Queen Mary
University of London

MTH 4104 Introduction to Algebra

Notes (version of February 12, 2020) **Spring 2020**

Contents

0 What is algebra?

Until around 1930, "algebra" meant the discipline of mathematics concerned with solving equations. An equation contains one or more symbols for unknowns, usually x, y, etc.; we have to find what numbers can be substituted for these symbols to make the equations valid. This is done by standard methods: rearranging the equation, applying the same operation to both sides, etc.

The word "algebra" is taken from the title of al Khwārizmī's algebra textbook *Ḥisāb al-jabr wa-l-muqābala*, circa 820. The word *al-jabr* means 'restoring', referring to the process of moving a negative quantity to the other side of an equation.
Al-Khwarizmi's name gives us the word "algorithm".

Sometimes we have to extend the number system to solve an equation. For example, there is no real number x such that $x^2 + 1 = 0$, so to solve this equation we must introduce complex numbers. Other times we may have equations to solve whose unknowns are not numbers at all but are objects of a different kind, perhaps vectors, matrices, functions, or sets.

In this way, attempting to solve equations leads one's attention to systems of mathematical objects and their abstract structure. The modern meaning of the word "algebra" (since van der Waerden's 1930 textbook *Moderne Algebra*) is the study of such abstract structure. In these new systems, we need to know whether the usual rules of arithmetic which we use to manipulate equations are valid. For example, if we are dealing with matrices, we cannot assume that AB is the same as BA.

So we will adopt what is known as the *axiomatic method*. We write down a set of rules called *axioms*; then anything we can *prove* from these axioms will be valid in all systems which satisfy the axioms. This leads us to the notion of *proof*, which is very important in mathematics.

What is mathematics about?

The short answer to this question: mathematics is about *proofs*. In any other subject, chemistry, history, sociology, or anything else, what one expert says can always be challenged by another expert. In mathematics, once a statement is proved, we are sure of it, and we can use it confidently, either to build the next part of mathematics on, or in an application of mathematics in another discipline.

In school teaching, this feature of mathematics does not get brought out; you are more likely to leave school thinking mathematics is about computation or formulae. One particularly bad habit instilled in school is the idea that if there are words in a mathematics question they are just window dressing, to be skipped over as you look for the numbers you need to start your workings. This is a terrible impulse when dealing with proofs and questions about proof, which are expressed in written prose in which every word is there for a mathematical reason. If you recognise this habit in yourself, you will need to break it!

In *Numbers, Sets, and Functions* you have seen your first examples of the techniques used for proofs. Most of them will come up in the course of this module. If you are not confident with words like "definition" or "theorem" or "to prove", I encourage you to refer to Appendix A at the end of these notes for a reminder of what these mean.

Conventions As you may already know, two different definitions are found for the set of *natural numbers*, $\mathbb{N}$. Some mathematicians say that $\mathbb{N} = \{0,1,2,3,\ldots\}$, including zero; others say that $\mathbb{N} = \{1,2,3,\ldots\}$, excluding zero[1]. For clarity, therefore, these notes will avoid the symbol $\mathbb{N}$, and distinguish the *nonnegative integers* $\mathbb{Z}_{\geq 0} = \{0,1,2,3,\ldots\}$ from the *positive integers* $\mathbb{Z}_{>0} = \{1,2,3,\ldots\}$.

I will write multiplication with $\cdot$, rather than $\times$. The $\times$ sign has other functions in this module, for example Cartesian product of sets, defined in Definition 3.1. Don't confuse this raised dot with the decimal point: $2 \cdot 3$ is not 2.3. (But there will not be many decimal numbers here. I prefer fractions.)

1 The integers

To study the integers from the point of view of modern algebra, a starting point is to understand how the basic arithmetic operations, addition, subtraction, multiplication, and division, behave in the context of integers. Integer addition, subtraction, and multiplication behave "normally", so they will not be our focus for now, though we will study what this "normal behaviour" is in Section 5.2.

Division is more interesting, because it is not always possible *within the integers*, and not just because of division by zero. We can say this another way without using division signs. If a and $b \neq 0$ are integers, there may not be an integer solution x to the equation $ax = b$ (for example, there is no integer solution to $2x = 1$).

[1] See `https://qmplus.qmul.ac.uk/mod/resource/view.php?id=871602` for my own feelings.

So we begin by making a closer study of the properties of division and divisibility in the integers.

1.1 Division with remainder

The *division rule* is the following property of the integers:

Proposition 1.1. *Let a and b be integers, and assume that $b > 0$. Then there exist integers q and r such that*

(a) $a = bq + r$;

(b) $0 \leq r \leq b - 1$.

Moreover, q and r are unique.

The numbers q and r are called the *quotient* and *remainder* when a is divided by b. The last part of the proposition (about uniqueness) means that, if q' and r' are another pair of integers satisfying $a = bq' + r'$ and $0 \leq r' \leq b - 1$, then $q = q'$ and $r = r'$.

Proof. We will show the uniqueness first. Let q' and r' be as above. If $r = r'$, then $bq = bq'$, so $q = q'$ (as $b > 0$). So suppose that $r \neq r'$. We may suppose that $r < r'$ (the case when $r > r'$ is handled similarly). Then $r' - r = b(q - q')$. This number is both a multiple of b, and also in the range from 1 to $b - 1$ (since both r and r' are in the range from 0 to $b - 1$ and they are unequal). This is not possible.

It remains to show that q and r exist. Let us first take the case that $a \geq 0$. Consider the multiples of b: 0, b, $2b$, ... Eventually these become greater than a. (Certainly $(a+1)b$ is greater than a.) Let qb be the last multiple of b which is not greater than a. Then $qb \leq a < (q+1)b$. So $0 \leq a - qb < b$. Putting $r = a - qb$ gives the result.

If $a < 0$, then instead we can let qb be the least multiple of $-b$ which is less than or equal to a, and let $r = a - qb$. (I leave it to you to check the details.) □

Since q and r are uniquely determined by a and b, we write them as $a \operatorname{div} b$ and $a \bmod b$ respectively. So, for example, $37 \operatorname{div} 5 = 7$ and $37 \bmod 5 = 2$.

The division rule is sometimes called the *division algorithm*. Most people understand the word "algorithm" to mean something like "computer program", but it really means a set of instructions which can be followed without any special knowledge or creativity and are guaranteed to lead to the result. A recipe is an algorithm for producing a meal. If I follow the recipe, I am sure to produce the meal. (But if I change things, for example by putting in too much chili powder, there is no guarantee about the result!) If I follow the recipe, and invite you to come and share the meal, I have to give you directions, which are an algorithm for getting from your house to mine.

The algorithm for long division by hand, which used to be taught in primary school (though this is out of fashion now), has been known and used for more than 3000 years. This algorithm is a set of instructions which, given two positive integers a and b, divides a by b and finds the quotient q and remainder r satisfying $a = bq + r$ and $0 \leq r \leq b - 1$. The example at right illustrates that if $a = 12345$ and $b = 6$, then $q = 2057$ and $r = 3$.

$$
\begin{array}{r}
2057 \\
\hline
6\,\big)\,12345 \\
12000 \\
\hline
345 \\
300 \\
\hline
45 \\
42 \\
\hline
3
\end{array}
$$

1.2 Greatest common divisor and least common multiple

Definition 1.2. Let a and b be integers. Then *a divides b* if and only if there exists an integer c such that $b = ac$. The notation for "a divides b" is $a \mid b$.

For example, $3 \mid 6$, but $6 \nmid 3$. The phrasing "a divides b" has several synonyms. We may also call a a *divisor* or *factor* of b, or call b a *multiple* of a.

Warning: You cannot substitute just any use of the word "divide" by $\mid$. The symbol $\mid$ is a relation symbol, like $=$ or $<$ (see Section 3.2 for more about relations and their symbols). This means that $a \mid b$ is a true-or-false statement, not a number. It is nonsense to write, for example[2], "$3 \mid 7$ has remainder 1". Another difference between $\mid$ and $/$ is which side of the symbol the divisor goes on: $a \mid b$ is true when b/a is an integer (as long as $a \neq 0$).

- Every integer, including zero, divides 0. This might seem odd, since we know that "you can't divide by zero"; but $0 \mid 0$ means simply that there exists a number c such that $0 = 0 \cdot c$, which is certainly true. On the other hand, zero doesn't divide any integer except zero.

- If a and b are nonnegative integers such that $a \mid b$ and $b \mid a$, then $a = b$. (In the language of relations, we say that $\mid$ is an *antisymmetric* relation on the nonnegative integers.) The same is not true if a and b could be any integers — why?

Definition 1.3. Let a and b be nonnegative integers. A *common divisor* of a and b is a nonnegative integer d with the property that $d \mid a$ and $d \mid b$. We call d the *greatest common divisor* if it is a common divisor, and if any other common divisor of a and b is smaller than d.

[2] If you write "$3 \mid 6 = 2$", this is legal mathematical syntax, but it means "$3 \mid 6$ and $6 = 2$". This is the same rule that lets you abbreviate e.g. "$0 \leq x$ and $x < 1$" to "$0 \leq x < 1$".

Thus, the common divisors of 12 and 18 are 1, 2, 3 and 6; and the greatest of these is 6. We write $\gcd(12,18) = 6$.

The remarks above about zero show that $\gcd(a,0) = a$ holds for any non-zero number a. What about $\gcd(0,0)$? Since every nonnegative integer divides zero, there is no greatest one. Later we will provide a corrected definition of gcd which addresses this flaw. See Proposition 1.9 and the discussion following.

Definition 1.4. The positive integer m is a *common multiple* of a and b if both $a \mid m$ and $b \mid m$. It is the *least common multiple* if it is a common multiple which is smaller than any other common multiple.

Thus the least common multiple of 12 and 18 is 36, written $\mathrm{lcm}(12,18) = 36$. Any two nonnegative integers a and b have a least common multiple. For there certainly exist common multiples, for example ab; and any non-empty set of nonnegative integers has a least element. (The least common multiple of 0 and a is 0, for any a.)

Is it true that any two nonnegative integers have a greatest common divisor? We will see that it is. Consider, for example, 8633 and 9167. Finding the gcd looks like a difficult job. But, if you know that $8633 = 89 \cdot 97$ and $9167 = 89 \cdot 103$, and that all the factors are prime, you can easily see that $\gcd(8633, 9167) = 89$.

Here is how this procedure works in general. We first recall a theorem on prime factorisation.

Theorem 1.5 (Fundamental Theorem of Arithmetic)**.** *Every positive integer n can be written as a product*

$$n = p_1^{e_1} \cdots p_k^{e_k}$$

where $p_1, \ldots, p_k$ are different prime numbers and $e_1, \ldots, e_k$ are positive integers. This expression is unique up to reordering of the factors.

- "Up to X", in mathematical prose, means that we are counting two things (in this case factorisations) to be the same if their only difference is X (in this case reordering). This makes sure we don't count 89×97 and 97×89 as different factorisations.

- What is the factorisation of the number 1? It's the *empty* product, where $k = 0$ and there are no factors. The product of no numbers is 1.

We can insert extra primes into the factorisation provided by the Fundamental Theorem of Arithmetic, as long as we give them the exponent 0. This is helpful when we want to compare multiple factorisations:

$$8633 = 89^1 \cdot 97^1 \cdot 103^0 \text{ and}$$
$$9167 = 89^1 \cdot 97^0 \cdot 103^1 \text{ have gcd}$$
$$89 = 89^1 \cdot 97^0 \cdot 103^0.$$

The following theorem supposes that we have done this, so that the same list of primes appears for two given integers.

Proposition 1.6. *Let a and b be positive integers, with factorisations $a = p_1^{e_1} \cdots p_k^{e_k}$ and $b = p_1^{f_1} \cdots p_k^{f_k}$.*

(i) $a \mid b$ if and only if $e_i \leq f_i$ for every $i = 1, \ldots, k$.

(ii)
$$\gcd(a,b) = p_1^{\min(e_1,f_1)} \cdots p_k^{\min(e_k,f_k)}.$$

Proof. To (i). Suppose $a \mid b$, so that there is an integer c such that $b = ac$. Because a and b are positive, c must also be. Therefore c has a prime factorisation, say $c = p_1^{g_1} \cdots p_k^{g_k}$. (Again, we may throw these primes into the earlier lists with their exponents set to 0, so that we can use the same list of primes every time.) It follows by laws of exponents that

$$p_1^{f_1} \cdots p_k^{f_k} = b = ac = p_1^{e_1+g_1} \cdots p_k^{e_k+g_k}.$$

Because of the uniqueness part of the Fundamental Theorem of Arithmetic, the left and right hand sides of this equation must be the same factorisation, implying that $f_i = e_i + g_i$ for every $i = 1, \ldots, k$. Therefore $e_i \leq e_i + g_i = f_i$ for every such i.

Conversely, suppose that $e_i \leq f_i$ for every i. Since a is not zero, b/a is a rational number, and we can test whether a divides b by testing whether it is an integer. By the laws of exponents,

$$\frac{b}{a} = \frac{p_1^{f_1} \cdots p_k^{f_k}}{p_1^{e_1} \cdots p_k^{e_k}} = \frac{p_1^{f_1}}{p_1^{e_1}} \cdots \frac{p_k^{f_k}}{p_k^{e_k}} = p_1^{f_1-e_1} \cdots p_k^{f_k-e_k}.$$

If $e_i \leq f_i$ for each i, then all of the exponents on the right hand side are greater than or equal to 0, which means the right hand side is an integer, since it is a product of integers (primes, with possible repetitions).

To (ii). By part (i), an integer d is a divisor of a if and only if its factorisation is $d = p_1^{g_1} \cdots p_k^{g_k}$, where $g_i \leq e_i$ for each i (and primes that don't divide a don't appear in d either). Since the same is true with b and f_i in place of a and e_i, we see that d is a common divisor of a and b if and only if $g_i \leq e_i$ and $g_i \leq f_i$ for each i. This gives two different upper bounds on g_i. Whichever one is greater is redundant, so it is

equivalent to keep only the lesser of the two, and require that $g_i \leq \min(e_i, f_i)$. Finally, to find the greatest of the common divisors d we can maximise all of these exponents independently. Therefore the gcd is $d = p_1^{g_1} \cdots p_k^{g_k}$, where each $g_i = \min(e_i, f_i)$ attains its upper bound. □

The downside of the method of Proposition 1.6 for finding the gcd of two numbers is that it is not *efficient*. Factorising a number into its prime factors is notoriously difficult. In fact, it is the difficulty of this problem which keeps internet commercial transactions secure!

Euclid discovered an efficient way to find the gcd of two numbers a long time ago. His method gives us much more information about the gcd as well. In the next section, we look at his method.

1.3 Euclid's algorithm

Euclid's algorithm is based on two simple rules:

Proposition 1.7.

$$\gcd(a,b) = \begin{cases} a & \text{if } b = 0, \\ \gcd(b, a \bmod b) & \text{if } b > 0. \end{cases}$$

Proof. We saw already that $\gcd(a,0) = a$, so suppose that $b > 0$. Let $r = a \bmod b = a - bq$, so that $a = bq + r$. If d divides a and b then it divides $a - bq = r$; and if d divides b and r then it divides $bq + r = a$. So the lists of common divisors of a and b, and common divisors of b and r, are the same, and the greatest elements of these lists are also the same. □

This really seems too slick to give us much information; but, if we look closely, it gives us an algorithm for calculating the gcd of a and b. If $b = 0$, the answer is a. If $b > 0$, calculate $a \bmod b = b_1$; our task is reduced to finding $\gcd(b, b_1)$, and $b_1 < b$. Now repeat the procedure; of $b_1 = 0$, the answer is b; otherwise calculate $b_2 = b \bmod b_1$, and our task is reduced to finding $\gcd(b_1, b_2)$, and $b_2 < b_1$. At each step, the second number of the pair whose gcd we have to find gets smaller; so the process cannot continue for ever, and must stop at some point. It stops when we are finding $\gcd(b_{n-1}, b_n)$, with $b_n = 0$; the answer is b_{n-1}.

This is *Euclid's Algorithm*. Here it is more formally:

To find $\gcd(a,b)$

Put $b_0 = a$ and $b_1 = b$.
As long as the last number b_n found is non-zero, put $b_{n+1} = b_{n-1} \bmod b_n$.
When the last number b_n is zero, then the gcd is b_{n-1}.

Example Find $\gcd(198,78)$.

$b_0 = 198$, $b_1 = 78$.

$198 = 2 \cdot 78 + 42$, so $b_2 = 42$.

$78 = 1 \cdot 42 + 36$, so $b_3 = 36$.

$42 = 1 \cdot 36 + 6$, so $b_4 = 6$.

$36 = 6 \cdot 6 + 0$, so $b_5 = 0$.

So $\gcd(198,78) = 6$.

Exercise Use Euclid's algorithm to find $\gcd(8633,9167)$.

1.4 Euclid's algorithm extended

The calculations that allow us to find the greatest common divisor of two numbers also do more.

Theorem 1.8. *Let a and b be nonnegative integers, and $d = \gcd(a,b)$. Then there are integers x and y such that $d = xa + yb$. Moreover, x and y can be found from Euclid's algorithm.*

Proof. The first, easy, case is when $b = 0$. Then $\gcd(a,0) = a = 1 \cdot a + 0 \cdot 0$, so we can take $x = 1$ and $y = 0$.

Now suppose that $r = a \bmod b$, so that $a = bq + r$. We saw that $\gcd(a,b) = \gcd(b,r) = d$, say. Suppose that we can write $d = ub + vr$. Then we have

$$d = ub + v(a - qb) = va + (u - qv)b,$$

so $d = xa + yb$ with $x = v$, $y = u - qv$.

Now, having run Euclid's algorithm, we can work back from the bottom to the top expressing d as a combination of b_i and b_{i+1} for all i, finally reaching $i = 0$. □

To make this clear, look back at the example. We have

$$\begin{aligned} 42 &= 1 \cdot 36 + 6, & 6 &= 1 \cdot 42 - 1 \cdot 36 \\ 78 &= 1 \cdot 42 + 36, & 6 &= 1 \cdot 42 - 1 \cdot (78 - 42) = 2 \cdot 42 - 1 \cdot 78 \\ 198 &= 2 \cdot 78 + 42, & 6 &= 2 \cdot (198 - 2 \cdot 78) - 1 \cdot 78 = 2 \cdot 198 - 5 \cdot 78. \end{aligned}$$

The final expression is $6 = 2 \cdot 198 - 5 \cdot 78$.

Euclid's algorithm proves that the greatest common divisor of two integers a and b is an integer d which can be written in the form $xa + yb$ for some integers x and y; and it proves this by giving us a recipe for finding d, x, y from the given values a and b. This is a *constructive* proof, in the sense discussed after Theorem 2.11.

We defined the greatest common divisor of a and b to be the largest nonnegative integer which divides both. Using the result of the extended Euclid's algorithm, we can say a bit more:

Proposition 1.9. *The greatest common divisor of the nonnegative integers a and b is the nonnegative integer $d \geq 0$ with the properties*

(a) $d \mid a$ and $d \mid b$;

(b) if e is a nonnegative integer satisfying $e \mid a$ and $e \mid b$, then $e \mid d$.

One of the assertions this proposition makes is that there is only *one* nonnegative integer d that has properties (a) and (b). This might escape you the first time you read the proposition, because it's conveyed in a subtle fashion: by the choice of the word "the", rather than "a", when we said "*the* integer d"!

Proof. Let $d = \gcd(a,b)$. Certainly condition (a) holds. Now suppose that e is a nonnegative integer satisfying $e \mid a$ and $e \mid b$. Euclid's algorithm gives us integers x and y such that $d = xa + yb$. Now $e \mid xa$ and $e \mid yb$; so $e \mid xa + yb = d$. □

Remark. With our earlier definition, we had to admit that $\gcd(0,0)$ doesn't exist, since every nonnegative integer divides 0 and there is no greatest one. But, with $a = b = 0$, there is a unique nonnegative integer satisfying the conclusion of Proposition 1.9, namely $d = 0$. So in fact this Proposition gives us a better way to define the greatest common divisor, which works for all pairs of nonnegative integers without exception!

The definition could be written word-for-word identically with the proposition, as follows.

Definition 1.10. The *greatest common divisor* of the nonnegative integers a and b is the integer $d \geq 0$ with the properties

(a) $d \mid a$ and $d \mid b$;

(b) if e is a nonnegative integer satisfying $e \mid a$ and $e \mid b$, then $e \mid d$.

But this definition cannot stand alone, since it is obvious neither that the number d it specifies exists, nor that it is unique, which is also implicitly being claimed when we say "*the* integer d". We still need Proposition 1.9 to establish that the definition works.

2 Polynomials and their roots

2.1 Polynomials

The equations at the historical heart of algebra are polynomial equations. We are familiar with polynomials as functions of a particular kind, e.g. $f_1(x) = x^2 + 1$ or $f_2(x) = 5x^3 - x + 1$ or $f_3(x) = \sqrt{2}x^4 - \pi x^3 - \sqrt{3}$. Let us start our study of polynomials by defining them carefully.

The polynomials f_1, f_2, and f_3 are *real*, because the powers of x appear multiplied by real numbers. But we will be referring to complex numbers as well in this section. We have seen at the close of *Numbers, Sets and Functions* that polynomials with complex coefficients are worthy of study. If you need a quick refresher on complex numbers, see Definition 5.3. If that isn't enough for you, please refer back to your *Numbers, Sets and Functions* notes and revise those.

The following definition is made to allow us to talk about either the real or the complex setting.

Definition 2.1. Let S be either the set $\mathbb{R}$ of real numbers or the set $\mathbb{C}$ of complex numbers. Let x be a variable.

A *polynomial in x with coefficients in S* is an expression

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

where $a_0, a_1, \ldots, a_{n-1}, a_n$ are elements of S. They are the *coefficients* of f.

The set of all such polynomials will be denoted by $S[x]$, that is, $\mathbb{R}[x]$ or $\mathbb{C}[x]$.

Here are some first remarks on this definition.

- When we do algebra in $\mathbb{R}[x]$ or $\mathbb{C}[x]$, we are working with expressions like "$x^2 + 1$" in their own right. We do not, by default, have in mind solving for x, or substituting in numbers for x. For example, in $\mathbb{R}[x]$, the answer to the question

  Does $x^2 + 1$ equal $4x - 2$?

  is just "no", not "$x = 1$ or $x = 3$". This is why I called the polynomial in my definition f, rather than $f(x)$.

- We may use a different symbol for the variable in place of x. For example, $t^4 + 6t^3 + 11t^2 + 6t$ is an element of $\mathbb{R}[t]$.

- Some coefficients may be zero. For example, $x^2 + 1$ would be written out in full as $1x^2 + 0x + 1$. This is a very different polynomial from $x^3 + 1 = 1x^3 + 0x^2 + 0x + 1$.

- *A polynomial is determined by its coefficients.* Compare this assertion to sentences like "a set is determined by its elements" or "a function is determined by its values": we mean that if you know all the coefficients of some polynomial, then you know everything about it.

  What about the converse? Do two different sequences of coefficients give two different polynomials? Yes, but there is one fly in the ointment. We don't want to say that a polynomial is changed by inclusion of extra zero terms, of the form $0x^n$. Therefore, we declare that two polynomials

  $$f = a_m x^m + a_{m-1}x^{m-1} + \cdots + a_1 x + a_0 \quad \text{and}$$
  $$g = b_n x^n + b_{n-1}x^{n-1} + \cdots + b_1 x + b_0$$

  are equal if and only if their sequences of coefficients are equal aside from leading zeroes. We can write this out formally by saying that there exists an integer p, with $p \leq n$ and $p \leq m$, so that and $a_i = b_i$ for all $i = 0, \ldots, p$, while $a_i = 0$ for all $i = p+1, \ldots, m$, and $b_i = 0$ for all $i = p+1, \ldots, n$. For example, $2x-4$ and $0x^3 + 0x^2 + 2x - 4$ are the same element of $\mathbb{R}[x]$.

Definition 2.2. The *degree* of a nonzero polynomial is the largest integer n for which its coefficient of x^n is non-zero.

That is, $x^2 + 1$ has degree 2, even though we could write it as $0x^{27} + x^2 + 1$. The zero polynomial doesn't have any non-zero coefficients, so its degree is not defined. The notation for the degree of f is $\deg f$.

We have special words for polynomials of low degree[3]:

degree	0	1	2	3	4	5	6	...
word	*constant*	*linear*	*quadratic*	*cubic*	*quartic*	*quintic*	*sextic*	...

By rights these words are adjectives, but except for "linear" they may also be used as nouns.

2.2 Roots of polynomials

Given an equation $f(x) = g(x)$ of two polynomials to be solved for x, collecting all the terms on one side lets us convert this to the equivalent equation $f(x) - g(x) = 0$, in which the left hand side is also a polynomial, $f - g$. So to solve polynomial equations it is enough to be able to find the *roots* or *zeroes* of a single polynomial, i.e. those values of its argument at which it evaluates to zero.

[3]Out in the mathematical world, the application of these words is not as cut and dried as I suggest. Every mathematician would call 0 a constant, but it is not a degree zero polynomial. Or, in some contexts, a "linear" function must have no constant term.

Remark. Being able to focus on roots is an example of the power of extending your number system: it is only possible due to the invention of *negative* numbers! Before negative numbers were accepted as legitimate – a slow process, not finished till the time of Leibniz in the 17th century – algebraists had to solve each of the three kinds of quadratic equation

$$ax^2 = bx + c; \qquad bx = ax^2 + c; \qquad c = ax^2 + bx$$

differently, since none of them could be converted to another.

Some polynomial equations can't be solved in $\mathbb{R}$. These include $x^2 = -1$, which has no real solution, and $x^3 = 2$, which has only one, though because of its degree we would like it to have three. Attempts to solve such equations[4] were what led mathematicians to invent larger number systems than $\mathbb{R}$.

The definition of the complex numbers expresses the insight that the first equation is the crucial one. That is, we invent a new number i, and declare that $\mathrm{i}^2 = -1$. We let $\mathbb{C} = \{a + b\mathrm{i} : a, b \in \mathbb{R}\}$, which is the smallest reasonable candidate for a number system that contains $\mathbb{R}$ as well as i, since we'd like to be able to add and multiply i with existing numbers. Then, wonderfully, *every* polynomial equation with real coefficients, or even with complex coefficients, can be solved inside $\mathbb{C}$! More precisely:

Theorem 2.3 (Fundamental Theorem of Algebra)**.** *Let $n \geq 1$, and let $a_0, a_1, \cdots, a_{n-1}, a_n$ be complex numbers, where $a_n \neq 0$. The polynomial equation*

$$a_n z^n + a_{n-1} z^{n-1} + \cdots + a_1 z + a_0 = 0$$

has at least one *solution inside* $\mathbb{C}$.

Despite the name, the proof of this theorem is beyond the scope of this module, because it relies on *analytic* properties of $\mathbb{R}$ or $\mathbb{C}$, that is, properties involving continuity and limits like the Intermediate Value Theorem. You will see a proof in the module *Complex Variables*.

I reassure you that, in this module, I will use real polynomials rather than complex ones for examples and exam questions wherever I can. The algebraic theory of complex polynomials is genuinely simpler, though, so it would be perverse to state the theorems for $\mathbb{R}[x]$ alone.

[4] There is a *cubic equation* for cubic polynomials, like the familiar quadratic equation, but even when using it on a real cubic with three real roots, complex numbers will sometimes turn up in intermediate steps. Phenomena like this are what really forced mathematicians of the sixteenth through eighteenth centuries to accept complex numbers.

2.3 How to find the roots

You already know how to solve real polynomials of low degree. Let's review these solutions, and see whether they still work when the polynomials might be complex.

Given two complex numbers α, β, we can consider the *linear equation*

$$\alpha z + \beta = 0,$$

to be solved for z. Provided α is non-zero, this equation has a unique solution, namely

$$z = -\frac{\beta}{\alpha}.$$

To see that this is true, we can solve the equation in the usual way, but taking care on the way to note what operations we are performing, and to make sure that our number system allows these operations, so that we're not doing anything illegal. Very briefly:

$$\begin{aligned} \alpha z + \beta &= 0 && \Rightarrow \\ (\alpha z + \beta) + (-\beta) &= -\beta && \Rightarrow \\ \alpha z &= -\beta && \Rightarrow \\ \alpha^{-1}(\alpha z) &= \alpha^{-1}(-\beta) && \Rightarrow \\ z &= \alpha^{-1}(-\beta) = -\tfrac{\beta}{\alpha}. \end{aligned}$$

For this argument to work, we need to be able to add the negative of β to both sides of the equation, and then we need to be able to divide the resulting equation by α, or put another way, multiply both sides by the multiplicative inverse $\alpha^{-1} = \frac{1}{\alpha}$ of α.

In $\mathbb{C}$ we can do both of these operations. Therefore, all linear equations with α nonzero have a solution in $\mathbb{C}$.

Foreshadowing. If you have already read Section 5 and know the definition of a "field": you can solve the linear equation $\alpha z + \beta = 0$ over any field, using exactly the same procedure as above. It is a worthwhile exercise for your revision to see which field laws we are using. For instance, can you spot the invocations of the associative laws?

What about *quadratic equations*? Let's consider the general quadratic equation

$$\alpha z^2 + \beta z + \gamma = 0$$

with complex coefficients $\alpha, \beta, \gamma \in \mathbb{C}$. Can we solve this equation inside the complex numbers?

The usual solution to the quadratic equation starts by *completing the square*, as follows:

$$\begin{aligned} \alpha z^2+\beta z+\gamma &= 0 && \Rightarrow \\ z^2+\left(\frac{\beta}{\alpha}\right)z+\left(\frac{\gamma}{\alpha}\right) &= 0 && \Rightarrow \\ z^2+\left(\frac{\beta}{\alpha}\right)z+\frac{\beta^2}{4\alpha^2}+\left(\frac{\gamma}{\alpha}\right) &= \frac{\beta^2}{4\alpha^2} && \Rightarrow \\ \left(z+\frac{\beta}{2\alpha}\right)^2 &= \frac{\beta^2}{4\alpha^2}-\frac{\gamma}{\alpha}=\frac{\beta^2-4\alpha\gamma}{4\alpha^2}. \end{aligned}$$

So far we have not done anything other than divide through by α (which is legal provided that $\alpha \neq 0$ — but if $\alpha = 0$ then we didn't truly have a quadratic), and add some constants to both sides of the equation. Since the usual laws of arithmetic hold for complex numbers, we can be confident that everything so far is correct in $\mathbb{C}$.

Now would come the extraction of the square roots of $\beta^2 - 4\alpha\gamma$, if we were working over the real numbers. (The $\sqrt{4\alpha^2} = 2\alpha$ in the denominator poses no problem.) Let us suppose for the moment that we knew how to find square roots. Using $\sqrt{\beta^2-4\alpha\gamma}$ to mean any complex number u satisfying the equation $u^2 = \beta^2 - 4\alpha\gamma$, we can complete the solution as follows:

$$\begin{aligned} \left(z+\frac{\beta}{2\alpha}\right)^2 &= \frac{\beta^2-4\alpha\gamma}{4\alpha^2} && \Rightarrow \\ z+\frac{\beta}{2\alpha} &= \pm\sqrt{\frac{\beta^2-4\alpha\gamma}{4\alpha^2}} && \Rightarrow \\ z &= \frac{-\beta\pm\sqrt{\beta^2-4\alpha\gamma}}{2\alpha}. \end{aligned}$$

The rest of these derivations work in $\mathbb{C}$ as well, so we see that we have reduced solving quadratic equations over $\mathbb{C}$ to the problem of extracting square roots inside $\mathbb{C}$.

Can we extract these square roots? This does *not* follow from the fact that $\mathbb{C}$ follows the usual laws of arithmetic: after all, $\mathbb{R}$ also does, but negative numbers have no square roots in $\mathbb{R}$. It does follow from the Fundamental Theorem of Algebra that the square roots exist, but that still doesn't help us find them. It turns out that there is a way to compute square roots of complex numbers, and indeed roots of any order. I have described the procedure in a set of supplementary notes, but will go no further with it here.

And what about polynomial equations of degree greater than two? For *cubic equations*, 16th century Italian algebraists Niccòlo Tartaglia, Scipione del Ferro and others discovered procedures for obtaining solutions similar to what we have just done for the quadratic, involving extraction of a cube root. Their procedure is sketched in another supplement to these notes. For *quartic equations* there is a procedure as well, usually credited to Lodovico Ferrari around the same time. But the quartic is the end of the line!

Theorem 2.4 (Abel-Ruffini Theorem)**.** *Let $n \geq 5$ be an integer. There is no expression built from the complex coefficients $a_0, a_1, \ldots, a_n$ using complex scalars, addition, subtraction, multiplication, division, and extraction of roots which evaluates, for all $a_0, a_1, \ldots, a_n \in \mathbb{C}$, to a complex solution to the equation*

$$a_n x^n + \cdots + a_1 x + a_0 = 0.$$

Of course, the Fundamental Theorem of Algebra guarantees that complex solutions exist to the polynomial in the Abel-Ruffini theorem. It is in writing these solutions down that the problem lies.

This theorem is another which we will not prove in this module. A proof will be presented in a course on *Galois theory* (at Queen Mary, the module title is "Further Topics in Algebra").

2.4 Roots and factors

The following proposition encapsulates the workings of the *polynomial long division* algorithm which you may be familiar with. We will discuss polynomial division in Section 2.6, together with a more general form of the proposition (Theorem 2.12).

Proposition 2.5. *Let R be either $\mathbb{R}$ or $\mathbb{C}$. Let $f \in R[x]$ and $\alpha \in R$. Then there exist $q \in R[x]$ and $r \in R$ such that*

$$f = (x - \alpha) \cdot q + r. \tag{1}$$

Proof. We prove this by induction on $\deg f$. The proof will be a *strong* induction: the $n+1$ case may not draw on the n case, but possibly on an earlier case, $n-1$ or $n-2$ or so on. To take care of this, we set up the inductive hypothesis to encompass not just polynomials of degree n, but polynomials of degree *at most* n. We also have to be mindful when writing the proof that the zero polynomial has undefined degree.

Base case. If $\deg f$ is zero or undefined then f is a constant (possibly zero), so we can write

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

Inductive hypothesis. Let n be a non-negative integer, and suppose that we know that any polynomial of degree at most n has an expression of the form (1).

Inductive step. Let f be a polynomial of degree at most $n+1$; we must show that f has an expression of the form (1). If f has degree less than $n+1$, we have already proven the claim for f. So we may assume that f has degree exactly $n+1$. That is,

$$f = a_{n+1} x^{n+1} + a_n x^n + \cdots + a_1 x + a_0$$

where $a_{n+1} \in R$ is not zero (but the remaining coefficients $a_n, \ldots, a_0$ may or may not be zero).

To apply the inductive hypothesis, we would like to pare f down to a polynomial of smaller degree. The first thing that might come to mind, perhaps, is to split f up as

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

The parenthesised summand is a polynomial of degree less than $n+1$, so the inductive hypothesis could be applied to it. But that would leave us no way to handle the $a_{n+1}x^{n+1}$. So instead we will split f up differently:

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

Let $f' = (a_n - \alpha a_{n+1})x^n + a_{n-1}x^{n-1}\cdots + a_1 x + a_0$. By the inductive hypothesis, there exist $q' \in R[x]$ and $r' \in R$ such that

$$f' = (x-\alpha)\cdot q' + r'.$$

It follows that

$$\begin{aligned} f &= a_{n+1}x^n(x-\alpha) + f' \\ &= (x-\alpha)\cdot a_{n+1}x^n + (x-\alpha)\cdot q' + r' \\ &= (x-\alpha)\cdot\big(a_{n+1}x^n + q'\big) + r'. \end{aligned}$$

Since $a_{n+1}x^n + q' \in R[x]$ and $r' \in R$, this completes the inductive step, and the proposition is proved. □

You are probably familiar with a corollary of this proposition, as the justification for having studied polynomial factorisation.

Corollary 2.6. *Let $f \in R[x]$ and $\alpha \in R$. The remainder obtained when dividing f by $x-\alpha$ is $f(\alpha)$.*

In particular, $x = \alpha$ is a solution of $f(x) = 0$ if and only if the polynomial $x-\alpha$ is a factor of f.

Proof. By Proposition 2.5, there exist a polynomial $q \in R[x]$ and a number $r \in R$ such that

$$f = (x-\alpha)\cdot q + r.$$

Substituting in $x = \alpha$, we get

$$f(\alpha) = (\alpha-\alpha)\cdot q(\alpha) + r = r.$$

Therefore if $f(\alpha) = 0$, we have $f(x) = (x-\alpha)\cdot q(x)$, i.e. $x-\alpha$ is a factor of $f(x)$. Conversely, if $x-\alpha$ is a factor of f, say $f = (x-\alpha)\cdot g$, then substitution gives

$$f(\alpha) = (\alpha-\alpha)\cdot g(\alpha) = 0\cdot g(\alpha) = 0.$$ □

Using polynomial factorisation, we can "stretch" the Fundamental Theorem of Algebra to tell us more. A typical complex polynomial equation of degree n has not just the one solution promised by the Theorem, but n of them. In fact, we can make every complex polynomial equation have its full complement of solutions by a sneaky bit of counting: we have to count some of the solutions multiple times.

Definition 2.7. Let k be a nonnegative integer. An element $\alpha \in R$ is a *solution of multiplicity k* to the equation $f(x) = 0$ if $(x-\alpha)^k$ is a factor of $f(x)$, but $(x-\alpha)^{k+1}$ is not.

For example, the solutions of $(x-1)^3(x-2)^4 = 0$ are 1 and 2, of which 1 has multiplicity 3 and 2 has multiplicity 4.

Note that the multiplicity is a function of a polynomial f *and* a number α. If we wanted to make a notation for multiplicity, a suitable one would be $m(f,\alpha)$, not just $m(\alpha)$ or $m(f)$ alone.

Theorem 2.8 (Fundamental Theorem of Algebra with multiplicities)**.** *Let $n \geq 1$, and let $a_0, a_1, \cdots, a_{n-1}, a_n$ be complex numbers, where $a_n \neq 0$. The polynomial equation*

$$a_n z^n + a_{n-1} z^{n-1} + \cdots + a_1 z + a_0 = 0$$

has exactly n solutions in $\mathbb{C}$, counted with multiplicity.

When we say there are "n solutions, counted with multiplicity", we mean that the sum of the multiplicities of the solutions is n.

Proof. First of all, to simplify the argument, we will divide through by the leading coefficient a_n, which is not zero. The resulting equation,

$$z^n + \frac{a_{n-1}}{a_n} z^{n-1} + \cdots + \frac{a_1}{a_n} z + \frac{a_0}{a_n} = 0.$$

has the same solutions as the original, so we will analyse it instead. Let $f = z^n + \cdots + (a_1/a_n)z + a_0/a_n$.

What we will show is that f factors completely as a product of n linear factors $z - \alpha_i$, possibly with repeats. This implies the statement of the theorem, because the sum of all the multiplicities is the total number of factors.

For the factorisation claim, we use induction. This induction argument displays a common feature: the case which "deserves" to be the base case, $n = 0$, would require us to work with the product of zero polynomials. That is actually unproblematic – the product of zero factors equals one – but it bothers many people encountering it for the first time, and so I will write the proof with $n = 1$ as the base case to avoid consternation.

Base case. If $n = 1$, then $f = z + b$ is already of the form $z - \alpha_1$, taking $\alpha_1 = -b$.

Inductive hypothesis. Assume that every monic polynomial of degree k factors as a product of k linear factors.

Inductive step. Let f be a monic polynomial of degree $k+1$. By the Fundamental Theorem of Algebra, $f(z) = 0$ has a complex solution $z = \alpha_{k+1}$. By Corollary 2.6, $z - \alpha_{k+1}$ is a factor of $f(z)$. Write $f(z) = (z - \alpha_{k+1}) \cdot q(z)$. Then q has degree k, so the inductive hypothesis applies, and q has a factorisation

$$q(z) = (z - \alpha_1) \cdots (z - \alpha_k)$$

into n linear factors. We conclude that

$$f(z) = q(z)(z - \alpha_{k+1}) = (z - \alpha_1) \cdots (z - \alpha_k)(z - \alpha_{k+1})$$

is a product of $k+1$ linear factors, as desired. This completes the induction, and the theorem is proved. □

2.5 Polynomial equations over $\mathbb{R}$

If $z = a + b\mathrm{i}$ is a complex number (where a and b are real), then the complex number $a - b\mathrm{i}$ is called the *complex conjugate* of z, and is written as $\overline{z}$. The following facts about complex conjugation are easy to check from the definitions, for any two complex numbers z and w:

- $\overline{z + w} = \overline{z} + \overline{w}$;
- $\overline{zw} = \overline{z} \cdot \overline{w}$;
- $\overline{\overline{z}} = z$;
- $z = \overline{z}$ if and only if z is a real number.

Lemma 2.9. *Let $f \in \mathbb{R}[x]$ be a real polynomial and z a complex number. If $f(z) = 0$, then also $f(\overline{z}) = 0$.*

Proof. Let $f = a_n x^n + \cdots + a_0$, where the a_i are real numbers. Conjugating both sides of the equation $f(z) = 0$ shows that

$$\begin{aligned}
\overline{f(z)} &= \overline{a_n z^n + \cdots + a_1 z + a_0} \\
&= \overline{a_n z^n} + \cdots + \overline{a_1 z} + \overline{a_0} \\
&= \overline{a_n} \cdot \overline{z}^n + \cdots + \overline{a_1} \cdot \overline{z} + \overline{a_0} \\
&= a_n \cdot \overline{z}^n + \cdots + a_1 \cdot \overline{z} + a_0 \\
&= f(\overline{z})
\end{aligned}$$

is equal to $\overline{0} = 0$. □

The next proposition is one equivalent to the Fundamental Theorem of Algebra (with multiplicities) for real polynomials.

Proposition 2.10. *Every real polynomial is a product of a real scalar and factors of the following two types:*

(a) linear factors $x-\alpha$, where α is a real number;

(b) quadratic factors x^2+cx+d, where c and d are real numbers with $c^2<4d$.

Proof. As in the proof of Theorem 2.8, once we show that every nonconstant real polynomial has at least one factor of type (a) or (b), we can produce a proof of the whole proposition using induction. I will prove that one factor exists, and leave the induction part as an exercise for you.

The Fundamental Theorem of Algebra shows that $f(x)=0$ has a complex solution $x=\alpha$, so that $x-\alpha$ is a factor of f. If α is a real number, then $x-\alpha$ is a linear factor of type (a).

If α is a complex number that is not real, then our last lemma shows that $x=\overline{\alpha}$ is a different solution to $f(x)=0$, and therefore a solution to $f/(x-\alpha)=0$. Therefore $(x-\overline{\alpha})$ divides $f/(x-\alpha)$, so that $(x-\alpha)(x-\overline{\alpha})$ divides f. Now write $\alpha=a+b\mathrm{i}$ where a and b are real, and $b\neq 0$ because α is not real. We have

$$\begin{aligned}(x-\alpha)(x-\overline{\alpha}) &= (x-a-b\mathrm{i})(x-a+b\mathrm{i})\\ &= x^2+(-2a)x+(a^2+b^2).\end{aligned}$$

This is a factor of f of our type (b), because if $c=-2a$ and $d=a^2+b^2$, then

$$c^2=4a^2<4a^2+4b^2=4d. \qquad \square$$

Theorem 2.11. *Let $f(x)$ be a real polynomial of odd degree. Then there is a real number α such that $f(\alpha)=0$.*

We will prove this in two ways. The first is as a corollary of Proposition 2.10.

Proof. Factor f as in Proposition 2.10. We cannot write f as a product of quadratic factors only (times a scalar), because the degree of any such product is even. So f must have a linear factor $x-\alpha$ for some real number α, and this α is the solution sought. $\square$

This theorem also permits a proof using your knowledge of Calculus that avoids the, so far unproved, Fundamental Theorem of Algebra.

Outline of proof. Let $f(x) = a_n x^n + a_{n-1}x^{n-1} + \cdots + a_0$, where n is odd. We can suppose that a_n is positive, since otherwise we can solve the equation $-f(x) = 0$ instead. Now, using calculus, we can show that $f(x) > 0$ for large positive values of x, because the term $a_n x^n$ is positive and much larger than the sum of the other terms. In the same way, $f(x) < 0$ for large negative values of x. By the Intermediate Value Theorem, there is a value α with $f(\alpha) = 0$. □

If you don't know or remember the Intermediate Value Theorem, the last sentence just says that the graph of $y = f(x)$ is above the x-axis for large positive x, and below it for large negative x, so it must cross the axis somewhere.

The above argument gives us no clue as to where to look for the number α. That means it is what is called a *non-constructive proof*.

Suppose we are trying to prove that an object having certain specified properties, such as a solution to some equation, exists. There are basically two ways we can go about it:

- We can give a "non-constructive proof". For example, we can suppose that the object doesn't exist, and deduce a contradiction. This is a valid argument, but it gives us absolutely no information about how to go about finding the object.
- We can give a constructive proof, which amounts to an algorithm or method for finding the object in question.

2.6 Polynomial division

There is a version of the division rule and Euclid's algorithm for polynomials. The method for long division for polynomials is similar to that for integers. Here is an example: Divide $x^4 + 4x^3 - x - 5$ by $x^2 + 2x - 1$.

$$
\begin{array}{rrrrrrr}
 & & & x^2 & +2x & -3 & \\
\cline{4-7}
x^2 & +2x & -1 \;\big)\; & x^4 & +4x^3 & -x & -5 \\
 & & & x^4 & +2x^3 & -x^2 & \\
\cline{4-5}
 & & & & 2x^3 & +x^2 & -x \\
 & & & & 2x^3 & +4x^2 & -2x \\
\cline{5-6}
 & & & & & -3x^2 & +x \quad -5 \\
 & & & & & -3x^2 & -6x \quad +3 \\
\cline{6-7}
 & & & & & & 7x \quad -8
\end{array}
$$

This calculation shows that when we divide $x^4 + 4x^3 - x + 5$ by $x^2 + 2x - 1$, the quotient is $x^2 + 2x - 3$ and the remainder is $7x - 8$.

Here is a general statement of the division rule for polynomials. Remember that we have defined the *degree* of a polynomial (Definition 2.2). Our earlier Proposition 2.5 was the special case of the next theorem where $\deg g = 1$.

Theorem 2.12. *Let f and g be two polynomials, with $g \neq 0$. Then there exist a quotient q and a remainder r which are polynomials such that*

- $f = gq + r$;
- *either $r = 0$ or the degree of r is smaller than the degree of g.*

The idea behind the proof of Theorem 2.12 is to follow the long division method that we used in the example. The method goes like this: we multiply g by a constant times a power of x and subtract that off of f, so that the difference has a smaller degree than f. Then we keep going, doing further subtractions. How do you make "then keep going" into a proof? The best way is *induction*. Let's begin:

Proof. Our proof will be by induction on the degree of f. Let g be a fixed nonzero polynomial (i.e. it will not change as we do the induction).

Base case. The base case is the case when $\deg(f) < \deg(g)$ or $f = 0$. This is legitimate as a base case because g is a fixed polynomial, so $\deg(g)$ is just some integer. Remember that we didn't define the degree of the polynomial 0, so we need to "manually" include it.

To prove the theorem in the base case, we set $q = 0$ and $r = f$.

Inductive hypothesis. Let n be a positive integer. The inductive hypothesis states that, if f^* is a polynomial such that $\deg(f^*) < n$, then there exist polynomials q^* and r^* such that

- $f^* = gq^* + r^*$;
- either $r^* = 0$ or the degree of r^* is smaller than the degree of g.

I have put stars in the names of these polynomials so that I can save the letters f, q, r without stars for the inductive case[5]. I didn't need to put a star on g, because it won't be changing.

Inductive case. We assume the inductive hypothesis is true for n, and prove it for $n+1$. If f is a polynomial such that $\deg(f) < n+1$, then either $\deg(f) < n$ or $\deg(f) = n$. The case $\deg(f) < n$ is covered by the inductive hypothesis for n. So the case we have to do some work to prove is $\deg(f) = n$.

[5] You could use primes too, like f'. I wanted to make sure no-one thought I meant the derivative.

Let

$$\begin{aligned} f &= a_n x^n + \text{l.d.t.}, \\ g &= b_m x^m + \text{l.d.t.}, \end{aligned}$$

where we have used the abbreviation l.d.t. for "lower degree terms". We have $a_n \neq 0$, $b_m \neq 0$, and, because we are not in the base case, $n \geq m$. Now let's "cancel the leading term". We have

$$(a_n/b_m)x^{n-m} \cdot g = a_n x^n + \text{l.d.t.},$$

and so the polynomial $f^* = f - (a_n/b_m)x^{n-m} \cdot g$ satisfies $\deg(f^*) < n$, because the $a_n x^n$ term is cancelled out in the subtraction. So by the induction hypothesis, there exist polynomials q^* and r^* such that

$$f^* = gq^* + r^*,$$

where $r^* = 0$ or $\deg(r^*) < \deg(g)$. Then

$$\begin{aligned} f &= f^* + (a_n/b_m)x^{n-m} \cdot g \\ &= g\Big((a_n/b_m)x^{n-m} + q^*\Big) + r^*, \end{aligned}$$

so we can put $q = (a_n/b_m)x^{n-m} + q^*$ and $r = r^*$ to complete the proof. □

Having proved a division rule for polynomials, we can now copy all the following stuff about division that we did for integers. Here is a summary of the definitions and results.

A non-zero polynomial is called *monic* if its leading coefficient is 1, that is, if it has the form

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

We also say that the zero polynomial is monic. If this sounds odd, you can regard it as a convention. But if there is no non-zero coefficient, it is vacuously correct to say that the non-zero coefficient with highest index is 1 (or indeed anything at all).

We say that g *divides* f if $f = gq$ for some polynomial q. In other words, g divides f if the remainder in the division rule is zero.

We define the greatest common divisor of two polynomials by the more advanced definition that we met at the end of the last section. The *greatest common divisor* of f and g is a polynomial d with the properties

(a) d divides f and d divides g;

(b) if h is any polynomial which divides both f and g, then h divides d;

(c) d is monic (this includes the possibility that it is the zero polynomial).

The last condition is put in because, for any non-zero scalar c, each of the polynomials f and cf divides the other. Without this condition, the gcd would not be uniquely defined, since any non-zero constant multiple of it would work just as well. In the world of nonnegative integers, the counterpart of this condition was the requirement that $\gcd(a,b) \geq 0$ (because each of d and $-d$ divides the other).

Theorem 2.13. *(a) Any two polynomials f and g have a greatest common divisor.*

(b) The g.c.d. of two polynomials can be found by Euclid's algorithm.

(c) If $\gcd(f,g) = d$, then there exist polynomials h and k such that

$$fh + gk = d;$$

these two polynomials can also be found from the extended version of Euclid's algorithm.

We will not prove this theorem in detail, since the proof works the same as that for integers.

Here is an example. Find the gcd of $x^4 + 2x^3 + x^2 - 4$ and $x^3 - 1$. By the division rule,

$$\begin{aligned} x^4 + 2x^3 + x^2 - 4 &= (x^3 - 1) \cdot (x+2) + (x^2 + x - 2), \\ x^3 - 1 &= (x^2 + x - 2) \cdot (x-1) + (3x-3), \\ x^2 + x - 2 &= (3x-3) \cdot \tfrac{1}{3}(x+2) + 0. \end{aligned}$$

The last divisor is $3x-3$; dividing by 3, we obtain the monic polynomial $x-1$, which is the required gcd.

Moreover, we have

$$\begin{aligned} 3x-3 &= (x^3-1) - (x-1)(x^2+x-2) \\ &= (x^3-1) - (x-1)((x^4+2x^3+x^2-4) - (x+2)(x^3-1)) \\ &= (x^2+x-1)(x^3-1) - (x-1)(x^4+2x^3+x^2-4), \end{aligned}$$

so

$$x-1 = -\tfrac{1}{3}(x-1) \cdot (x^4+2x^3+x^2-4) + \tfrac{1}{3}(x^2+x-1) \cdot (x^3-1).$$

3 Relations

You have briefly met relations in *Numbers, Sets and Functions*, but they were defined in a relatively informal fashion. In this module we will define them formally, and also introduce the most important kind of relations, the *equivalence relations*, which will be the cornerstone of several algebraic constructions.

3.1 Ordered pairs and Cartesian product

We write $\{x, y\}$ to mean a set containing just the two elements x and y. More generally, $\{x_1, x_2, \ldots, x_n\}$ is a set containing just the n elements x_1, x_2, $\ldots$, x_n.

The order in which elements come in a set is not important. So $\{y, x\}$ is the same set as $\{x, y\}$. This set is sometimes called an *unordered pair*.

Often, however, we want the order of the elements to matter, and we need a different construction. We write the *ordered pair* with first element x and second element y as (x, y). This is not the same as (y, x) unless x and y are equal. You have seen this notation used for the coordinates of points in the plane. The point with coordinates $(2, 3)$ is not the same as the point with coordinates $(3, 2)$. The rule for equality of ordered pairs is:

$$(x, y) = (u, v) \text{ if and only if } x = u \text{ and } y = v.$$

This notation can be extended to ordered n-tuples for larger n. For example, a point in three-dimensional space is given by an *ordered triple* (x, y, z) of coordinates.

The idea of coordinatising the plane or three-dimensional space by ordered pairs or triples of real numbers was invented by Descartes. In his honour, we call the system "Cartesian coordinates". This great idea of Descartes allows us to use algebraic methods to solve geometric problems, as you are learning in *Vectors and Matrices* this term.

By means of Cartesian coordinates, the set of all points in the plane is matched up with the set of all ordered pairs (x, y), where x and y are real numbers. We call this set $\mathbb{R} \times \mathbb{R}$, or $\mathbb{R}^2$. This notation works much more generally, as we now explain.

Definition 3.1. Let X and Y be any two sets. We define their *Cartesian product* $X \times Y$ to be the set of all ordered pairs (x, y), with $x \in X$ and $y \in Y$; that is, all ordered pairs which can be made using an element of X as first coordinate and an element of Y as second coordinate.

We write this as follows:

$$X \times Y = \{(x, y) : x \in X, y \in Y\}.$$

You should read this formula exactly as in the explanation. The notation

$$\{x : P\} \quad \text{or} \quad \{x \mid P\}$$

means "the set of all elements x for which P holds". This is a very common way of specifying a set.

If $Y = X$, we write $X \times Y$ more briefly as X^2. Similarly, if we have sets $X_1, \ldots, X_n$, we let $X_1 \times \cdots \times X_n$ be the set of all ordered n-tuples $(x_1, \ldots, x_n)$ such that $x_1 \in X_1$, ..., $x_n \in X_n$. If $X_1 = X_2 = \cdots = X_n = X$, say, we write this set as X^n.

If the sets are finite, we can do some counting. Remember that we use the notation $|X|$ for the number of elements of the set X (not to be confused with $|z|$, the modulus of the complex number z, for example).

Proposition 3.2. *Let X and Y be sets with $|X| = p$ and $|Y| = q$. Then*

(a) $|X \times Y| = pq$;

(b) $|X^n| = p^n$.

Proof. (a) In how many ways can we choose an ordered pair (x, y) with $x \in X$ and $y \in Y$? There are p choices for x, and q choices for y. Each choice of x can be combined with each choice for y, so we multiply the numbers. We don't miss any ordered pairs this way, nor do we count any of them more than once. Thus there are pq different ordered pairs.[6]

(b) This is an exercise for you. □

The "multiplicative principle" used in part (a) of the above proof is very important. For example, if $X = \{1, 2\}$ and $Y = \{a, b, c\}$, then we can arrange the elements of $X \times Y$ in a table with two rows and three columns as follows:

$$\begin{array}{ccc} (1,a) & (1,b) & (1,c) \\ (2,a) & (2,b) & (2,c) \end{array}$$

3.2 Relations

Suppose we are given a set of people $P_1, \ldots, P_n$. What does the relation of being sisters mean? For each ordered pair (P_i, P_j), either P_i and P_j are sisters, or they are not; so we can think of the relation as being a rule of some kind which answers "true" or "false" for each pair (P_i, P_j).

But to say that a relation is "a rule of some kind" is not amenable to careful mathematical reasoning about the properties of relations. We want to *formalise* relations. That is, we want to build a structure that will let us contain the data of a relation using

[6] In case you find the proof of part (a) unsatisfying, Prof. Peter Cameron has a blog post at `https://cameroncounts.wordpress.com/2011/09/21/the-commutative-law/` showing two approaches which you could use to do it more rigorously.

the mathematical building-blocks we know about already: functions, sets, sequences, and so forth.

One perfectly workable way to encode the data would be as a function from a Cartesian product $\{(P_i, P_j) : P_i, P_j \text{ people}\}$ to a special set $\{\text{true}, \text{false}\}$. If relations had only been invented this year, this might indeed be the definition mathematicians would settle on. But the accepted definition of relations dates back to the early twentieth century, when the great projects of trying to put all of mathematics on rigorous foundations were in progress, and set theory was at the core of the endeavour. So relations are defined as a kind of set.

Definition 3.3. A *relation* R on a set X is a subset of the Cartesian product $X^2 = X \times X$. That is, it is a set of ordered pairs of elements of X.

We think of the relation R as saying "true" about x and y if the pair (x, y) is in R, and saying "false" otherwise. So, in our example, the sisterhood relation is set up as the *set* of all ordered pairs (P_i, P_j) of people who are sisters.

Here is a mathematical example. Let $X = \{1, 2, 3, 4\}$, and let R be the relation "less than" (this means, the relation that holds between x and y if and only if $x < y$). Then we can write R as a set by listing all the pairs for which this is true:

$$R = \{(1,2), (1,3), (1,4), (2,3), (2,4), (3,4)\}.$$

Here is another relation on X:

$$S = \{(1,1), (1,2), (2,3), (3,1), (3,4), (4,2), (4,4)\}.$$

I don't know any simple rule describing S, the way R can be described as "less than". But that's no problem. Just as I can specify a function by giving a table of values, with no formula, I can write down a relation as a set without having a rule in mind[7].

An example of a relation on an infinite set is the *divisibility* relation $|$ on the set $\mathbb{Z}_{\geq 0}$ which we defined in Section 1.2.

How many different relations are there on the set $X = \{1, 2, 3, 4\}$? A relation on X is a subset of $X \times X$. There are $4 \times 4 = 16$ elements in $X \times X$, by Proposition 3.2. How many subsets does a set of size 16 have? For each element of the set, we can decide to include that element in the subset, or to leave it out. The two choices can be made independently for each of the sixteen elements of X^2, so the number of subsets is

$$2 \cdot 2 \cdot \cdots \cdot 2 = 2^{16} = 65536.$$

[7]For more on the similarity between functions and relations, see the appendix "Functions as relations" on QMPlus.

So there are 65536 relations. Of course, most of them don't have simple rules like "less than".

When you want to write that a number x is less than another number y, you are used to writing $x < y$. In other words, you put the symbol for the relation between the names of the two elements concerned. We allow ourselves to use a similar notation for any relation. That is, if R is a relation, we can write $x\,R\,y$ to mean $(x,y) \in R$.

3.3 Equivalence relations and partitions

Just as there are certain laws that operations like multiplication may or may not satisfy, so there are laws that relations may or may not satisfy. Here are some important ones.

Let R be a relation on a set X. We say that R is

reflexive if $(x,x) \in R$ for all $x \in X$;

symmetric if $(x,y) \in R$ implies that $(y,x) \in R$;

transitive if $(x,y) \in R$ and $(y,z) \in R$ together imply that $(x,z) \in R$.

For example, the relation "less than" is not reflexive (since no element is less than itself); is not symmetric (since $x < y$ and $y < x$ cannot both hold); but is transitive (since $x < y$ and $y < z$ do imply that $x < z$). The relation of being sisters, where x and y satisfy the relation if each is the sister of the other, is not reflexive: it is debatable whether a woman can be her own sister (we will say no), but a man certainly cannot! It is obviously symmetric, though. Is it transitive? Nearly: if x and y are sisters, and y and z are sisters, then x and z are sisters **unless** it happens that $x = z$. But this is certainly a possible case. So we conclude that the relation is not transitive. [Remember that, to be transitive, the condition has to hold without exception; any exception would be a counterexample which would disprove the transitivity.]

A very important class of relations are called equivalence relations. An *equivalence relation* is a relation which is reflexive, symmetric, and transitive.

Before seeing the job that equivalence relations do in mathematics, we need another definition.

Definition 3.4. Let X be a set. A *partition* of X is a set $\mathscr{P}$ of subsets of X, whose elements are called its *parts*, having the following properties:

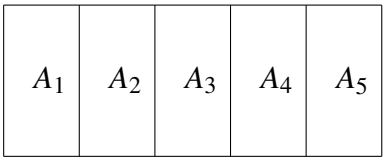
(a) $\emptyset$ is not a part of $\mathscr{P}$;

(b) if A and B are distinct parts of $\mathscr{P}$, then $A \cap B = \emptyset$;

(c) The union of all parts of $\mathscr{P}$ is X.

So each set is non-empty; no two sets have any element in common; and between them they cover the whole of X. The name "partition" arises because the whole set X is divided up into disjoint parts.

For example, $\{\{a,e\},\{b,d\},\{c\}\}$ is a partition of $\{a,b,c,d,e\}$ with three parts, whereas $\{\{a,b,c\},\{c,d\}\}$ is not a partition (of any set) because c is in two different parts, violating property (b). More abstractly, the figure below represents a partition $\mathscr{P} = \{A_1,\ldots,A_5\}$ of a set $X = A_1 \cup \cdots \cup A_5$.



The statement and proof of the next theorem are quite long, but the message is very simple. The job of an equivalence relation on X is to produce a partition of X; every equivalence relation gives a partition, and every partition comes from an equivalence relation. This result is called the *Equivalence Relation Theorem*.

First we need one piece of notation. Let R be a relation on a set X, and let x be an element of X. We write $[x]_R$ for the set of elements of X which are related to x; that is,

$$[x]_R = \{y \in X : (x,y) \in R\}.$$

For example, if R is the relation of being sisters, then $[x]_R$ is the set of all sisters of x.

Definition 3.5. If R is an equivalence relation, then the sets $[x]_R$ are called the *equivalence classes* of R.

If R is not an equivalence relation, then there is no name in general use for the set $[x]_R$.

Theorem 3.6 (Equivalence Relation Theorem)**.** *(a) Let R be an equivalence relation on X. Then the sets $[x]_R$, for $x \in X$, form a partition of X.*

(b) Conversely, given any partition $\mathscr{P}$ of X, there is a unique equivalence relation R on X such that the parts of $\mathscr{P}$ are the same as the sets $[x]_R$ for $x \in X$.

Proof. (a) We have to show that the sets $[x]_R$ satisfy the conditions in the definition of a partition of X.

- For any x, we have $(x,x) \in R$ (since R is reflexive), so $x \in [x]_R$; thus $[x]_R \neq \emptyset$.
- We have to show that, if $[x]_R \neq [y]_R$, then $[x]_R \cap [y]_R = \emptyset$. The contrapositive of this is: if $[x]_R \cap [y]_R \neq \emptyset$, then $[x]_R = [y]_R$; we prove this. Suppose that $[x]_R \cap [y]_R \neq \emptyset$; this means that there is some element, say z, lying in both $[x]_R$ and

$[y]_R$. By definition, $(x,z) \in R$ and $(y,z) \in R$; hence $(z,y) \in R$ by symmetry and $(x,y) \in R$ by transitivity.

We have to show that $[x]_R = [y]_R$; this means showing that every element in $[x]_R$ is in $[y]_R$, and every element of $[y]_R$ is in $[x]_R$. For the first claim, take $u \in [x]_R$. Then $(x,u) \in R$. Also $(y,x) \in R$ by symmetry; and we know that $(x,y) \in R$; so $(y,u) \in R$ by transitivity, and $u \in [y]_R$. Conversely, if $u \in [y]_R$, a similar argument (which you should try for yourself) shows that $u \in [x]_R$. So $[x]_R = [y]_R$, as required.

- Finally we have to show that the union of all the sets $[x]_R$ is X, in other words, that every element of X lies in one of these sets. But we already showed in the first part that x belongs to the set $[x]_R$.

(b) Suppose that $\mathscr{P}$ is a partition of x. We define a relation R as follows:

$$R = \{(x,y) : x \text{ and } y \text{ lie in the same part of } \mathscr{P}\}.$$

Now

- x and x lie in the same part of the partition, so R is reflexive.
- If x and y lie in the same part of the partition, then so do y and x; so R is symmetric.
- Suppose that x and y lie in the same part A of the partition, and y and z lie in the same part B. Then $y \in A$ and $y \in B$, so $y \in A \cap B$; so we must have $A = B$, since different parts are disjoint. Thus x and z both lie in A. So R is transitive.

Thus R is an equivalence relation. By definition $[x]_R$ consists of all elements lying in the same part of the partition $\mathscr{P}$ as x; so, if $x \in A$, then $[x]_R = A$. So the partition $\mathscr{P}$ consists of the sets $[x]_R$.

We leave it as an exercise to check the uniqueness claim of the theorem, that is, that R is the *only* equivalence relation whose parts are the sets $[x]_R$. □

Here is an example. There are five partitions of the set $\{1,2,3\}$. One has a single part; three of them have one part of size 1 and one of size 2; and one has three parts of size 1. Here are the partitions and the corresponding equivalence relations.

Partition	Equivalence relation
$\{\{1,2,3\}\}$	$\{(1,1),(1,2),(1,3),(2,1),(2,2),(2,3),(3,1),(3,2),(3,3)\}$
$\{\{1\},\{2,3\}\}$	$\{(1,1),(2,2),(2,3),(3,2),(3,3)\}$
$\{\{2\},\{1,3\}\}$	$\{(1,1),(1,3),(2,2),(3,1),(3,3)\}$
$\{\{3\},\{1,2\}\}$	$\{(1,1),(1,2),(2,1),(2,2),(3,3)\}$
$\{\{1\},\{2\},\{3\}\}$	$\{(1,1),(2,2),(3,3)\}$

Since partitions and equivalence relations amount to the same thing, we can use whichever is more convenient.

Example Let $X = \mathbb{Z}$, and define a relation $\equiv_4$, called "congruence mod 4", by the rule

> $a \equiv_4 b$ if and only if $b - a$ is a multiple of 4, that is, $b - a = 4m$ for some $m \in \mathbb{Z}$.

Don't be afraid of the notation; "$\equiv_4$" is a different kind of symbol to "R", but we can use them the same way.

We check that this is an equivalence relation.

reflexive? $a - a = 0 = 4 \cdot 0$, so $a \equiv_4 a$.

symmetric? If $a \equiv_4 b$, then $b - a = 4m$, so $a - b = -4m = 4 \cdot (-m)$, so $b \equiv_4 a$.

transitive? If $a \equiv_4 b$ and $b \equiv_4 c$, then $b - a = 4m$ and $c - b = 4n$, so $c - a = 4m + 4n = 4(m+n)$, so $a \equiv_4 c$.

What are its equivalence classes?

- $[0]_{\equiv_4} = \{b : b - 0 = 4m\} = \{\ldots, -8, -4, 0, 4, 8, 12, \ldots\}$, the set of multiples of 4.
- $[1]_{\equiv_4} = \{b : b - 1 = 4m\} = \{\ldots, -7, -3, 1, 5, 9, \ldots, \}$, the set of numbers which leave a remainder of 1 when divided by 4.
- Similarly $[2]_{\equiv_4}$ and $[3]_{\equiv_4}$ are the sets of integers which leave a remainder of 2 or 3 respectively when divided by 4.
- At this point we have caught every integer in one of these four parts, so we have a complete partition of $\mathbb{Z}$. The other equivalence classes repeat the ones we have already seen: $[4]_{\equiv_4} = [0]_{\equiv_4}$, $[5]_{\equiv_4} = [1]_{\equiv_4}$, etc.

We are about to start modular arithmetic, where we will be doing computations with these equivalence classes. For our proofs, one small part of the picture in Theorem 3.6 will be constantly useful.

Corollary 3.7. *Let R be an equivalence relation on a set X, and $x, y \in X$. Then $[x]_R = [y]_R$ if and only if xRy.*

Proof. Assume $[x]_R = [y]_R$. As before, reflexivity of R implies that $y \in [y]_R$. So also $y \in [x]_R$, which by definition of $[x]_R$ is the same assertion as xRy.

For the converse, we assume xRy, that is $y \in [x]_R$. Since $y \in [y]_R$ also, we have two parts $[x]_R$ and $[y]_R$ of the partition in Theorem 3.6 that are not disjoint. So these parts must be equal, that is, $[x]_R = [y]_R$. □

4 Modular arithmetic

You are probably familiar with rules of parity like "odd + odd = even" and "odd · even = even". These rules are a first example of *modular arithmetic*, which is a form of algebra based on remainders. The rule "odd + odd = even" says that if a and b are integers which both have remainder 1 when divided by 2, then $a+b$ has remainder 0 when divided by 2. Similar rules exist for dividing by integers other than 2. They are the subject of this section.

4.1 Congruence mod m

The formalisation of modular arithmetic is based on a very important equivalence relation. Let $X = \mathbb{Z}$, the set of integers.

Definition 4.1. We define a relation $\equiv_m$ on $\mathbb{Z}$, called *congruence mod m*, where m is a positive integer, as follows:

$$a \equiv_m b \text{ if and only if } b-a \text{ is a multiple of } m.$$

We read $a \equiv_m b$ are "a is congruent to b mod m". Some people write the relation $a \equiv_m b$ as $a \equiv b \pmod{m}$.

We check the conditions for it to be an equivalence relation.

reflexive: $x - x = 0 \cdot m$, so $x \equiv_m x$.

symmetric: if $x \equiv_m y$, then $y - x = cm$ for some integer c, so $x - y = (-c)m$, so $y \equiv_m x$.

transitive: if $x \equiv_m y$ and $y \equiv_m z$, then $y - x = cm$ and $z - y = dm$, so $z - x = (c+d)m$, so $x \equiv_m z$.

So $\equiv_m$ is an equivalence relation.

This means that the set of integers is partitioned into equivalence classes of the relation $\equiv_m$. These classes are called *congruence classes mod m*. We write $[x]_m$ for the congruence class mod m containing the integer x. (This is the set we wrote as $[x]_R$ in the Equivalence Relation Theorem, where R was the name of the relation. So we should really write $[x]_{\equiv_m}$. But this looks a bit odd, so we abbreviate it to $[x]_m$ instead.)

For example, when $m = 4$, we have

$$\begin{aligned}
[0]_4 &= \{\ldots, -8, -4, 0, 4, 8, 12, \ldots\}, \\
[1]_4 &= \{\ldots, -7, -3, 1, 5, 9, 13, \ldots\}, \\
[2]_4 &= \{\ldots, -6, -2, 2, 6, 10, 14, \ldots\}, \\
[3]_4 &= \{\ldots, -5, -1, 3, 7, 11, 15, \ldots\},
\end{aligned}$$

and then the pattern repeats: $[4]_4$ is the same set as $[0]_4$ (since $0 \equiv_4 4$). So there are just four equivalence classes. More generally:

Proposition 4.2. *The equivalence relation $\equiv_m$ has exactly m equivalence classes, namely $[0]_m, [1]_m, [2]_m, \ldots, [m-1]_m$.*

Proof. Given any integer n, we can divide it by m to get a quotient q and remainder r, so that $n = mq + r$ and $0 \leq r \leq m-1$. Then $n - r = mq$, so $r \equiv_m n$, and $n \in [r]_m$. So every integer lies in one of the classes $[0]_m, [1]_m, [2]_m, \ldots, [m-1]_m$.

We must also check that these classes are all different, because if not there would be fewer than m of them. Let i and j be integers in the range $0, \ldots, m-1$. We wish to prove that $[i]_m \neq [j]_m$. By Corollary 3.7, it is equivalent to prove $i \not\equiv_m j$. But our assumption implies $-m+1 \leq j-i \leq m-1$, so $j-i$ cannot be a multiple of m unless it equals 0, that is unless $i = j$. □

To give a practical example, what is the time on the 24-hour clock if 298 hours have passed since midnight on 1 January this year? Since two events occur at the same time of day if their times are congruent mod 24, we see that the time is $[298]_{24} = [10]_{24}$, that is, 10:00am, or 10 in the morning.

Notation. We use the notation $\mathbb{Z}_m$ for the set of congruence classes mod m. Thus, $|\mathbb{Z}_m| = m$. Remember that vertical bars around a *set* mean the number of elements in the set.

4.2 Operations on congruence classes

We define addition, subtraction, and multiplication of congruence classes as follows:

$$\begin{aligned} [a]_m + [b]_m &:= [a+b]_m, \\ [a]_m - [b]_m &:= [a-b]_m, \\ [a]_m \cdot [b]_m &:= [a \cdot b]_m. \end{aligned}$$

Look carefully at these supposed definitions. First, notice that the symbols for addition, subtraction, and multiplication on the left are the things being defined. On the right we take the ordinary addition etc. of integers.

The second important thing is that we have to do some work to show that we have defined anything at all. The inputs to the addition operation we have defined are congruence classes—that is, sets— but we have done it by writing the sets as $[a]_m$ and $[b]_m$, and then working with a and b. Remember that there are lots of ways to write the same congruence class in the form $[x]_m$!

Suppose a' and b' are different integers such that $[a]_m = [a']_m$ and $[b]_m = [b']_m$. What guarantee have we that $[a+b]_m = [a'+b']_m$? If this is not true, then our definition

is worthless, because the same pair of congruence classes could have two different sums, depending on whether we happened to pick a and b, or a' and b', from the classes.

So let's try to prove it. Corollary 3.7 helps with this proof again. The assumptions $[a]_m = [a']_m$ and $[b]_m = [b']_m$ unravel to $a \equiv_m a'$ and $b \equiv_m b'$, and we would like to prove $a+b \equiv_m a'+b'$. Now we know that there are integers c and d such that

$$\begin{aligned} a'-a &= cm, \text{ and} \\ b'-b &= dm. \text{ So} \\ (a'+b')-(a+b) &= (c+d)m, \end{aligned}$$

so indeed $a+b \equiv_m a'+b'$. Similarly, with the same assumption,

$$(a'-b')-(a-b) = (c-d)m,$$

so $a-b \equiv_m a'-b'$. And

$$\begin{aligned} a'b'-ab &= (cm+a)(dm+b)-ab \\ &= m(cdm+cm+ad) \end{aligned}$$

so $ab \equiv_m a'b'$. So our definition is valid.

For example, here are an "addition table" and "multiplication table" for the integers mod 4. To make the tables easier on the eyes, I have written $0,1,2,3$ instead of the correct forms $[0]_4, [1]_4, [2]_4, [3]_4$.

$+$	0	1	2	3
0	0	1	2	3
1	1	2	3	0
2	2	3	0	1
3	3	0	1	2

$\cdot$	0	1	2	3
0	0	0	0	0
1	0	1	2	3
2	0	2	0	2
3	0	3	2	1

4.3 Modular inverses

We have just defined addition, subtraction, and multiplication in modular arithmetic. What about division?

If a and b are real numbers, then

$$\frac{a}{b} = a \cdot \frac{1}{b}.$$

Therefore if we know how to multiply and how to compute reciprocals, we can divide by combining these two ingredients.

We will approach the question of division in modular arithmetic the same way, and ask for a reciprocal, or *multiplicative inverse*, of a single element. That is, given the element $[a]_m$, we seek an element $[b]_m$ such that

$$[a]_m[b]_m = [1]_m.$$

If we find it, we write $[b]_m = [a]_m^{-1}$.

But we find that not every element in $\mathbb{Z}_m$ has a multiplicative inverse. For example, $[2]_4$ has no inverse. If you look at row 2 of the multiplication table for $\mathbb{Z}_4$, you see that it contains only the entries 0 and 2, so there is no element $[b]_4$ such that $[2]_4[b]_4 = [1]_4$. However, $[1]_4$ and $[3]_4$ do have inverses, which are unique.

In $\mathbb{Z}_5$ we are luckier. Every non-zero element has an inverse, since

$$[1]_5[1]_5 = [1]_5, \quad [2]_5[3]_5 = [1]_5, \quad [4]_5[4]_5 = [1]_5.$$

This is the best that can be hoped for. In $\mathbb{Z}_m$, just like in $\mathbb{R}$, you can't divide by zero.

Theorem 4.3. *The element $[a]_m$ of $\mathbb{Z}_m$ has a multiplicative inverse if and only if* $\gcd(a,m) = 1$.

Proof. We have two things to prove: if $\gcd(a,m) = 1$, then $[a]_m$ has an inverse; if $[a]_m$ has an inverse, then $\gcd(a,m) = 1$.

First we translate the fact that $[a]_m$ has an inverse. If $[b]_m$ is the inverse, this means that

$$[ab]_m = [a]_m[b]_m = [1]_m,$$

so $ab \equiv_m 1$; in other words,

$$ab - 1 = xm \qquad (*)$$

for some integer x. So $[a]_m$ has an inverse if and only if we can solve this equation.

Let $d = \gcd(a,m)$. Suppose first that $[a]_m$ has an inverse $[b]_m$, so that the equation has a solution. Then d divides a and d divides m, so d divides $ab - xm = 1$, whence $d = 1$.

In the other direction, suppose that $\gcd(a,m) = 1$. The *extended Euclid's algorithm*, Theorem 1.8, shows that there exist integers u and v such that $ua + vm = 1$. This rearranges to $ua - 1 = -vm$, so we can solve equation (*) with $b = u$ and $x = -v$. □

Example What is the inverse of $[4]_{21}$? First we find $\gcd(4,21)$ by Euclid's algorithm:

$$\begin{aligned} 21 &= 4 \cdot 5 + 1, \\ 4 &= 4 \cdot 1, \end{aligned}$$

so $\gcd(4,21) = 1$. This shows that there is an inverse. Now the calculation gives

$$1 = 21 - 5 \cdot 4,$$

so the inverse of $[4]_{21}$ is $[-5]_{21} = [16]_{21}$.

Note that if p is a prime number, then $\gcd(a,p) = 1$ for all $0 < a < p$, which means we may divide by any nonzero element of $\mathbb{Z}_p$. We take this idea up again in Theorem 5.7.

5 Algebraic structures

We will now embark on the programme I promised at the start of the module, the *axiomatic method*. By now we have seen several examples of sets whose elements can be added and multiplied, including long familiar sets of numbers like $\mathbb{Z}$ and $\mathbb{R}$ and new sets like $\mathbb{R}[x]$ and $\mathbb{Z}_m$. We would like to make a single definition that encompasses all of them. That way, if we can write a proof of some algebraic fact that uses only assumptions in this single definition, our proof will automatically be valid in every one of these systems.

What kind of objects are addition and multiplication? They are a special kind of function, which we call operations.

Definition 5.1. A (binary[8]) *operation* on a set X is a function whose domain is $X \times X$ and whose codomain is X.

In other words, the input to this function consists of a pair (x,y) of elements of X, and the output is a single element of X. So we can think of the operation as a rule that "combines" two inputs from X in some way to produce an output in X. Recall that we can use the notation $f : X \times X \to X$ for such a function.

So we might invent the following definition.

Draft definition. An *algebraic structure* is a set X that comes with two operations $+$ and $\cdot$ on X.

But this is no good. If we have an "algebraic structure" in this sense, we can't do any algebra with it. Nothing in the draft definition ensures that the procedures we like to use in algebraic manipulations, such as collecting like terms or expanding parentheses, are logically correct inferences in X. There is no guarantee that the "$+$"

[8] I will just say "operation" in this module, but the more explicit name *binary operation* distinguishes them from *unary* operations $f : X \to X$ and *ternary* operations $f : X \times X \times X \to X$ and so on.

and "$\cdot$" in X behave how we expect addition and multiplication to behave. So our actual definitions will include some laws that addition and multiplication must satify.

It is an important point that we will not include in the definition a rule for how to work out sums and products, only laws restricting them. (By their deeds shall ye know them.) How could we give a rule when we don't even know what the set X is? We would have to give the rules for complex numbers, polynomials, matrices, ... separately. And this would spoil our hopes of generality: when we encountered a new algebraic system it wouldn't be on the list, so it wouldn't fit the definition.

5.1 Fields

Here is our first actual definition.

Definition 5.2. A *field* is a set K of elements that comes with[9] two operations on K, *addition* (written $+$) and *multiplication* (written $\cdot$ or just by juxtaposing the factors), which satisfies the following *axioms*.

Additive laws:

(A0) Closure law: For all $a, b \in K$, we have $a + b \in K$.

(A1) Associative law: For all $a, b, c \in K$, we have $a + (b + c) = (a + b) + c$.

(A2) Identity law: There exists an element $0 \in K$ such that for all $a \in K$, we have $a + 0 = 0 + a = a$.

(A3) Inverse law: For all $a \in K$, there exists an element $b \in K$ such that $a + b = b + a = 0$. We write b as $-a$.

(A4) Commutative law: For all $a, b \in K$, we have $a + b = b + a$.

Multiplicative laws:

(M0) Closure law: For al $a, b \in K$, we have $ab \in K$.

(M1) Associative law: For all $a, b, c \in K$, we have $a(bc) = (ab)c$.

(M2) Identity law: There exists an element $1 \in K$ such that for all $a \in K$, we have $a1 = 1a = a$.

[9] What is "comes with", rigorously? A completely formal definition of a field would say that it is a triple $(K, +, \cdot)$ where K is a set, $+$ and $\cdot$ and are operations on K. I haven't cast my definition this way because the language is less cumbersome if we get to say that the field *is* the set: for example, we can then speak of "elements of a field".

(M3) Inverse law: For each $a \in K$ which is *not equal to* 0, there exists an element $b \in K$ such that $ab = ba = 1$. We write b as a^{-1}.

(M4) Commutative law: For all $a, b \in K$, we have $ab = ba$.

Mixed laws:

(D) Distributive law: For all $a, b, c \in K$, we have

(LD) $a(b+c) = ab + ac$ (the "left distributive law") and

(RD) $(b+c)a = ba + ca$ (the "right distributive law").

(NT) Nontriviality law: $1 \neq 0$.

Many of these axioms deserve some explanation. You might want to come back to the following commentary after reading the examples.

- Strictly speaking, the closure laws are not necessary, since to say that $+$ is an operation on R means that when we input a and b to the function "+", the output belongs to R. We put the closure laws in as a reminder that, when we are checking that something is a field, we have to be sure that this holds.[10]

- We have to be careful about what the identity and inverse laws mean. The identity law for multiplication, e.g., means that there is a particular element e in our system such that $ea = a$ for every element a. In the case of number systems, this element e is the number 1, and it is on this account that we used the symbol "1" for the identity element, not "e". But other algebraic systems need not literally contain the real number 1, so e, or "1", may have to be some other element. The same goes for "0" in the additive identity law.

- The elements "0" and "1" are given their meaning by the identity laws, and they are later referred to in the inverse laws. If the 0 and 1 weren't unique, this would be a problem with the definition: which 0 and which 1 are the inverse laws talking about? But we will prove shortly (Propositions 5.8 and 5.9) that these identity elements are unique.

- We do not bother to try to check the inverse laws unless the corresponding identity law holds. If (say) the multiplicative identity law does not hold, then there is no element "1", and without this the rest of the inverse law doesn't make sense.

[10] For example, checking the closure law for a group will become very essential in Section 8.6.

- We have stated the identity and inverse laws and the distributive law in a redundant way. Since we go on to state commutative laws, we could simply have said in e.g. the multiplicative identity law that $1a = a$. We'll see the reason soon, when we define rings.

- If $0 = 1$ in K, then for every element a of K we have

$$a = 1a = 0a = 0.$$

  So the only algebraic systems ruled out from being fields by the nontriviality law are sets with one element. But note that the equation $0a = 0$ is not a field axiom! See Proposition 5.11 for why this equation is true.

The sets $\mathbb{Q}$ of rational numbers and $\mathbb{R}$ of real numbers are two familiar examples of fields. In this module we will take it on trust that the laws of algebra we have laid out above hold for these sets.

The set $\mathbb{C}$ of complex numbers is also a field, but here we don't have to take the laws on trust. We can prove them from the way $\mathbb{C}$ was defined. We repeat the definition here, laid out to match our definition of "field".

Definition 5.3. The field $\mathbb{C}$ of complex numbers has set of elements

$$\{a + b\mathrm{i} : a, b \in \mathbb{R}\},$$

addition and mutiplication operations defined by

$$\begin{aligned}(a + b\mathrm{i}) + (c + d\mathrm{i}) &:= (a + c) + (b + d)\mathrm{i},\\ (a + b\mathrm{i}) \cdot (c + d\mathrm{i}) &:= (ac - bd) + (ad + bc)\mathrm{i},\end{aligned}$$

and identity elements $0 = 0 + 0\mathrm{i}$ and $1 = 1 + 0\mathrm{i}$.

To prove that $\mathbb{C}$ is a field, we have to prove that all twelve of the field axioms are true. Here, for example, is a proof of the left distributive law. Let $z_1 = a_1 + b_1\mathrm{i}$, $z_2 = a_2 + b_2\mathrm{i}$, and $z_3 = a_3 + b_3\mathrm{i}$. Now

$$\begin{aligned}z_1(z_2 + z_3) &= (a_1 + b_1\mathrm{i})((a_2 + a_3) + (b_2 + b_3)\mathrm{i})\\ &= (a_1(a_2 + a_3) - b_1(b_2 + b_3)) + a_1(b_2 + b_3) + b_1(a_2 + a_3))\mathrm{i},\end{aligned}$$

and

$$\begin{aligned}z_1z_2 + z_1z_3 &= ((a_1a_2 - b_1b_2) + (a_1b_2 + a_2b_1)\mathrm{i}) + ((a_1a_3 - b_1b_3) + (a_1b_3 + a_3b_1)\mathrm{i})\\ &= (a_1a_2 - b_1b_2 + a_1a_3 - b_1b_3) + (a_1b_2 + a_2b_1 + a_1b_3 + a_3b_1)\mathrm{i},\end{aligned}$$

and a little bit of rearranging, using the laws of algebra we have granted for *real* numbers, shows that the two expressions are the same.

And here is a proof of the multiplicative inverse law. Let $z = a + b\mathrm{i}$ be a complex number which is not zero. Then at least one of a and b is a nonzero real number. This implies that $a^2 + b^2 > 0$: since squares of real numbers are never negative, $a^2 + b^2$ is greater than or equal to 0, and the only way it could be equal is if $a^2 = b^2 = 0$, which was ruled out by assumption. This means the complex number

$$w = \left(\frac{a}{a^2+b^2}\right) + \left(\frac{-b}{a^2+b^2}\right)\mathrm{i}$$

is well-defined; we have not divided by zero. Now w is the multiplicative inverse of z, because

$$\begin{aligned} zw &= \left(a \cdot \frac{a}{a^2+b^2} - b \cdot \frac{-b}{a^2+b^2}\right) + \left(a \cdot \frac{-b}{a^2+b^2} + b \cdot \frac{a}{a^2+b^2}\right)\mathrm{i} \\ &= \frac{a^2+b^2}{a^2+b^2} + \frac{-ab+ab}{a^2+b^2} \cdot \mathrm{i} \\ &= 1 + 0i = 1 \end{aligned}$$

and

$$\begin{aligned} wz &= \left(\frac{a}{a^2+b^2} \cdot a - \frac{-b}{a^2+b^2} \cdot b\right) + \left(\frac{a}{a^2+b^2} \cdot b + \frac{-b}{a^2+b^2} \cdot a\right)\mathrm{i} \\ &= \frac{a^2+b^2}{a^2+b^2} + \frac{ab-ab}{a^2+b^2} \cdot \mathrm{i} \\ &= 1 + 0i = 1. \end{aligned}$$

5.2 Rings

Fields are the "best behaved" algebraic structures: they are the structures in which the greatest number of rules of algebra from school continue to hold true. For example, the way we solved the linear equation in Section 2.3 works in any field.

But being a field is very restrictive. Some of our algebraic structures, like $\mathbb{Z}$ and $\mathbb{R}[x]$, are not fields, and so we will not be able to prove results about them if we start from the field axioms. Our solution to this will be to make a new definition, that of a *ring*, with fewer laws, so that all of the systems we have encountered will be rings, and we can handle them all with the axiomatic method.

Definition 5.4. A *ring* R is defined to be a set with two operations, $+$ and $\cdot$, satisfying the following axioms:

- the additive closure, associative, identity, inverse, and commutative laws;
- the multiplicative closure and associative laws;
- and the distributive law.

We also have special names for algebraic structures which satisfy more laws than a ring but not as many as a field. Let R be a ring. We say that R is a *ring with identity* if it satisfies the multiplicative identity law. We say that R is a *skewfield* if it is a ring with identity and also satisfies the multiplicative inverse and nontriviality laws. We say that R is a *commutative ring* if it satisfies the multiplicative commutative law. (Note that the word "commutative" here refers to the multiplication; the addition in a ring is always commutative.)

Putting these three definitions together – and illustrating some of the grammatical flexibility in the terminology – we could say that a field is the same thing as a commutative skewfield with identity.

Here is the reason for the "redundancy" in the axioms we mentioned last section: In a non-commutative ring, we need to assume both parts of the identity and multiplicative inverse laws, since one does not follow from the other in the absence of a commutative law. Similarly, we do need both the left and right parts of the distributive law.

Examples 5.5. Let's apply this new terminology to familiar rings of numbers.

- $\mathbb{Q}$, $\mathbb{R}$ and $\mathbb{C}$ are fields. Therefore, they are commutative rings, skewfields, and rings with identity.
- $\mathbb{Z}$ is a commutative ring with identity. However, it is not a skewfield, and therefore not a field. This is because it does not satisfy the multiplicative inverse law: for example, the integer 2 has no multiplicative inverse in $\mathbb{Z}$.

  You may object that the multiplicative inverse of 2 is $\frac{1}{2}$. But $\frac{1}{2}$ is not an integer, and when we are testing the field axioms for the set $\mathbb{Z}$, we are not allowed to use numbers that are not elements of $\mathbb{Z}$.
- Where have the natural numbers gone? The set $\mathbb{Z}_{\geq 0}$ is not even a ring, because it does not satisfy the additive inverse law: there is no nonnegative integer b such that $b+1=0$. The set $\mathbb{Z}_{>0}$ does even worse, failing to satisfy the additive identity law.

5.3 Rings from modular arithmetic

Theorem 5.6. *The set $\mathbb{Z}_m$, with addition and multiplication mod m, is a commutative ring with identity.*

Proof. To prove a theorem like this, we must prove each one of the axioms for rings. In these notes, I will only write down some parts of the proof, because the rest are similar and I expect you will see how to do them.

Here is a proof of the left distributive law. We are trying to prove that

$$[a]_m([b]_m + [c]_m) = [a]_m[b]_m + [a]_m[c]_m.$$

The left-hand side is equal to $[a]_m[b+c]_m$ (by the definition of addition mod m), which in turn is equal to $[a(b+c)]m$ (by the definition of multiplication mod m. Similarly the right-hand side is equal to $[ab]_m + [ac]_m$, which is equal to $[ab+ac]_m$. Now $a(b+c) = ab+ac$, by the distributive law for integers; so the two sides are equal.

Now let's check the additive identity law. This law asserts that there should exist an additive identity element (a "zero"); choosing $[0]_m$ for this element will make the proof work. Having done so, the equation that we must prove is

$$[a]_m + [0]_m = [0]_m + [a]_m = [a]_m.$$

By the definition of addition mod m, the two quantities on the left are $[a+0]_m = [a]_m$ and $[0+a]_m = [a]_m$, which is equal to the right hand side.

The other proofs are much the same. To show that two expressions involving congruence classes are equal, just show that the corresponding integers are congruent. The multiplicative identity element in $\mathbb{Z}_m$ will be seen to be $[1]_m$. □

Unlike all the examples of rings we have seen so far, $\mathbb{Z}$ and $\mathbb{R}$ and the rest, the rings $\mathbb{Z}_m$ are *finite* sets. Personally, I find finite rings very useful to have in one's stock of mental examples. You can write down the entire addition and multiplication tables and have the whole ring laid out in front of you. If push comes to shove, you can even solve equations completely by brute force, by trying every possible value for each variable!

Remark on notation. In any ring, x^2 is short for $x \cdot x$, and x^3 for $x \cdot x \cdot x$, and so on.

Example. Find all solutions in $\mathbb{Z}_6$ to the equation $x^2 = x$.

Solution. We compute the square of every element of $\mathbb{Z}_6$:

x	$[0]_6$	$[1]_6$	$[2]_6$	$[3]_6$	$[4]_6$	$[5]_6$
x^2	$[0]_6$	$[1]_6$	$[4]_6$	$[9]_6 = [3]_6$	$[16]_6 = [4]_6$	$[25]_6 = [1]_6$

So $x = [0]_6$, $[1]_6$, $[3]_6$, and $[4]_6$ are all the solutions to $x^2 = x$.

Does $\mathbb{Z}_m$ satisfy the multiplicative inverse law? We can give a tidy answer using Theorem 4.3.

Theorem 5.7. *Suppose that p is a prime number. Then $\mathbb{Z}_p$ is a field.*

Proof. Building on Theorem 5.6, we have two properties left to prove. One is the nontriviality law, that $[1]_p \neq [0]_p$. This is true: $p \nmid 1 - 0 = 1$ when p is a prime, because 1 is not prime.

The other is the multiplicative inverse law. To prove this, we must show that every non-zero element of $\mathbb{Z}_p$ has an inverse. If p is prime, then every number a with $1 \leq a < p$ satisfies $\gcd(a, p) = 1$. (For the gcd divides p, so can only be 1 or p; but p clearly doesn't divide a.) Then Theorem 4.3 implies that $[a]_p$ has an inverse in $\mathbb{Z}_p$. □

5.4 Properties of rings

We now give a few properties of rings. Since we only use the ring axioms in the proofs, and not any special properties of the elements, these are valid for all rings. This is the advantage of the axiomatic method.

Proposition 5.8. *In a ring R,*

(a) there is a unique zero element;

(b) any element has a unique additive inverse.

Proof. (a) Suppose that z and z' are two zero elements. This means that, for any $a \in R$,

$$a + z = z + a = a,$$
$$a + z' = z' + a = a.$$

Now we have $z + z' = z'$ (putting $a = z'$ in the first equation) and $z + z' = z$ (putting $a = z$ in the second). So $z = z'$.

This justifies us in calling the unique zero element 0.

(b) Suppose that b and b' are both additive inverses of a. This means that

$$a + b = b + a = 0,$$
$$a + b' = b' + a = 0.$$

Hence

$$b = b + 0 = b + (a + b') = (b + a) + b' = 0 + b' = b'.$$

(Here the first and last equalities hold because 0 is the zero element; the second and second last are our assumptions about b and b'; and the middle equality is the associative law.

This justifies our use of $-a$ for the unique inverse of a. □

Proposition 5.9. *Let R be a ring.*

(a) If R has an identity, then this identity is unique.

(b) If $a \in R$ has a multiplicative inverse, then this inverse is unique.

The proof is almost identical to that of the previous proposition, and is left as an exercise.

The next result is called the *cancellation law*.

Proposition 5.10. *Let R be a ring. If $a+b = a+c$, then $b = c$.*

Proof.

$$b = 0+b = (-a+a)+b = -a+(a+b) = -a+(a+c) = (-a+a)+c = 0+c = c.$$

Here the third and fifth equalities use the associative law, and the fourth is what we are given. To see where this proof comes from, start with $a+b = a+c$, then add $-a$ to each side and work each expression down using the associative, inverse and zero laws. □

Remark. Try to prove that, if R is a field and $a \neq 0$, then $ab = ac$ implies $b = c$.

The next result is something you might have expected to find amongst our basic laws. But it is not needed there, since we can prove it!

Proposition 5.11. *Let R be a ring. For any element $a \in R$, we have $0a = a0 = 0$.*

Proof. We have $0+0 = 0$, since 0 is the zero element. Multiply both sides by a:

$$a0+a0 = a(0+0) = a0 = a0+0,$$

where the last equality uses the zero law again. Now from $a0+a0 = a0+0$, we get $a0 = 0$ by the cancellation law. The other part $0a = 0$ is proved similarly; try it yourself. □

There is one more fact we need. This fact uses only the associative law in its proof, so it holds for both addition and multiplication. To state it, we take $\diamond$ to be a binary operation on a set X, which satisfies the associative law. That is,

$$a \diamond (b \diamond c) = (a \diamond b) \diamond c$$

for all $a, b, c \in X$. This means that we can write $a \diamond b \diamond c$ without ambiguity.

What about applying the operation to four elements? We have to put in brackets to specify the order in which the operation is applied. There are five possibilities:

$$\begin{gathered} a \diamond (b \diamond (c \diamond d)) \\ a \diamond ((b \diamond c) \diamond d) \\ (a \diamond b) \diamond (c \diamond d) \\ (a \diamond (b \diamond c)) \diamond d \\ ((a \diamond b) \diamond c) \diamond d \end{gathered}$$

Now the first and second are equal, since $b \diamond (c \diamond d) = (b \diamond c) \diamond d$. Similarly the fourth and fifth are equal. Consider the third expression. If we put $x = a \diamond b$, then this expression is $x \diamond (c \diamond d)$, which is equal to $(x \diamond c) \diamond d$, which is the last expression. Similarly, putting $y = c \diamond d$, we find it is equal to the first. So all five are equal.

The same works for any number of elements.

Proposition 5.12. *Let $\diamond$ be an operation on a set X which satisfies the associative law. Then the value of the expression*

$$a_1 \diamond a_2 \diamond \cdots \diamond a_n$$

is the same, whatever (legal) way $n-2$ pairs of brackets are inserted.

We will not prove this proposition, but you are encouraged to try to prove it yourself (one way to approach the proof is mathematical induction on n).

6 New rings from old

6.1 Polynomial rings

In the first week of the module we discussed polynomials whose coefficients are real or complex numbers. In fact, Definition 2.1 still works when the set R is allowed to be any ring. Let's repeat the definition with this substitution.

Definition 6.1. Let R be a ring and x a formal symbol. A *polynomial in x with coefficients in R* is an expression

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

where $a_0, a_1, \ldots, a_{n-1}, a_n$ all lie in R. They are the *coefficients* of f.

The set of all such polynomials will be denoted by $R[x]$.

All of the remarks that followed Definition 2.1 are still true when R is a ring.

With this definition, however, we have changed our point of view on polynomials. Polynomials will no longer be functions, in which a number is to be substituted for x; instead they will be expressions to be manipulated algebraically, just like the expressions "$a+b$i" that we call complex numbers. Therefore we have declared x to be a *formal symbol*. This means that the symbol x, and expressions involving it, are assumed to be inert and have no meaning other than the meaning given to them by definitions. The imaginary unit i is another example of a formal symbol[11].

In Definition 6.1, the powers $x^2, x^3, \ldots$ are formal symbols as well. In particular, the definition does not tell us that x times x is x^2! But we wish to make $R[x]$ into a ring. In pursuit of this we are about to define addition and multiplication operations on it, and the latter *will* tell us that x times x is x^2.

Let

$$\begin{aligned} f &= a_m x^m + a_{m-1}x^{m-1} + \cdots + a_1 x + a_0 \quad \text{and} \\ g &= b_n x^n + b_{n-1}x^{n-1} + \cdots + b_1 x + b_0 \end{aligned}$$

be two polynomials in $R[x]$. To define their sum, it is most convenient to assume $m = n$, which we are free to do by supplying leading zero coefficients. Then

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

The product of f and g is defined by

$$\begin{aligned} fg = a_m b_n x^{m+n} + (a_m b_{n-1} + a_{m-1} b_n)x^{m+n} + \cdots \\ \cdots + (a_2 b_0 + a_1 b_1 + a_0 b_2)x^2 + (a_1 b_0 + a_0 b_1)x + a_0 b_0; \end{aligned}$$

the coefficient of the general term x^k is the sum of the products $a_i b_j$ for all pairs of indices i, j with $i + j = k$. Don't be put off by the formidable look of this definition. It simply expresses the usual procedure for multiplying polynomials, namely to expand, multiply the terms pairwise, and then collect like terms.

Note that the formal symbol x commutes with each element of R, that is $x \cdot r = rx = r \cdot x$ for all $r \in R$, even if R is not a commutative ring.

[11] Another word is often used: an *indeterminate* is a formal symbol that plays the role of a variable. So the x in $R[x]$ is an indeterminate, but the imaginary unit i is not.

Theorem 6.2. *If R is a ring, then so is $R[x]$.*

If R is a ring with identity, then so is $R[x]$. If R is commutative, then so is $R[x]$.

The proof is long because of the number of axioms to check, so it will be postponed. But it not difficult.

Proposition 6.3. *If R is a ring, then $R[x]$ is not a skewfield.*

Proof. If R has no nonzero elements, then neither does $R[x]$, so $R[x]$ is not a skewfield because it does not satisfy the nontriviality law.

Otherwise, let b be a nonzero element of R. Then there is no polynomial $f \in R[x]$ such that

$$f \cdot bx = b,$$

because if $f = a_n x^n + \cdots + a_0$ we have

$$f \cdot bx = a_n b x^{n+1} + \cdots + a_1 bx$$

whose constant term is zero, not b. This means that bx cannot have a multiplicative inverse g, because if it did, we could take $f = b \cdot g$ and have

$$f \cdot bx = b \cdot g \cdot bx = b. \qquad \square$$

6.2 Matrix rings

Let R be a ring. An *$m \times n$ matrix with coefficients in R* is an array

$$a = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \cdots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{pmatrix}.$$

We frequently write $a = (a_{ij})_{m \times n}$ in shorthand notation.

The set of all $n \times n$ matrices with coefficients in R is denoted by $M_n(R)$. These matrices, which have the same number of rows and columns, are known as *square* matrices. We will only consider square matrices for the rest of this section. We are about to define addition and multiplication: this can in fact be done for all matrices, but matrix multiplication only gives an *operation* on a set, as defined at the start of Section 5, for square matrices.

Define operations $+$ and $\cdot$ on $M_n(R)$ as follows:

$$(a+b)_{ij} = a_{ij} + b_{ij}, \quad \text{and} \quad (a \cdot b)_{ij} := a_{i1}b_{1j} + a_{i2}b_{2j} + \cdots + a_{in}b_{nj} = \sum_{k=1}^{n} a_{ik}b_{kj}$$

for all $i, j = 1, \ldots, n$.

Theorem 6.4. *If R is a ring, then so is $M_n(R)$.*

If R is a ring with identity, then so is $M_n(R)$.

The proof is not difficult, but quite long, and is therefore deferred until *Algebraic Structures I* next year. The point is that in order to do algebra with matrices, it is not necessary for the entries to be numbers. All that is required is that the entries can be added and multiplied and the results of these operations are again things of the same kind.

Proposition 6.5. *If R is a ring in which not all products of two elements equal zero, and $n \geq 2$, then $M_n(R)$ is neither a commutative ring nor a skewfield.*

Proof. We will write the proof here for $n = 2$ only. The proof for general n is no harder, it's just more irritating to write down the matrices.

Let $ab \neq 0$ in R. Note that a and b cannot equal zero in R either, by Proposition 5.11. Then

$$\begin{pmatrix} a & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} 0 & b \\ 0 & 0 \end{pmatrix} = \begin{pmatrix} 0 & ab \\ 0 & 0 \end{pmatrix}$$

is not equal to

$$\begin{pmatrix} 0 & b \\ 0 & 0 \end{pmatrix} \begin{pmatrix} a & 0 \\ 0 & 0 \end{pmatrix} = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix},$$

proving that $M_2(R)$ is not commutative.

We also use the second equation to show that $M_2(R)$ does not satisfy the multiplicative inverse law. Suppose that $\begin{pmatrix} 0 & b \\ 0 & 0 \end{pmatrix}$ had a multiplicative inverse; call it C. Then $C \begin{pmatrix} 0 & b \\ 0 & 0 \end{pmatrix} = I$, the (multiplicative) identity matrix. We can use these two facts together to reach a contradiction:

$$C \begin{pmatrix} 0 & b \\ 0 & 0 \end{pmatrix} \begin{pmatrix} a & 0 \\ 0 & 0 \end{pmatrix} = C \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix} = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}$$

by Proposition 5.11, while working in the other order gives

$$C \begin{pmatrix} 0 & b \\ 0 & 0 \end{pmatrix} \begin{pmatrix} a & 0 \\ 0 & 0 \end{pmatrix} = I \begin{pmatrix} a & 0 \\ 0 & 0 \end{pmatrix} = \begin{pmatrix} a & 0 \\ 0 & 0 \end{pmatrix}$$

which is not the zero matrix because $a \neq 0$. □

Examples 6.6. (a) Let $R = \mathbb{C}$, let $a = \begin{pmatrix} i & 0 \\ 0 & -i \end{pmatrix}$ and $b = \begin{pmatrix} 1 & i \\ 0 & -1 \end{pmatrix}$. Then

$$a^2 = \begin{pmatrix} i & 0 \\ 0 & -i \end{pmatrix} \cdot \begin{pmatrix} i & 0 \\ 0 & -i \end{pmatrix} = \begin{pmatrix} i^2 & 0 \\ 0 & i^2 \end{pmatrix} = \begin{pmatrix} -1 & 0 \\ 0 & -1 \end{pmatrix} = -I_{2\times 2}$$

and similarly

$$b^2 = \begin{pmatrix} 1 & i \\ 0 & -1 \end{pmatrix} \begin{pmatrix} 1 & i \\ 0 & -1 \end{pmatrix} = \begin{pmatrix} 1 & i-i \\ 0 & 1 \end{pmatrix} = I_{2\times 2}$$

(b) Now take $R = \mathbb{Z}_2$ to be integers mod 2. Then $R = \{[0]_2, [1]_2\}$ by Proposition 6.1; here $[0]_2$ is the zero element 0 of R and $[1]_2$ is the identity element 1 of R. If $a = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} \in M_2(R)$ then

$$a^2 = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} 1 & 1+1 \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} = I_{2\times 2}$$

because $1+1=0$ in R. Similarly, if $b = \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}$ then $b^2 = \begin{pmatrix} 1+1 & 1+1 \\ 1+1 & 1+1 \end{pmatrix} = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}$ is the zero matrix. Since

$$ab = \begin{pmatrix} 0 & 0 \\ 1 & 1 \end{pmatrix} \quad \text{and} \quad ba = \begin{pmatrix} 1 & 0 \\ 1 & 0 \end{pmatrix}$$

we see that $M_2(\mathbb{Z}_2)$ is not commutative.

(c) Let R be a ring. Then so is $M_2(R)$ by the above Theorem. But now we can apply the Theorem again to the ring $M_2(R)$ in place of R to deduce that $M_2(M_2(R))$ is again a ring! Its elements are matrices of the form

$$\begin{pmatrix} \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} & \begin{pmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{pmatrix} \\ \begin{pmatrix} c_{11} & c_{12} \\ c_{21} & c_{22} \end{pmatrix} & \begin{pmatrix} d_{11} & d_{12} \\ d_{21} & d_{22} \end{pmatrix} \end{pmatrix}$$

where the a_{ij}, b_{ij}, c_{ij} and d_{ij} all lie in R.

Can you see how this ring relates to $M_4(R)$?

7 Permutations

So far, we have done algebra on numbers, polynomials, matrices, and sets. In this chapter we turn our eye to another type of object: permutations, which are certain special functions.

7.1 Definition and notation

A *permutation* of a set X is a function $f : X \to X$ which is a bijection (one-to-one and onto).

In this module we will focus on the case when X is a finite set. When there's no reason to use a different set, we will take X to be the set $\{1, 2, \ldots, n\}$ for convenience. As an example of a permutation, we will take $n = 8$ and let f be the function which maps $1 \mapsto 4$, $2 \mapsto 7$, $3 \mapsto 3$, $4 \mapsto 8$, $5 \mapsto 1$, $6 \mapsto 5$, $7 \mapsto 2$, and $8 \mapsto 6$.

We can represent a permutation in *two-line notation*. We write a matrix with two rows and n columns. In the first row we put the numbers $1, \ldots, 8$; under each number x we put its image under the permutation f. In our example, we have

$$f = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 \\ 4 & 7 & 3 & 8 & 1 & 5 & 2 & 6 \end{pmatrix}.$$

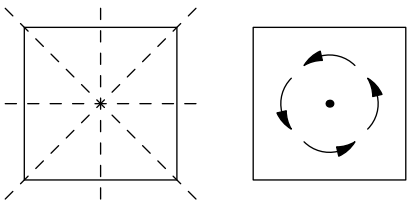
How many permutations of the set $\{1, \ldots, n\}$ are there? We can ask this question another way? How many matrices are there with two rows and n columns, such that the first row has the numbers $1, \ldots, n$ in order, and the second contains these n numbers in an arbitrary order? There are n choices for the first element in the second row; then $n-1$ choices for the second element (since we can't re-use the element in the first column); then $n-2$ for the third; and so on until the last place, where the one remaining number has to be put. So altogether the number of permutations is

$$n \cdot (n-1) \cdot (n-2) \cdots 1.$$

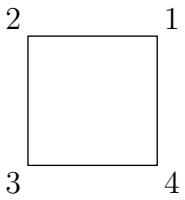
This number is called $n!$, read "n factorial", the product of the integers from 1 to n. Thus we have proved:

Proposition 7.1. *The number of permutations of the set* $\{1, \ldots, n\}$ *is* $n!$.

One of the first uses of permutations in mathematics was as a unified language for *symmetries*. For example, as you know, a square has four axes of reflection symmetry, and fourfold rotational symmetry around its centre.



Each of these symmetries describes some way that the square could be moved so that it lines back up with itself. Let's number the corners of the square, say like this.



Now each symmetry, of whatever kind it is (reflection, rotation, ...), gives rise to a permutation f, by declaring $f(i)$ to be the label of the position where corner i ends up after carrying out the symmetry. Thus an anticlockwise rotation by 90° yields the permutation $\begin{pmatrix}1&2&3&4\\2&3&4&1\end{pmatrix}$, because corner 1 ends up where corner 2 started out, etcetera. Reflection across the vertical line yields the permutation $\begin{pmatrix}1&2&3&4\\2&1&4&3\end{pmatrix}$. And so on. Here's a question to hold in the back of your mind as you read on: what special properties does the *set* of all symmetries of a shape have?

7.2 Composition

Let f and g be permutations. We define the *composition* of f and g, written $f \circ g$, to be the permutation defined by

$$(f \circ g)(x) = f(g(x)).$$

Note that the permutation on the right, g, is the innermost and therefore applies to x first. Do not confuse $f \circ g$ with "apply f and then g", which is $g \circ f$ instead.

You should be aware that some mathematicians (including some who may be your lecturers for further modules in algebra![12]) use a different notation, in which functions are written on the right hand side of their arguments, that is, they write xf rather than $f(x)$. To go with this notation, composition is also done the other way round, as $x(fg) = (xf)g2$.

Here is a fact which we will need later.

Proposition 7.2. *If f and g are permutations of $\{1,\ldots,n\}$, then $f \circ g$ is as well.*

Proof. The domain of $f \circ g$ is the domain of f, and its codomain is the codomain of g. Both are $\{1,\ldots,n\}$.

So we must show that $f \circ g$ is a bijection. First we prove injectivity. Suppose $(f \circ g)(x) = (f \circ g)(y)$ for $x, y \in \{1,\ldots,n\}$, that is,

$$f(g(x)) = f(g(y)).$$

[12] In my impression, in this country, this is basically a generational divide: $f(x)$ is the young algebraist's choice, xf the old.

Because f is injective, this implies $g(x) = g(y)$. Then because g is injective, we conclude $x = y$. Therefore $f \circ g$ is injective.

Next, surjectivity. Let $z \in \{1,\ldots,n\}$. We want to show that there is an $x \in \{1,\ldots,n\}$ so that $(f \circ g)(x) = z$, that is $f(g(x)) = z$. Because f is surjective, there is a y such that $f(y) = z$. And because g is surjective, there is an x such that $g(x) = y$. Then $f(g(x)) = f(y) = z$ as required, so $f \circ g$ is surjective. □

In practice, how do we compose permutations? (Practice is the right word here: you should practise composing permutations until you can do it without stopping to think.) Let f be the permutation we used as an example in the last section, and let

$$g = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 \\ 6 & 3 & 1 & 8 & 7 & 2 & 5 & 4 \end{pmatrix}.$$

The easiest way to calculate $f \circ g$ is to take each of the numbers $1,\ldots,8$, map it by g, map the result by f, and write down the result to get the bottom row of the two-line form for $f \circ g$. Thus, g maps 1 to 6, and f maps 6 to 5, so $f \circ g$ maps 1 to 5. Next, g maps 2 to 3, and f maps 3 to 3, so $f \circ g$ maps 2 to 3. And so on.

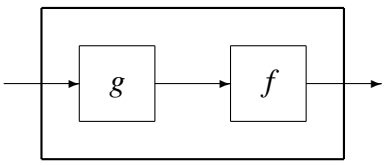
Another way to do it is to re-write the two-line form for f by shuffling the columns around so that the first row agrees with the second row of g. Then the second row will be the second row of $f \circ g$. Thus,

$$f = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 \\ 4 & 7 & 3 & 8 & 1 & 5 & 2 & 6 \end{pmatrix} = \begin{pmatrix} 6 & 3 & 1 & 8 & 7 & 2 & 5 & 4 \\ 5 & 3 & 4 & 6 & 2 & 7 & 1 & 8 \end{pmatrix};$$

so

$$f \circ g = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 \\ 5 & 3 & 4 & 6 & 2 & 7 & 1 & 8 \end{pmatrix}.$$

To see what is going on, remember that a permutation is a function, which can be thought of as a black box. The black box for $f \circ g$ is a composite containing the black boxes for f and g with the output of g connected to the input of f:



Now to calculate the result of applying $f \circ g$ to 1, we feed 1 into the input; the first inner black box outputs 6, which is input to the second inner black box, which outputs 5.

We define a special permutation, the *identity permutation*, which leaves everything where it is:

$$e = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 \\ 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 \end{pmatrix}.$$

Then we have $e \circ f = f \circ e = f$ for any permutation f.

Given a permutation f, we define the *inverse permutation* of f to be the permutation which "puts everything back where it came from" – thus, if f maps x to y, then f^{-1} maps y to x. This is the inverse function in the usual sense, the same way the square root function is the inverse of squaring.

f^{-1} can be worked out using the definition: find x_1 such that $f(x_1) = 1$ and then set $f^{-1}(1) = x_1$; then do the same for 2, and so on. A method to speed this up is to take the two-line form for f, shuffle the columns so that the bottom row is $1\,2\,\ldots\,n$, and then interchange the top and bottom rows. For our example,

$$f = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 \\ 4 & 7 & 3 & 8 & 1 & 5 & 2 & 6 \end{pmatrix} = \begin{pmatrix} 5 & 7 & 3 & 1 & 6 & 8 & 2 & 4 \\ 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 \end{pmatrix},$$

so

$$f^{-1} = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 \\ 5 & 7 & 3 & 1 & 6 & 8 & 2 & 4 \end{pmatrix}.$$

We then see that $f \circ f^{-1} = f^{-1} \circ f = e$.

7.3 Cycles

We come now to a way of representing permutations which is more compact than the two-line notation described earlier, but (after a bit of practice!) just as easy to calculate with: this is *cycle notation*.

Let $a_1, a_2, \ldots, a_k$ be distinct numbers chosen from the set $\{1, 2, \ldots, n\}$. The *cycle* $(a_1, a_2, \ldots, a_k)$ denotes the permutation which maps $a_1 \mapsto a_2$, $a_2 \mapsto a_3$, $\ldots$, $a_{k-1} \mapsto a_k$, and $a_k \mapsto a_1$. If you imagine $a_1, a_2, \ldots, a_k$ written around a circle, then the cycle is the permutation where each element moves to the next place round the circle. Any number not in the set $\{a_1, \ldots, a_k\}$ is fixed by this manoeuvre.

Notice that the same permutation can be written in many different ways as a cycle, since we may start at any point:

$$(a_1, a_2, \ldots, a_k) = (a_2, \ldots, a_k, a_1) = \cdots = (a_k, a_1, \ldots, a_{k-1}).$$

If $(a_1, \ldots, a_k)$ and $(b_1, \ldots, b_l)$ are cycles with the property that no element lies in both of the sets $\{a_1, \ldots, a_k\}$ and $\{b_1, \ldots, b_l\}$, then we say that the cycles are *disjoint*. In this case, their composition is the permutation which acts as the first cycle on the

as, as the second cycle on the bs, and fixes the other elements (if any) of $\{1,\ldots,n\}$. The composition of any set of pairwise disjoint cycles can be understood in the same way.

When working in cycle notation, to save space, we often omit the symbol $\circ$ for composition, just like we usually leave out the multiplication sign $\cdot$.

Theorem 7.3. *Any permutation can be written as a composition of disjoint cycles. The representation is unique, up to the facts that the cycles can be written in any order, and each cycle can be started at any point.*

Proof. Our proof is an algorithm to find the *cycle decomposition* of a permutation. We will consider first our running example:

$$f = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 \\ 4 & 7 & 3 & 8 & 1 & 5 & 2 & 6 \end{pmatrix}.$$

Now we do the following. Start with the first element, 1. Follow its successive images under f until it returns to its starting point:

$$f : 1 \mapsto 4 \mapsto 8 \mapsto 6 \mapsto 5 \mapsto 1.$$

This gives us a cycle $(1,4,8,6,5)$.

If this cycle contains all the elements of the set $\{1,\ldots,n\}$, then stop. Otherwise, choose the smallest unused element (in this case 2, and repeat the procedure:

$$f : 2 \mapsto 7 \mapsto 2,$$

so we have a cycle $(2,7)$ disjoint from the first.

We are still not finished, since we have not seen the element 3 yet. Now $f : 3 \mapsto 3$, so (3) is a cycle with a single element. Now we have the cycle decomposition:

$$f = (1,4,8,6,5)(2,7)(3).$$

The general procedure is the same. Start with the smallest element of the set, namely 1, and follow its successive images under f until we return to something we have seen before. This can only be 1. For suppose that $f : 1 \mapsto a_2 \mapsto \cdots \mapsto a_k \mapsto a_s$, where $1 < s < k$. Then we have $f(a_{s-1}) = a_s = f(a_k)$, contradicting the fact that f is one-to-one. So the cycle ends by returning to its starting point.

Now continue this procedure until all elements have been used up. We cannot ever stray into a previous cycle during this procedure. For suppose we start at an element b_1, and have $f : b_1 \mapsto \cdots \mapsto b_k \mapsto a_s$, where a_s lies in an earlier cycle. Then as before, $f(a_{s-1}) = a_s = f(b_k)$, contradicting the fact that f is one-to-one. So the cycles we produce really are disjoint.

The uniqueness is hopefully clear. □

Here is a notational shortcut. Any cycle of length 1 is the identity permutation, and composing with the identity permutation does nothing. So our example permutation could be written simply as $f = (1,4,8,6,5)(2,7)$, leaving out (3). The fact that 3 is not mentioned means that it is fixed. (You may notice that there is a problem with this convention: the identity permutation fixes everything, and so would be written just as a blank space! We get around this either by leaving in one cycle (1) to represent it, or by just calling it e.)

Example Write the permutation $(1,4,2,3,5)(1,6,3,2,5,4) \in S_6$ in disjoint cycle notation.

Solution. The first thing I want to make clear is that $f = (1,3,5,2,4)(1,5,4,2,6,3)$ is a legitimate permutation! It is not in disjoint cycle notation, because there are numbers repeated between the cycles, but it's still meaningful.

Using the method from Theorem 7.3, we find the image of 1 (call it a_2), then the image of a_2, and so on until the cycle closes. Now f is a composition of two cycles $g = (1,3,5,2,4)$ and $h = (1,5,4,2,6,3)$. So $f(1) = g(h(1)) = g(5) = 2$. Next, $f(2) = g(h(2)) = g(6) = 6$, where g fixes 6 because it does not appear. Continuing this way, we find

$$f : 1 \mapsto 2 \mapsto 6 \mapsto 5 \mapsto 1.$$

As for the other cycles, $f : 3 \mapsto 3$ and $f : 4 \mapsto 4$ are fixed points, and as above we may leave them out. So the answer is $f = (1,2,6,5)$.

You should practise composing and inverting permutations in disjoint cycle notation. Finding the inverse is particularly simple: all we have to do to find f^{-1} is to write each cycle of f in reverse order!

Cycle notation makes it easy to get some information about a permutation. For instance, how many times must one compose f with itself, $f \circ f \circ f \cdots$, to first get back to the identity? We call this number the *order* of f. As for notation, by $f^{\circ n}$ we mean $f \circ \cdots \circ f$, with n repeats of f.

Proposition 7.4. *The order of a permutation is the least common multiple of the lengths of the cycles in its disjoint cycle representation.*

To see what is going on, return to our running example:

$$f = (1,4,8,6,5)(2,7)(3).$$

Now elements in the first cycle return to their starting position after 5 steps, and again after 10, 15, ... steps. So, if $f^{\circ n} = e$, then n must be a multiple of 5. But also the elements 2 and 7 swap places if f is applied an odd number of times, and return to their original positions after an even number of steps. So if $f^{\circ n} = e$, then n must also

be even. Hence if $f^{\circ n} = e$ then n is a multiple of 10. The point 3 is fixed by any number of applications of f so doesn't affect things further. Thus, the order of n is a multiple of 10. But $f^{10} = e$, since applying f ten times takes each element back to its starting position; so the order is exactly 10.

Proof. For the proof we use a general permutation. If the cycle lengths are $k_1, k_2, \ldots, k_r$, then elements of the ith cycle are fixed by $f^{\circ n}$ if and only if n is a multiple of k_i; so $f^{\circ n} = e$ if and only if n is a multiple of all of $k_1, \ldots, k_r$, that is, a multiple of $\mathrm{lcm}(k_1, \ldots, k_r)$. So this lcm is the order of f. □

8 Groups

In this section we study a new algebraic structure, *groups*, and their properties. We have seen two motivations for groups so far. For one, the additive and multiplicative axioms for rings are very similar, and this similarity suggests considering a structure (a group) with only a single operation, that might be either addition or multiplication. The other is that the set of symmetries of any shape will form a group under composition. We treat the first of these below, but we will not formally define symmetries in this module so a proper treatment of the second will have to wait for another time.

8.1 Definition

A *group* is a set G with an operation $\diamond$ on G satisfying the following axioms:

(G0) Closure law: for all $a, b \in G$, we have $a \diamond b \in G$.

(G1) Associative law: for all $a, b, c \in G$, we have $a \diamond (b \diamond c) = (a \diamond b) \diamond c$.

(G2) Identity law: there is an element $e \in G$ (called the *identity*) such that $a \diamond e = e \diamond a = a$ for any $a \in G$.

(G3) Inverse law: for all $a \in G$, there exists $b \in G$ such that $a \diamond b = b \diamond a = e$, where e is the identity. The element b is called the *inverse* of a, written a^*.

If in addition the following law holds:

(G4) Commutative law: for all $a, b \in G$ we have $a \diamond b = b \diamond a$

then G is called a *commutative group*, or more usually an *abelian group* (after the Norwegian mathematician Niels Abel).

If G is a group, then the size of the set $|G|$ is known as the **order** of G.

The resemblance of the axioms for addition in a ring to the group axioms gives us our first ready-made examples of groups.

Theorem 8.1. *Let R be a ring. Take $G = R$, with operation $+$. Then G is an abelian group.*

The group G is called the *additive group* of the ring R. Its identity is 0, and the inverse of a is $-a$.

Proof. Each of the group axioms (G0) through (G3), as well as the commutative law (G4), is the same assertion about the behaviour of the operation $+$ on the set $G = R$ as the corresponding ring axiom (A0) through (A4). Because we have assumed R is a ring, all of these properties hold of the operation $+$. □

If you have encountered the definition of a vector space, you should be able to prove along similar lines that any vector space V, with the operation of vector addition, is an abelian group. The identity is the zero vector $\mathbf{0}$, and the inverse of a vector $\mathbf{v}$ is $-\mathbf{v}$.

What about the multiplication in R: does it yield a group? Expecting the set R with the operation $\cdot$ to be a group turns out to be too naïve. The additive identity element 0 in a ring never has a multiplicative inverse, and unlike the inverse law for rings, the inverse law (G3) for groups contains no proviso that lets us overlook this. But it turns out a group can be cooked up from the multiplication in a ring; we will see how in section 8.5 below.

As another example, the operations on permutations we saw in Section 7 make them into a group.

Theorem 8.2. *The set of all permutations of $\{1,\ldots,n\}$, with the operation of composition, is a group.*

Proof. The closure, identity and inverse laws have been verified in Section 7.2. So the only other law we have to worry about is the associative law. We have

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

for all x; so the associative law, $f\circ(g\circ h) = (f\circ g)\circ h$, holds.

(Essentially, this last argument shows that the result of applying $f\circ g\circ h$ is "h, then g, then f", regardless of how brackets are inserted.) □

We call this group the *symmetric group* on n symbols, and write it S_n. Note that S_n is a group of order $n!$.

Proposition 8.3. *S_n is an abelian group if $n \leq 2$, and is non-abelian if $n \geq 3$.*

Proof. S_1 has order 1, and S_2 has order 2; it is easy to check that these groups are abelian, for example by writing down their Cayley tables.

For $n \geq 3$, S_n contains elements $f = (1,2)$ and $g = (2,3)$. Now check that $f\circ g = (1,2,3)$ does not equal $g\circ f = (1,3,2)$. □

8.2 Elementary properties

Many of the simple properties work in the same way as for rings.

Proposition 8.4. *Let G be a group.*

(a) The identity of G is unique.

(b) Each element has a unique inverse.

(c) For any $a, b \in G$, we have $(a \diamond b)^* = b^* \diamond a^*$.

(d) Cancellation law: if $a \diamond b = a \diamond c$ then $b = c$.

Here is how Proposition 8.4(d), the statement that $(a \diamond b)^* = b^* \diamond a^*$, is explained by Hermann Weyl in his book *Symmetry*, published by Princeton University Press.

> With this rule, although perhaps not with its mathematical expression, you are all familiar. When you dress, it is not immaterial in which order you perform the operations; and when in dressing you start with the shirt and end up with the coat, then in undressing you observe the opposite order; first take off the coat and the shirt comes last.

Proof. (a) If e and e' are identities then

$$e = e \diamond e' = e'.$$

(b) If b and b' are inverses of a then

$$b = b \diamond e = b \diamond a \diamond b' = e \diamond b' = b'.$$

(c) We have:

$$(a \diamond b) \diamond (b^* \diamond a^*) = a \diamond (b \diamond b^*) \diamond a^* = a \diamond e \diamond a^* = a \diamond a^* = e,$$

and similarly

$$(b^* \diamond a^*) \diamond (a \diamond b) = b^* \diamond (a^* \diamond a) \diamond b = b^* \diamond e \diamond b = b^* \diamond b = e.$$

Thus, by the uniqueness of the inverses proved in part (b), we conclude that $b^* \diamond a^* = (a \diamond b)^*$.

(d) If $a \diamond b = a \diamond c$, multiply on the left by the inverse of a to get $b = c$. □

8.3 Cayley tables

If a group is finite, it can be represented by its operation table. In the case of groups, this table is more usually called the *Cayley table*, after Arthur Cayley who pioneered its use. Here, for example, is the Cayley table of the additive group of $\mathbb{Z}_4$.

$+$	0	1	2	3
0	0	1	2	3
1	1	2	3	0
2	2	3	0	1
3	3	0	1	2

Notice that, like the solution to a Sudoku puzzle, the Cayley table of a group contains each symbol exactly once in each row and once in each column (ignoring row and column labels). Why? Suppose we are looking for the element b in row a. It occurs in column x if $a \diamond x = b$. This equation has the unique solution $x = a^{-1} \diamond b$, where a^{-1} is the inverse of a. A similar argument applies to the columns.

8.4 Units

Let R be a ring with identity element 1. An element $u \in R$ is called a *unit* if there is an element $v \in R$ such that $uv = vu = 1$. The element v is called the *inverse* of u, written u^{-1}. By Proposition 8.4, a unit has a unique inverse.

Here are some properties of units.

Proposition 8.5. *Let R be a nontrivial ring with identity.*

(a) 0 *is not a unit.*

(b) 1 *is a unit; its inverse is* 1.

(c) If u is a unit, then so is u^{-1}; its inverse is u.

(d) If u and v are units, then so is uv; its inverse is $v^{-1}u^{-1}$.

Proof. (a) Since $0v = 0$ for all $v \in R$ and $0 \neq 1$, there is no element v such that $0v = 1$.

(b) The equation $1 \cdot 1 = 1$ shows that 1 is the inverse of 1.

(c) The equation $u^{-1}u = uu^{-1} = 1$, which holds because u^{-1} is the inverse of u, also shows that u is the inverse of u^{-1}.

(d) Suppose that u^{-1} and v^{-1} are the inverses of u and v. Then

$$\begin{aligned}(uv)(v^{-1}u^{-1}) &= u(vv^{-1})u^{-1} = u1u^{-1} = uu^{-1} = 1,\\ (v^{-1}u^{-1})(uv) &= v^{-1}(u^{-1}u)v = v^{-1}1v = v^{-1}v = 1,\end{aligned}$$

so $v^{-1}u^{-1}$ is the inverse of uv. □

Here are some examples of units in familiar rings.

- In a field, every non-zero element is a unit.
- In $\mathbb{Z}$, the only units are 1 and -1.
- Let F be a field and n a positive integer. An element A of the ring $M_{n\times n}(F)$ is a unit if and only if the determinant of A is non-zero. In particular, $\begin{pmatrix} a & b \\ c & d \end{pmatrix}$ is a unit in $M_{2\times 2}(\mathbb{R})$ if and only if $ad - bc \neq 0$; if this holds, then its inverse is

$$\frac{1}{ad-bc}\begin{pmatrix} d & -b \\ -c & a \end{pmatrix}.$$

- Which elements are units in the ring $\mathbb{Z}_m$ of integers mod m? The next result gives the answer.

Proposition 8.6. *Suppose that $m > 1$.*

(a) An element $[a]_m$ of $\mathbb{Z}_m$ is a unit if and only if $\gcd(a,m) = 1$.

(b) If $\gcd(a,m) > 1$*, then there exists $b \not\equiv_m 0$ such that $[a]_m[b]_m = [0]_m$.*

Proof. Suppose that $\gcd(a,m) = 1$; we show that a is a unit. By Euclid, there exist integers x and y such that $ax + my = 1$. This means $ax \equiv_m 1$, so that $[a]_m[x]_m = [1]_m$, and $[a]_m$ is a unit.

Now suppose that $\gcd(a,m) = d > 1$. Then a/d and m/d are integers, and we have

$$a\left(\frac{m}{d}\right) = \left(\frac{a}{d}\right) \equiv_m 0,$$

so $[a]_m[b]_m = [0]_m$, where $b = m/d$. Since $0 < b < m$, we have $[b]_m \neq [0]_m$.

But this equation shows that a cannot be a unit. For, if $[x]_m[a]_m = [1]_m$, then

$$[b]_m = [1]_m[b]_m = [x]_m[a]_m[b]_m = [x]_m[0]_m = [0]_m,$$

a contradiction. □

Example The table shows, for each non-zero element $[a]_{10}$ of $\mathbb{Z}_{10}$, an element $[b]_{10}$ such that the product is either 0 or 1. To save space we write a instead of $[a]_{10}$.

a	1	2	3	4	5	6	7	8	9
ab	$1\cdot 1=1$	$2\cdot 5=0$	$3\cdot 7=1$	$4\cdot 5=0$	$5\cdot 2=0$	$6\cdot 5=0$	$7\cdot 3=1$	$8\cdot 5=0$	$9\cdot 9=1$
Unit?	$\surd$	$\times$	$\surd$	$\times$	$\times$	$\times$	$\surd$	$\times$	$\surd$

So the units in $\mathbb{Z}_{10}$ are $[1]_{10}$, $[3]_{10}$, $[7]_{10}$, and $[9]_{10}$. Their inverses are $[1]_{10}$, $[7]_{10}$, $[3]_{10}$ and $[9]_{10}$ respectively.

Euler's function $\phi(m)$, sometimes called *Euler's totient function*, is defined to be the number of integers a satisfying $0 \leq a \leq m-1$ and $\gcd(a,m) = 1$. Thus $\phi(m)$ is the number of units in $\mathbb{Z}_m$.

8.5 The group of units

If R is a ring with identity. we let $R^\times$ denote the set of units of R, with the operation of multiplication in R. On account of the following theorem, we name $R^\times$ the *group of units* of R.

Theorem 8.7. *$R^\times$ is a group.*

Proof. The associative law in $R^\times$ follows from the ring axiom (M1). For the remaining laws, closure, identity and inverse, the important thing to check is that the elements of R provided by the ring axioms themselves lie in $R^\times$. This follows from Proposition 8.5. □

Groups of units are a particularly important example of groups; in particular, they provide our first examples of nonabelian groups. We list some special cases.

- If F is a field, then the group $F^\times$ of units of F consists of all the non-zero elements of F. This is called the *multiplicative group* of F.
- Let F be a field and n a positive integer. The set $M_{n\times n}(F)$ of all $n \times n$ matrices with elements in F is a ring. The group $M_{n\times n}(F)^\times$ is called the *general linear group* of dimension n over F, written $\mathrm{GL}(n,F)$. The general linear group is not abelian if $n \geq 2$.

We will meet another very important class of groups in the next chapter.

Remark on notation I have used the symbol $\diamond$ for the group operation in a general group, because it has relatively little baggage from previous use. In books, you will often see the group operation written as multiplication, or for abelian groups as addition. Here is a table comparing a few different notations.

Notation	Operation	Identity	Inverse
General	$a \diamond b$	e	a^*
Multiplicative	ab or $a \cdot b$	1	a^{-1}
Additive	$a+b$	0	$-a$

In order to specify the notation, instead of saying, "Let G be a group", we often say, "Let $(G, \diamond)$ be a group", or "$(G,+)$" or whichever symbol we want to use for the binary operation. The rest of the notation should then be fixed as in the table.

Sometimes the notations get a bit mixed up. For example, even with the general notation, it is common to use a^{-1} instead of a^* for the inverse of a. I will do so from now on.

8.6 Subgroups

Here is the Cayley table of the group $\mathbb{Z}_{12}^{\times}$.

$\cdot$	1	5	7	11
1	1	5	7	11
5	5	1	11	7
7	7	11	1	5
11	11	7	5	1

Consider the elements $[1]_{12}$ and $[5]_{12}$; forget the other rows and columns of the table. We get a small table

$\cdot$	1	5
1	1	5
5	5	1

Is this a group? Just as for the full table, we can check the axioms (G0), (G2) and (G3) very easily. What about the associative law? Do we have to check all $2 \times 2 \times 2 = 8$ cases? No, because these 8 cases are among the 64 cases in the larger group, and we know that all instances of the associative law hold there. So the small table is a group. We call it a subgroup of the larger group, since we have chosen some of the elements which happen to form a group.

Definition 8.8. Let $(G, \diamond)$ be a group, and H a sub*set* of G, that is, a selection of some of the elements of G. We say that H is *subgroup* of G if H, with the same operation (addition in our example) is itself a group.

How do we decide if a subset H is a subgroup? It has to satisfy the group axioms.

(G0) We require that, for all $h_1, h_2 \in H$, we have $h_1 \diamond h_2 \in H$.

(G1) H should satisfy the associative law; that is, $(h_1 \diamond h_2) \diamond h_3 = h_1 \diamond (h_2 \diamond h_3)$, for all $h_1, h_2, h_3 \in H$. But since this equation holds for any choice of three elements of G, it is certainly true if the elements belong to H.

(G2) H must contain an identity element. If e_H is the identity element of H, then $e_H \diamond e_H = e_H$, and the cancellation law in G then implies that e_H equals the identity element of G. So this condition requires that H should contain the identity of G.

(G3) Each element of H must have an inverse. Again by the uniqueness, this must be the same as the inverse in G. So the condition is that, for any $h \in H$, its inverse h^{-1} belongs to H.

So we get one axiom for free and have three to check. But the amount of work can be reduced. The next result is called the *Subgroup Test*.

Proposition 8.9. *A non-empty subset H of a group $(G, \diamond)$ is a subgroup if and only if, for all $h_1, h_2 \in H$, we have $h_1 \diamond h_2^{-1} \in H$.*

Proof. If H is a subgroup and $h_1, h_2 \in H$, then $h_2^{-1} \in H$, and so $h_1 \diamond h_2^{-1} \in H$.

Conversely suppose this condition holds. Since H is non-empty, we can choose some element $h \in H$. Taking $h_1 = h_2 = h$, we find that $e = h \diamond h^{-1} \in H$; so (G2) holds. Now, for any $h \in H$, we have $h^{-1} = e \diamond h^{-1} \in H$; so (G3) holds. Then for any $h_1, h_2 \in H$, we have $h_2^{-1} \in H$, so $h_1 \diamond h_2 = h_1 \diamond (h_2^{-1})^{-1} \in H$; so (G0) holds. As we saw, we get (G1) for free. □

Example Let $G = (\mathbb{Z}, +)$, the additive group of $\mathbb{Z}$, and $H = 4\mathbb{Z}$ (the set of all integers which are multiples of 4). Take two elements h_1 and h_2 of H, say $h_1 = 4a_1$ and $h_2 = 4a_2$ for some $a_1, a_2 \in \mathbb{Z}$. Since the group operation is $+$, the inverse of h_2 is $-h_2$, and we have to check whether $h_1 + (-h_2) \in H$. The answer is yes, since $h_1 + (-h_2) = 4a_1 - 4a_2 = 4(a_1 - a_2) \in 4\mathbb{Z} = H$. So $4\mathbb{Z}$ is a subgroup of $(\mathbb{Z}, +)$.

8.7 Cosets and Lagrange's Theorem

In our example above, we saw that $4\mathbb{Z}$ is a subgroup of $\mathbb{Z}$. Now $\mathbb{Z}$ can be partitioned into four congruence classes mod 4, one of which is the subgroup $4\mathbb{Z}$. We now generalise this to any group and any subgroup.

Let G be a group and H a subgroup of G. Define a relation $\sim$ on G by

$$g_1 \sim g_2 \text{ if and only if } g_2 \diamond g_1^{-1} \in H.$$

We claim that $\sim$ is an equivalence relation.

reflexive: $g_1 \diamond g_1^{-1} = e \in H$, so $g_1 \sim g_1$.

symmetric: Let $g_1 \sim g_2$, so that $h = g_2 \diamond g_1^{-1} \in H$. Then $h^{-1} = g_1 \diamond g_2^{-1} \in H$, so $g_2 \sim g_1$.

transitive: Suppose that $g_1 \sim g_2$ and $g_2 \sim g_3$. Then $h = g_2 \diamond g_1^{-1} \in H$ and $k = g_3 \diamond g_2^{-1} \in H$. Then

$$k \diamond h = (g_3 \diamond g_2^{-1}) \diamond (g_2 \diamond g_1^{-1}) = g_3 \diamond g_1^{-1} \in H,$$

so $g_1 \sim g_3$.

Now since we have an equivalence relation on G, the set G is partitioned into equivalence classes for the relation. These equivalence classes are called *cosets* of H in G, and the number of equivalence classes is the *index* of H in G, written $|G : H|$.

What do cosets look like?

For any $g \in G$, let

$$H \diamond g = \{h \diamond g : h \in H\}.$$

We claim that any coset has this form. Take $g \in G$, and let X be the equivalence class of $\sim$ containing g. That is, $X = \{x \in G; g \sim x\}$.

- Take $x \in X$. Then $g \sim x$, so $x \diamond g^{-1} \in H$. Let $h = x \diamond g^{-1}$. Then $x = h \diamond g \in H \diamond g$.
- Take an element of $H \diamond g$, say $h \diamond g$. Then $(h \diamond g) \diamond g^{-1} = h \in H$, so $g \sim h \diamond g$; thus $h \diamond g \in X$.

So every equivalence class is of the form $H \diamond g$. We have shown:

Theorem 8.10. *Let H be a subgroup of G. Then the cosets of H in G are the sets of the form*

$$H \diamond g = \{h \diamond g : h \in H\}$$

and they form a partition of G.

Example Let $G = \mathbb{Z}$ and $H = 4\mathbb{Z}$. Since the group operation is $+$, the cosets of H are the sets $H + a$ for $a \in G$, that is, the congruence classes. There are four of them, so $|G : H| = 4$.

Remark. We write the coset as $H \diamond g$, and call the element g the *coset representative*. But **any** element of the coset can be used as its representative. In the above example,

$$4\mathbb{Z} + 1 = 4\mathbb{Z} + 5 = 4\mathbb{Z} - 7 = 4\mathbb{Z} + 100001 = \cdots$$

If G is finite, the *order* of G is the number of elements of G. (If G is infinite, we sometimes say that it has infinite order.) We write the order of G as $|G|$.

Now the partition into cosets allows us to prove an important result, *Lagrange's Theorem*:

Theorem 8.11. *Let G be a finite group, and H a subgroup of G. Then $|H|$ divides $|G|$. In fact, $|G| = |G : H| \cdot |H|$, where $|G : H|$ is the index of H in G.*

Proof. We know that G is partitioned into exactly $|G : H|$ cosets of H. If we can show that each coset has the same number as elements as H does, then the theorem will be proved.

So let $H \diamond g$ be a coset of H. We define a function $f : H \to H \diamond g$ by the rule that $f(h) = h \diamond g$. We show that f is one-to-one and onto. Then the conclusion that $|H \diamond g| = |H|$ will follow.

f is one-to-one: suppose that $f(h_1) = f(h_2)$, that is, $h_1 \diamond g = h_2 \diamond g$. By the Cancellation Law, $h_1 = h_2$.

f is onto: take an element $x \in H \diamond g$, say $x = h \diamond g$. Then $x = f(h)$, as required. □

A The vocabulary of proposition and proof

There are many specialised terms in mathematics used to talk about the nature of proof, its ingredients, and its results. For reference we discuss some of them here.

Theorem, Proposition, Lemma, Corollary These words all mean the same thing: a statement which we can prove. We use them for slightly different purposes.

A *theorem* is an important statement which we can prove. A *proposition* is like a theorem but less important. A *corollary* is a statement which follows easily from a theorem or proposition. For example, if I have proved this statement, call it statement A:

Let n be an integer. Then n^2 is even if and only if n is even.

then statement B

Let n be an integer. Then n^2 is odd if and only if n is odd.

follows easily, so I could call statement B a corollary of statement A. Finally, a *lemma* is a statement which is proved as a stepping stone to some more important theorem. Statement A above is used in Pythagoras' proof of the theorem that $\sqrt{2}$ is irrational, so in this context I could call it a lemma.

Of course these words are not used very precisely. It is a matter of judgment whether something is a theorem, proposition, or whatever, and some statements have traditional names which use these words in an unusual way. For example, there is a very famous theorem called *Fermat's Last Theorem*, which is the following:

Theorem A.1. *Let n be an integer bigger than* 2*. Then there are no positive integers x, y, z satisfying $x^n + y^n = z^n$.*

This was proved in 1994 by Andrew Wiles, so why do we attribute it to Fermat?

Pierre de Fermat wrote the statement of this theorem in the margin of one of his books in 1637. He said, "I have a truly wonderful proof of this theorem, but this margin is too small to contain it." No such proof was ever found, and today we don't believe he had a proof; but the name stuck.

Conjecture The proof of Fermat's Last Theorem is rather complicated, and I will not give it here! Note that, for the roughly 350 years between Fermat and Wiles, "Fermat's Last Theorem" wasn't a theorem, since we didn't have a proof! A statement that we think is true but we can't prove is called a *conjecture*. So we should really have called it *Fermat's Conjecture*.

An example of a conjecture which hasn't yet been proved is *Goldbach's conjecture*:

> Every even number greater than 2 is the sum of two prime numbers.

To prove this is probably very difficult. But to disprove it, a single counterexample (an even number which is not the sum of two primes) would do.

Prove, show, demonstrate These words all mean the same thing. We have discussed how to give a mathematical **proof** of a statement. These words all ask you to do that.

Converse The converse of the statement "A implies B" (or "if A then B") is the statement "B implies A". They are not logically equivalent, as we saw when we discussed "if" and "only if". You should regard the following conversation as a warning! Alice is at the Mad Hatter's Tea Party and the Hatter has just asked her a riddle: 'Why is a raven like a writing-desk?'

> 'Come, we shall have some fun now!' thought Alice. 'I'm glad they've begun asking riddles.—I believe I can guess that,' she added aloud.
>
> 'Do you mean that you think you can find out the answer to it?' said the March Hare.
>
> 'Exactly so,' said Alice.
>
> 'Then you should say what you mean,' the March Hare went on.
>
> 'I do,' Alice hastily replied; 'at least—at least I mean what I say—that's the same thing, you know.'
>
> 'Not the same thing a bit!' said the Hatter. 'You might just as well say that "I see what I eat" is the same thing as "I eat what I see"!' 'You might just as well say,' added the March Hare, 'that "I like what I get" is the same thing as "I get what I like"!' 'You might just as well say,' added the Dormouse, who seemed to be talking in his sleep, 'that "I breathe when I sleep" is the same thing as "I sleep when I breathe"!'
>
> 'It is the same thing with you,' said the Hatter, and here the conversation dropped, and the party sat silent for a minute, while Alice thought over all she could remember about ravens and writing-desks, which wasn't much.

Definition To take another example from Lewis Carroll, recall Humpty Dumpty's statement: "When I use a word, it means exactly what I want it to mean, neither more nor less".

In mathematics, we use a lot of words with very precise meanings, often quite different from their usual meanings. When we introduce a word which is to have a special meaning, we have to say precisely what that meaning is to be. Once we have done so, every time we use the word in future, we are invoking this new precise meaning.

Usually, the word being defined is written in italics. For example, in *Geometry I*, you met the definition

> An $m \times n$ *matrix* is an array of numbers set out in m rows and n columns.

From that point, whenever the lecturer uses the word "matrix", it has this meaning, and has no relation to the meanings of the word in geology, in medicine, and in science fiction.

If you are trying to solve a coursework question containing a word whose meaning you are not sure of, check your notes to see if you can find a definition of that word. Many students develop the habit of working out mathematical problems using previous familiar examples as a model. This is a good way to build intuition, but when it comes to dealing with words that have been given definitions, it can lead you astray. If asked whether something is (say) a matrix, the right thing to do is not to see whether it is like other examples of matrices you know, but to turn to the definition!

Define To define is to give a definition, in the sense just discussed. If I ask you to define some term X, I have asked a more specific question than "what is an X?". I want you to tell me the precise mathematical meaning that X was given, in the notes or the lectures. To return to the example of matrices, a sentence like

> A matrix is what you use to write a system of linear equations as a single vector equation.

is a perfectly fine answer to "what is a matrix", but it does not *define* "matrix".

Axiom Axioms are special parts of certain definitions. They are basic rules which we assume, and prove other things from. For example, we *define* a ring to be a set of elements with two operations, addition and multiplication, satisfying a list of axioms which we have seen in Section 5.2. Then we prove that any ring has certain properties, and we can be sure that any system which satisfies the axioms (including systems of numbers, matrices, polynomials or sets) will have all these properties. In that way, one theorem can be applied in many different situations.

The Greek alphabet

When mathematicians run out of symbols, they often turn to the Greek alphabet for more. You don't need to learn this; keep it for reference. Apologies to Greek students: you may not recognise this, but it is the Greek alphabet that mathematicians use!

Name	Capital	Lowercase
alpha	A	α
beta	B	β
gamma	Γ	γ
delta	Δ	δ
epsilon	E	ε
zeta	Z	ζ
eta	H	η
theta	Θ	θ
iota	I	ι
kappa	K	κ
lambda	Λ	λ
mu	M	μ
nu	N	ν
xi	Ξ	ξ
omicron	O	o
pi	Π	π
rho	P	ρ
sigma	Σ	σ
tau	T	τ
upsilon	Υ	υ
phi	Φ	ϕ or φ
chi	X	χ
psi	Ψ	ψ
omega	Ω	ω