_{n-1} + a_0x + a_1x^2 + \cdots + a_{n-2}x^{n-2}]_f\end{aligned}$$

since $x^n - 1 = 0 \bmod f$. The important point to note here is that the coefficients of the powers of x have cycled round.

Exercises 6.4

1. Let f be the irreducible quadratic $x^2 + x + 2$ over $\mathbb{Z}_3$. Write down the (9) standard representatives of the congruence classes modulo f. Draw up the multiplication table of the (8) non-zero congruence classes. Find a polynomial g such that every congruence class modulo f is a power of the class $[g]_f$.
2. Let f be any non-constant polynomial and let r, s and t be any polynomials. Suppose that 1 is a greatest common divisor for f and t. Show that if $[rt]_f = [st]_f$ then $[r]_f = [s]_f$.
3. Find the inverses of the following polynomial congruence classes:
   (i) $[g]_f$ over $\mathbb{Z}_2$ when $f(x) = x^2 + x + 1$ and $g(x) = x + 1$;
   (ii) $[g]_f$ over $\mathbb{Z}_3$ when $f(x) = x^3 + x^2 + x + 2$ and $g(x) = x^2 + x$;
   (iii) $[g]_f$ over $\mathbb{R}$ when $f(x) = x^2 + 1$ and $g(x) = x + 1$.

6.5 Cyclic codes

Definition A code C of **length** n (that is, whose words are of length n) over $\mathbf{B} = \mathbb{Z}_2$ is said to be **cyclic** if

(1) C is a linear code, and
(2) if $c_0c_1 \ldots c_{n-1}$ is a codeword, then so is $c_{n-1}c_0 \ldots c_{n-2}$ (it will be clear later why it is more convenient to label the entries from 0 to $n - 1$ rather than from 1 to n).

Examples We consider two examples from Section 5.4.

First, consider the code with codewords

0000 0011 0101 1001 0110 1010 1100 1111.

When we cycle 1010, we obtain 0101 which is not a codeword. Thus this code is not cyclic.

For our second example take the 3-repetition code with codewords in $\mathbf{B}^6$. This has 4 codewords, namely

$$000000 \quad 101010 \quad 010101 \quad 111111.$$

It is clear that this is a cyclic code.

Our first aim is to give a fairly concrete description of all cyclic codes of length n. To do this we start by representing a general vector of length n, say $(a_0, a_1, \ldots, a_{n-1})$ by a polynomial over $\mathbf{B} = \mathbb{Z}_2$. This is done by using the individual entries in the vector as 'markers' for the appropriate power of x: thus $(a_0, a_1, \ldots, a_{n-1})$ corresponds to the polynomial of degree $n-1$ which is

$$f(x) = a_0 + a_1 x + \cdots + a_{n-1}x^{n-1}.$$

Then multiplication of this polynomial by x gives a polynomial of degree n,

$$a_0 x + a_1 x^2 + \cdots + a_{n-2}x^{n-1} + a_{n-1}x^n.$$

If now we consider polynomial congruence classes modulo $f(x) = x^n - 1$ (so $x^n \equiv 1 \mod f$), this is congruent to the polynomial

$$a_{n-1} + a_0 x + a_1 x^2 + \cdots + a_{n-2}x^{n-1},$$

which corresponds to the n-tuple $(a_{n-1}, a_0, a_1, \ldots, a_{n-2})$. Note that this corresponds to the 'cycling' operation which occurs in the second clause of the definition of cyclic code. We actually have a bijection with the set of polynomial congruence classes modulo $x^n - 1$ because each such class has a unique standard representative polynomial of degree less than n and so corresponds to a unique vector of length n. We will frequently move between these two interpretations of codeword as vectors or as polynomial congruence classes. It is not difficult from this point of view to establish our first result.

Proposition 6.5.1 *A code C of length n, regarded as a code of polynomial congruence classes with respect to $f(x) = x^n - 1$ is a cyclic code if and only if*

(i) *C is a linear code, and*
(ii) *for any polynomial t, if $[g]_f$ is a codeword in C, then so is $[tg]_f$.*

Proof Suppose first that C is a cyclic code. Then C is linear. Also, as we have seen, the word obtained from the cyclic permutation of a codeword is another codeword. Now, cyclic permutation corresponds precisely to multiplication by

$[x]_f$ in the set of polynomial congruence classes with respect to $f(x) = x^n - 1$. It follows (by induction) that multiplication by $[x^i]_f$ corresponds to this cyclic permutation being performed i times and hence, if the polynomial congruence class $[g]_f$ corresponds to a codeword, then so does $[x^i g]_f$. Since our code is linear any sum of such terms is a codeword. Therefore if $[g]_f$ is a codeword of a cyclic code, then for any polynomial t, the polynomial congruence class of the product tg is another codeword (remember that we are working over $\mathbb{Z}_2$ so every coefficient of a polynomial is either 0 or 1).

For the converse, suppose that C satisfies conditions (i) and (ii) of the proposition. Then C is linear and, since we can multiply a codeword by $[x]_f$ and obtain another codeword, we deduce that the cyclic condition is satisfied and so we do have a cyclic code. □

Proposition 6.5.2 *Let C be a cyclic code. Then there is a polynomial congruence class $[g]_f$ such that every codeword in C is equivalent to the product of g and a polynomial.*

Proof Suppose that C is a cyclic code. Among all the non-zero codewords of C (regarded as polynomial congruence classes) choose one, g say, of smallest degree. Suppose that g has degree d. If now $[s]_f$ is any element in C, use the division algorithm to write $s = qg + r$ where either r is the zero polynomial or the degree of r is less than d. In this latter case, since $r = s - qg$ and since $[s]_f$ is a codeword, as is, by 6.5.1, $[qg]_f$, we deduce that $[r]_f$ is also in C. This contradicts the definition of d unless $r = 0$. We deduce that g divides s. Thus every codeword in C is equivalent to a product qg for some polynomial q. □

A non-zero polynomial g of least degree in a cyclic code is called a **generator polynomial** for the code and we say that the code C is **generated** by g.

Example Our cyclic code above with codewords

$$000000 \quad 101010 \quad 010101 \quad 111111$$

corresponds to congruence classes modulo $f(x) = x^6 - 1$ of the polynomials, 0, $f_1(x) = 1 + x^2 + x^4$, $f_2(x) = x + x^3 + x^5$ and $f_3(x) = 1 + x + x^2 + x^3 + x^4 + x^5$. Clearly, f_1 is one choice of generator polynomial with

$$0 = 0 \cdot f(x), \qquad f_2(x) \equiv x f_1(x) \quad \text{and} \quad f_3(x) \equiv (1 + x) f_1(x) \bmod f.$$

Note that, by 6.5.1 and 6.5.2, the codewords of a cyclic code are exactly those corresponding to congruence classes of the form $[tg]_f$ as t ranges over

polynomials (of degree less than n), where g is a generator for the code. Conversely, if h is any non-zero polynomial of degree less than n, then the set of codewords corresponding to congruence classes of the form $[th]_f$ (as t ranges over polynomials of degree less than n) is easily checked to be a cyclic code. (It is linear since $[sh]_f + [th]_f = [(s+t)h]_f$ and it is cyclic because cycling corresponds to multiplying by x.) It need not be the case, however, that h is a generator polynomial for this code, because h might not satisfy the property of having least degree. But there will be some generator polynomial for the code defined in this way. In fact the next result tells us that any generator polynomial for a cyclic code of length n must be a factor of $x^n - 1$.

Corollary 6.5.3 *Let C be a cyclic code of length n. If g is a generator polynomial for C then g divides $x^n - 1$.*

Proof Suppose that g does not divide $x^n - 1$, then we use the division algorithm to write $x^n - 1$ in the form $qg + r$ with the degree of r less than that of g. Since, note, $x^n - 1$ represents the zero word and we know that qg is a codeword, we deduce that r represents a codeword, contrary to the definition of g as having minimal degree, unless $r = 0$. Thus $x^n - 1 = q(x)g(x)$ and so g divides $x^n - 1$. □

For instance, in our example above, we have $x^6 - 1 = (x^2 - 1)f_1(x)$.

Once we have a generator polynomial, g, of degree m for a cyclic code C of length n, we can write down an $n - m$ by m generator matrix by placing the coefficients of $g(x)$ in its first row, those of $xg(x)$ in its second, those of $x^2g(x)$ in the third row, and so on, with the coefficients of $x^{n-m-1}g(x)$ in the last row. Notice that this is a different type of generator matrix from those we used in Chapter 5, but it is nevertheless the case (it follows by 6.5.2) that we may obtain any codeword by taking linear combinations of the rows of our matrix.

As an example, the generator matrix associated with the generator polynomial $1 + x^2 + x^4$ in our previous example (so $n = 6, m = 4$) is the 2×6 matrix

$$\begin{pmatrix} 1\,0\,1\,0\,1\,0 \\ 0\,1\,0\,1\,0\,1 \end{pmatrix}.$$

Here is one further piece of terminology. We refer to a polynomial $p(x)$ which satisfies $x^n - 1 = p(x)g(x)$ as a **parity polynomial** for the code generated by g. The parity polynomial for our example above is $x^2 - 1$.

We now illustrate these ideas by discussing cyclic codes of length 6.

Example The polynomial $x^6 - 1$ factorises as $(x^3 - 1)(x^3 + 1)$. Over $\mathbf{B} = \mathbb{Z}_2$ this is the same as $(x^3 + 1)^2$. Also, over $\mathbb{Z}_2$, $x + 1$ is a factor of $x^3 + 1$, so we have the factorisation $x^6 - 1 = x^6 + 1 = (x + 1)^2(x^2 + x + 1)^2$. Thus we can list the generator polynomials for the various cyclic codes of length 6. They are:

$$1, \quad x + 1, \quad x^2 + 1, \quad x^2 + x + 1, \quad x^3 + 1, \quad (x^2 + x + 1)^2,$$
$$(x + 1)^2(x^2 + x + 1), \quad (x + 1)(x^2 + x + 1)^2, \quad x^6 + 1.$$

We consider in turn the codes with these generator polynomials.

(1) The polynomial 1 clearly generates the whole of $\mathbf{B}^6$, so the set of codewords is all 64 vectors of length 6 over $\mathbf{B}$. Clearly this code has no error detecting or correcting properties.

(2) The polynomial $1 + x$ gives 32 codewords given as linear combinations of the rows of the generator matrix

$$\begin{pmatrix} 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 \end{pmatrix}.$$

With just a little thought one sees that the 32 codewords will be all those vectors in $\mathbf{B}^6$ with an even number of 1s. This means that the least weight of a non-zero codeword for this code is 2 so it detects one error. (With one error, we obtain a word with an odd number of 1s, but we have no way of knowing where the error lies.)

(3) The polynomial $1 + x^2$ gives the generator matrix

$$\begin{pmatrix} 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 & 1 \end{pmatrix}.$$

It is again clear that the least weight of a non-zero codeword for this code is 2. Any single error will give rise to an odd number of 1s and so will be immediately detected.

(4) We next consider the irreducible quadratic $1 + x + x^2$, with associated matrix

$$\begin{pmatrix} 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 \end{pmatrix}.$$

In this case, we write out the 16 codewords:

000000	111000	011100	100100	001110	110110	010101	101010
000111	111111	011011	100011	001001	110001	010010	101101

We note that for this code the least weight of a non-zero codeword is 2 so again the code detects a single error.

In fact we can always detect a single error in a code with a non-constant generator polynomial g; let p be the parity polynomial for C (so $[gp]_f = [0]_f$). Then if r is any codeword we have $r = tg$ for some t, so $[rp]_f = [tgp]_f = [t \cdot 0]_f = [0]_f$. In fact (as we shall see in Exercise 6.5.5) the converse of this result holds. Hence the only polynomial congruence classes $[r]_f$ with $[rp]_f$ being zero are codewords. Thus we can tell whether we have a codeword by multiplying by p.

(5) Now consider the cubic polynomial $(x+1)(x^2+x+1) = x^3+1$. This gives the matrix

$$\begin{pmatrix} 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 & 1 \end{pmatrix}.$$

Thus there are 8 codewords. Clearly at least one of these (given by the first row of the matrix) has weight 2, so again the code detects (by multiplying by the parity polynomial) one error.

(6) For the polynomial $(x+1)^2(x^2+x+1) = x^4+x^3+x+1$ the corresponding matrix is

$$\begin{pmatrix} 1 & 1 & 0 & 1 & 1 & 0 \\ 0 & 1 & 1 & 0 & 1 & 1 \end{pmatrix}.$$

There are now just 4 codewords:

$$000000 \quad 110110 \quad 011011 \quad 101101.$$

Hence the least weight is 4. In fact this code is a two-word repetition code (every codeword has the form ww with w a word of even weight in $\mathbf{B}^3$). Hence this code detects up to 3 errors and it corrects one error, since with one error, one half of the word will be of odd weight so we can tell which is the correct half and hence recover the correct codeword.

(7) When the generator polynomial is $(x^2+x+1)^2 = x^4+x^2+1$, the generator matrix is

$$\begin{pmatrix} 1 & 0 & 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 & 1 \end{pmatrix}$$

and so the codewords are

$$000000 \quad 101010 \quad 010101 \quad 111111.$$

This has least weight 3, so detects two errors, since a two-error word will have weight 1, 2, 4 or 5. It also corrects one error (indeed, this code is the three repetition code we have already met).

(8) For the generator polynomial $(x+1)(x^2+x+1)^2 = x^5+x^4+x^3+x^2+x+1$, the generator matrix is

$$\begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 \end{pmatrix}.$$

The only codewords are 000000 and 111111, so the least distance is 6. This code detects up to 5 errors and corrects 2. With only two codewords of length 6 this code is certainly not efficient.

(9) In the final case, the generator polynomial is $x^n - 1$, so the only codeword is the zero vector (therefore this code cannot be used to transmit information).

Cyclic codes are extensively studied. Among the cyclic codes of special interest are the so-called quadratic residue codes. These have excellent error-correcting properties and so are of great practical use. More about cyclic codes can be found in [Hill].

Exercises 6.5

1. Let g be a polynomial over $\mathbb{Z}_2$. Show that if g is irreducible then the number of non-zero coefficients of powers of x (including the constant term) is an odd integer.
2. Factorise $x^5 - 1$ over $\mathbb{Z}_2$. Hence write down generator polynomials for all the cyclic codes of length 5 over $\mathbf{B}$, and state how many errors each cyclic code detects and how many errors it corrects.
3. Let C be the cyclic code of length 7 with generator polynomial $x^3 + x^2 + 1$. List the codewords of C, and show that every 7-vector is within one error of a codeword of C.
4. Use the polynomial from Exercise 6.5.3 to determine all cyclic codes of length 7 over $\mathbf{B}$, stating how many errors each cyclic code detects and how many errors it detects.
5. Let p be a parity polynomial for a cyclic code with generator polynomial g. Use the division algorithm to show that if c is a polynomial with $[cp]_f = [0]_f$ then c is a codeword.

Summary of Chapter 6

In this chapter, we defined polynomials and gave their basic 'arithmetic' properties. The exposition of our material was closely modelled on that of Chapter 1 for integers. This showed how essentially the same arguments from the earlier chapter can be used to produce a very similar theory for polynomials. In Section 6.1, we discussed the basic operations (addition, subtraction and multiplication) for polynomials. Section 6.2 was concerned with division for polynomials and included the Euclidean algorithm, one of the direct generalisations from Chapter 1. We considered factorisation of polynomials in Section 6.3. The results in this section depend on the ring of coefficients of our polynomials. Section 6.4 was a development of the idea of polynomial congruence classes (a direct generalization of congruence classes of integers). An important application of these was a method for, in principle, constructing any finite field. Finally, in Section 6.5, we showed how we could make use of polynomial congruence classes to produce an important special class of linear codes – the cyclic codes.

Appendix on complex numbers

The reader will be accustomed, from an early age, to the idea of extending number systems. The natural numbers are used for counting, but it soon becomes clear that questions involving natural numbers may not have answers which are natural numbers. For example: 'On a winter's day, the temperature is 6°C. At night the temperature falls by 10 degrees. What is the overnight temperature'? To deal with this problem, we extend the natural numbers to the set of integers, by including negatives. However, one soon meets integer equations with non-integer solutions. For example: 'Share 3 cookies between two people' (that is, 'Solve $2x - 3 = 0$'). Again, we extend our number system from integers to rationals (by including fractions). Even after extending to rationals, there are still unanswered questions. 'Find the ratio of the length of a diagonal of a square to the length of a side' (that is, 'Find x such that $x^2 = 2$'). This time we extend the rationals to the real numbers. It is usual to make do with the real number system for everyday life and for a good part of school life. As we have seen, however, a polynomial like $x^2 + 1$ cannot have real number zeros, since the square of a real number is never negative.

The way to meet this difficulty is to require the existence of a new number i for which $\mathrm{i}^2 = -1$. Then the set of numbers of the form $z = a + \mathrm{i}b$ (as a, b vary over the set of real numbers) is referred to as the set of **complex numbers**, with a being the **real part** of z and b being its **imaginary part**. We write $\mathbb{C}$ for the set of complex numbers:

$$\mathbb{C} = \{a + \mathrm{i}b : a, b \in \mathbb{R} \text{ where } \mathrm{i}^2 = -1\}.$$

Then we add two complex numbers by adding their real parts and their imaginary parts separately:

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

To multiply two complex numbers, we use the usual rules of algebra to simplify

brackets. In addition, we regard the symbol i as commuting with any real number (so $a(\mathrm{i}d) = \mathrm{i}(ad)$) and recall that $\mathrm{i}^2 = -1$:

$$(a + \mathrm{i}b)(c + \mathrm{i}d) = ac + \mathrm{i}bc + \mathrm{i}ad + \mathrm{i}^2bd = (ac - bd) + \mathrm{i}(bc + ad).$$

It may occur to the reader that this process is, potentially, a never-ending one. Maybe, in order to solve equations involving complex numbers, we need to invent other number systems. However this is not the case. A result known as the Fundamental Theorem of Algebra says that every non-constant polynomial over the complex numbers has a complex number as a zero. As we saw in Chapter 6, this is sufficient information to be able to prove that every zero of a complex polynomial is a complex number. This fact is sometimes expressed by saying that the field of complex numbers is **algebraically closed**. This result is not without its controversy. It took more than two centuries for complex numbers to become widely accepted, even among mathematicians. One feature disliked by some, is that although the Fundamental Theorem shows that every complex polynomial has a complex zero, its proof does not tell us how to find such a zero.

Complex numbers are often illustrated by the Argand diagram. This is nothing more than the ordinary plane from Cartesian geometry, but with the usual x-axis now being thought of as the 'real'-axis (corresponding to the real part, a, of the complex number $z = a + \mathrm{i}b$) and the usual y-axis being thought of as the 'imaginary' axis (corresponding to $\mathrm{i}b$ where b is the imaginary part of $a + \mathrm{i}b$). For some purposes, it is convenient to think of points in the Argand diagram using polar coordinates. Thus a point is specified by giving its distance, d, from the origin (this distance is known as its **modulus**) and the angle, θ, between the positive real axis and the line joining the point to the origin (this angle is known as its **argument**). This gives the representation $z = d\mathrm{e}^{\mathrm{i}\theta}$ of a complex number, where $\mathrm{e} = 2.71828\ldots$ is the base of natural logarithms and where θ is measured in radians. For an explanation of this notation you should consult a book which deals with complex numbers. This way of considering complex numbers is particularly useful when one wishes to find powers and roots of a given complex number.

Given a complex number $z = a + \mathrm{i}b$, its **complex conjugate** is the complex number $a - \mathrm{i}b$ (so this is the reflection of the given complex number in the real axis of the Argand diagram). Thus a complex number which is equal to its complex conjugate has no (i.e. zero) imaginary part and is purely real. It is usual to denote the complex conjugate of z by $\bar{z}$. If we add a complex number to its complex conjugate, we clearly get a real number (twice the real part of the complex number). It is also true that if we multiply a complex number by its complex conjugate we obtain a real number. To see this, consider the product

of a complex number z with its complex conjugate $\bar{z}$:

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

which is clearly a real number.

There are some very simple rules for dealing with conjugates.

Rules for complex conjugates *Let z_1 and z_2 be complex numbers. Then*

(i) $\overline{z_1 + z_2} = \overline{z_1} + \overline{z_2}$ *and*
(ii) $\overline{z_1 z_2} = \overline{z_1}\,\overline{z_2}$.

Proof (i) Consider the two sides of our claimed equality. First $\overline{z_1 + z_2}$ is the complex conjugate of $z_1 + z_2$. So, if $z_1 = a_1 + \mathrm{i}b_1$ and $z_2 = a_2 + \mathrm{i}b_2$, we have

$$\begin{aligned}\overline{z_1 + z_2} &= \overline{(a_1 + \mathrm{i}b_1) + (a_2 + \mathrm{i}b_2)} \\ &= \overline{(a_1 + a_2) + \mathrm{i}(b_1 + b_2)} \\ &= (a_1 + a_2) - \mathrm{i}(b_1 + b_2).\end{aligned}$$

The right-hand side is

$$\overline{z_1} + \overline{z_2} = (a_1 - \mathrm{i}b_1) + (a_2 - \mathrm{i}b_2) = (a_1 + a_2) - \mathrm{i}(b_1 + b_2).$$

Since these two expressions are equal, we have proved (i). The proof of (ii) may now be safely left to the reader, although it is slightly more complicated because one needs to expand expressions for products of complex numbers, then gather together their real and imaginary parts. □

Once these basic rules have been established, we can apply (ii) with $z_1 = z_2 = z$ to obtain $\overline{z^2} = \overline{z}^2$. It is then easy to prove (by mathematical induction) that, for all positive integers n, $\overline{z^n} = \overline{z}^n$. Suppose now that we have a complex polynomial

$$f(z) = a_0 + a_1 \cdot z + a_2 \cdot z^2 + \cdots + a_n \cdot z^n,$$

so the coefficients $a_0, a_1, \ldots, a_n$ are complex numbers. Using our rules for complex conjugates (and induction again), gives

$$\overline{f(z)} = \overline{a_0} + \overline{a_1} \cdot \overline{z} + \cdots + \overline{a_n} \cdot \overline{z^n}.$$

Now use the fact that $\overline{z^n} = \overline{z}^n$, to see that

$$\overline{f(z)} = \overline{a_0} + \overline{a_1} \cdot \overline{z} + \cdots + \overline{a_n} \cdot \overline{z}^n.$$

If in addition, each coefficient, a_i, is actually real (so that it is its own complex conjugate, $a_i = \overline{a_i}$), we obtain

$$\overline{f(z)} = a_0 + a_1\overline{z} + \cdots a_n\overline{z}^n \text{ which equals } f(\overline{z}).$$

Therefore if z is a zero of the polynomial $f(z)$ with real coefficients, then

$$0 = \overline{0} = \overline{f(z)} = f(\overline{z}).$$

We have therefore proved the statement used in Chapter 6: if $f(z)$ is a complex polynomial with real coefficients, then whenever z is a zero of f, so is $\overline{z}$.

Answers

Chapter 1

Exercises 1.1

1. (i) $2 \cdot 11 + (-3) \cdot 7 = 1$; (ii) $2 \cdot (-28) + (-1)(-63) = 7$; (iii) $7 \cdot 91 + (-5) \cdot 126 = 7$; (iv) $(-9)630 + 43 \cdot 132 = 6$; (v) $35 \cdot 7245 + (-53)4784 = 23$; (vi) $(-31)6499 + 47 \cdot 4288 = 67$. Note that there are many ways of expressing the gcd.
2. $20 \cdot 6 + (-10) \cdot 14 + 21 = 1$.
4. One example occurs when a is 4, b is 2 and c is 6.
5. Take $a = 2$, $b = 4$ and $c = 12$ for an example.
7. We start with both jugs empty, which we write as (0, 0); fill the larger jug, (0, 17); fill the smaller from the larger, (12, 5); empty the smaller, (0, 5); pour the remains into the smaller, (5, 0); then continue as (5, 17), (12, 10), (0, 10), (10, 0), (10, 17), (12, 15), (0, 15), (12, 3), (0, 3), (3, 0), (3, 17), (12, 8). Note that we add units of 17 and subtract units of 12 and, in effect, produce 8 as the linear combination $4 \cdot 17 - 5 \cdot 12$ of 17 and 12.

Exercises 1.2

1. $a_1 = 1$, $a_2 = 3$, $a_3 = 7$, $a_4 = 15$ and $a_5 = 31$. Now to prove, using induction, that a_{n+1} is a power of 2, first notice (for the base case) that when $n = 1$, $a_1+1 = 2$ which is a power of 2. Next suppose that $a_k + 1$ is a power of 2, say $a_k + 1 = 2^n$. Then

$$a_{k+1} + 1 = (2a_k + 1) + 1 = 2a_k + 2 = 2(a_k + 1) = 2.2^n = 2^{n+1}$$

which is also a power of 2 as required.

4. $1^3 + 2^3 + \cdots + n^3 = \{n(n+1)/2\}^2 = \frac{1}{4}n^4 + \frac{1}{2}n^3 + \frac{1}{4}n^2$.
6. $1 + 3 + \cdots + (2n-1) = n^2$.
11. $2^{11} - 1 = 23 \times 89$.
12. (a) False, the given argument breaks down on a set with 2 elements.
    (b) False, the base case $n = 1$ is untrue.

Exercises 1.3

1. The primes less than 250 are
   2, 3, 5, 7, 11, 13, 17, 19, 23, 29, 31, 37, 41, 43, 47, 53, 59, 61, 67, 71, 73, 79, 83, 89, 97, 101, 103, 107, 109, 113, 127, 131, 137, 139, 149, 151, 157, 163, 167, 173, 179, 181, 191, 193, 197, 199, 211, 223, 227, 229, 233, 239, 241.
3. (a)
$$136 = 2 \cdot 68 = 2^2 \cdot 34 = 2^3 \cdot 17$$
$$150 = 2 \cdot 75 = 2 \cdot 5 \cdot 15 = 2 \cdot 3 \cdot 5^2$$
$$255 = 5 \cdot 51 = 3 \cdot 5 \cdot 17.$$

   After trying small primes, we see that
$$713 = 23 \cdot 31$$
$$3549 = 3 \cdot 1183 = 3 \cdot 7 \cdot 169 = 3 \cdot 7 \cdot 13^2.$$

   Checking divisibility for all primes less than 70 shows that 4591 is prime.
   (b) Thus $(136, 150) = 2$, lcm $(136, 150) = 2^3 \times 3 \times 5^2 \times 17 = 10200$
   $(255, 3549) = 3$, lcm $(255, 3549) = 3 \times 5 \times 7 \times 13^2 \times 17 = 301665$.
4. If $c_n = p_1 \times \cdots \times p_n + 1$ then $c_1 = 3$, $c_2 = 7$, $c_3 = 31$, $c_4 = 211$ and $c_5 = 2311$ are all, as you may check, prime. But $c_6 = 30031 = 59 \cdot 509$ is not prime.

Exercises 1.4

1. (i) $48 - 8 = 40$ which is not divisible by 14 so the assertion is false.
   (ii) $48 - (-8) = 56$ which is divisible by 14, so $-8 \equiv 48 \bmod 14$.
   (iii) $10 - 0 = 10$ which is not divisible by 100, so the assertion is false.
   (iv) $357482 - 7754 = 349728$ which is divisible by 3643, so $357482 \equiv 7754 \bmod 3643$.
   (v) $135227 - 16023 = 1309204$ which is divisible by 25177 so the congruence holds.
   (vi) $33303 - 4015 = 29288$. Since 1295 does not divide 29288, the congruence does not hold.

2. When n is 6, we obtain

+	0	1	2	3	4	5
0	0	1	2	3	4	5
1	1	2	3	4	5	0
2	2	3	4	5	0	1
3	3	4	5	0	1	2
4	4	5	0	1	2	3
5	5	0	1	2	3	4

×	0	1	2	3	4	5
0	0	0	0	0	0	0
1	0	1	2	3	4	5
2	0	2	4	0	2	4
3	0	3	0	3	0	3
4	0	4	2	0	4	2
5	0	5	4	3	2	1

When n is 7, we obtain

+	0	1	2	3	4	5	6
0	0	1	2	3	4	5	6
1	1	2	3	4	5	6	0
2	2	3	4	5	6	0	1
3	3	4	5	6	0	1	2
4	4	5	6	0	1	2	3
5	5	6	0	1	2	3	4
6	6	0	1	2	3	4	5

×	0	1	2	3	4	5	6
0	0	0	0	0	0	0	0
1	0	1	2	3	4	5	6
2	0	2	4	6	1	3	5
3	0	3	6	2	5	1	4
4	0	4	1	5	2	6	3
5	0	5	3	1	6	4	2
6	0	6	5	4	3	2	1

3. (i) The inverse of 7 modulo 11 is 8;
   (ii) the inverse of 10 modulo 26 does not exist;
   (iii) the inverse of 11 modulo 31 is 17;
   (iv) the inverse of 23 modulo 31 is 27; and
   (v) the inverse of 91 modulo 237 is 112.
4. When n is 16, G_n has 8 elements, 1, 3, 5, 7, 9, 11, 13 and 15. The multiplication table is

	1	3	5	7	9	11	13	15
1	1	3	5	7	9	11	13	15
3	3	9	15	5	11	1	7	13
5	5	15	9	3	13	7	1	11
7	7	5	3	1	15	13	11	9
9	9	11	13	15	1	3	5	7
11	11	1	7	13	3	9	15	5
13	13	7	1	11	5	15	9	3
15	15	13	11	9	7	5	3	1

When n is 15 the elements of G_n are 1, 2, 4, 7, 8, 11, 13 and 14. The table is

	1	2	4	7	8	11	13	14
1	1	2	4	7	8	11	13	14
2	2	4	8	14	1	7	11	13
4	4	8	1	13	2	14	7	11
7	7	14	13	4	11	2	1	8
8	8	1	2	11	4	13	14	7
11	11	7	14	2	13	1	8	4
13	13	11	7	1	14	8	4	2
14	14	13	11	8	7	4	2	1

9. (i) The first calculation is in error; we can draw no conclusions about the second or third (in fact the second is wrong, but the third is correct).
   (ii) The underlined digit should be 3.

Exercises 1.5

1. (i) no solution;
   (ii) $[4]_{11}$;
   (iii) $[11]_{21}$ or $[11]_{84}$, $[32]_{84}$, $[53]_{84}$, $[74]_{84}$;
   (iv) $[6]_{17}$;
   (v) no solution;
   (vi) $[7]_{20}$ or $[7]_{100}$, $[27]_{100}$, $[47]_{100}$, $[67]_{100}$, and $[87]_{100}$;
   (vii) $[10]_{107}$.
2. (i) $[172]_{264}$;
   (ii) $[7]_{20}$;
   (iii) $[123]_{280}$.
3. 1944
4. (i) $x^4 + x^2 + 1 = (x^2 + 2)(x^2 + 2) = (x + 1)(x + 1)(x + 2)(x + 2)$.
   (ii) Reduce modulo 3.
5. The minimum number of gold pieces was 408.

Exercises 1.6

1. (i) 5; (ii) 6; (iii) 16; (iv) 20.
2. (i) $5^{20} \equiv 4 \bmod 7$; (ii) $2^{16} \equiv 0 \bmod 8$; (iii) $7^{1001} \equiv 7 \bmod 11$;
   (iv) $6^{76} \equiv 9 \bmod 13$.
5. $\phi(32) = 16$; $\phi(21) = 12$; $\phi(120) = 32$; $\phi(384) = 128$.
6. (i) $2^{25} \equiv 2 \bmod 21$;
   (ii) $7^{66} \equiv 7^2 \equiv 49 \bmod 120$.
   (iii) $\phi(100) = 40$. So the last two digits of 7^{162} are 49. Note that, since $(5, 100) \neq 1$, Euler's Theorem cannot be applied to calculate the last

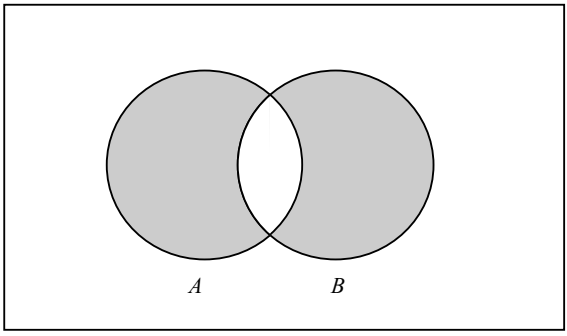

Fig. A1

two digits of 5^{121}. It can be seen directly that $5^k \equiv 25 \bmod 100$ for $k \geq 2$. Using this plus $3^{40} \equiv 1 \bmod 100$ and, say, $5^2 \cdot 3^4 \equiv 25{\cdot}81 \equiv 25 \bmod 100$, it follows that the last two digits of $5^{143} \cdot 3^{312}$ are 25. So the answer is 75.

10. $2^{37} - 1 = 223 \cdot 616318177$.
11. $2^{32} + 1 = 4294967297 = 641 \cdot 6700417$.
12. The message is FOOD.
13. The message is JOHN.

Chapter 2

Exercises 2.1

1. $X = W = Z; Y = V$.
2. Subsets of {a, b, c} are Ø {a}, {b}, {c}, {a, b}, {a, c}, {b, c} and {a, b, c}. Subsets of {a, b, c, d} are Ø, {a}, {b}, {c}, {d}, {a, b}, {a, c}, {a, d}, {b, c}, {b, d}, {c, d}, {a, b, c}, {a, b, d}, {a, c, d}, {b, c, d} and {a, b, c, d}. If X has n elements, the set of subsets of X has 2^n elements.
4. The symmetric difference is shaded in Fig. A1.
6. $X \times Y = \{(0,2), (0,3), (1,2), (1,3)\}$. This means that the set has $2^4 = 16$ subsets: Ø, {(0,2)}, {(0,3)}, {(1,2)}, {(1,3)}, {(0,2), (0,3)}, {(0,2), (1,2)}, {(0,2), (1,3)}, {(0,3), (1,2)}, {(0,3), (1,3)}, {(1,2), (1,3)}, {(0,2), (0,3), (1,2)}, {(0,2) (0,3), (1,3)} {(0,2), (1,2), (1,3)}, {(0,3), (1,2), (1,3)} and {(0,2), (0,3), (1,2), (1,3)}.
7. (i) True.
   (ii) False. One possible counterexample is given by taking A to be {1}, B to be {2}, C = {a} and D = {b}. Then (1, b) is in the right-hand term but not in the left-hand term.

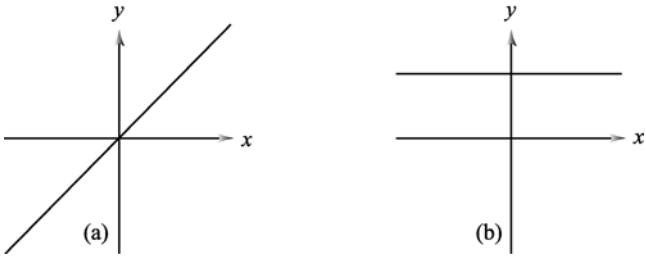

Fig. A2 (a) The graph of the identity function $y = f(x) = x$. (b) The graph of the constant function $y = f(x) = 1$.

8. $X \times Y$ has mn elements.
9. Take $A = B = \{1, 2\}$ and $X = \{(1, 1), (2, 2)\}$.

Exercises 2.2

1. A function is given by specifying $f(0)$ (two possibilities 0 or 5 for this), $f(1)$ (again 0 or 5) and $f(2)$ (again 0 or 5). There are $2 \times 2 \times 2 = 8$ such functions.
2. (i) bijective; (ii) not injective but surjective; (iii) neither injective nor surjective; (iv) surjective, not injective; (v) injective, not surjective.
3. See Fig. A2.
4. $fg(x) = x^2 - 1$; $gf(x) = x^2 + 2x - 1$; $f^2(x) = x + 2$; $g^2(x) = x^4 - 4x^2 + 2$.
5. For instance:
   (i) $f(x) = \log(x)$; (ii) $f(x) = \tan(x)$; (iii) $f(2k) = k$ and $f(2k - 1) = -k$.
6. (Compare Theorem 4.1.1.) There are 6 bijections.
7. (i) The inverse is $(4 - 3x)$;
   (ii) the inverse of $f(x) = (x - 1)^3$ is $1 + (x)^{1/3}$.

Exercises 2.3

1. (a) is R (reflexive); not S (symmetric); not WA (weakly antisymmetric); is not T (transitive) − consider $a = 2, b = 1, c = 0$.
   (b) R, S, not WA, T.
   (c) not R, S, not WA, not T.
   (d) not R, S, not WA, not T.
   (e) R, not S, not WA, T.
   (f) R, S, not WA, T.
   (g) R, not S, WA, not antisymmetric, T.

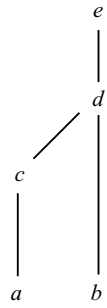

Fig. A3

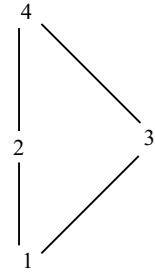

Fig. A4

2. (c) The relation of equality on any non-empty set is an example.
   (f) For instance, take
   $X = \{a, b,\}$ and
   $R = \{(b, b)\}$.
3. There is an unjustified hidden assumption that for each $x \in X$ there exists some $y \in X$ with xRy.
5. The Hasse diagram is as shown in Fig. A3.
6. The adjacency matrix is given below and Hasse diagram is as shown in Fig. A4.

	1	2	3	4
1	1	1	1	1
2	0	1	0	1
3	0	0	1	1
4	0	0	0	1

7. The equivalence classes are {1,2} and {3,4}.

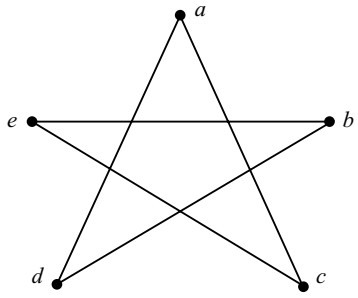

Fig. A5

8. The equivalence classes are {(1,1)}, {(1,2), (2,1)}, {(1,3), (2,2), (3,1)}, {(1,4), (2,3), (3,2), (4,1)}, {(2,4), (3,3), (4,2)} {(3,4), (4,3)}, {(4,4)}.
9. Not transitive (e.g. aRd, dRb but not aRb). The digraph requires no arrows since the relation is symmetric. See Fig. A5.

Exercises 2.4

1. The state diagrams are as shown in Fig. A6.
2. The tables are:

(1)

	a	b
0	1	1
1	1	2
2	0	2

(2)

	a	b
0	1	0
1	1	1
2	1	2

(3)

	a	b
0	0	1
1	2	1
2	2	3
3	3	3

3. (a) The words accepted by the machine are those in which the number of b's is of the form $1 + 3k$;
   (b) words with at least one a;
   (c) no words are accepted;
   (d) words of the form *b*a*b* where each * can denote any sequence (possibly empty) of letters.
4. The words accepted are those with an odd number of letters. The state diagram is as shown in Fig. A7 with $F = \{1\}$;
5. See Fig. A8.
6. The state diagram is as shown in Fig. A9 with $F = \{4\}$.

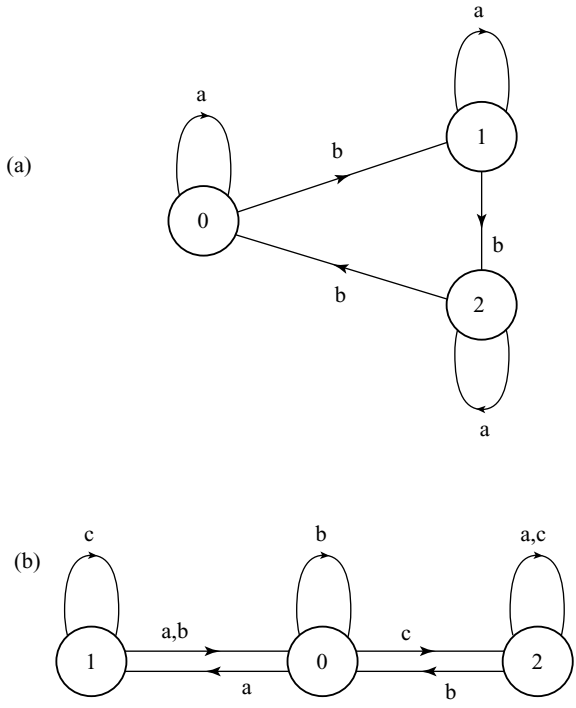

Fig. A6

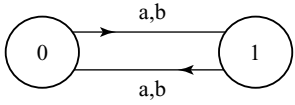

Fig. A7

Chapter 3

Exercises 3.1

1. (a) (i) $(p \wedge q) \rightarrow r$; (ii) $(\neg t \wedge p) \rightarrow ((s \vee q) \wedge \neg (s \wedge q))$.
   (b) (i) Either it is raining on Venus and the Margrave of Brandenburg carries his umbrella, or the umbrella will dissolve;
   (ii) It is raining on Venus, and either the Margrave of Brandernburg carries his umbrella or the umbrella will dissolve;
   (iii) The fact that it is not raining on Venus implies both that X loves Y and also that if the umbrella will dissolve then Y does not love Z;