Queen Mary
University of London

MTH 4104 Introduction to Algebra

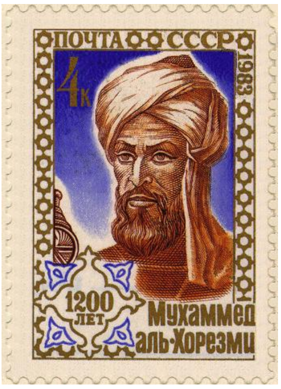
Notes (version of April 12, 2023) **Spring 2023**

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 (which we will do in Section 3.3). 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 meaning of the word “algebra” in modern mathematics, ever since the 1930 textbook *Moderne Algebra* by van der Waerden, 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.

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.

In this module we will be revisiting some mathematical concepts you have seen before, but using new notation or new definitions in terms of sets and functions, so that we can prove facts about them using *Numbers, Sets, and Functions* techniques. In the notes I say that the new definitions are *formal*; changing to the formal style is called *formalisation*. When mathematicians say "formal" they don't mean "according to etiquette", but "allowing you to manipulate the symbols using definitions, without having to think about their meaning", as we must when working with abstract structures.

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 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!

0.1 Notation

New symbols are defined throughout these notes. In this section I talk about a couple symbols that aren't new but that I might be using differently to your expectation.

As you may 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; this is what I do when I write research papers[1]. Others say that $\mathbb{N} = \{1, 2, 3, \ldots\}$, excluding zero, this is what your first-semester textbooks do. In these notes I will avoid taking a side and not use the symbol $\mathbb{N}$. Instead I will write about the *nonnegative integers* $\mathbb{Z}_{\geq 0} = \{0, 1, 2, 3, \ldots\}$ or the *positive integers* $\mathbb{Z}_{>0} = \{1, 2, 3, \ldots\}$. If *you* use $\mathbb{N}$ in your coursework, that's fine. I will assume you mean $\mathbb{N} = \mathbb{Z}_{>0}$ unless you say otherwise.

When I want a multiplication sign I will use $\cdot$ rather than $\times$. The $\times$ sign has other uses in this module, for example Cartesian product of sets, defined in Definition 1.1. Don't confuse the mutiplication dot with the decimal point: $2 \cdot 3$ is not 2.3. (But there will not be many decimal numbers here. Algebraists prefer fractions.)

[1] I explain my preference in the document on QMPlus called *On the meaning of "natural numbers"*.

1 Relations

You have seen relations in *Numbers, Sets and Functions*. In this module we will make a lot of use of relations in our proofs. To help us with this task, we will give a new definition of relations, as a kind of set. These relations-as-sets will do exactly the same job as *Numbers, Sets and Functions* relations, but using sets will make it easier to use our existing mathematical tools to talk about them.

Then we will introduce the most important kind of relations, the *equivalence relations*. Equivalence relations will be the cornerstone of several algebraic constructions because of their connection to *partitions*.

1.1 Ordered pairs and Cartesian product

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

$(x,y) = (u,v)$ if and only if $x = u$ 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 1.1. 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 $X_1, \dots, X_n$, we let $X_1 \times \cdots \times X_n$ be the set of all ordered n-tuples $(x_1, \dots, x_n)$ 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 1.2. *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.[2]

(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:

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

[2] 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.

1.2 Relations as sets

You have been taught to think of a relation as being a rule of some kind which answers "true" or "false" for each ordered pair (x_i, x_j) of elements from a set $X = \{x_1, \ldots, x_n\}$. For example, suppose X is a set of people $\{P_1, \ldots, P_n\}$. One relation is the relation of being sisters. For each ordered pair (P_i, P_j), either P_i and P_j are sisters, or they are not.

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 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 $\{(x_i, x_j) : x_i, x_j \in X\}$ to a special set $\{\text{true}, \text{false}\}$. If I were just inventing relations now for the first time, I might use that definition. But the accepted definition of relations dates back to the early twentieth century, when the great project of trying to put all of mathematics on rigorous foundations were in progress, and at the core of the endeavour was set theory. So the definition mathematicians use of relations says that they are a kind of set.

Definition 1.3. 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 an example with numbers. 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:

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

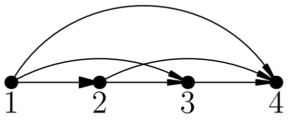
I don't know any simple rule describing S, the way R can be described as "less than". But that's no problem. Just as I can specify a function by giving a table of values, with no formula, I can write down a relation as a set without having a rule in mind[3].

If you like thinking in pictures, here is a way to draw a relation R on a set X that I will use in lectures. Start by drawing a dot for each element of X. Then for each

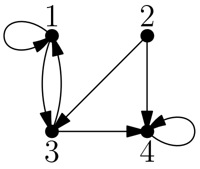
[3] For more on the similarity between functions and relations, see the appendix "Functions as relations" on QMPlus.

pair (x,y) in R, draw an arrow from x to y. Using an arrow rather than just a line ensures that you can tell (x,y) and (y,x) apart in the drawing. Below are drawings of the examples R and S.

R:



S:



It's OK to have a relation on an infinite set, for example the "less than" relation $<$ on the set $\mathbb{R}$.

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 1.2. 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

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

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 $(x,y) \in R$.

1.3 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.

Definition 1.4. 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 $(x,z) \in R$.

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

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.

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

Definition 1.5. Let X be a set. A *partition* of X is a set P of subsets of X, whose elements are called its *parts*, having the following properties:

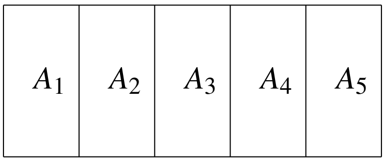
(a) $\emptyset$ is not a part of P;

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

(c) The union of all parts of 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. Don't confuse "partition" with "part": a part is an element of a partition.

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 $P = \{A_1,\ldots,A_5\}$ 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 1.6. 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 1.7 (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 P of X, there is a unique equivalence relation R on X such that the parts of P are the same as the sets $[x]_R$ for $x \in X$. This equivalence relation is

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

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 P is a partition of x, and R is as defined in the theorem. 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 P as x; so, if $x \in A$, then $[x]_R = A$. So the partition 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.

In the next chapter on modular arithmetic, we will be doing arithmetic with these equivalence classes as if they were numbers. One thing we'll do a lot is ask whether two equivalence classes are equal. Here is what Theorem 1.7 has to say about that question.

Corollary 1.8. *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 1.7 that are not disjoint. So these parts must be equal, that is, $[x]_R = [y]_R$. □

2 Modular arithmetic

You are probably familiar with the rules for arithmetic on odd and even numbers, like "odd + odd = even". Here are some tables:

$+$	even	odd
even	even	odd
odd	odd	even

$\cdot$	even	odd
even	even	even
odd	even	odd

These 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 the remainders of division by integers other than 2. They are the subject of this section.

How should we formalise the symbols "even" and "odd" that appear in the above tables? Since they refer to remainders 0 and 1 on division by 2, we could just represent them by the integers 0 and 1. But we're actually going to represent them by the sets of *all* even integers and *all* odd integers. These two sets are equivalence classes of a suitable equivalence relation. The reason for doing it this way is generality. As you study algebra further, you will see that the equivalence relation setup can be done in any ring (see Section 3.1) for a large class of relations, giving you a way to make various smaller rings.

2.1 Integer division

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

Proposition 2.1. *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) $0 \leq r < b$.

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 sentence of the proposition, about uniqueness, means that, if (q_1, r_1) and (q_2, r_2) are two pairs of integers so that $a = bq_1 + r_1$ and $0 \leq r_1 < b$ but also $a = bq_2 + r_2$ and $0 \leq r_2 < b$, then $q_1 = q_2$ and $r_1 = r_2$.

Proof. We will show the uniqueness first. Let q_1, q_2, r_1, and r_2 be as above. If $r_1 = r_2$, then $bq_1 = bq_2$, so $q_1 = q_2$ (as $b > 0$). So suppose that $r_1 \neq r_2$. We may suppose that $r_1 < r_2$ (the case when $r_1 > r_2$ is handled similarly). Then $r_2 - r_1 = b(q_2 - q_1)$. This difference is both a multiple of b, and also in the range from 1 to $b-1$, since both r_1 and r_2 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 $r = 3$.

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

Definition 2.2. 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 1.2 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[4], "$3 \mid 7$ 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 non-

[4] 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 "$0 \leq x < 1$".

negative integers.) The same is not true if a and b could be any integers — why?

2.2 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 2.3. 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", or "…modulo 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 2.4. *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 $0 \le r \le m-1$. Then $n - r = mq$, so $r \equiv_m n$, and $n \in [r]_m$. So every integer lies in one of the classes $[0]_m, [1]_m, [2]_m, \dots, [m-1]_m$.

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 $0, \dots, m-1$. We wish to prove that $[i]_m \neq [j]_m$. By Corollary 1.8, it is equivalent to prove $i \not\equiv_m j$. But our assumption implies $-m+1 \le j - i \le m-1$, 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.

2.3 Arithmetic with congruence classes

Definition 2.5. We define addition, subtraction, and multiplication of congruence classes in $\mathbb{Z}_m$ 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 1.8 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.

$+$	$[0]_4$	$[1]_4$	$[2]_4$	$[3]_4$
$[0]_4$	$[0]_4$	$[1]_4$	$[2]_4$	$[3]_4$
$[1]_4$	$[1]_4$	$[2]_4$	$[3]_4$	$[0]_4$
$[2]_4$	$[2]_4$	$[3]_4$	$[0]_4$	$[1]_4$
$[3]_4$	$[3]_4$	$[0]_4$	$[1]_4$	$[2]_4$

$\cdot$	$[0]_4$	$[1]_4$	$[2]_4$	$[3]_4$
$[0]_4$	$[0]_4$	$[0]_4$	$[0]_4$	$[0]_4$
$[1]_4$	$[0]_4$	$[1]_4$	$[2]_4$	$[3]_4$
$[2]_4$	$[0]_4$	$[2]_4$	$[0]_4$	$[2]_4$
$[3]_4$	$[0]_4$	$[3]_4$	$[2]_4$	$[1]_4$

2.4 gcd and Euclid's algorithm

Definition 2.6. Let a and b be nonnegative integers, not both zero. 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.

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

If you are reading this definition with a careful mathematical eye, when you get to "the *greatest common divisor*" you should take notice of the word "the"! By saying "the" I'm asserting that d exists and there's only one number with this property. When you see "the" before a term being defined in your mathematical reading, and the definition doesn't just give a formula or rule, you should ask yourself these questions.

For Definition 2.6, here's an argument that there's one and only one d. Some common divisor exists because 1 is always a common divisor, and no common divisor is greater than a (if $a > 0$; otherwise we can use b). So the set of common divisors is finite and nonempty, and therefore has a greatest element.

Our divisibility facts about zero show that $\gcd(a,0) = a$ holds for any non-zero number a. What about $\gcd(0,0)$? If we try to use the definition to work out $\gcd(0,0)$ despite its warning not to, we find that it provides no answer: every nonnegative integer divides zero, so there is no greatest one. We will define that $\gcd(0,0) = 0$. You can simply accept this as a convention, or see Extra Questions 2.5.4 for a reason it is a good choice.

Definition 2.7. 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.)

When you first encountered the gcd and lcm, you probably learned to compute them by prime factorisation[5]. The downside of this method 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. Euclid's algorithm is based on two simple rules:

Proposition 2.8.

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

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

[5] If you need to revise how this is done, see the document on QMPlus called *Greatest common divisors by prime factorisation*.

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 cleanly:

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 $\gcd(8633, 9167)$.

2.5 Euclid's algorithm extended

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

Theorem 2.9. *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. Therefore this is a *constructive* proof.

2.6 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 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 2.10. *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,

$$ab - 1 = xm \tag{*}$$

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 2.9, 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 3.9.

By definition (and commutativity), if $[a]_m[b]_m = [1]_m$ then $[a]_m^{-1}$ exists, namely, it is $[b]_m$. The next proposition shows us a similar multiplication statement which implies that $[a]_m^{-1}$ does *not* exist.

Proposition 2.11. *Suppose that $m > 1$. The element $[a]_m$ of $\mathbb{Z}_m$ has no multiplicative inverse if and only if there exists $b \not\equiv_m 0$ such that $[a]_m[b]_m = [0]_m$.*

Proof. If $[a]_m$ has no multiplicative inverse, Theorem 2.10 implies 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 $[b]_m \neq [0]_m$.

Conversely, this equation shows that a cannot have a multiplicative inverse. 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 $[0]_{10}$ or $[1]_{10}$. By Theorem 2.10 and Proposition 2.11, this answers the question of which elements have inverses. To save space we write a instead of $[a]_{10}$.

a	1	2	3	4	5	6	7	8	9
ab	$1\cdot1=1$	$2\cdot5=0$	$3\cdot7=1$	$4\cdot5=0$	$5\cdot2=0$	$6\cdot5=0$	$7\cdot3=1$	$8\cdot5=0$	$9\cdot9=1$
invertible?	$\surd$	$\times$	$\surd$	$\times$	$\times$	$\times$	$\surd$	$\times$	$\surd$

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

Let's return to division. We haven't given a definition of division, but we do know what it means to give the solution to an equation involving multiplication. If we were to declare that "$[a]_m/[b]_m = [c]_m$" for some particular congruence classes, then we expect also $[b]_m[c]_m = [a]_m$. It would be misleading for division to mean anything else.

So how do we solve such an equation for $[c]_m$, given $[a]_m$ and $[b]_m$? Our training from school in solving equations suggests what to do, as in the next example.

Example Solve the equation $[4]_{21}[c]_{21} = [11]_{21}$ for $[c]_{21} \in \mathbb{Z}_{21}$.

Solution?? Isolate $[c]_{21}$ by multiplying both sides by the inverse of $[4]_{21}$, which we computed to be $[4]_{21}^{-1} = [16]_{21}$ in the last example.

$$[c]_{21} = [1]_{21}[c]_{21} = [16]_{21}[4]_{21}[c]_{21} = [16]_{21}[11]_{21} = [176]_{21} = [8]_{21}.$$

Is this a valid method of solution? Can we prove that it works?

As we will see soon, this *is* a valid method, but a few of the steps do things that we have not yet proved to be correct. These steps are the ones that assume that the multiplication function in modular arithmetic has the same properties as usual multiplication.

- In the solution?? above, must $[1]_{21}[c]_{21}$ really be equal to $[c]_{21}$ for every integer c? Yes, and this isn't hard to prove using the definition, but this is still something we have to check.

- The second thing we didn't check is subtle, so well done if you spotted it. Is it true that
$$\big([16]_{21}[4]_{21}\big) \cdot [c]_{21} = [16]_{21} \cdot \big([4]_{21}[c]_{21}\big)?$$
When we use the computation $[1]_{21} = [16]_{21}[4]_{21}$ we end up with the left hand side, but when we want to use the given equation $[4]_{21}[c]_{21} = [11]_{21}$ we can only substitute it into the right hand side. So we have to prove that these two are equal as well.

By contrast, we do not have to prove anything extra to justify the step

$$[16]_{21} \cdot \big([4]_{21}[c]_{21}\big) = [16]_{21}[11]_{21}.$$

We got this by multiplying both sides of the equation $[4]_{21}[c]_{21} = [11]_{21}$ by $[16]_{21}$. But even if instead of multiplying we used some other function f, it would still be true that $f([16]_{21}, [4]_{21}[c]_{21}) = f([16]_{21}, [11]_{21})$, by well-definedness of f.

If we wanted to solve more complicated equations in modular arithmetic, there would be more of these steps that need a separate proof. And I'm not saying that they need proof just to be difficult! Sometimes the usual rules *don't* work. For example, by substituting in all the values $a = 0, 1, \ldots, 5$, you can check that the equation

$$[a]_6^2 = [a]_6$$

has four solutions: $[a]_6 = [0]_6$, $[1]_6$, $[3]_6$ and $[4]_6$. ($[a]_6^2$ is just a shorthand for $[a]_6[a]_6$.) But our usual methods show that $x^2 = x$ has just two real solutions $x \in \mathbb{R}$, namely $x = 0$ and $x = 1$. Why do these methods go wrong in $\mathbb{Z}_6$?

Answering questions like these, and taking care of all the little proofs for individual steps, is one good reason to look at $\mathbb{Z}_m$ using the framework of the *axiomatic method.* This is the subject of the next chapter.

3 Algebraic structures

In this chapter we 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, both long familiar sets of numbers like $\mathbb{Z}$ and $\mathbb{R}$ and new sets like $\mathbb{Z}_m$. We would like to make a single definition that encompasses all of them. That way, we can talk about algebraic facts in all these contexts at once. 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 are addition and multiplication? They are a special kind of function, which we call operations.

Definition 3.1. A (binary[6]) *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 an operation 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 start the definition this way.

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

Next, we need to say what the functions $+$ and $\cdot$ should be. But it is an important point that the definition can't itself include a rule for how to work out sums and products. 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, and so on, all 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.

Instead, we add conditions to the definition that require that the procedures we use in simplifying and manipulating algebraic manipulations, such as collecting like terms or expanding parentheses, are logically correct in X when using the new $+$ and $\cdot$ operations. These conditions are called *axioms*. Roughly, there will be one axiom for each of the different basic steps of our manipulations. If a law of algebra can be broken down and proved using smaller steps, then we leave it out as an axiom: see Proposition 3.13 for an example.

[6] 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.

3.1 Rings and fields

Here is our first actual definition.

Definition 3.2. A *ring* is a set R that comes with[7] two operations on R, *addition* (written $+$) and *multiplication* (written $\cdot$ or just by juxtaposing the factors), which satisfies the following *axioms*.

Additive laws:

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

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

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

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

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

Multiplicative laws:

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

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

Mixed laws:

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

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

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

Rings are a sort of "bare minimum" definition, meant to be easy for algebraic structures to satisfy. Some of the algebraic laws of $\mathbb{R}$ that weren't chosen as ring axioms will instead be axioms in the following more restrictive definition of *field*.

Definition 3.3. A *field* K is defined to be a ring satisfying the following additional axioms:

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

Multiplicative laws:

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

(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 $ab = ba$.

and the

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

Let's start by applying these new definitions to familiar sets of numbers.

Examples 3.4.

- 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.
- $\mathbb{Z}$ is a ring but 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 when the field axioms are applied to the set $\mathbb{Z}$, all the variables must be elements of $\mathbb{Z}$, and $\frac{1}{2}$ is not an integer. This is an important part of the perspective of modern algebra. We think about each set of numbers (or other elements) on its own, and don't "automatically" switch e.g. from $\mathbb{Z}$ to $\mathbb{Q}$ when it would help us solve an equation.
- What about the natural numbers? 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.

3.2 Understanding the axioms

In the definition, "+" and "·" are just names for two functions on R. The axioms ensure that they behave like addition and multiplication, but they need not literally be addition and multiplication of real numbers If this reuse of symbols bothers you, bear two things in mind.

- We have given new definitions to "+" and "$\cdot$" before. For example, addition of matrices is not literally addition of single real numbers either.

- Using familiar symbols makes the axioms look familiar, which helps when working with them. I could have written $a(x,y)$ and $m(x,y)$ for the addition and multiplication in R, but then e.g. the left distributive law would be the long and unfamiliar
$$m(x,a(y,z)) = a(m(x,y),m(x,z)).$$

Many of the axioms deserve some explanation.

- 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.[8]

- 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. For the real numbers, this element e is the number 1, and it is in imitation of this 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.

  An alternate version of the notation is to write 0_R and 1_R for the identity elements of R. This makes it clear that the 0_R and 1_R are different from the real numbers 0 and 1.

- 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 3.10 and 3.11) 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.

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

[8] For example, checking the closure law for a group will become very essential in Section 7.5.

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 3.13 for why this equation is true.

- From the point of view of a field, we have stated the identity and inverse laws and the distributive law in a redundant way, because the commutative law makes the two versions equivalent. But in the definition of a ring there is no multiplicative commutative law, so it's important to have both the left and right distributive laws. In the same way, we might want to talk about rings which satisfy the multiplicative identity and/or inverse law but not the commutative law, and then we need both equations.

We have special names for rings which satisfy some, but maybe not all, of the field axioms. 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.

Examples 3.5.

- $\mathbb{Q}$ and $\mathbb{R}$ are fields. Therefore, they are commutative rings, skewfields, and rings with identity.

- $\mathbb{Z}$ is a commutative ring with identity. However, Example 3.4 showed it is not a skewfield, and for this reason it is not a field.

3.3 The complex numbers

Complex numbers were briefly defined in *Numbers, Sets and Functions*, and you may have seen them before that. The set $\mathbb{C}$ of complex numbers is another example of 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 will start by repeating the definition of complex numbers here, with all the bits we need to match our definition of "field".

Definition 3.6. The set $\mathbb{C}$ of complex numbers is[9]

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

We define addition and mutiplication operations on $\mathbb{C}$ by

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

Theorem 3.7. *$\mathbb{C}$ is a field.*

To prove that $\mathbb{C}$ is a field, we have to prove that all twelve of the field axioms are true. I won't write all twelve proofs down in the notes, just a few. 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.

Next, let's do a proof of the multiplicative inverse law. We can't actually jump straight to the multiplicative inverse law, since it mentions the identity elements "0" and "1", and we don't know which complex numbers these should be yet. So let's imagine we have already proved that "0"$= 0+0\mathrm{i}$ and "1"$= 1+0\mathrm{i}$ satisfy the respective identity laws (this is an exercise for you).

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

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

[9] Some texts use an even more formal definition, with (a,b) in place of $a+b\mathrm{i}$. The reason is to remind us that the addition operation on $\mathbb{C}$ is not being invoked when we write out an element $a+b\mathrm{i}$, only when we do a sum $(a+b\mathrm{i})+(c+d\mathrm{i})$.

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

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

3.4 Rings from modular arithmetic

Theorem 3.8. *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. The law states that, for all elements $A, B, C \in \mathbb{Z}_m$, the equation

$$A(B+C) = AB + AC \tag{1}$$

must hold. The addition and multiplication operations in $\mathbb{Z}_m$ are defined in terms of representatives, so to evaluate either side of equation (1) we need to choose representatives of the congruence classes A, B, and C. So let $A = [a]_m$, $B = [b]_m$, and $C = [c]_m$ for some integers a, b, c. Now what we are trying to prove is 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 $[ab+ac]_m$. We know $a(b+c) = ab+ac$, 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. The equation that we must prove is

$$A+[0]_m = [0]_m + A = A,$$

for all $A \in \mathbb{Z}_m$ — or, after letting $A = [a]_m$ for an integer a,

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

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 2.10.

Theorem 3.9. *Suppose that p is a prime number. Then $\mathbb{Z}_p$ is a field.*

Proof. Building on Theorem 3.8, 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 2.10 implies that $[a]_p$ has an inverse in $\mathbb{Z}_p$. □

3.5 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 3.10. *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 $a \in R$,

$$\begin{aligned} a+z &= z+a = a, \\ a+z' &= z'+a = a. \end{aligned}$$

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

$$\begin{aligned} a+b &= b+a = 0, \\ a+b' &= b'+a = 0. \end{aligned}$$

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 3.11. *Let R be a ring.*

(a) If R has a multiplicative 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 3.12. *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 skewfield 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 3.13. *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 will find useful. 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 3.14. *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).

4 Polynomials

In this section and the next we explore two further examples of rings, polynomial rings and matrix rings. The special feature of these is that you can "build them on top of" another ring: if R is a ring then you can take the coefficients of a polynomial, or the entries of a matrix, to be elements of R. We start with polynomials.

4.1 Defining 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 make the definition more general.

Definition 4.1. Let R be a ring. Let x be a variable.

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

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

Our point of view on polynomials is a little different to the one you may be used to. We will not be thinking of polynomials primarily as functions, as processes waiting for a number to be substituted in for x. We will think of them as expressions to be manipulated algebraically, just like the expressions "$a + b\mathrm{i}$" that we call complex numbers. For example, the answer to the question

Does x^2+1 equal $4x-2$, as elements of $\mathbb{R}[x]$?

is just "no", not "if $x = 1$ or $x = 3$".

Sometimes the x in Definition 4.1 is called 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[10].

Here are some further remarks on this definition.

- Make sure not to confuse the notations $\mathbb{R}[x]$, the set of polynomials with real coefficients, and $R[x]$, the set of polynomials with coefficients in some ring R.

- 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 $\mathbb{Z}[t]$.

- 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? Mostly 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

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

  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 $\mathbb{R}[x]$.

Any polynomial $f \in R[x]$ determines a function $R \to R$, given by substituting a value for x. You're used to calling this function f, and thinking of f as a function itself. Why didn't I define polynomials as a kind of function? The reason is that, over some rings, different polynomials can give you the same function. For example, in $\mathbb{Z}_2[x]$, both the polynomials $f = x$ and $g = x^2$ determine the identity function.

[10] 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.

Definition 4.2. 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[11]:

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.

4.2 Polynomial rings

To give a full definition of $R[x]$ as a ring, we need to define its addition and multiplication operations. 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

$$\begin{aligned} fg = a_m b_n x^{m+n} &+ (a_m b_{n-1} + a_{m-1}b_n)x^{m+n} + \cdots \\ &\cdots + (a_2b_0 + a_1b_1 + a_0b_2)x^2 + (a_1b_0 + a_0b_1)x + a_0b_0; \end{aligned}$$

the coefficient of the general term x^k is the sum of the products a_ib_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] 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. And in many contexts, for a function to qualify as "linear" it must have no constant term.

Theorem 4.3. *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 $R[x]$.

Like previous proofs of this kind, the proof of Theorem 4.3 is long because of the number of axioms to check, so these notes will not include the whole thing. But it is not difficult to see what to do in the proofs. The task that needs care is handling polynomials which may have any number of terms, and keeping track of all the subscripts.[12] Here's one part of the proof.

Proof of the left distributive law. Let

$$\begin{aligned} f &= a_m x^m + a_{m-1}x^{m-1} + \cdots + a_1 x + a_0, \\ g &= b_n x^n + b_{n-1}x^{n-1} + \cdots + b_1 x + b_0, \\ h &= c_n x^n + c_{n-1}x^{n-1} + \cdots + c_1 x + c_0 \end{aligned}$$

be polynomials in $R[x]$. We have supplied leading zero terms in g or h as needed so that we can use coefficients indexed from 0 to n in both. Later in this proof it will be helpful to insert even more zeroes, so define $a_i = 0$ if $i > m$, and $b_i = c_i = 0$ if $i > n$.

The left distributive law assers that $f \cdot (g+h) = fg + fh$. Using the definitions, the left hand side is

$$\begin{aligned} f \cdot (g+h) &= f \cdot ((b_n + c_n)x^n + \cdots + (b_1 + c_1)x + (b_0 + c_0)) \\ &= a_m(b_n + c_n)x^{m+n} + (a_m(b_{n-1} + c_{n-1}) + a_{m-1}(b_n + c_n))x^{m+n-1} \\ &\quad + \cdots + (a_1(b_0 + c_0) + a_0(b_1 + c_1))x + a_0(b_0 + c_0) \end{aligned}$$

and the right hand side is

$$\begin{aligned} fg + fh &= \big(a_m b_n x^{m+n} + (a_m b_{n-1} + a_{m-1} b_n)x^{m+n-1} + \cdots + (a_1 b_0 + a_0 b_1)x + a_0 b_0\big) \\ &\quad + \big(a_m c_n x^{m+n} + (a_m c_{n-1} + a_{m-1} c_n)x^{m+n-1} + \cdots + (a_1 c_0 + a_0 c_1)x + a_0 c_0\big) \\ &= (a_m b_n + a_m c_n)x^{m+n} + (a_m b_{n-1} + a_{m-1} b_n + a_m c_{n-1} + a_{m-1} c_n)x^{m+n-1} \\ &\quad + \cdots + (a_1 b_0 + a_0 b_1 + a_1 c_0 + a_0 c_1)x + (a_0 b_0 + a_0 c_0) \end{aligned}$$

We must prove these two results are equal. To do this, we prove that the coefficient of x^k on the left hand side equals the coefficient of x^k on the right hand side, for every $k = 0, \ldots, m+n$. For any k, we can write the coefficient of x^k on the left hand side as

$$\ell = a_k(b_0 + c_0) + a_{k-1}(b_1 + c_1) + \cdots + a_1(b_{k-1} + c_{k-1}) + a_0(b_k + c_k) \qquad (*)$$

[12] If you are comfortable with the sigma notation for sums, like $\sum_{i=0}^{n} a_i x^i$, I think it makes the bookkeeping for these proofs easier.

because the terms in this expression that weren't actually present in $f \cdot (g+h)$ have been defined to be zero.[13] In the same way, the coefficient of x^k on the right hand side is

$$r = a_k b_0 + a_{k-1} b_1 + \cdots + a_1 b_{k-1} + a_0 b_k + a_k c_0 + a_{k-1} c_1 + \cdots + a_1 c_{k-1} + a_0 c_k.$$

To prove these equal we use the ring axioms in R. Using the left distributive law on each of the $k+1$ terms in ℓ shows that

$$\ell = a_k b_0 + a_k c_0 + a_{k-1} b_1 + a_{k-1} c_1 + \cdots + a_1 b_{k-1} + a_1 c_{k-1} + a_0 b_k + a_0 c_k.$$

Now, repeatedly using the commutative (and associative) laws for addition in R, we can rearrange these summands to prove that

$$\ell = a_k b_0 + a_{k-1} b_1 + \cdots + a_1 b_{k-1} + a_0 b_k + a_k c_0 + a_{k-1} c_1 + \cdots + a_1 c_{k-1} + a_0 c_k = r.$$

□

Proposition 4.4. *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

$$f \cdot bx = b \cdot g \cdot bx = b.$$ □

4.3 Polynomial division

There is a version of the division rule and Euclid's algorithm for polynomials with coefficients in a field. The method for long division for polynomials is similar to

[13] This is not necessary for the proof but it saves us from having to keep track of where this sum should start and end.

that for integers. You probably know it from A-levels. Here is an example: Divide x^4+4x^3-x-5 by x^2+2x-1.

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

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 the statement of the division rule for polynomials. In the integer division rule (Proposition 2.1), we asked for the remainder to be less than the divisor; for polynomials we compare *degrees* instead.

Theorem 4.5. *Let K be a field. Let f and g be two polynomials in $K[x]$, with $g \neq 0$. Then there exist a quotient q and a remainder r in $K[x]$ such that*

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

We have to include the exception "either $r = 0$ or ..." because the degree of the polynomial 0 is undefined.

The idea behind the proof of Theorem 4.5 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*. In the inductive step of the proof, we perform one subtraction, and then use the inductive hypothesis on the resulting polynomial to show that, "after doing the rest of the subtractions", q and r have the properties we want.

In fact we will use *strong* induction. Remember that, in a strong induction proof with the inductive hypothesis $P(k)$ and base case $n = a$ (usually $a = 1$ or $a = 0$ or similar), you may assume all of the facts $P(a), P(a+1), \dots, P(n-1)$ when proving $P(n)$. The inductive variable k will be a degree of a polynomial; strong induction accounts for the fact that the degree might decrease by more than just 1 in one step of long division.[14]

[14] You are probably used to "assume $P(n)$ and prove $P(n+1)$", but this change of variables makes no difference, and it will be less annoying if the statement we have to prove has n in it, not $n+1$.

If you are confused by the proof below, I recommend reading "Division by a linear polynomial" on QMPlus. This is a proof of the special case of Theorem 4.5 when $g = x - \alpha$, and it's simpler than the proof below in a few ways. Once you understand that proof, give this one another go.

With no further ado:

Proof. Our proof will be by strong 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 k be a positive integer. The inductive hypothesis states that, if f^* is a polynomial such that $\deg(f^*) = k$, 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[15]. 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 $k < n$, and prove it for $k = n$.

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

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

[15] The real reason I use a star f^* and not a prime f' in this section of the notes is because I wrote it later, after some students asked me not to use the prime so they don't confuse it with the derivative.

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. □

4.4 Roots and factors

Like integers, polynomials have a divisibility relation.

Definition 4.6. Let f and g be polynomials in $K[x]$. We say that f is a *factor* of, or *divides*, g if $fh = g$ for some $h \in K[x]$.

The notation for "f is a factor of g" is $f \mid g$.

The factors of a polynomial are closely related to its roots, as you know from school when $K = \mathbb{R}$. To discuss the relationship, let's first write down a special case of the polynomial division rule (Theorem 4.5).

Corollary 4.7. *Let K be a field. Let $f \in K[x]$ and $\alpha \in K$. Then there exist $q \in K[x]$ and $r \in K$ such that*

$$f = (x - \alpha) \cdot q + r. \tag{2}$$

Proof. This is the case of Theorem 4.5 when $g = x - \alpha$. Apart from replacing g, the only difference between this corollary and Theorem 4.5 is that r is now an element of K, not of $K[x]$. This is because of the second bullet point in the conclusion of Theorem 4.5, which says that either $r = 0$ or $\deg r < \deg(x - \alpha) = 1$. If $r \neq 0$ and $\deg r < 1$ then $\deg r = 0$, i.e. r is a nonzero scalar in K. □

For a free-standing proof, one which doesn't use Theorem 4.5, see the file "Division by a linear polynomial" on QMPlus.

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

Corollary 4.8. *Let $f \in K[x]$ and $\alpha \in K$. The remainder obtained when dividing f by $x - \alpha$ is $f(\alpha)$.*

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

Proof. By Corollary 4.8, 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

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

To be fully careful about the above proof, we need to check that substituting a value for the variable of a polynomial "works right" in sums and products of polynomials. See the coursework for these facts and their proofs.

You will also have seen repeated roots in your school studies. Here is our definition to handle these.

Definition 4.9. Let k be a nonnegative integer. An element $b \in K$ is a *root of multiplicity* k of the polynomial $f \in K[x]$ if $(x-b)^k$ is a factor of f, but $(x-b)^{k+1}$ is not.

For example, the roots of $(x-1)^3(x-2)^4 \in \mathbb{R}[x]$ are 1 and 2, of which 1 has multiplicity 3 and 2 has multiplicity 4.

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

4.5 The Fundamental Theorem of Algebra

Let's return to the topic of solving polynomial equations, a central interest of historical algebraists. 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 values of its argument at which it evaluates to zero.

Definition 4.10. Let K be a field[16]. We say that the element $b \in K$ is a *root*, or *zero*, of the polynomial $f = a_n x^n + \cdots + a_1 x + a_0 \in K[x]$ if

$$a_n b^n + \cdots + a_1 b + a_0 = 0$$

[16] Most of what these notes say about roots is also true for skewfields, but for simplicity we will stick to fields.

in K. We write the sum on the left hand side as $f(b)$.

Some real polynomials have no roots in $\mathbb{R}$. These include x^2-1, which has no real root, and x^3-2, which has only one, though because of its degree we would like it to have three. Polynomials like these are what historically led mathematicians to look for larger fields than $\mathbb{R}$ and to invent the complex numbers[17]. Wonderfully, *every* polynomial with real coefficients, or even with complex coefficients, has a root inside $\mathbb{C}$!

Theorem 4.11 (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*

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

has at least one *root 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*.

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. With one trick, we can make every complex polynomial equation have its full complement of solutions. The trick is counting repeated roots.

Theorem 4.12 (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*

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

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

When we say there are "n roots, counted with multiplicity", we mean that the sum of the multiplicities of the roots 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 polynomial,

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

[17] There is a *cubic equation* for cubic polynomials, like the familiar quadratic equation, but even when using it to solve a real cubic with three real roots, complex cube roots will sometimes be required in intermediate steps. Phenomena like this are what really forced mathematicians of the sixteenth through eighteenth centuries to give up their philosophical unease and *accept* complex numbers.

has the same roots as the original, so we will analyse it instead.

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 4.8, $z-\alpha_{k+1}$ 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. □

The Fundamental Theorem of Algebra is not constructive. How do we find the roots of polynomials? You already know how to solve real *linear* polynomials. We will review this procedure below and see whether it still works in $K[x]$, when K is a field.

Given two elements $\alpha,\beta\in K$, 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

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

(Note that β/α is just a more familiar way to write $\alpha^{-1}\beta$. We allow ourselves to use the division sign because multiplication is commutative so we don't have to tell $\alpha^{-1}\beta$ and $\beta\alpha^{-1}$ apart.)

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 the field axioms justify these operations, so that we're not doing anything that's not allowed. Very briefly:

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

In this argument to work, the first implication uses the additive identity law, together with the innocuous step of adding the negative of β to both sides of the equation. The second implication uses the additive associative, inverse and identity laws. In the third, α^{-1} exists by the multiplicative inverse law, since we assumed $\alpha \neq 0$; once we know α^{-1} exists there is no problem multiplying both sides by it. Finally, the fourth inplication uses the multiplicative associative, inverse and identity laws, as well as the proposition that $x\cdot(-y) = -xy$ which we proved on the coursework is true in any ring. Therefore, the argument is valid in all fields K, so all linear equations with α nonzero have a unique solution in K.

What about polynomials of degree 2 or greater? These do not always have roots in an arbitrary field: think again of $x^2+1 \in \mathbb{R}[x]$. Even if they do, the roots can't be found using only the field operations. Think about the familiar quadratic equation: it uses a square root, and there's no field axiom about square roots. On QMPlus I have put up a section of supplementary notes on solving higher-degree equations, if you assume you know how to take roots.

4.6 gcds of polynomials

[This section is not examinable.]

Since we have a division rule for polynomials, we can use it to redo polynomial versions of all the facts about division that we did for integers using the integer division rule. 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

$$f = x^n + a_{n-1}x^{n-1} + \cdots + a_1x + a_0.$$

We also say, by convention, that the zero polynomial is monic. (Note that if f has no non-zero coefficients, it is vacuously correct to say that every non-zero coefficient of f with only zero coefficients to its left is 1.)

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 4.13. *(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

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

5 Matrices

Because matrices are one of the main characters of the module *Vectors and Matrices*, in this module I will give them a much briefer treatment than I did with polynomials in the previous chapter. Refer to your *Vectors and Matrices* notes for further basic examples with matrices.

5.1 Defining matrices

Let R be a ring. An *$m \times n$ matrix with entries 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.

Sums and products of matrices don't exist for all matrices; their sizes have to be compatible. The sum of two matrices only exists if the two matrices have the same size. If $a = (a_{ij})_{m\times n}$ and $b = (b_{ij})_{m\times n}$, then we define $a+b$ to be the $m \times n$ matrix with entries

$$(a+b)_{ij} = a_{ij} + b_{ij}$$

for all $i = 1,\ldots,m$ and all $j = 1,\ldots,n$. For products the condition is different: ab exists when the number of columns of a equals the number of rows of b. So if $a = (a_{ij})_{m\times n}$ and $b = (b_{ij})_{n\times p}$, we define ab to be the $m \times p$ matrix with entries

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

for all $i = 1,\ldots,m$ and all $j = 1,\ldots,p$.

So in order for but matrix multiplication to be an *operation* on a set, as defined in Definition 3.1, that set cannot be the set of all matrices, but must contain only matrices all of the same size. These must also have the same number of rows and columns: such matrices are known as *square* matrices. The set of all $n \times n$ (square) matrices with entries in R is denoted by $\mathrm{M}_n(R)$.

5.2 Matrix rings

Theorem 5.1. *If R is a ring, then so is $\mathrm{M}_n(R)$.*
If R is a ring with identity, then so is $\mathrm{M}_n(R)$.

The proof is not difficult, but again is 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.

But so as to not leave you with nothing, here is a quick proof of the multiplicative identity law when R is a ring with identity and $n = 2$. We will show that the 2×2 identity matrix is $\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$, where 0 and 1 are the additive and multiplicative identity elements in R. Any matrix $A \in \mathrm{M}_2(R)$ can be written[18] $A = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$ for $a, b, c, d \in R$. Now

$$\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} a & b \\ c & d \end{pmatrix} = \begin{pmatrix} 1a + 0c & 1b + 0d \\ 0a + 1c & 0b + 1d \end{pmatrix}.$$

Looking at the upper-left entry, $1a = a$ by the multiplicative identity law for R and $0c = 0$ by Proposition 3.13; therefore $1a + 0c = a + 0 = a$ by the additive identity law for R. Proceeding in the same way for every entry of the product, we see that it simplifies to

$$\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} a & b \\ c & d \end{pmatrix} = \begin{pmatrix} a & b \\ c & d \end{pmatrix} = A.$$

In the same way,

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} a \cdot 1 + b \cdot 0 & a \cdot 0 + b \cdot 1 \\ c \cdot 1 + d \cdot 0 & c \cdot 0 + d \cdot 1 \end{pmatrix} = \begin{pmatrix} a & b \\ c & d \end{pmatrix} = A.$$

Proposition 5.2. *If R is a ring in which not all products of two elements equal zero, and $n \geq 2$, then $\mathrm{M}_n(R)$ is neither a commutative ring nor a skewfield.*

[18] I could have used as with indices like a_{22} instead of separate letters like d. But when there are a fixed number of variables, I find this way easier to read.

Proof. I 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 3.13. 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 $\mathrm{M}_2(R)$ is not commutative.

We also use the second equation to show that $\mathrm{M}_2(R)$ 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 3.13, 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 5.3. (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 $a = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} \in \mathrm{M}_2(R)$ 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 $\mathrm{M}_2(\mathbb{Z}_2)$ is not commutative.

6 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.

6.1 Definition and notation

Definition 6.1. 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 $\{1,2,\dots,n\}$ for some natural number n. We use the notation S_n for the set of permutations of this set $\{1,\dots,n\}$.

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 $8 \mapsto 6$. That is, f is an element of S_8.

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 $1,\dots,8$; under each number x we put its image under the permutation f. In our example, we have

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

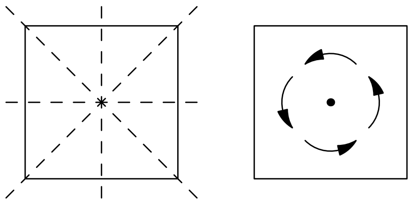
How many permutations of the set $\{1,\dots,n\}$ are there? We can ask this question another way? How many matrices are there with two rows and n columns, such that the first row has the numbers $1,\dots,n$ in order, and the second contains these n numbers in an arbitrary order? There are n choices for 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

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

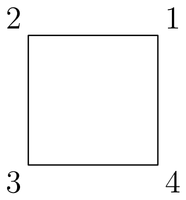
This number is called $n!$, read "n factorial", the product of the integers from 1 to n. Thus we have proved:

Proposition 6.2. $|S_n| = n!$.

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 $f \in S_4$, 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?

6.2 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![19]) 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, to preserve the fact that $x(fg) = (xf)g$.

Here is a fact which we will need later.

Proposition 6.3. *If f and g are elements of S_n, then the composite function $f \circ g$ is in S_n 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,

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

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

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

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

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.

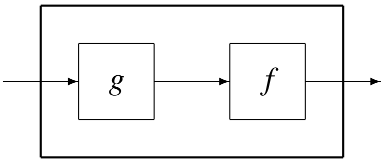
Another way to do it is to re-write the two-line form for f by shuffling the columns around so that the first row agrees with the second row of g. Then the second row will be the second row of $f \circ g$. Thus,

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

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.

The identity function is a permutation in S_n for every n, the *identity permutation*, which leaves everything where it is. We will denote it by e. In S_8 the identity permutation 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.

Proposition 6.4. *If $f \in S_n$, then the inverse function f^{-1} exists and is an element of S_n as well.*

Instead of a proof I will illustrate the truth of this proposition with the running example permutation. 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 $f \circ f^{-1} = f^{-1} \circ f = e$.

6.3 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 elements of $\{1, 2, \ldots, n\}$. The *cycle* $(a_1\ a_2\ \ldots\ a_k)$ is the permutation in S_n which maps $a_1 \mapsto a_2$, $a_2 \mapsto a_3$, ..., $a_{k-1} \mapsto a_k$, 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.

Note: when working in cycle notation, to save space, some authors omit the symbol $\circ$ for composition, just like the multiplication sign $\cdot$ is usually left out.

Theorem 6.5. *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;*

- *each cycle can be started at any point;*
- *cycles of length 1 can be left out.*

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:

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

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 $f = (1\ 4\ 8\ 6\ 5) \circ (2\ 7)$, 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 $(1\ 3\ 5\ 2\ 4) \circ (1\ 5\ 4\ 2\ 6\ 3) \in S_6$ in disjoint cycle notation.

Solution. The first thing I want to make clear is that $f = (1\,3\,5\,2\,4) \circ (1\,5\,4\,2\,6\,3)$ 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 6.5, 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!

[In 2022/23, the rest of this section is not examinable.]

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, $f^{\circ n}$ means $f \circ \cdots \circ f$, with n repeats of f.

Proposition 6.6. *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:

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

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 $\operatorname{lcm}(k_1, \ldots, k_r)$. So this lcm is the order of f. □

7 Groups

In this section we study a new algebraic structure, *groups*, and take a brief look at 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.

7.1 Definition

Definition 7.1. 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 7.2. *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 7.4 below.

As another example, the operations on permutations we saw in Section 6 make them into a group.

Theorem 7.3. *The set S_n of all permutations of $\{1,\dots,n\}$, with the operation of composition, is a group.*

Proof. The closure, identity and inverse laws have been verified in Section 6.2. 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; the letter S in the notation S_n stands for "symmetric". Note that S_n is a group of order $n!$.

Proposition 7.4. *S_n is an abelian group if $n \le 2$, and is non-abelian if $n \ge 3$.*

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 brute force.

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

Remark on notation I have used the symbol $\diamond$ for the group operation in a general group, because it has no baggage from previous uses. In books, you will often see the group operation written as $\cdot$, or for abelian groups as $+$, even when the operation is not meant to be multiplication or 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 chosen 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.

7.2 Cayley tables

If I have a finite group G, I can define it for you by writing down its "times table". The elements of the set G are the labels of the rows and columns, and the values of the group operation can be looked up in the table.

Since the operation of a group might not be "times", 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 $a \diamond x = b$. Using the group axioms, this implies

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

is the unique solution, where a^* is the inverse of a. That is, the single place that a b shows up in row a is in column $a^* \diamond b$. A similar argument applies to the columns.

7.3 Elementary properties

[In 2022/23, everything that follows is not examinable *except* that Definition 7.6 *is* examinable.]

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

Proposition 7.5. *Let G be a group.*

(a) The identity of G is unique.

(b) Each element has a unique inverse.

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

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

Here is how Proposition 7.5(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) If $a \diamond b = a \diamond c$, multiply on the left by the inverse of a to get $b = c$. For the second statement multiply on the right.

(d) 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 $b^* \diamond a^* = (a \diamond b)^*$. □

7.4 Units

Definition 7.6. 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}.

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.
- Theorem 2.10 is a test for being a unit in $\mathbb{Z}_m$.
- Let F be a field and n a positive integer. An element A of the ring $\mathrm{M}_n(F)$ is a unit if and only if the determinant of A is non-zero. (We will not prove this! I think a later *Linear Algebra* module does.) In particular, $\begin{pmatrix} a & b \\ c & d \end{pmatrix}$ is a unit in $\mathrm{M}_2(\mathbb{R})$ if and only if $ad - bc \neq 0$; if this holds, then its inverse is
$$\frac{1}{ad-bc}\begin{pmatrix} d & -b \\ -c & a \end{pmatrix}.$$

By Proposition 7.5, a unit has a unique inverse. Here are some properties of units.

Proposition 7.7. *Let R be a ring with identity.*

(a) If R satisfies the nontrivial law, then 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. □

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 7.8. *$R^\times$ 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 7.7. □

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 $\mathrm{M}_n(F)$ of all $n \times n$ matrices with elements in F is a ring. The group $\mathrm{M}_n(F)^\times$ is called the *general linear group* of dimension n over F, written $\mathrm{GL}(n,F)$. The general linear group is not abelian if $n \geq 2$.
- The elements of the group $\mathbb{Z}_m^\times$ are described in Theorem 2.10.

7.5 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 7.9. Let $(G,\diamond)$ be a group, and H a subset of G. We say that H is a *subgroup* of G if H, with the same operation $\diamond$, is itself a group.

So not every subset is a subgroup. How do we decide if a subset H is a subgroup? It has to satisfy the group axioms for the operation $\diamond$ that G came with.

(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^* 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 7.10. *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 $h_1 \diamond h_2^* \in H$.*

Proof. If H is a subgroup and $h_1,h_2 \in H$, then $h_2^* \in H$, and so $h_1 \diamond h_2^* \in H$.

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 $e = h \diamond h^* \in H$; so (G2) holds. Now, for any $h \in H$, we have $h^* = e \diamond h^* \in H$; so (G3) holds. Then for any $h_1,h_2 \in H$, we have $h_2^* \in H$, so $h_1 \diamond h_2 = h_1 \diamond (h_2^*)^* \in H$; 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 $(\mathbb{Z},+)$.

We close with one important result about subgroups, *Lagrange's Theorem*.

Theorem 7.11. *Let G be a finite group, and H a subgroup of G. Then $|H|$ divides $|G|$.*

I could have proved Lagrange's Theorem if I had one or two lectures more. Instead I will post an outline of the proof as a supplementary document. I heartily recommend the exercise of filling in the details, which nicely ties together several topics from this module.

A The vocabulary of proposition and proof

This appendix is not examinable per se, but the words it defines are used throughout the examinable material and may appear on the exam.

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. It does not have to be in the exact same words but it should convey the same precise idea. 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 an 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 3.1. 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: this table may look wrong to you, 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