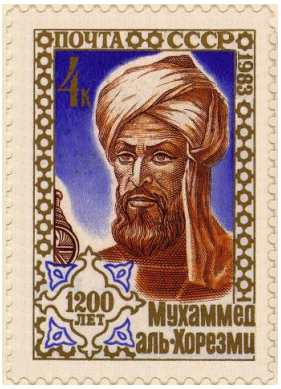
Ten Chapters of the Algebraical Art

Peter J. Cameron

Preface

Abu Ja'far Muhammad ibn Musa al-Khwarizmi (whose name gives us the word 'algorithm') wrote an algebra textbook which included much of what is still regarded as elementary algebra today. The title of his book was *Hisab al-jabr w'al-muqabala*. The word *al-jabr* means 'restoring', referring to the process of moving a negative quantity to the other side of an equation; the word *al-muqabala* means 'comparing', and refers to subtracting equal quantities from both sides of an equation. Both processes are familiar to anyone who has to solve an equation! The word *al-jabr* has, of course, been incorporated into our language as 'algebra'.



In a similar vein, Doctor Johnson gave this definition of "algebra" in his *Dictionary* of 1755:

> This is a peculiar kind of arithmetick, which takes the quantity sought, whether it be a number or a line, or any other quantity, as if it were granted, and by means of one or more quantities given, proceeds by consequence, till the quantity at first only supposed to be known, or at least some power thereof, is found to be equal to some quantity or quantities which are known, and consequently itself is known.

Since the time of Al-Khwarizmi and Johnson, the subject of algebra has changed considerably. Firstly, we no longer restrict ourselves to considering just numbers; the variables and symbols in our equations may be vectors, matrices, polynomials, sets, or permutations. Secondly, the way we look at these equations has also changed. As far as possible, we don't care what the variables stand for, but only the "laws" that they obey (associative, distributive, etc.); so that we can prove something about a system satisfying certain laws which will apply to systems of numbers, matrices, polynomials, etc. We sometimes refer to this as "abstract algebra".

These notes are intended for the course MAS117, *Introduction to Algebra*, at Queen Mary, University of London. The course is to be given for the first time in the spring semester of 2007.

The course is intended as a first introduction to the ideas of proof and abstraction in mathematics, as well as to the concepts of abstract algebra (groups and rings). The *Undergraduate Studies Handbook* says:

> This module is an introduction to the basic notion of algebra, such as sets, numbers, matrices, polynomials and permutations. It not only introduces the topics, but shows how they form examples of abstract mathematical structures such as groups, rings, and fields and how algebra can be developed on an axiomatic foundation. Thus, the notions of definitions, theorem and proof, example and counterexample are described. The course is an introduction to later modules in algebra.

The course replaces the earlier course *Discrete Mathematics*, with which it shares some material. But since it is a new course, I have re-written the notes from scratch. Of course, these notes are *not* a substitute for the lectures!

The exercises at the ends of the chapters vary in difficulty from routine to challenging. To a first approximation, the easier exercises come first.

If you enjoyed this course, the next step is MAS201, *Algebraic Structures I*. You can also find a set of notes for this course on my web page.

This set of notes is a slightly revised version of the notes which were available during the course. I am grateful to Matilda Okungbowa for a number of corrections.

Note: The pictures and information about mathematicians in these notes are taken from the St Andrews *History of Mathematics* website:
`http://www-groups.dcs.st-and.ac.uk/~history/index.html`

Peter J. Cameron
25 June 2007

Contents

1 What is mathematics about? 1
- 1.1 Some examples of proofs . . . 1
- 1.2 Some proof techniques . . . 6
- 1.3 Proof by induction . . . 7
- 1.4 Some more mathematical terms . . . 10

2 Numbers 15
- 2.1 The natural numbers . . . 16
- 2.2 The integers . . . 17
- 2.3 The rational numbers . . . 18
- 2.4 The real numbers . . . 19
- 2.5 The complex numbers . . . 19
- 2.6 The complex plane, or Argand diagram . . . 21

3 Other algebraic systems 25
- 3.1 Vectors . . . 25
- 3.2 Matrices . . . 28
- 3.3 Polynomials . . . 31
- 3.4 Sets . . . 32

4 Relations and functions 35
- 4.1 Ordered pairs and Cartesian product . . . 35
- 4.2 Relations . . . 37
- 4.3 Equivalence relations and partitions . . . 37
- 4.4 Functions . . . 40
- 4.5 Operations . . . 43
- 4.6 Appendix: Relations and functions . . . 44

5 Division and Euclid's algorithm 47
- 5.1 The division rule . . . 47
- 5.2 Greatest common divisor and least common multiple . . . 48

5.3 Euclid's algorithm 49
5.4 Euclid's algorithm extended 50
5.5 Polynomials 51

6 Modular arithmetic **55**
6.1 Congruence mod *m* 55
6.2 Operations on congruence classes 56
6.3 Inverses 57
6.4 Fermat's Little Theorem 58

7 Polynomials revisited **61**
7.1 Polynomials over other systems 61
7.2 Division and factorisation 62
7.3 "Modular arithmetic" for polynomials 63
7.4 Finite fields 66
7.5 Appendix: Laws for polynomials 67

8 Rings **69**
8.1 Rings 69
8.2 Examples of rings 71
8.3 Properties of rings 71
8.4 Units 74
8.5 Appendix: The associative law 76

9 Groups **79**
9.1 Definition 79
9.2 Elementary properties 79
9.3 Examples of groups 80
9.4 Cayley tables 81
9.5 Subgroups 82
9.6 Cosets and Lagrange's Theorem 83
9.7 Orders of elements 85
9.8 Cyclic groups 87

10 Permutations **89**
10.1 Definition and representation 89
10.2 The symmetric group 90
10.3 Cycles 92
10.4 Transpositions 94
10.5 Even and odd permutations 96

Chapter 1

What is mathematics about?

There is a 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.

1.1 Some examples of proofs

In this part of the course we are going to talk about how to prove things. Let us start with an easy theorem.

Theorem 1.1 *Let n be a natural number. Then n^2 is odd if and only if n is odd.*

If you know what the words in the theorem mean, you might try a few cases, to get a feel for what the theorem is about:

1 is odd $1^2 = 1$ is odd
2 is even $2^2 = 4$ is even
3 is odd $3^2 = 9$ is odd

and so on. It seems to work. But this is not yet a proof; we are not convinced that if you went on far enough, you might find a number for which the theorem was not true.

First let us read the theorem more carefully.

Natural number This means one of the counting numbers, $0, 1, 2, 3, 4, \ldots$. (Arguments still occur among mathematicians about whether 0 should count as a natural number or not. This is just a matter of names, and doesn't affect the theorem very much. We will say that 0 is a natural number.)

If and only if We will come back to this later. For now, it means that, for any value of n, either the two statements "n is odd" and "n^2 is odd" are both true, or they are both false. In other words,

- if n is odd, then n^2 is odd;
- if n^2 is odd, then n is odd.

This shows us that we have two things to show, in order to prove the theorem. The first one looks fairly straightforward, but the second seems more difficult. But we can turn it round into something simpler. The statement

$$\text{if } n^2 \text{ is odd, then } n \text{ is odd}$$

is logically the same as the statement

$$\text{if } n \text{ is even, then } n^2 \text{ is even.}$$

So we have to prove the two statements:

- if n is odd, then n^2 is odd;
- if n is even, then n^2 is even.

So let's try to prove them.

We have one more thing to consider. What are even and odd numbers, mathematically speaking? An even number is one which is divisible by 2 exactly; in other words, n is even if it can be written as $n = 2k$ for some natural number k. An odd number is one which leaves a remainder of 1 when divided by 2; in other words, n is odd if it can be written as $n = 2k+1$ for some number k.

So to prove the first statement, we assume that n is an odd number, and have to show that n^2 is an odd number. That is, we assume that $n = 2k+1$ for some natural number k. Then

$$n^2 = (2k+1)^2 = 4k^2 + 4k + 1 = 2(2k^2 + 2k) + 1 = 2m + 1,$$

where $m = 2k^2 + 2k$. So n^2 is odd.

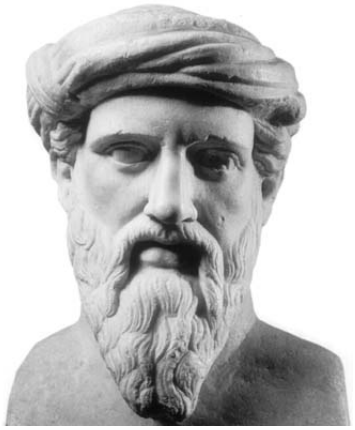
For the second statement, assume that n is even, that is, $n = 2k$ for some natural number k. Then

$$n^2 = (2k)^2 = 4k^2 = 2(2k^2) = 2m,$$

where $m = 2k^2$; so n^2 is even.

Now we have finished the proof, and we are sure that the theorem is true for all natural numbers n.

Now let's use this theorem as a building block in a very famous theorem, proved by Pythagoras, who has some claim to be the first mathematician ever (that is, the first person to insist that mathematical statements must have proofs). It was Pythagoras who invented the words "mathematics" and "theorem".



Theorem 1.2 *The number $\sqrt{2}$ is irrational.*

First we have to examine what the theorem means. The number $\sqrt{2}$ is a positive real number x such that $x^2 = 2$. A rational number is a number that can be expressed as a fraction a/b, where a and b are integers, that is, natural numbers or their negatives.

Now my calculator tells me that $\sqrt{2} = 1.414213562$. If this is right, then Pythagoras is wrong, because this means that

$$\sqrt{2} = \frac{1414213562}{1000000000}.$$

But it turns out that the calculator is wrong, because it also tells me that

$$(1.414213562)^2 = 1.999999998944727844,$$

which is close to 2 but not exactly 2. Pythagoras claims that, no matter how accurately the calculator does the sum and to how many places of decimals it expresses the answer, it will never get the exact value of $\sqrt{2}$.

So how did Pythagoras prove his theorem? He used another important technique:

Proof by contradiction If I am trying to prove a statement P, I have succeeded if I can show that the assumption that P is false leads to a contradiction, a logical absurdity. For this shows that P is not false, that is, it is true.

So we prove Pythagoras's theorem by contradiction; we assume the falsity of what we are trying to prove and head for a contradiction. That is, we assume that

$$\sqrt{2} \text{ is rational.}$$

That is,

$$\sqrt{2} = \frac{m}{n}$$

for some natural numbers m and n. Now, in a fraction like this, if there is a common factor of m and n, we can divide it out, and assume that they have no common factor. (For example, $\frac{15}{10} = \frac{3}{2}$.)

Now take our equation. Square roots are awkward; it usually simplifies an equation if you can get rid of them. We can easily do this by squaring both sides of the equation, to get

$$2 = \frac{m^2}{n^2},$$

or in other words,

$$m^2 = 2n^2.$$

This equation tells us that m^2 is even, since it is $2k$ where $k = n^2$. Now we are able to use Theorem 1.1, since we already proved this. Since m^2 is even, necessarily m is even; say $m = 2p$ for some natural number p. Substituting this into the equation gives

$$4p^2 = 2n^2,$$

and cancelling a factor 2 gives

$$2p^2 = n^2.$$

Now we can "do it again". The last equation shows that n^2 is even, so that n is even, say $n = 2q$. So our original fraction for $\sqrt{2}$ is

$$\sqrt{2} = \frac{m}{n} = \frac{2p}{2q}.$$

We can cancel the 2 to get a simpler fraction for $\sqrt{2}$.

But stop and remember what we are doing. We started off by saying that we can assume that m and n have no common factor, and we ended up with their having a common factor of 2. So we have reached a contradiction.

According to the principle of proof by contradiction, our assumption that $\sqrt{2}$ is rational must be wrong, so that $\sqrt{2}$ is irrational, as Pythagoras claimed.

Now let us have another famous example of a proof by contradiction. We will prove that the prime numbers go on for ever; there is no largest prime. (Recently, a computer search found a previously unknown prime number bigger than any others found so far. A journalist got the idea that they had found "the largest prime number", and phoned one of my colleagues for a comment. What would you say if this happened to you?)

This beautiful proof was discovered by the Greek geometer Euclid, who wrote one of the world's most successful textbooks ever, which was used for nearly two thousand years.

Theorem 1.3 *There are infinitely many prime numbers.*

A prime number is a natural number which is divisible only by itself and 1. So 2, 3, 5 and 7 are prime numbers; 4 is not, since $4 = 2 \times 2$. By convention, we say that 1 is not a prime number, even though it satisfies the condition of having no divisors except itself and 1; this is just a convention, and we will see the reason for it later. Now if the number n is not prime, it must be divisible by some prime number smaller than n. (Again, we will see why later. This is not meant to be obvious!)

We prove Euclid's theorem by contradiction. That is, we assume that there are only finitely many prime numbers. Then we can make a list of prime numbers:

$$p_1, p_2, p_3, \ldots, p_k$$

are all the prime numbers.

Let n be the number that we get when we multiply all of these primes together and add 1:

$$n = p_1 p_2 p_3 \cdots p_k + 1.$$

Now there are two cases to consider: either n is prime, or it is not. We need to show that either case leads us to a contradiction.

Case n is prime: In this case, since $p_1, \ldots, p_k$ are all the primes, n must be one of them. But this is impossible, since n is bigger than any of these primes. (Remember how we formed n.)

Case n is not prime: Then n must have a prime factor, which must be one of the primes $p_1, \ldots, p_k$. But n is the product of all the primes plus one; so if we divide it by any of the primes $p_1, \ldots, p_k$, we get a remainder of one. So this case is also contradictory.

So, again according to the principle of proof by contradiction, the assumption that there are only finitely many primes must be wrong; so there must be infinitely many primes.

1.2 Some proof techniques

Here are some words you'll find in statements you are asked to prove.

If, implies, sufficient The three statements

If A, then B

A implies B

A is a sufficient condition for B

all have the same meaning. They mean, "if A is true, then B is true".

Look more closely at this. How could this statement fail to be true? The only way it could fail is if A is true and B is false. (If A is false, then the statement is correct no matter whether B is true or false.) This seems a bit odd, sometimes, so let us take an everyday example. Suppose I say to you, "If it is fine tomorrow, we will go for a picnic." The only situation in which my statement is false is if it is fine tomorrow and we don't go for a picnic; if it rains tomorrow, my statement is technically correct (though maybe not helpful!)

So how do we prove "if A, then B"? The obvious way is to assume that A is true, and deduce that B must be true. Look back at our proof of "if n is even, then n^2 is even" in the last section. We assume that n is even and prove that n^2 is even.

Only if, is implied by, necessary This is exactly the reverse. The three statements

B only if A

A is implied by B

A is a necessary condition for B

all mean the same as "if B, then A".

The proof strategy, then, is to assume that B is true, and deduce that A must be true.

If and only if, equivalent, necessary and sufficient We saw earlier that to say "A if and only if B" means that either A and B are both true, or they are both false. We also saw that there are two things we have to do to show this: "if A, then B" and "if B, then A". This agrees with what we just learned about "if" and "only if". We sometimes also say "the statements A and B are equivalent", or "A is a necessary and sufficient condition for B".

Now we turn to some proof techniques.

Proof by contradiction We already met this idea. In order to prove A, we can assume that A is false and deduce a contradiction (a statement that is logically impossible). We saw two examples of this: the proofs of Pythagoras' Theorem on the irrationality of $\sqrt{2}$, and Euclid's theorem that there are infinitely many primes.

Proof by contrapositive This is a fancy way of saying that "A implies B" is logically equivalent to "not-B implies not-A". We saw an example of this on page 2. In order to prove the statement "if n^2 is even, then n is even", we proved instead its contrapositive, the statement "if n is odd, then n^2 is odd".

Counterexamples Sometimes you will be given a general proposition, and asked whether it is true or false.

Suppose for example you are trying to prove that some property holds for every natural number n. Let us call the property $A(n)$. Now:

- If $A(n)$ is true, then we have to give a general proof for it.
- If $A(n)$ is false, we only have to give one value of n for which it is not true.

For example, suppose we are considering the statement "every odd number is prime". So $A(n)$ would be, "if n is odd, then n is prime". If this happened to be true, we would have to give a proof of it. But it is false, and all we need to say is "the number 9 is odd, but is not prime since it is equal to 3×3". In this case, we say that 9 is a *counterexample* to the statement that, if n is odd, then n is prime.

1.3 Proof by induction

This is a more specialised technique but is very important, so we give it a section to itself.

Suppose that we are trying to prove a statement about all natural numbers. Suppose that $A(n)$ is the statement about the particular natural number n. The strategy of proof by induction is to do the following:

(a) Prove the statement $A(0)$, that is, the case when $n = 0$.

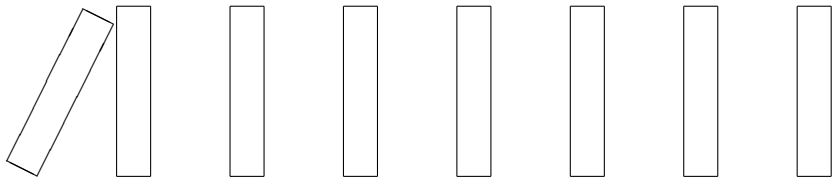
(b) Prove that, if $A(n)$ is true, then $A(n+1)$ is true. In other words, *assume* $A(n)$ and *prove* $A(n+1)$.

Here (a) is called "starting the induction", and (b) is "the inductive step".

This is a bit confusing at first, since in part (b) we seem to be assuming the thing we are trying to prove, namely $A(n)$; an argument where you assume what you are trying to prove can't be valid, right? Well, in this case the argument is

right. By (a), we know that $A(0)$ is true. Now by (b) (in the case $n = 0$), we know that $A(0)$ implies $A(1)$, so $A(1)$ must be true. By (b) again (with $n = 1$), we know that $A(1)$ implies $A(2)$, so $A(2)$ must be true. And so on. Given any number n, we can count up to n; and at each step of the way, (b) allows us to get from the truth of each statement to the truth of the next.

Suppose that we have a line of dominos, as shown in the diagram.



If we push over the first domino, what will happen? It will knock over the second, which will knock over the third, and so on; eventually all the dominos will fall. This is like induction. The inductive step is the fact that each domino knocks over the next one, and starting the induction is giving the first domino a push.

We have a bit of freedom about starting the induction. Instead of 0, it might be more convenient to start by proving $A(1)$; this and the inductive step show that $A(n)$ is true for all $n \geq 1$. We'll see an example soon where we start with $A(2)$.

Here is an example. What is the sum of the first n positive integers? Induction doesn't help us *guess* the answer, but if we can guess it, induction will let us *prove* that our guess is correct.

Theorem 1.4 *The sum of the first n positive integers is* $n(n+1)/2$.

Again we can check this for small values: for example,

$$1+2+3+4+5 = 15 = 5 \times 6/2.$$

Here is the proof by induction. Let $A(n)$ be the statement

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

Starting the induction For $n = 1$, the left hand side is 1, and the right-hand side is $1 \times 2/2 = 1$; so $A(1)$ is true.

The inductive step Suppose that $A(n)$ is true; that is,

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

We have to prove that $A(n+1)$ is true.

Now the left-hand side of $A(n+1)$ is $1+2+\cdots+n+(n+1)$. Since we are assuming that $A(n)$ is true, this is equal to

$$\begin{aligned}\frac{n(n+1)}{2}+(n+1) &= \frac{n(n+1)}{2}+\frac{2(n+1)}{2}\\ &= \frac{(n+1)(n+2)}{2},\end{aligned}$$

after a little bit of algebraic manipulation. But this is exactly the right-hand side of $A(n+1)$; it is what we get from the expression $n(n+1)/2$ if we substitute $n+1$ in place of n. So the left and right sides of $A(n+1)$ are equal, and $A(n+1)$ is true.

By induction, we have proved that $A(n)$ is true for all $n \geq 1$.

Unfinished business I told you earlier that if a natural number n is greater than 1 and is not prime, then it is divisible by some prime number less than n. In other words,

Theorem 1.5 *Every natural number $n > 1$ has a prime factor.*

We prove this theorem by induction. Take $A(n)$ to be the statement "every natural number k satisfying $1 < k \leq n$ has a prime factor". We prove $A(n)$ by induction.

Starting the induction We can conveniently start the induction with $n = 2$: there is only one number k satisfying $1 < k \leq 2$, namely $k = 2$, and it has a prime factor, namely 2. [**Note:** We could start the induction with $n = 1$: there are no numbers k satisfying $1 < k \leq 1$, and so any statement at all is true for all of them! But you may feel uncomfortable with this sort of argument!]

The inductive step We assume that $A(n)$ is true, and we have to prove $A(n+1)$. In other words, we assume that every natural number k satisfying $1 < k \leq n$ has a prime factor, and we have to prove that every natural number k satisfying $1 < k \leq n+1$ has a prime factor. Well, we don't have to prove it for *all* these numbers, since the hypothesis $A(n)$ shows that it is true for $k = 2, 3, \ldots, n$; we only have to prove it for $k = n+1$.

Case 1: $n+1$ is prime. If it is prime, it certainly has a prime factor, namely itself.

Case 2: $n+1$ is not prime; so $n+1 = ab$ for some natural numbers a and b, where neither factor is 1. Then each factor must be smaller than $n+1$. So, for example, $1 < a \leq n$. By $A(n)$, we know that a has a prime factor p. Then p is also a factor of $n+1$, and we have finished.

This completes the proof by induction.

Here is a variant on the principle of induction. Sometimes you might find this easier to apply.

Suppose that we are trying to prove a statement $A(n)$. We begin by arguing by contradiction: we assume that $A(n)$ isn't true for all values of n, that is, there is some value of n for which it is false. So there must be a smallest value of n for which $A(n)$ is false. Now this n has the property that $A(n)$ is false but $A(m)$ is true for all numbers m smaller than n – so we call n the "minimal counterexample" to the statement we are trying to prove. (Some people call n the "least criminal".) If we can show that no minimal counterexample can exist, then we have proved that $A(n)$ is true for all n.

Why is this the same as induction? Well, let n be the minimal counterexample, and remember we are trying to get a contradiction. Maybe $n = 0$. To show a contradiction, we have to show that $A(0)$ is true. Or maybe $n > 0$. Now $A(n)$ is false and $A(n-1)$ is true, so if we could show that $A(n-1)$ implies $A(n)$, we would have a contradiction in this case too. So the two things we have to prove are precisely the same as starting the induction and doing the inductive step in a proof by induction. But sometimes it is easier to think about a minimal counterexample.

Take an induction proof and try writing it out in the "minimal counterexample" style, and see which you prefer.

1.4 Some more mathematical terms

There are many other specialised terms in mathematics.

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 a statement which is less important. (Of the five theorems we've seen so far, I would normally call two of them "theorems" and the other three "propositions"; can you guess which are which?) A *corollary* is a statement which follows easily from a theorem or proposition. For example, the statement

Let n be a natural number. Then n^2 *is odd if and only if n is odd.*

follows easily from Theorem 1 in the notes, so I could call it a corollary of Theorem 1. Finally, a *lemma* is a statement which is proved as a stepping stone to some more important theorem. So I could have called Theorem 1 a lemma for the proof of Theorem 2. (Remember how we used Theorem 1 in the proof of Theorem 2.)

Of course these words are not used very precisely; it is a matter of judgment whether something is a theorem, proposition, or whatever. For example, there is a very famous theorem called *Fermat's Last Theorem*, which is the following:

Theorem 1.6 *Let n be a natural number bigger than 2. Then there are no positive integers x, y, z satisfying $x^n + y^n = z^n$.*

This was proved fairly recently 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. 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 about 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. 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.

Exercises

1.1 Write down and prove the contrapositive of the statement

If x is an irrational number then $1-x$ is an irrational number.

1.2 Find counterexamples to the statements

(a) Every odd number is prime.

(b) Every prime number is odd.

1.3 Prove by induction that

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

1.4 Let n be a positive natural number, and suppose that n has the property that every positive natural number smaller than $n/2$ divides n. Prove that $n\leq 6$, and hence find all numbers n with this property.

1.5 Define the *binomial coefficient* $\binom{n}{k}$ for natural numbers n and k by the rule

$$\binom{n}{k}=\frac{n\cdot(n-1)\cdots(n-k+1)}{k\cdot(k-1)\cdots 1}=\begin{cases}\dfrac{n!}{k!\,(n-k)!} & \text{if } 0\leq k\leq n,\\ 0 & \text{if } k>n.\end{cases}$$

(Here $n!$ is the product of the natural numbers from 1 to n.)

(a) I have given you two definitions here. Prove that they are equivalent.

(b) Prove that

$$\binom{n}{k}+\binom{n}{k-1}=\binom{n+1}{k}.$$

(c) Using this and induction on n, prove the *Binomial Theorem*:

$$(a+b)^n=\sum_{k=0}^{n}\binom{n}{k}a^k b^{n-k}$$

for positive integers n.

1.6 Prove that

$$\sum_{i=0}^{n}\binom{i}{k}=\binom{n+1}{k+1}.$$

1.7 Find the mistake in the following proof of the "Theorem": *All triangles are isosceles*. (You will need to draw a figure!)

Proof Given any triangle ABC, let D be the point inside the triangle where the bisector of the angle A meets the perpendicular bisector of the side BC. Now let DM be the perpendicular from D to AB and DN be the perpendicular from D to AC.

Step 1 The triangles ADN and ADM are congruent (since they have the same angles and they also have the side AD in common).

Step 2 The triangles CDN and BDM are congruent (since $DN = DM$ from Step 1, and $DC = DB$ as DL is the perpendicular bisector of BC by construction, and the angles $C\hat{N}D$ and $B\hat{M}D$ are both right angles).

Step 3 From Step 1 we have $AN = AM$, and from Step 2 we have $NC = MB$. Hence $AC = AB$.

Chapter 2

Numbers

Algebra begins by considering numbers and their properties, and moves on to other kinds of mathematical objects. In this section of the notes, we will look at numbers.

The important sets of numbers are:

- the natural numbers, denoted by $\mathbb{N}$;
- the integers, denoted by $\mathbb{Z}$;
- the rational numbers, denoted by $\mathbb{Q}$;
- the real numbers, denoted by $\mathbb{R}$;
- the complex numbers, denoted by $\mathbb{C}$.

The notation we use for them is a special typeface called "blackboard bold". Originally, number systems were printed in bold type: $\mathbf{N}, \mathbf{Z}$, etc.; lecturers writing on the blackboard couldn't write in bold, so invented a different way of doing it; then the printers had to catch up by designing a typeface.

The notation $\mathbb{N}$, $\mathbb{R}$ and $\mathbb{C}$ for natural, real and complex numbers is easy to remember; but what about the others? If the real numbers are called $\mathbb{R}$, then we need a different letter for the rational numbers; we choose $\mathbb{Q}$ for "quotients", since every rational number has the form a/b where a and b are integers. The $\mathbb{Z}$ comes from the German word *Zahlen*, meaning numbers.

In this section, you will not learn *definitions* of numbers. I will assume that you know what numbers are; we will revise some of their properties.

2.1 The natural numbers

The German mathematician Leopold Kronecker (pictured) said, "God made the natural numbers; all the rest is the work of man." In the same spirit, the French mathematician Émil Borel said, "All of mathematics can be deduced from the sole notion of an integer; here we have a fact universally acknowledged today."

The important properties of the natural numbers are:

(a) They are used in counting. We can start from zero and, in principle, count up a step at a time to reach any natural number. (Of course there are practical limits!) This is the basis of proof by induction, as we saw in the last chapter.

(b) We can add and multiply natural numbers. These operations satisfy a number of familiar laws that you probably never stopped to think about. These include:

$$\begin{aligned} a+b &= b+a, & ab &= ba, \\ (a+b)+c &= a+(b+c), & (ab)c &= a(bc), \\ a(b+c) &= ab+ac, & & \\ 0+a &= a, & 1a &= a. \end{aligned}$$

These laws are important to us, and they have been given names, which you will need to know. The first two are the *commutative laws* (for addition and multiplication respectively), the next two are the *associative laws* (for addition and multiplication), the fifth is the *distributive law*, and the last two are the *identity laws* (for addition and multiplication).

(c) Although we can add and multiply, we cannot always subtract or divide natural numbers. There is no natural number x such that $4+x=2$, and no y such that $3y=5$.

The facts that subtraction and division are not possible in the natural numbers can be viewed another way. Since we can think of subtraction as "adding the negative" and division as "multiplying by the reciprocal", we can formulate two further laws known as the *inverse laws* to describe the situation. These are laws which *do not hold* for the natural numbers!

Additive inverse law: For any element a, there exists an element $-a$ such that $a+(-a)=0$.

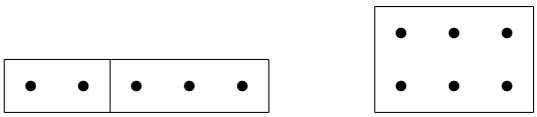
Multiplicative inverse law: For any element $a \neq 0$, there exists an element a^{-1} such that $a \cdot a^{-1} = 1$.

Notice the exclusion in the multiplicative inverse law; we can't divide by zero!

The laws for the natural numbers can be interpreted in terms of counting. This depends on two obvious principles:

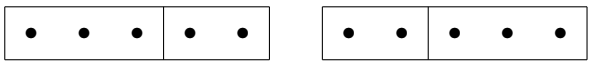
- a row of a dots, followed by a row of b dots, contains $a+b$ dots.
- a rectangle of dots with sides a and b contains ab dots.

The figure illustrates this for $a=2$ and $b=3$.

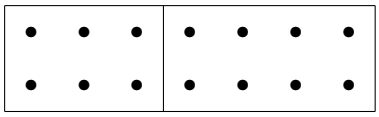


$2+3=5$ $\qquad$ $3\times 2=6$

Now the laws of algebra can be explained by geometric transformations. For example, the picture below shows the commutative law for addition and the distributive law. In the first case, we have reflected the figure left-to-right.



$3+2=2+3$



$(3+4)\times 2 = 3\times 2+4\times 2$

You are invited to produce similar geometric explanations of the commutative law for multiplication and the associative laws.

2.2 The integers

We enlarge the number system because we are trying to solve equations which can't be solved in the original system. At every stage in the process, people first

thought that the new numbers were just aids to calculating, and not "proper" numbers. The names given to them reflect this: negative numbers, improper fractions, irrational numbers, imaginary numbers! Only later were they fully accepted. You may like to read the book *Imagining Numbers* by Barry Mazur, about the long process of accepting imaginary numbers.

Anyway, we can't always subtract natural numbers, so we add negative numbers to make it possible. The *integers* are the natural numbers together with their negatives. So addition, subtraction, and multiplication are all possible for integers. The laws we met for natural numbers all continue to hold for integers. Also, the additive inverse law (but not the multiplicative inverse law) holds for integers.

The natural numbers $1, 2, \ldots$ are positive, while $-1, -2, \ldots$ are negative. Integers satisfy the law of signs: the product of a positive and a negative number is negative, while the product of two negative numbers is positive.

2.3 The rational numbers

In a similar way, rational numbers are introduced because we cannot always divide integers. A rational number is a number which can be written as a fraction

$$\frac{a}{b}$$

where a and b are integers and $b \neq 0$. We require that multiplying or dividing numerator and denominator (top and bottom) of a fraction by the same thing doesn't change the fraction. So, if the denominator is negative, we can multiply by -1 to make it positive; and if numerator and denominator have a common factor, we can divide by it. (We say that a fraction a/b is *in its lowest terms* if the highest common factor of a and b is 1.)

We can write rules for adding and multiplying rational numbers:

$$\frac{a}{b} + \frac{c}{d} = \frac{ad+bc}{bd}, \qquad \frac{a}{b} - \frac{c}{d} = \frac{ad-bc}{bd},$$
$$\frac{a}{b} \times \frac{c}{d} = \frac{ac}{bd}, \qquad \frac{a}{b} \div \frac{c}{d} = \frac{ad}{bc} \text{ if } c \neq 0.$$

The last rule says: to divide by a fraction, turn it upside down and multiply.

So, for rational numbers, addition, subtraction, multiplication, and division (except by 0) are all possible. The rules we met for natural numbers all hold for rational numbers, and so do the two inverse laws.

2.4 The real numbers

There are still many equations we can't solve with rational numbers. One such equation is $x^2 = 2$. (we saw Pythagoras' proof of this in the last chapter.) Other equations involve functions from trigonometry (such as $\sin x = 1$, which has the irrational solution $x = \pi/2$) and calculus (such as $\log x = 1$, which has the irrational solution $x = \mathrm{e}$).

So, we take a larger number system in which these equations can be solved, the *real numbers*. A real number is a number that can be represented as an infinite decimal. This includes all the rational numbers and many more, including the solutions of the three equations above; for example,

$$\begin{aligned} \tfrac{2}{5} &= 0.4 \\ \tfrac{1}{7} &= 0.142857142857\ldots, \\ \sqrt{2} &= 1.41421356237\ldots, \\ \tfrac{\pi}{2} &= 1.57079632679\ldots, \\ \mathrm{e} &= 2.71828182846\ldots \end{aligned}$$

In the last three cases, we cannot write out the number exactly as a decimal, but we assume that the approximation gets better as the number of digits increases.

We can add, subtract, multiply, and divide (except by zero) in the system of real numbers, and the laws we met earlier (including the inverse laws) all hold here too.

2.5 The complex numbers

The final extension arises because there are still equations we can't solve, such as $x^2 = -1$ (which has no real solution) or $x^3 = 2$ (which has only one, though for various reasons we would like it to have three). It turns out that the first equation is the crucial one.

A *complex number* is a number of the form $a + b\mathrm{i}$, where a and b are real numbers, and i is a mysterious symbol which will have the property that $\mathrm{i}^2 = -1$. The rules for addition and multplication are

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

You can work out the rule for subtraction. How do we divide? You can check that the rule above gives

$$(a+b\mathrm{i})(a-b\mathrm{i}) = a^2 + b^2,$$

which is a positive number unless $a = b = 0$. So, to divide by $a + b\mathrm{i}$, we multiply by

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

Thus, in the complex numbers, we can add, subtract, multiply, and divide (except by zero), and the laws we met earlier (including the inverse laws) all apply here too.

Complex numbers are not called complex because they are complicated: a modern advertising executive would certainly have come up with a different name! They are called "complex" because each complex number is built of two parts, each of which is simpler (being a real number).

Here, unlike for the other forms of numbers, we don't have to take on trust that the laws hold; we can prove them. Here, for example, is the 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 shows that the two expressions are the same.

If $z = a + b\mathrm{i}$ is a complex number (where a and b are real), we say that a and b are the *real part* and *imaginary part* of z respectively. The complex number $a - b\mathrm{i}$ is called the *complex conjugate* of z, and is written as $\bar{z}$. So the rules for addition and subtraction can be put like this:

> To add or subtract complex numbers, we add or subtract their real parts and their imaginary parts.

The rule for multiplication looks more complicated as we have written it out. There is another representation of complex numbers which makes it look simpler. Let $z = a + b\mathrm{i}$. We define the *modulus* and *argument* of z by

$$\begin{aligned} |z| &= \sqrt{a^2+b^2}, \\ \arg(z) = \theta \quad &\text{where} \quad \cos\theta = a/|z| \text{ and } \sin\theta = b/|z|. \end{aligned}$$

In other words, if $|z| = r$ and $\arg(z) = \theta$, then

$$z = r(\cos\theta + \mathrm{i}\sin\theta).$$

Now the rules for multiplication and division are:

To multiply two complex numbers, multiply their moduli and add their arguments. To divide two complex numbers, divide their moduli and subtract their arguments.

2.6 The complex plane, or Argand diagram

The complex numbers can be represented geometrically, by points in the Euclidean plane (which is usually referred to as the *Argand diagram* or the *complex plane* for this purpose. The complex number $z = a + b\mathrm{i}$ is represented as the point with coordinates (a,b). Then $|z|$ is the length of the line from the origin to the point z, and $\arg(z)$ is the angle between this line and the x-axis. See Figure 2.1.

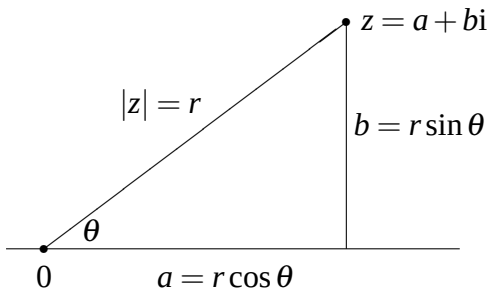


Figure 2.1: The Argand diagram

In terms of the complex plane, we can give a geometric description of addition and multiplication of complex numbers. The addition rule is the same as you learned for adding vectors in Geometry I, namely, the *parallelogram rule* (see Figure 2.2).

Multiplication is a little bit more complicated. Let z be a complex number with modulus r and argument θ, so that $z = r(\cos\theta + \mathrm{i}\sin\theta)$. Then the way to multiply an arbitrary complex number by z is a combination of a stretch and a rotation: first we expand the plane so that the distance of each point from the origin is multiplied by r; then we rotate the plane through an angle θ. See Figure 2.3, where we are multiplying by $1+\mathrm{i} = \sqrt{2}(\cos(\pi/4)+\mathrm{i}\sin(\pi/4))$; the dots represent the stretching out by a factor of $\sqrt{2}$, and the circular arc represents the rotation by $\pi/4$.

Now let's check the correctness of our rule for multiplying complex numbers. Remember that the rule is: to multiply two complex numbers, we multiply the moduli and add the arguments. To see that this is correct, suppose that z_1 and

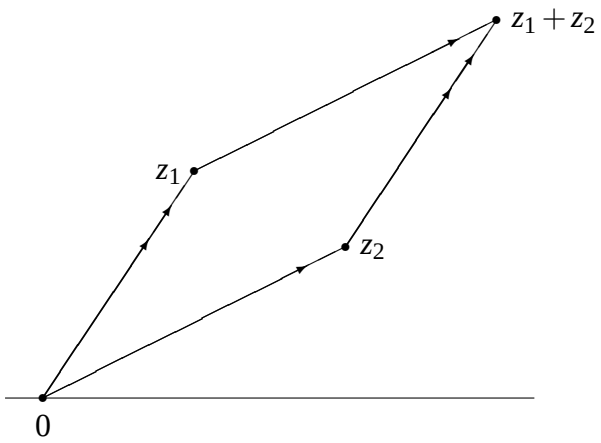


Figure 2.2: Addition of complex numbers

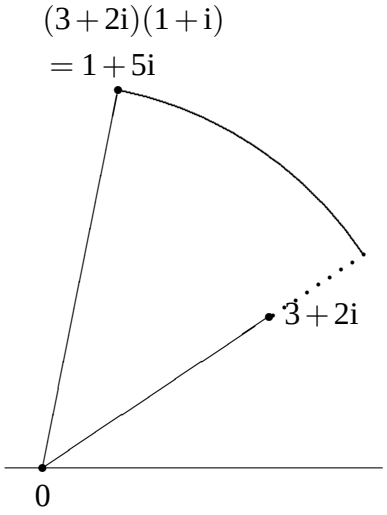


Figure 2.3: Multiplication of complex numbers

z_2 are two complex numbers; let their moduli be r_1 and r_2, and their arguments $\theta_1 + \theta_2$. Then

$$\begin{aligned} z_1 &= r_1(\cos\theta_1 + \mathrm{i}\sin\theta_1), \\ z_2 &= r_2(\cos\theta_2 + \mathrm{i}\sin\theta_2). \end{aligned}$$

Then

$$\begin{aligned} z_1 z_2 &= r_1 r_2(\cos\theta_1 + \mathrm{i}\sin\theta_1)(\cos\theta_2 + \mathrm{i}\sin\theta_2) \\ &= r_1 r_2((\cos\theta_1\cos\theta_2 - \sin\theta_1\sin\theta_2) + (\cos\theta_1\sin\theta_2 + \sin\theta_1\cos\theta_2)\mathrm{i}) \\ &= r_1 r_2(\cos(\theta_1+\theta_2) + \mathrm{i}\sin(\theta_1+\theta_2)), \end{aligned}$$

which is what we wanted to show.

From this we can prove *De Moivre's Theorem*:

Theorem 2.1 *For any natural number n, we have*

$$(\cos\theta + \mathrm{i}\sin\theta)^n = \cos n\theta + \mathrm{i}\sin n\theta.$$

Proof The proof is by induction. Starting the induction is easy since $(\cos\theta + \mathrm{i}\sin\theta)^0 = 1$ and $\cos 0 + \mathrm{i}\sin 0 = 1$.

For the inductive step, suppose that the result is true for n, that is,

$$(\cos\theta + \mathrm{i}\sin\theta)^n = \cos n\theta + \mathrm{i}\sin n\theta.$$

Then

$$\begin{aligned}(\cos\theta + \mathrm{i}\sin\theta)^{n+1} &= (\cos\theta + \mathrm{i}\sin\theta)^n \cdot (\cos\theta + \mathrm{i}\sin\theta)\\ &= (\cos n\theta + \mathrm{i}\sin n\theta)(\cos\theta + \mathrm{i}\sin\theta)\\ &= \cos(n+1)\theta + \mathrm{i}\sin(n+1)\theta,\end{aligned}$$

which is the result for $n+1$. So the proof by induction is complete.

Note that, in the second line of the chain of equations, we have used the inductive hypothesis, and in the third line, we have used the rule for multiplying complex numbers.

The argument is clear if we express it geometrically. To multiply by the complex number $(\cos\theta + \mathrm{i}\sin\theta)^n$, we rotate n times through an angle θ, which is the same as rotating through an angle $n\theta$.

De Moivre's Theorem is useful in deriving trigonometrical formulae. For example,

$$\begin{aligned}\cos 3\theta + \mathrm{i}\sin 3\theta &= (\cos\theta + \mathrm{i}\sin\theta)^3\\ &= (\cos^3\theta - 3\cos\theta\sin^2\theta) + (3\cos^2\theta\sin\theta - \sin^3\theta)\mathrm{i},\end{aligned}$$

so

$$\begin{aligned}\cos 3\theta &= \cos^3\theta - 3\cos\theta\sin^2\theta,\\ \sin 3\theta &= 3\cos^2\theta\sin\theta - \sin^3\theta.\end{aligned}$$

These can be converted into the more familiar forms $\cos 3\theta = 4\cos^3\theta - 3\cos\theta$ and $\sin 3\theta = 3\sin\theta - 4\sin^3\theta$ by using the equation $\cos^2\theta + \sin^2\theta = 1$.

Exercises

2.1 Prove by induction or otherwise that

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

2.2 Use De Moivre's Theorem to express $\cos 4x$ as a polynomial in $\cos x$, and to express $\sin 4x$ as a polynomial in $\sin x$.

2.3 Find

$$\sqrt{2+\sqrt{2+\sqrt{2+\cdots}}}.$$

2.4 The *quaternions* form a number system discovered by Hamilton. They have the form $a+b\mathrm{i}+c\mathrm{j}+d\mathrm{k}$, where $a,b,c,d\in\mathbb{R}$ and i, j, k are new symbols which satisfy

$$\mathrm{i}^2=\mathrm{j}^2=\mathrm{k}^2=\mathrm{ijk}=-1.$$

(a) Write down rules for the sum and product of two quaternions.

(b) Show that the associative law for multiplication holds for quaternions.

(c) Show that $(a+b\mathrm{i}+c\mathrm{j}+d\mathrm{k})(a-b\mathrm{i}-c\mathrm{j}-d\mathrm{k})=(a^2+b^2+c^2+d^2)$, and hence show that the quaternions satisfy the inverse law for multiplication (that is, every non-zero quaternion has a multiplicative inverse).

Chapter 3

Other algebraic systems

In this section, we will look at other algebraic systems which have operations which resemble addition and multiplication for number systems. These operations satisfy some of the laws which hold for numbers, but not necessarily all of them. A reminder: we are interested in the following laws:

Commutative laws: $a+b=b+a$, $ab=ba$

Associative laws: $(a+b)+c=a+(b+c)$, $(ab)c=a(bc)$

Distributive law: $a(b+c)=ab+ac$

Identity laws: $0+a=a$, $1a=a$

Inverse laws: For all a there exists $(-a)$ such that $a+(-a)=0$; for all $a\neq 0$, there exists a^{-1} such that $a\cdot a^{-1}=1$.

We have to be a bit careful about what the identity laws mean, since in other algebraic systems there will not be numbers 0 and 1 to use here. The identity law for multiplication should mean that there is a particular element e (say) in our system such that $ea=a$ for every element a. In the case of number systems, the number 1 has this property. Similarly we have to be careful about the interpretation of $-a$ and a^{-1} in the inverse laws. But notice that we don't even have to try to check the additive or multiplicative inverse laws unless the additive or multiplicative identity laws hold.

3.1 Vectors

In Geometry I, you learned how to add 3-dimensional vectors, and two different ways to multiply them: the scalar product or dot product, and the vector product

or cross product. Given two vectors $\mathbf{u},\mathbf{v}$, we denote their sum by $\mathbf{u}+\mathbf{v}$, their dot product by $\mathbf{u}\cdot\mathbf{v}$, and their cross product by $\mathbf{u}\times\mathbf{v}$.

(We can't do something like $\mathbf{u}$ in handwriting, or writing on the blackboard. So you should write the vector $\mathbf{u}$ as $\underline{u}$, as you did in the Geometry I course.)

Remember that we can represent a vector by a column consisting of three numbers; for example,

$$\mathbf{u} = 2\mathbf{i}+\mathbf{j}-5\mathbf{k} = \begin{pmatrix} 2 \\ -1 \\ 5 \end{pmatrix}.$$

Addition The commutative and associative laws hold for vector addition; so does the zero and inverse laws, if we take the vector $\mathbf{0} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$ to be the zero element:

$$\begin{aligned} \mathbf{u}+\mathbf{v} &= \mathbf{v}+\mathbf{u}, \\ (\mathbf{u}+\mathbf{v})+\mathbf{w} &= \mathbf{u}+(\mathbf{v}+\mathbf{w}), \\ \mathbf{0}+\mathbf{v} &= \mathbf{v}, \\ \mathbf{v}+(-\mathbf{v}) &= \mathbf{0}. \end{aligned}$$

These can all be proved by a calculation. For example, here is a proof of the associative law. Let

$$\mathbf{u} = \begin{pmatrix} a \\ b \\ c \end{pmatrix}, \qquad \mathbf{v} = \begin{pmatrix} p \\ q \\ r \end{pmatrix}, \qquad \mathbf{w} = \begin{pmatrix} x \\ y \\ z \end{pmatrix}.$$

Then

$$\begin{aligned} (\mathbf{u}+\mathbf{v})+\mathbf{w} = \begin{pmatrix} a+p \\ b+q \\ c+r \end{pmatrix} + \begin{pmatrix} x \\ y \\ z \end{pmatrix} &= \begin{pmatrix} (a+p)+x \\ (b+q)+y \\ (c+r)+z \end{pmatrix}, \\ \mathbf{u}+(\mathbf{v}+\mathbf{w}) = \begin{pmatrix} a \\ b \\ c \end{pmatrix} + \begin{pmatrix} p+x \\ q+y \\ r+z \end{pmatrix} &= \begin{pmatrix} a+(p+x) \\ b+(q+y) \\ c+(r+z) \end{pmatrix}, \end{aligned}$$

and $(a+p)+x = a+(p+x)$, etc., since the associative law holds for addition of real numbers. So the two expressions are equal.

Notice what we have done here: we used the associative law for the real numbers to prove it for 3-dimensional vectors.

Scalar product Asking about the associative law or other laws for the scalar product doesn't really make sense, since the scalar product of two vectors is a number, not a vector! So $(\mathbf{u}\cdot\mathbf{v})\cdot\mathbf{w}$ is meaningless.

The lesson is that the operations we will be studying must take two objects of some kind and combine them into another object of the same kind.

Vector product Remember the formula for the vector product:

$$\begin{pmatrix} a \\ b \\ c \end{pmatrix} \times \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{vmatrix} \mathbf{i} & a & x \\ \mathbf{j} & b & y \\ \mathbf{k} & c & z \end{vmatrix},$$

or, to put it another way,

$$(a\mathbf{i}+b\mathbf{j}+c\mathbf{k})\times(x\mathbf{i}+y\mathbf{j}+z\mathbf{k}) = (bz-cy)\mathbf{i}+(cx-az)\mathbf{j}+(ay-bx)\mathbf{k}.$$

(This was not the *definition*, but it was proved in Part 5 of the notes that this formula holds.)

What properties does it have? You also met these properties in the Geometry I course.

Associative law: This does not hold. Remember that $\mathbf{v}\times\mathbf{v}=\mathbf{0}$ for any vector v. Now

$$\begin{aligned} (\mathbf{i}\times\mathbf{i})\times\mathbf{j} &= \mathbf{0}\times\mathbf{j}=\mathbf{0}, \\ \mathbf{i}\times(\mathbf{i}\times\mathbf{j}) &= \mathbf{i}\times\mathbf{k}=-\mathbf{j}. \end{aligned}$$

(Remember that to *disprove* something like the associative law, a single counterexample is enough!)

Commutative law: This does not hold either. In fact, I hope you remember from Geometry I that

$$\mathbf{u}\times\mathbf{v} = -(\mathbf{v}\times\mathbf{u})$$

for any two vectors u and v. To get a specific counterexample, we could observe that

$$\mathbf{i}\times\mathbf{j}=\mathbf{k}, \qquad \mathbf{j}\times\mathbf{i}=-\mathbf{k}.$$

Distributive law: This one is true:

$$\mathbf{u}\times(\mathbf{v}+\mathbf{w}) = (\mathbf{u}\times\mathbf{v})+(\mathbf{u}\times\mathbf{w}).$$

How do you prove this?

Identity law: This one also fails. There cannot be a vector $\mathbf{e}$ with the property that $\mathbf{e} \times \mathbf{v} = \mathbf{v}$ for any choice of $\mathbf{v}$, because $\mathbf{e} \times \mathbf{v}$ is always perpendicular to $\mathbf{v}$!

The lesson here is that even nice operations might fail to satisfy the usual laws for numbers.

3.2 Matrices

Matrices form another class of objects which can be added and multiplied. We will consider just 2×2 matrices, as these illustrate the general principles. Recall the rules, that you learned in Geometry I. Let $A = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$ and $B = \begin{pmatrix} e & f \\ g & h \end{pmatrix}$ be two matrices. We will take the entries $a, \ldots, h$ to be arbitrary real numbers.

Addition The sum of two matrices A and B is the matrix obtained by adding corresponding elements of A and B:

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix} + \begin{pmatrix} e & f \\ g & h \end{pmatrix} = \begin{pmatrix} a+e & b+f \\ c+g & d+h \end{pmatrix}.$$

Multiplication The rule for multiplication is more complicated:

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix} \begin{pmatrix} e & f \\ g & h \end{pmatrix} = \begin{pmatrix} ae+bg & af+bh \\ ce+dg & cf+dh \end{pmatrix}.$$

It works like this. To work out the entry in the first row and second column of the product AB, we take the first row of A (which is $\begin{pmatrix} a & b \end{pmatrix}$), and the second column of B (which is $\begin{pmatrix} f \\ h \end{pmatrix}$; multiply corresponding elements (a by f, and b by h), and add the products, to get $af+bh$. The rule for the other entries in AB is similar.

Do these operations satisfy the laws we wrote down earlier?

Addition The commutative, associative, identity, and inverse laws all hold.

To verify that $A+B = B+A$, we have to show that corresponding entries of these matrices are equal. These entries are obtained by adding corresponding entries in A and B in either order; the results are equal. In detail,

$$\begin{aligned} \begin{pmatrix} a & b \\ c & d \end{pmatrix} + \begin{pmatrix} e & f \\ g & h \end{pmatrix} &= \begin{pmatrix} a+e & b+f \\ c+g & d+h \end{pmatrix}, \\ \begin{pmatrix} e & f \\ g & h \end{pmatrix} + \begin{pmatrix} a & b \\ c & d \end{pmatrix} &= \begin{pmatrix} e+a & f+b \\ g+c & h+d \end{pmatrix}, \end{aligned}$$

and the matrices on the right are equal because $a+e=e+a$ etc.

The associative law is true, and the argument to prove it is similar. If we define the *zero matrix* to be

$$O_{2\times 2}=\begin{pmatrix}0&0\\0&0\end{pmatrix},$$

then we have $O_{2\times 2}+A=A$ for any matrix A; for, to work out $O_{2\times 2}+A$, we add zero to each entry of A, which doesn't change it. Similarly, for any matrix A, we let $-A$ be the matrix whose entries are the negatives of the entries of A; then $A+(-A)=O$.

Multiplication Here we find our first surprise: The commutative law for multiplication fails! Remember that to disprove a general assertion, we only need one counterexample:

$$\begin{pmatrix}1&2\\3&4\end{pmatrix}\begin{pmatrix}5&6\\7&8\end{pmatrix}=\begin{pmatrix}19&22\\43&50\end{pmatrix}\neq\begin{pmatrix}23&34\\31&46\end{pmatrix}=\begin{pmatrix}5&6\\7&8\end{pmatrix}\begin{pmatrix}1&2\\3&4\end{pmatrix}.$$

[How did I find this example? Trial and error; I wrote down the first two matrices I could think of, multiplied them both ways round, and found that the results were different.]

Despite this, the associative law and the identity law do both hold for matrix multiplication. For the associative law, there is no alternative but to multiply it out and see:

$$\begin{aligned}\begin{pmatrix}a&b\\c&d\end{pmatrix}\left(\begin{pmatrix}e&f\\g&h\end{pmatrix}\begin{pmatrix}i&j\\k&l\end{pmatrix}\right)&=\begin{pmatrix}a&b\\c&d\end{pmatrix}\begin{pmatrix}ei+fk&ej+fl\\gi+hk&gj+hl\end{pmatrix}\\&=\begin{pmatrix}a(ei+fk)+b(gi+hk)&\dots\\\cdots&\dots\end{pmatrix},\\\left(\begin{pmatrix}a&b\\c&d\end{pmatrix}\begin{pmatrix}e&f\\g&h\end{pmatrix}\right)\begin{pmatrix}i&j\\k&l\end{pmatrix}&=\begin{pmatrix}ae+bg&af+bh\\ce+dg&cf+dh\end{pmatrix}\begin{pmatrix}i&j\\k&l\end{pmatrix}\\&=\begin{pmatrix}(ae+bg)i+(af+bh)k&\dots\\\cdots&\dots\end{pmatrix}.\end{aligned}$$

Algebraic manipulation shows that

$$a(ei+fk)+b(gi+hk)=(ae+bg)i+(ce+dg)k.$$

[Take a look at this manipulation. We first expand the brackets on the left, using the distributive law. This gives $a(ei)+a(fk)+b(gi)+b(hk)$. Now use the associative law for multiplication to switch this into $(ae)i+(af)k+(bg)i+(bh)k$. Then the commutative law for addition changes this to $(ae)i+(bg)i+(af)k+(bh)k$,

and the distributive law once more turns this into the right-hand expression. Oh, and I forgot to mention that I used the associative law for addition without telling you, when I wrote down the sum of four terms without telling you where the brackets go! So almost all the laws for real numbers get used.]

To prove the identity law for multiplication, we have to know what the identity matrix is. Since the zero matrix has every entry zero, you might guess that the identity matrix has every entry 1, but it doesn't:

$$\begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix} = \begin{pmatrix} 4 & 6 \\ 4 & 6 \end{pmatrix}.$$

In fact the identity matrix has ones on the main diagonal and zeros elsewhere:

$$I_2 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}.$$

We have $I_2 A = A$ for any 2×2 matrix A:

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

Now another possible problem might occur to you. Since multiplication is not commutative, is it true that $AI_2 = A$ for any A? Well, yes it is:

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

as you can check. [You may also notice that, as well as the identity law for multiplication, we use the fact that $0a = 0$ for any real number a and the zero law for addition.]

The inverse law for multiplication does not hold. For example, if $A = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}$, then there is no matrix B such that $AB = I_2$. You learned in Geometry I that the condition for a matrix to have an inverse is that its *determinant* is not zero.

Distributive law: I leave it to you to check that

$$A(B+C) = AB + AC$$

for any matrices A, B, C. You might even want to check which laws for real numbers are used in the proof. Because multiplication is not commutative, we can also check the other way round:

$$(B+C)A = BA + CA$$

for any matrices A, B, C.

We use the notation $M_{2\times 2}(\mathbb{R})$ for the set of all 2×2 matrices with real numbers as entries. (We call these matrices "real matrices" for short.) As you can see, we can easily generalise this notation. By changing the subscript, we can talk about the set of matrices of different size, say 3×3; and by putting $\mathbb{Z}$, $\mathbb{Q}$ or $\mathbb{C}$ in place of $\mathbb{R}$, we can talk about matrices whose entries are integers, rational numbers, or complex numbers.

3.3 Polynomials

You can think of a polynomial as a function which can be written as a sum of terms, each of which is a power of x multiplied by a constant. So "the polynomial x^2" should really be "the polynomial $1x^2$". We write x^1 as x, and leave out x^0 altogether (just writing the constant). If the coefficient of a power of x is zero, we usually don't bother writing it: so we write $2x^2+3$ rather than $2x^2+0x+3$. Of course, if all the terms are zero, we have to write something; so we just write 0.

So a typical polynomial has the form

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

Note that a constant a_0 is a special kind of polynomial called a *constant polynomial*.

The *degree* of a polynomial is the largest number n such that the polynomial contains a term $a_n x^n$ with $a_n \neq 0$. Thus, a non-zero constant polynomial has degree 0, since it has the form $a_0 x^0$. The zero polynomial 0 doesn't have a degree, since it doesn't have any non-zero terms! [Be warned: some people say that it has degree -1; others say that it has degree $-\infty$. Of course, these are merely conventions.]

Addition and multiplication You already know how to add and multiply polynomials. But it is difficult to give a proper mathematical definition. For example,

$$\left(2x^2+3\right)+\left(x^3+x-5\right) = x^3+2x^2+x-2.$$

We can't just say "add corresponding terms", since some terms may be missing; we have first to put the missing terms in with coefficients 0. For multiplication, we multiply each term of the first factor by each term of the second, and then gather up terms involving the same power of x:

$$\begin{aligned}\left(2x^2+3\right)\left(x^3+x-5\right) &= 2x^5+(2x^3+3x^3)-10x^2+3x-15\\ &= 2x^5+5x^3-10x^2+3x-15.\end{aligned}$$

I ask you to take on trust for now that it is possible to give good definitions of addition and multiplication of polynomials, and to show that they do satisfy the commutative, associative and identity laws for both addition and multiplication, the inverse law for addition, and the distributive law.

We use the notation $\mathbb{R}[x]$ for the set of all polynomials with real numbers as coefficients. (We call them "real polynomials" for short.) As you can see, this notation can be generalised: $\mathbb{Q}[x]$ and $\mathbb{C}[x]$ denote the sets of polynomials with rational or complex numbers as coefficients. These sets satisfy the same rules for addition and multiplication as the real polynomials.

3.4 Sets

Here is another example where we have an operation or rule of combination for objects which are nothing like numbers.

Let $\mathscr{S}$ be a set. We regard it as a "universal set"; in Probability I, it was called the *sample space*. Our objects will be subsets of $\mathscr{S}$.

Two operations which can be performed on sets are union and intersection, defined as follows:

Union: the *union* of two sets A and B is the set of all elements lying in either A or B:

$$A \cup B = \{x : x \in A \text{ or } x \in B\}.$$

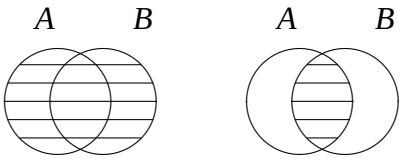
We read $A \cup B$ as "A union B", or "A or B".

Intersection: the *intersection* of two sets A and B is the set of all elements lying in both A and B:

$$A \cap B = \{x : x \in A \text{ and } x \in B\}.$$

We read $A \cap B$ as "A intersection B", or "A and B".

We can represent sets by *Venn diagrams*, and show these two operations in a diagram as follows:



$A \cup B$ $A \cap B$

Here are some laws they satisfy.

Commutative laws	$A \cup B = B \cup A$	$A \cap B = B \cap A$
Associative laws	$A \cup (B \cup C) = (A \cup B) \cup C$	$A \cap (B \cap C) = (A \cap B) \cap C$
Identity laws	$A \cup \emptyset = A$	$A \cap \mathscr{S} = A$

When we come to the distributive law, there is a small surprise. To write down the distributive law for numbers, we have to distinguish between addition and multiplication. It is true that

$$a \times (b+c) = (a \times b) + (a \times c),$$

but it is *not* true that

$$a + (b \times c) = (a+b) \times (a+c).$$

For sets, which of our two operations should play the role of addition, and which should be multiplication?

It turns out that it works both ways round. We can replace "plus" by "or" and "times" by "and", or *vice versa*:

Distributive laws $A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$ $A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$

All of these assertions have similar proofs: draw a Venn diagram to convince yourself, and then give a mathematical argument. Here is the proof of the first distributive law. I leave the Venn diagram to you.

$$\begin{aligned} x \in A \cap (B \cup C) &\Leftrightarrow x \in A \text{ and } x \in B \cup C \\ &\Leftrightarrow (x \in A \text{ and } x \in B) \text{ or } (x \in A \text{ and } x \in C) \\ &\Leftrightarrow x \in (A \cap B) \cup (A \cap C). \end{aligned}$$

So the two sets $A \cap (B \cup C)$ and $(A \cap B) \cup (A \cap C)$ have the same members, and hence are equal.

The inverse laws are not true. For example, we saw that the zero element for the operation of union is the empty set $\emptyset$; and, given a set A which is not the empty set, it is impossible to find a set B such that $A \cup B = \emptyset$, since $A \cup B$ is at least as large as A. The failure of the inverse law for intersection is similar.

In Probability I, you saw several other operations on sets: *difference*, *symmetric difference*, and *complement*. You might like to check which of our laws are satisfied by difference, or by symmetric difference, for example.

Exercises

3.1 (a) Find $\begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix} \begin{pmatrix} 1 & -1 \\ 3 & -5 \end{pmatrix}$.

(b) Find the inverse of $\begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix}$.

3.2 Find two matrices having entries 0 and 1 only which do not commute with each other.

3.3 Show that the symmetric difference of sets satisfies the associative, commutative, identity and inverse laws, where the identity element is the empty set $\emptyset$ and the inverse of any set A is equal to A.

3.4 Recall the definition of the quaternions from the last chapter.

(a) Show that any quaternion can be formally written as $a+\mathbf{v}$, where $a \in \mathbb{R}$ and $\mathbf{v}$ is a 3-dimensional real vector.

(b) Show that

$$\begin{aligned}(a+\mathbf{v})+(b+\mathbf{w}) &= (a+b)+(\mathbf{v}+\mathbf{w}),\\ (a+\mathbf{v})(b+\mathbf{w}) &= (ab-\mathbf{v}\cdot\mathbf{w})+(a\mathbf{w}+b\mathbf{v}+\mathbf{v}\times\mathbf{w}),\end{aligned}$$

where $\cdot$ and $\times$ denote the dot and cross product of vectors.

Chapter 4

Relations and functions

4.1 Ordered pairs and Cartesian product

We write $\{x,y\}$ to mean a set containing just the two elements x and y. More generally, $\{x_1,x_2,\ldots,x_n\}$ is a set containing just the n elements x_1, x_2, ..., 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, the order of the elements does matter, and we need a different construction. We write the *ordered pair* with first element x and second element y as (x,y); this is not the same as (y,x) unless x and y are equal. You have seen this notation used for the coordinates of points in the plane. The point with coordinates $(2,3)$ is not the same as the point with coordinates $(3,2)$. The rule for equality of ordered pairs is:

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

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

The idea of coordinatising the plane or three-dimensional space by ordered pairs or triples of real numbers was invented by Descartes. In his honour, we call the system "Cartesian coordinates". This great idea of Descartes allows us to use algebraic methods to solve geometric problems, as you saw in the Geometry I course last 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.

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

$$\{x : P\}$$

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

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

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

Proposition 4.1 *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.

(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:

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

4.2 Relations

Suppose we are given a set of people $P_1, \ldots, P_n$. What does the relation of being sisters mean? For each ordered pair (P_i, P_j), either P_i and P_j are sisters, or they are not; so we can think of the relation as being a rule of some kind which answers "true" or "false" for each pair (P_i, P_j). Mathematically, there is a more abstract way of saying the same thing; the relation of sisterhood is the *set* of all ordered pairs (P_i, P_j) for which the relation is true. (When I say that P_i and P_j are sisters, I mean that each of them is the sister of the other.)

So we define a *relation* R on a set X to be a subset of the Cartesian product $X^2 = X \times X$; that is, a set of ordered pairs. We think of the relation as holding between x and y if the pair (x,y) is in R, and not holding otherwise.

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

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

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 4.1. 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

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

So there are 65536 relations. Of course, not all of them have simple names like "less than".

You will see that a relation like "less than" is written $x < y$; in other words, we put the symbol for the relation between the names of the two elements making up the ordered pair. We could, if we wanted, invent a similar notation for any relation. Thus, if R is a relation, we could write $x\,R\,y$ to mean $(x,y) \in R$.

4.3 Equivalence relations and partitions

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

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

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

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

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

For example, the relation "less than" is not reflexive (since no element is less than itself); is not symmetric (since $x < y$ and $y < x$ cannot both hold); but is transitive (since $x < y$ and $y < z$ do imply that $x < z$). The relation of being sisters is not reflexive (it is debatable whether a girl can be her own sister, but a boy certainly cannot!), but it is symmetric. It is "almost" transitive: if x and y are sisters, and y and z are sisters, then x and z are sisters except in the case when $x = z$. But this case can actually occur, so the relation is not transitive. (For it to be transitive, the transitive law would have to hold without any exceptions.)

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

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

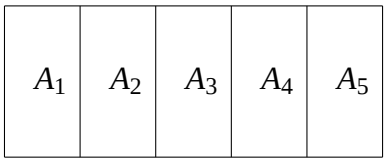
Let X be a set. A *partition* of X is a collection $\{A_1, A_2, \ldots\}$ of subsets of X having the following properties:

(a) $A_i \neq \emptyset$;

(b) $A_i \cap A_j = \emptyset$ for $i \neq j$;

(c) $A_1 \cup A_2 \cup \cdots = X$.

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 arises because the set X is divided into disjoint parts $A_1, A_2, \ldots$.



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. We write $R(x)$ for the set of elements of X which are related to R; that is,

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

Theorem 4.2 *(a) Let R be an equivalence relation on X. Then the sets $R(x)$, for $x \in X$, form a partition of X.*

(b) Conversely, given any partition $\{A_1, A_2, \ldots\}$ of X, there is an equivalence relation R on X such that the sets A_i are the same as the sets $R(x)$ for $x \in X$.

Proof (a) We have to show that the sets $R(x)$ 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 $x \in R(x)$; thus $R(x) \neq \emptyset$.

- We have to show that, if $R(x) \neq R(y)$, then $R(x) \cap R(y) = \emptyset$. The contrapositive of this is: if $R(x) \cap R(y) \neq \emptyset$, then $R(x) = R(y)$; we prove this. Suppose that $R(x) \cap R(y) \neq \emptyset$; this means that there is some element, say z, lying in both $R(x)$ and $R(y)$. 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 $R(x) = R(y)$; this means showing that every element in $R(x)$ is in $R(y)$, and every element of $R(y)$ is in $R(x)$. For the first claim, take $u \in R(x)$. Then $(x,u) \in R$. Also $(y,x) \in R$ (by symmetry; we know that $(x,y) \in R$; so $(y,u) \in R$ by transitivity, and $u \in R(y)$. Conversely, if $u \in R(y)$, a similar argument (which you should try for yourself) shows that $u \in R(x)$. So $R(x) = R(y)$, as required.

- Finally we have to show that the union of all the sets $R(x)$ 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 $R(x)$.

(b) Suppose that $\{A_1, A_2, \ldots\}$ is a partition of x. We define a relation R as follows:

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

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_i of the partition, and y and z lie in the same part A_j. Then $y \in A_i$ and $y \in A_j$; and so we have $A_i = A_j$ (since different parts are disjoint). Thus x and z both lie in A_i. So R is transitive.

Thus R is an equivalence relation. But clearly $R(x)$ consists of all elements lying in the same part of the partition as x; so, if $x \in A_i$, then $R(x) = A_i$. So the partition consists of the sets $R(x)$.

If R is an equivalence relation, then the sets $R(x)$ (the parts of the partition corresponding to R) are called the *equivalence classes* of R.

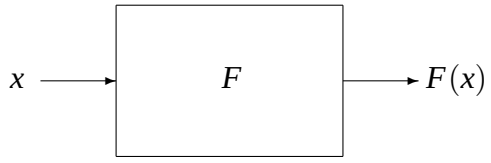
Here is an example. There are five partitions of the set $\{1,2,3\}$. One has a single part; three of them have one part of size 1 and one of size 2; and one has three parts of size 1. Here are the partitions and the corresponding equivalence relations.

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.

4.4 Functions

What is a function? This is a question that has given mathematicians a lot of trouble over the ages. People used to think that a function had to be given by a formula, such as x^2 or $\sin x$. We don't require this any longer. All that is important is that you put in a value for the argument of the function, and out comes a value. Think of a function as a kind of black box:



The name of the function is F; we put x into the black box and $F(x)$ comes out. Be careful not to confuse F, the name written on the black box, with $F(x)$, which is what comes out when x is put in. Sometimes the language makes it hard to keep this straight. For example, there is a function which, when you put in x, outputs x^2. We tend to call this "the function x^2", but it is really "the squaring

function", or "the function $x \mapsto x^2$". (You see that we have a special symbol $\mapsto$ to denote what the black box does.)

Black boxes are not really mathematical notation, so we re-formulate this definition in more mathematical terms. We have to define what we mean by a function F. Now there will be a set X of allowable inputs to the black box; X is called the *domain* of F. Similarly, there will be a set Y which contains all the possible outputs; this is called the *codomain* of F. (We don't necessarily require that every value of Y can come out of the black box. For the squaring function, the domain and the codomain are both equal to $\mathbb{R}$, even though none of the outputs can be negative.)

The important thing is that every input $x \in X$ produces exactly one output $y = F(x) \in Y$. The ordered pair (x,y) is a convenient way of saying that the input x produces the output y. Then we can take all the possible ordered pairs as a description of the function. Thus we come to the formal definition:

> Let X and Y be sets. Then a *function* from X to Y is a subset F of $X \times Y$ having the property that, for every $x \in X$, there is exactly one element $y \in Y$ such that $(x,y) \in F$. We write this unique y as $F(x)$. We write $F : X \to Y$ (read "F from X to Y") to mean that F is a function with domain X and codomain Y.

The set of all elements $F(x)$, as x runs through X, is called the *range* of the function F. It is a subset of the codomain, but (as we remarked) it need not be the whole codomain.

Here is an example. Let $X = Y = \{1,2,3,4,5\}$, and let

$$F = \{(1,1),(2,4),(3,5),(4,4),(5,1)\}.$$

Then F is a function from X to Y, with $F(1) = 1$, $F(2) = 4$, and so on. (In this particular case, it happens that F is given by a fairly simple formula: $F(x) = 6x - x^2 - 4$.)

A function $F : X \to Y$ is called

> *injective*, or *one-to-one*, if different elements of X have different images under F: $x_1 \neq x_2$ implies $F(x_1) \neq F(x_2)$ (or equivalently, $F(x_1) = F(x_2)$ implies $x_1 = x_2$).
>
> *surjective*, or *onto*, if its range is equal to Y: that is, for every $y \in Y$, there is some $x \in X$ such that $F(x) = y$.
>
> *bijective*, or *a one-to-one correspondence*, if it is both injective and surjective.

A bijective function from X to Y matches up the two sets: for each $x \in X$ there is a unique $y = F(x) \in Y$; and for each $y \in Y$ there is a unique $x \in X$ such that $F(x) = y$. This can only happen if X and Y have the same number of elements.

If F is a bijective function from X to Y, then there is an *inverse function* G from Y to X which takes every element $y \in Y$ to the unique $x \in X$ for which $F(x) = y$. In other words, the black box for G is the black box for F in reverse:

$$x = G(y) \text{ if and only if } y = F(x).$$

The inverse function G is also bijective. Thus a bijective function F and its inverse G satisfy

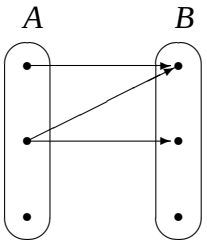
- $G(F(x)) = x$ for all $x \in X$;
- $F(G(y)) = y$ for all $y \in Y$.

Notice that F is the inverse function of G.

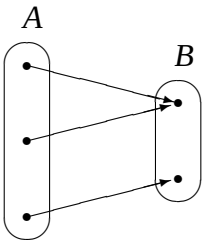
Sometimes we represent a function $F : A \to B$ by a picture, where we show the two sets A and B, and draw an arrow from each element a of A to the element $b = F(a)$ of B. For such a picture to show a function, each element of A must have exactly one arrow leaving it. Now

- F is one-to-one (injective) if no point of B has two or more arrows entering it;
- F is onto (surjective) if every point of B has at least one arrow entering it;
- F is one-to-one and onto (bijective) if every point of B has exactly one arrow entering it; in this case, the arrows match up the points of A with the points of B.

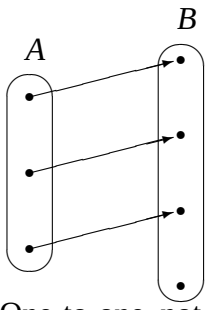
Here are some illustrations. The first is not a function because some elements of A have more than one arrow leaving them while some have none.



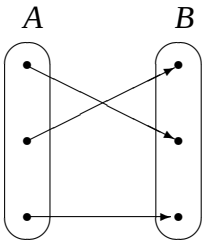
Not a function



Onto, not one-to-one



One-to-one, not onto



One-to-one and onto

4.5 Operations

An *operation* is a special kind of function: its domain is $X \times X$ and its codomain is X, where X is a set. In other words, the input to the black box for F 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 function as "combining" the two inputs in some way.

There is a different notation often used for operations. Rather than write the function as F, so that $z = F(x,y)$ is the output when x and y are input, instead we choose a symbol like $+$, $\times$, $*$, $\div$, $\circ$ or $\bullet$, and place it between the two inputs: that is, we write $x+y$, or $x \times y$, or $\ldots$, instead of $F(x,y)$. This is called *infix notation*.

Many of the operations we have already met (addition, subtraction, multiplication for numbers; addition and vector product for vectors; addition and multiplication for matrices or polynomials; union and intersection for sets) are binary operations.

An operation on a finite set can be represented by an operation table. This is a square table with elements of the set X labelling the rows and columns of the table. To calculate $x \circ y$ (if $\circ$ is the operation), we look in the row labelled x and column labelled y; the element in the table in this position is $x \circ y$. Here is an example:

$\circ$	a	b	c
a	a	b	c
b	b	b	c
c	c	c	c

Given an operation, we can ask whether it satisfies the laws of algebra that we have met several times already. Consider the above example.

Commutative? Yes, since the table is symmetric about the main diagonal, so $x \circ y$ is always the same as $y \circ x$.

Associative? Yes, though this is harder to show. You are invited either to prove it

by considering all cases of the associative law, or to find a nicer proof using a description of what the operation actually does.

Identity? Yes, a is an identity, since

$$a \circ a = a, \qquad a \circ b = b, \qquad a \circ c = c.$$

Inverse? No, there is no element x such that $c \circ x = a$, since $c \circ x$ is always equal to c, whatever x is.

4.6 Appendix: Relations and functions

In this section we will see that, given an arbitrary function, we can turn it into a bijective function. If $F : X \to Y$ is not onto, we can throw away the points of the codomain Y which are not in the range of F. Making it one-to-one is more difficult. The theorem below shows how to do it.

Theorem 4.3 *Let $F : X \to Y$ be a function.*

(a) The range of F, the set $\{y \in Y : y = F(x)$ for some $x \in X$, is a subset B of Y.

(b) Define a relation R on X by the rule that $(x_1, x_2) \in R$ if and only if $F(x_1) = F(x_2)$. Then R is an equivalence relation on X.

(c) Let A be the set of equivalence classes of R. Then the function $\tilde{F} : A \to B$ defined by $\tilde{F}(R(x)) = F(x)$ for all $x \in X$, is a bijective function from A to B.

Proof Part (a) is clear. Part (b) is quite easy: R is

reflexive because $F(x) = F(x)$ for all $x \in X$;

symmetric because $F(x_1) = F(x_2)$ implies $F(x_2) = F(x_1)$;

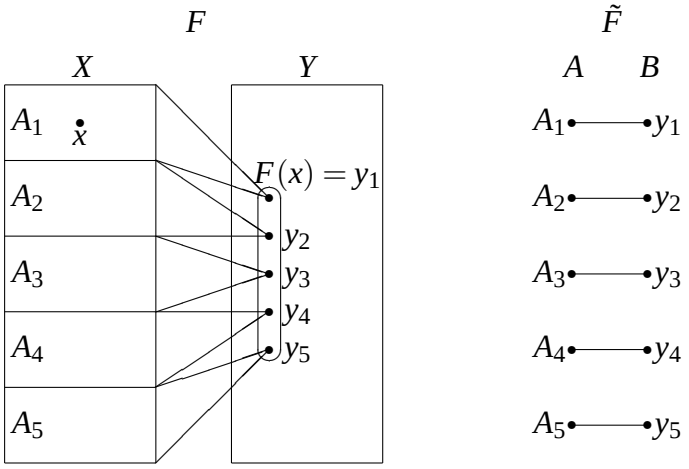
transitive because $F(x_1) = F(x_2)$ and $F(x_2) = F(x_3)$ implies $F(x_1) = F(x_3)$.

Look at part (c). There is one important thing we have to do before we even have a function $\tilde{F}$: to show that it is well-defined. How could this go wrong? If x_1 and x_2 are equivalent (that is, if $(x_1, x_2) \in R$, then $R(x_1) = R(x_2)$. What guarantee do we have that $\tilde{F}(R(x_1)) = \tilde{F}(R(x_2))$, as we need? This means that $F(x_1) = F(x_2)$; but that is exactly the condition that ensures $(x_1, x_2) \in R$. So $\tilde{F}$ is a well-defined function.

Is it one-to-one? Suppose that $\tilde{F}(R(x_1)) = \tilde{F}(R(x_2))$. Then by definition, $F(x_1) = F(x_2)$; so $(x_1, x_2) \in R$, so $R(x_1) = R(x_2)$.

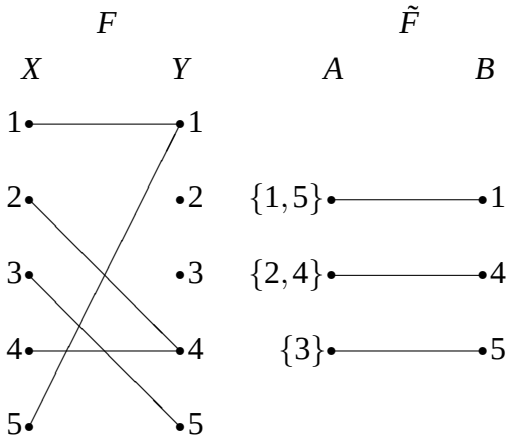
Is it onto? Take any $y \in B$. Since B is the range of F, there exists some $x \in X$ with $F(x) = y$. Then $\tilde{F}(R(x)) = y$, so y is in the range of $\tilde{F}$.

If that seems complicated, here is a picture.



The five slabs on the left are the equivalence classes of the relation R; each point in the top slab is mapped by F to the same point $F(x)$ on the right. The five points in the oval on the right make up the range of F. It is clear that equivalence classes on the left are matched up with points of the range on the right by a bijective function.

In our earlier example, the equivalence classes of the relation R are $\{1,5\}$, $\{2,4\}$ and $\{3\}$; the range of F is $\{1,4,5\}$; and the one-to-one correspondence $\tilde{F}$ maps $\{1,5\}$ to 1, $\{2,4\}$ to 4, and $\{3\}$ to 5.



Exercises

4.1 Which of the following relations R on sets X are (i) reflexive, (ii) symmetric, and (iii) transitive? For any relation which is an equivalence relation (that is, satisfies all three conditions), describe its equivalence classes.

(a) X is the set of positive integers, $R = \{(x,y) : x+y = 100\}$.

(b) X is the set of integers, $R = \{(x,y) : x = y\}$.

(c) X is the set of railway stations in Great Britain, R is the set of pairs (x,y) of stations for which there is a scheduled direct train from x to y.

4.2 For each of the following functions F, describe the image of F, and state whether F is (i) one-to-one and (ii) onto:

(a) $F : \{0,1,2,3,4,5\} \to \{0,1,2,3,4,5\}$, $F(x) = \lfloor x/2 \rfloor$ (the greatest integer not exceeding $x/2$).

(b) $F : \mathbb{R} \to \mathbb{R}$, $F(x) = e^x$.

(c) $F : \mathbb{R} \to \mathbb{R}$, $F(x) = x^3 + x$.

4.3 How many operations are there on the set $\{a,b\}$ with two elements? How many of them satisfy (i) the associative law, (ii) the identity law?

4.4 The *Fundamental Theorem of Algebra* says that a polynomial of degree n over the complex numbers has n complex roots.

Define a "function" $F : \mathbb{C}^2 \to \mathbb{C}^2$ by the rule that $F(a,b) = (c,d)$ if c and d are the roots of the quadratic equation $x^2 + ax + b = 0$. (So, for example, $F(-3,2) = (1,2)$.)

Show that F is not in fact a function.

Can you suggest a way to fix the definition?

Chapter 5

Division and Euclid's algorithm

5.1 The division rule

The *division rule* is the following property of natural numbers:

Proposition 5.1 *Let a and b be natural numbers, and assume that $b > 0$. Then there exist natural numbers q and r such that*

(a) $a = bq + r$;

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

Moreover, q and r are unique.

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

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

It remains to show that q and r exist. Consider the multiples of b: $0, b, 2b, \ldots$. 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.

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.

You learned in primary school an algorithm for long division which 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 $0 \leq r \leq b-1$.

5.2 Greatest common divisor and least common multiple

We write $a \mid b$ to mean that a divides b, or b is a multiple of a. *Warning:* Don't confuse $a \mid b$ with a/b, which means a divided by b; this is the opposite way round! So $a \mid b$ is a relation on the natural numbers which holds if $b = ac$ for some natural number c.

Every natural number, 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 natural number except zero.

Let a and b be natural numbers. A *common divisor* of a and b is a number 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 $\gcd(12,18) = 6$. We write gcd as shorthand for "greatest common divisor".

The remarks above about zero show that $\gcd(a,0) = a$ holds for any non-zero number a. What about $\gcd(0,0)$? Since every natural number divides zero, there is no greatest one.

The number 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 natural numbers a and b have a least common multiple. For there certainly exist common multiples, for example ab; and any

non-empty set of natural numbers has a least element. (The least common multiple of 0 and a is 0, for any a.) We write lcm as shorthand for "least common multiple".

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

But this is *not* an efficient way to find the gcd of two numbers. Factorising a number into its prime factors is notoriously difficult. In fact, it is the difficulty of this problem which keeps internet commercial transactions secure!

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

5.3 Euclid's algorithm

Euclid's algorithm is based on two simple rules:

Proposition 5.2

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

Proof We saw already that $\gcd(a,0) = a$, so suppose that $b > 0$. Let $r = a \operatorname{div} b = a - bq$, 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 is so slick that it doesn't tell us much. But looking more closely we see that it gives us an algorithm for calculating the gcd of a and b. If $b = 0$, the answer is a. If $b > 0$, calculate $a \bmod b = b_1$; our task is reduced to finding $\gcd(b, b_1)$, and $b_1 < b$. Now repeat the procedure; of $b_1 = 0$, the answer is b; otherwise calculate $b_2 = b \bmod b_1$, and our task is reduced to finding $\gcd(b_1, b_2)$, and $b_2 < b_1$. At each step, the second number of the pair whose gcd we have to find gets smaller; so the process cannot continue for ever, and must stop at some point. It stops when we are finding $\gcd(b_{n-1}, b_n)$, with $b_n = 0$; the answer is b_{n-1}.

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

To find $\gcd(a,b)$

Put $b_0 = a$ and $b_1 = b$.

As long as the last number b_n found is non-zero, put $b_{n+1} = b_{n-1} \bmod$

b_n.
When the last number b_n is zero, then the gcd is b_{n-1}.

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

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

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

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

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

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

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

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

5.4 Euclid's algorithm extended

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

Theorem 5.3 *Let a and b be natural numbers, 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 $6 = 2 \cdot 198 - 5 \cdot 78$.

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

Proposition 5.4 *The greatest common divisor of the natural numbers a and b is the natural number d with the properties*

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

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

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

Remark Recall that, with our earlier definition, we had to admit that $\gcd(0,0)$ doesn't exist, since every natural number divides 0 and there is no greatest one. But, with $a = b = 0$, there is a unique natural number satisfying the conclusion of Proposition 5.4, namely $d = 0$. So in fact this Proposition gives us a better way to define the greatest common divisor, which works for all pairs of natural numbers without exception!

5.5 Polynomials

Now we leave integers for a while, and turn to the set $\mathbb{R}[x]$ of all polynomials with real coefficients.

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

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

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 $7x + 2$.

In general, let $f(x)$ and $g(x)$ be two polynomials, with $g(x) \neq 0$. Then the division rule produces a quotient $q(x)$ and a remainder $r(x)$ such that

- $f(x) = g(x)q(x) + r(x)$;
- either $r(x) = 0$ or the degree of $r(x)$ is smaller than the degree of $g(x)$.

(Remember that we didn't define the degree of the zero polynomial.)

Let us prove that the division rule works. The proof follows the method that we used in the example: we multiply $g(x)$ by a constant times a power of x so that, when we subtract it, the degree of the result is smaller than it was. Our proof will be by induction on the degree of $f(x)$.

So let $f(x)$ and $g(x)$ be polynomials, with $g(x) \neq 0$.

Case 1: Either $f(x) = 0$, or $\deg(f(x)) < \deg(g(x))$. In this case we have nothing to do except to put $q(x) = 0$ and $r(x) = f(x)$.

Case 2: $\deg(f(x)) \geq \deg(g(x))$. We let $\deg(f(x)) = n$, and assume (as induction hypothesis) that the result is true if $f(x)$ is replaced by a polynomial of degree less than n. Let

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

where we have used the abbreviation l.d.t. for "lower degree terms". We have $a_n \neq 0$, $b_m \neq 0$, and (by the case assumption) $n \geq m$. Then

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

and so the polynomial $f^*(x) = f(x) - (a_n/b_m)x^{n-m} \cdot g(x)$ satisfies $\deg(f^*(x)) < \deg(f(x))$: the subtraction cancels the leading term of $f(x)$. So by the induction hypothesis, we have

$$f^*(x) = g(x)q^*(x) + r^*(x),$$

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

$$f(x) = g(x)\left((a_n/b_m)x^{n-m} + g^*(x)\right) + r^*(x),$$

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

Having got a division rule for polynomials, we can now copy everything that we did for integers. Here is a summary of the definitions and results.

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

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

We say that $g(x)$ *divides* $f(x)$ if $f(x) = g(x)q(x)$ for some polynomial $q(x)$; in other words, 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 $f(x)$ and $g(x)$ is a polynomial $d(x)$ with the properties

(a) $d(x)$ divides $f(x)$ and $d(x)$ divides $g(x)$;

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

(c) $d(x)$ is monic (if it is not the zero polynomial).

The last condition is put in because, for any non-zero real number c, each of the polynomials $f(x)$ and $cf(x)$ 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.

Theorem 5.5 *(a) Any two polynomials $f(x)$ and $g(x)$ have a greatest common divisor.*

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

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

$$f(x)h(x) + g(x)k(x) = d(x);$$

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 is the same as that for integers.

Exercises

5.1 Find the greatest common divisor of 2047 and 2323, and write it in the form $2047x + 2323y$ for some integers x and y.

5.2 Find the least common multiple of 2047 and 2323.

5.3 Find the greatest common divisor of $x^4 - 1$ and $x^3 + 3x^2 + x + 3$, and write it in the form $(x^4 - 1)u(x) + (x^3 + 3x^2 + x + 3)v(x)$ for some polynomials $u(x)$ and $v(x)$.

5.4 Prove that, for any two positive integers m and n,

$$\gcd(m,n) \cdot \mathrm{lcm}(m,n) = mn.$$

Does any similar result hold for three positive integers?

Chapter 6

Modular arithmetic

Modular arithmetic is an important example of an algebraic system with only a finite number of elements, unlike most of our examples (the number systems, matrices, polynomials, etc.) which have infinitely many elements.

6.1 Congruence mod m

Here is a very important example of an equivalence relation.

Let $X = \mathbb{Z}$, the set of integers. We define a relation $\equiv_m$ on $\mathbb{Z}$, called *congruence mod m*, where m is a positive integer, as follows:

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

You will often see this relation written as $x \equiv y \pmod{m}$. The meaning is exactly the same.

We check the conditions for 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 what we called $R(x)$ in the Equivalence Relation Theorem, where R is the name of the relation; so we should really call it $\equiv_m(x)$. But this looks a bit odd, so we say $[x]_m$ 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 6.1 *The equivalence relation $\equiv_m$ has exactly m equivalence classes, namely $[0]_m, [1]_m, [2]_m, \ldots, [m-1]_m$.*

Proof Given any integer n, we can divide it by m to get a quotient q and remainder r, so that $n = mq + r$ and $0 \leq r \leq m-1$. Then $n - r = mq$, so $r \equiv_m n$, and $n \in [r]_m$. So every integer lies in one of the classes in the proposition. These classes are all different, since if i, j both lie in the range $0, \ldots, m-1$, then $j - i$ cannot be a multiple of m unless $i = j$.

We use this in everyday life. Consider time on the 24-hour clock, for example. What is the time 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, 10:00, or 10am in the morning.

6.2 Operations on congruence classes

Now we can add and multiply congruence classes as follows:

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

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

The second important thing is that we have to do some work to show that we have defined anything at all. Suppose that $[a]_m = [a']_m$ and $[b]_m = [b']_m$. What guarantee have we that $[a+a']_m = [b+b']_m$? If this is not true, then our definition is worthless; so let's try to prove it. We have

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

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

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

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

For example, here are "addition table" and "multiplication table" for the integers mod 4. I have been lazy and written $0,1,2,3$ instead of the correct forms $[0]_4,[1]_4,[2]_4,[3]_4$.

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

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

We denote the set of congruence classes mod m, with these operations of addition and multiplication, by $\mathbb{Z}_m$. Note that $\mathbb{Z}_m$ is a set with m elements. We call the operations "addition and multiplication mod m".

Theorem 6.2 *The set $\mathbb{Z}_m$, with addition and multiplication mod m, satisfies the commutative, associative, and identity laws for both addition and multiplication, the inverse law for addition, and the distributive law.*

Proof We won't prove the whole thing; here is a proof of the distributive law. We are trying to prove that

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

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

6.3 Inverses

What about multiplicative inverses? Not every element in $\mathbb{Z}_m$ has an 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 $[2]_4[b]_4 = [1]_4$. On the other hand, in $\mathbb{Z}_5$, every non-zero element has an inverse, since

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

Theorem 6.3 *The element $[a]_m$ of $\mathbb{Z}_m$ has an 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 $\gcd(a,m) = 1$.

Suppose that $\gcd(a,m) = 1$. As we saw in the last chapter, Euclid's algorithm shows that there exist integers x and y such that $xa + ym = 1$. This says that $xa - 1 = ym$ is a multiple of m, so that $xa \equiv_m 1$. This means $[x]_m[a]_m = [1]_m$, so $[x]_m$ is the inverse of $[a]_m$.

Now suppose that $[x]_m$ is the inverse of $[a]_m$, so that $xa \equiv_m 1$. This means that $xa + ym = 1$ for some integer y. Now let $d = \gcd(a,m)$. Then $d \mid xa$ and $d \mid ym$, so $d \mid xa + ym = 1$; so we must have $d = 1$, as required.

Corollary 6.4 *Suppose that p is a prime number. Then the multiplicative inverse law holds in $\mathbb{Z}_p$; that is, every non-zero element of $\mathbb{Z}_p$ has an inverse.*

Proof If p is prime, then every number a with $1 \leq a \leq p$ 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 the Theorem implies that $[a]_p$ has an inverse in $\mathbb{Z}_p$.

6.4 Fermat's Little Theorem

We already met Fermat, whose "Last Theorem" gave mathematicians so much trouble for so many years. In this section, we will prove a theorem which Fermat did succeed in establishing. First, two results about $\mathbb{Z}_p$ for p prime.

Lemma 6.5 *Let p be a prime number and suppose that $p \mid ab$. Then $p \mid a$ or $p \mid b$.*

Proof Suppose that p divides ab but p does not divide a. Since p is prime, we see that $\gcd(a,p) = 1$. By Euclid's algorithm, there exist x and y such that $xa + yp = 1$. Then $xab + ypb = b$. Now p divides xab (since it divides ab, and clearly p divides ypb; so p divides b.

Lemma 6.6 *Let p be a prime number.*

(a) If $[a]_p[b]_p = [0]_p$, then either $[a]_p = [0]_p$ or $[b]_p = [0]_p$.

(b) If $[ab]_p = [ac]_p$ and $[a]_p \neq [0]_p$, then $[b]_p = [c]_p$.

Proof (a) Since $[a]_p[b]_p = [ab]_p$, the assumption $[a]_p[b]_p = [0]_p$ means that $ab \equiv_p 0$, that is, $p \mid ab$. Then $p \mid a$ or $p \mid b$ by the preceding Lemma; so $[a]_p = [0]_p$ or $[b]_p = [0]_p$.

(b) We have $[a]_p[b-c]_p = [0]_p$; so, if $[a]_p \neq [0]_p$, then $[b-c]_p = [0]_p$, so that $[b]_p = [c]_p$.

So we come to *Fermat's Little Theorem:*

Theorem 6.7 *Let p be a prime number. If a is any integer not divisible by p, then $a^{p-1} \equiv_p 1$.*

So, for example, $3^6 \equiv_7 1$, as you can check.

Proof Consider the non-zero elements $[1]_p, [2]_p, \ldots, [p-1]_p$. Multiply them all by a, to get $[a]_p, [2a]_p, \ldots, [(p-1)a]_p$. Now by the preceding Lemma, none of these elements is equal to $[0]_p$, and no two of them are equal; so we have the same list of elements in a different order. So their product is the same:

$$[a]_p[2a]_p \cdots [(p-1)a]_p = [1]_p[2]_p \cdots [p-1]_p,$$

from which we see that

$$[a^{p-1}]_p[1]_p[2]_p \cdots [p-1]_p = [1]_p[2]_p \cdots [p-1]_p.$$

Since $[1]_p[2]_p \cdots [p-1]_p \neq [0]_p$, we conclude from the lemma that

$$[a^{p-1}]_p = [1]_p,$$

in other words, $a^{p-1} \equiv_p 1$, as required.

For example, if $p = 7$ and $a = 3$, then the multiples of 3 mod 7 are

$$[3]_7, [6]_7, [9]_7 = [2]_7, [12]_7 = [5]_7, [15]_7 = [1]_7, [18]_7 = [4]_7,$$

so we do obtain all the non-zero congruence classes in a different order.

Exercises

6.1 Find the units in $\mathbb{Z}_{30}$ and their inverses.

6.2 Calculate $\dfrac{2}{3} + \dfrac{3}{4}$ in $\mathbb{Z}_{29}$.

6.3 Solve the quadratic equation $x^2 + 2x + 2 = 0$

(a) in $\mathbb{Z}_{17}$,

(b) in $\mathbb{Z}_{19}$.

6.4 Prove that $(p-1)! \equiv_p -1$ if and only if p is prime. (This is *Wilson's Theorem*.)

Chapter 7

Polynomials revisited

We will look at three further aspects of polynomials. First, we have only considered real polynomials so far, but this can be generalised: as long as we can add and multiply the coefficients, we can do the same with polynomials. Second, we look at factorisation and show that, under the right conditions, the division rule, Euclid's algorithm, and the Remainder and Factor Theorems hold. Finally, the construction of the integers mod m by means of congruence classes can be extended to polynomials. This gives us a less *ad hoc* construction of the complex numbers, as well as other finite systems having addition and multiplication.

7.1 Polynomials over other systems

Let R be a set on which two operations (called *addition* and *multiplication* are defined. Suppose that R satisfies the following laws. (We call this collection of laws CRI).

- the commutative, associative, identity and inverse laws for addition (the identity for addition is called 0, and the inverse of a is $-a$);
- the commutative, associative, and identity laws for multiplication (the identity for multiplication is called 1);
- the distributive law.

Later, we will study such systems formally under the name "commutative rings with identity". In this section, we will put them to use.

The examples we have met already include:

- $\mathbb{Z}$, $\mathbb{Q}$, $\mathbb{R}$, $\mathbb{C}$;
- $\mathbb{R}[x]$, the polynomials with real coefficients;

- $\mathbb{Z}_m$, the integers mod m.

We can define a *polynomial* over R to be an expression of the form

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

where n is a non-negative integer and $a_n, a_{n-1}, \ldots, a_1, a_0 \in R$. We adopt the same rules as we discussed earlier for when two expressions represent the same polynomial (we can insert or remove terms with coefficient zero, and we can replace $1x^n$ by x^n). Now we can add and multiply polynomials by the same rules as before.

Let $R[x]$ be the set of all polynomials with coefficients in R.

Proposition 7.1 *If R satisfies the laws (CRI) above, then so does $R[x]$.*

Proof As usual, we don't give a detailed verification of all the laws. The appendix to this chapter gives some of the details.

We end this section with a warning. We informally defined a real polynomial to be a function on the real numbers given by an expression of the right form. This no longer works for more general polynomials.

Example Let $R = \mathbb{Z}_2$, the integers mood 2. The set R contains two elements, $[0]_2$ and $[1]_2$, which we will write more briefly as 0 and 1. The laws (CRI) are satisfied.

Now consider the two polynomials x and x^2. Since $0^2 = 0$ and $1^2 = 1$, these two polynomials give rise to the same function on $\mathbb{Z}_2$. However, we really do want to regard them as different polynomials! Hence we regard polynomials as being formal expressions, not the functions they define.

7.2 Division and factorisation

The division rule and Euclid's algorithm work in almost the same way for polynomials as for integers. So we can mimic the definition of the integers mod m.

We need one more property for the coefficients, beyond the laws (CRI) we assumed before. The extra law is the inverse law for multiplication, which states that every element a of R except 0 has a multiplicative inverse a^{-1}. A system satisfying (CRI) and the inverse law for multiplication is called a *field*. The examples we know so far are $\mathbb{Q}$, $\mathbb{R}$, $\mathbb{C}$, and $\mathbb{Z}_p$ for prime numbers p.

[Why do we need this extra law? Look back at the proof of the division rule for polynomials. To divide $f(x) = a_n x^n + \cdots$ by $g(x) = b_m x^m + \cdots$, where $b_m \neq 0$ and $n > m$, we first subtract $(a_n/b_m)x^{n-m}g(x)$ from $f(x)$ to obtain a polynomial of smaller degree. So we need to be able to divide a_n by b_m, that is, we need a multiplicative inverse for b_m.]

Theorem 7.2 *If R is a field, then the set $R[x]$ of polynomials with coefficients in R satisfies the division rule, and Euclid's algorithm works in $R[x]$.*

A polynomial $g(x)$ of degree greater than zero is called *irreducible* if it cannot be written in the form $g(x) = h(x)k(x)$, where $h(x)$ and $k(x)$ are polynomials with degrees smaller than the degree of $g(x)$.

We will not treat irreducible polynomials in detail, but simply look at one technique for recognising them. Let $f(x)$ be a polynomial over the field R. For any $c \in R$, we let $f(c)$ denote the result of "substituting c into $f(x)$"; that is,

$$\text{if } f(x) = a_n x^n + \cdots + a_0, \text{ then } f(c) = a_n c^n + \cdots + a_0.$$

The next theorem combines two familiar results about polynomials, the *Remainder Theorem* and the *Factor Theorem*. Notice that, if we divide $f(x)$ by a polynomial of degree 1, the remainder has degree zero, that is, it is a constant polynomial (which we regard as an element of R).

Theorem 7.3 *Let $f(x)$ be a polynomial over a field R, and $c \in R$.*

(a) The remainder when $f(x)$ is divided by $x - c$ is $f(c)$.

(b) $f(x)$ is divisible by $x - c$ if and only if $f(c) = 0$.

Proof (a) Write $f(x) = (x-c)q(x) + r$, where r is a constant polynomial. Substituting c into this equation we find $f(c) = r$.

(b) If $f(c) = 0$ then $f(x) = (x-c)q(x)$, so $x - c$ divides $f(x)$. The converse is clear from the uniqueness of the remainder in the division rule.

Example The polynomial $f(x) = x^3 - 2$ is irreducible in $\mathbb{Q}[x]$. For if it factorises, it must be a product of polynomials of degrees 1 and 2. The polynomial of degree 1 has the form $x - c$, where c is a rational number; by the Factor Theorem, $f(c) = 0$, that is, $c^3 = 2$. But, following Euclid's proof, it can be shown that $\sqrt[3]{2}$ is irrational (this is an exercise for you); so this is impossible.

7.3 "Modular arithmetic" for polynomials

Now let R be a field, and let $g(x)$ be a fixed non-zero polynomial in $R[x]$. To make things easier, we assume that $g(x)$ is monic. We say that two polynomials $f_1(x)$ and $f_2(x)$ are *congruent mod $g(x)$* if $g(x)$ divides $f_1(x) - f_2(x)$, that is, if $f_1(x) = g(x)h(x) + f_2(x)$ for some polynomial $h(x)$.

Proposition 7.4 *Congruence mod $g(x)$ is an equivalence relation, and each equivalence class contains a unique polynomial $r(x)$ such that $r(x) = 0$ or $\deg(r(x)) < \deg(g(x))$.*

This is just a re-statement of the division rule. We will denote the equivalence class of $f(x)$ by $[f(x)]$, and call it a *congruence class mod $g(x)$*.

Let E be the set of congruence classes mod $g(x)$. Just as we did for congruence classes mod m in the integers, we are going to give rules for adding and multiplying elements of E. The rules are the obvious ones:

- $[f_1(x)] + [f_2(x)] = [f_1(x) + f_2(x)]$,
- $[f_1(x)] \cdot [f_2(x)] = [f_1(x) f_2(x)]$.

Just as before, we have first to do some work to show that our definition is a good one. That is, if $f_1(x) \equiv f_1'(x)$ and $f_2(x) \equiv f_2'(x)$, then $f_1(x) + f_2(x) \equiv f_1'(x) + f_2'(x)$ and $f_1(x) f_2(x) \equiv f_1'(x) f_2'(x)$. (All congruences are modulo $g(x)$.) Here is the proof of the first statement; try the second for yourself. We are given that $f_1(x) - f_1'(x) = g(x)h_1(x)$ and $f_2(x) - f_2'(x) = g(x)h_2(x)$. Then we find

$$(f_1(x) + f_2(x)) - (f_1'(x) + f_2'(x)) = g(x)(h_1(x) + h_2(x)),$$

which shows the required congruence.

Proposition 7.5 *If R is a field, then the set E of congruence classes mod $g(x)$ also satisfies (CRI).*

The proof of this simply involves routine checking of laws.

For integers, we saw that $\mathbb{Z}_p$ is a field if p is prime. Something very similar happens here; in place of primes, we use irreducible polynomials.

Theorem 7.6 *Suppose that R is a field and $g(x)$ an irreducible polynomial in $R[x]$. Then the set E of equivalence classes mod $g(x)$ is also a field, and contains the field R.*

Proof We have to show that a non-zero congruence class has a multiplicative inverse.

Suppose that the equivalence class $[f(x)]$ is not zero. This means that $g(x)$ doesn't divide $f(x)$. So $\gcd(f(x), g(x)) = 1$. (For the gcd is a monic polynomial dividing $g(x)$; since $g(x)$ is irreducible, it cannot have positive degree.)

By Euclid's algorithm, there are polynomials $h(x)$ and $k(x)$ such that

$$f(x)h(x) + g(x)k(x) = 1.$$

This equation says that $f(x)h(x) \equiv 1 \bmod g(x)$, so $[f(x)] \cdot [h(x)] = [1]$. Thus we have found an inverse for $[f(x)]$.

To find a copy of the field R inside E, we just take the equivalence classes of the constant polynomials; they add and multiply just like elements of R:

$$[c] + [d] = [c+d], \qquad [c] \cdot [d] = [cd],$$

This is all very good, but a bit too abstract for practical use. Here is a description which is easier to calculate with.

Proposition 7.7 *Suppose that the hypotheses of the previous theorem are satisfied, and let m be the degree of $g(x)$. Then E is the set*

$$\{c_{m-1}\alpha^{m-1} + \cdots + c_0 : c_0, \ldots, c_{m-1} \in R\},$$

where α is a new symbol satisfying $g(\alpha) = 0$.

Proof We saw that the constant polynomials $[c]$ are just like elements of R, so we ignore the difference and identify them with elements of R. Let $\alpha = [x]$, the congruence class containing the polynomial x. Now we saw that each equivalence class contains a unique polynomial $r(x)$ of degree less than m (or zero). If $r(x) = c_{m-1}x^{m-1} + \cdots + c_0$, then

$$\begin{aligned} [r(x)] &= [c_{m-1}x^{m-1} + \cdots + c_0] \\ &= [c_{m-1}][x]^{m-1} + \cdots + [c_0] \\ &= c_{m-1}\alpha^{m-1} + \cdots + c_0. \end{aligned}$$

(In the second line we used the rules for adding and multiplying equivalence classes; in the third, we put $[c] = c$ and $[x] = \alpha$.

Finally, $g(x) \equiv 0 \bmod g(x)$, so $[g(x)] = [0]$. By the same argument, this gives $g(\alpha) = 0$.

Time for a (very important) example. Let R be the field $\mathbb{R}$ of real numbers. We take $g(x)$ to be the polynomial $x^2 + 1$. (This is irreducible; for its factors, if any, would have degree 1, but if, say,

$$x^2 + 1 = (x-a)(x-b),$$

then $a^2 = -1$, which is impossible since the square of any real number is positive.

Now the field E of our construction consists of all expressions of the form $c + d\alpha$, where c and d are real numbers and α is a new symbol satisfying $\alpha^2 + 1 = 0$. Thus α is the symbol usually called i. So the complex numbers are not just a fluke; they are a special case of a very general construction!

7.4 Finite fields

We saw that $\mathbb{Z}_m$ is a field with m elements if m is prime, but is not a field if m is composite. Is there a finite field with four elements? We cannot use $\mathbb{Z}_4$, since $[2]_4$ has no multiplicative inverse.

We apply the construction of the preceding section. First we find an irreducible polynomial.

Lemma 7.8 *The polynomial x^2+x+1 is irreducible over $\mathbb{Z}_2$.*

Proof If not, it has a factor of the form $x-c$ for some $c \in \mathbb{Z}_2$. By the Factor Theorem, this would mean that $f(c)=0$. But, writing 0 and 1 instead of the more cumbersome $[0]_2$ and $[1]_2$,

$$\begin{aligned} 0^2+0+1 &= 1, \\ 1^2+1+1 &= 1, \end{aligned}$$

so no c satisfying $c^2+c+1=0$ exists, and there is no factor $(x-c)$.

So there is a field E consisting of the elements $c\alpha+d$, with $c,d \in \mathbb{Z}_2$ and $\alpha^2+\alpha+1=0$. The elements of E are $0,1,\alpha,\alpha+1$. The elements 0 and 1 comprise the field $\mathbb{Z}_2$, so that $1+1=0$. Then $x+x=(1+1)x=0$ for any x, and so $\alpha^2=-\alpha-1=\alpha+1$. Then we can do calculations like

$$(\alpha+1)^2=\alpha^2+\alpha+\alpha+1=\alpha+1+1=\alpha.$$

In general, any expression involving α can be calculated to be one of the four elements $0,1,\alpha,\alpha+1$.

The addition and multiplication tables for the field E can now be worked out:

$+$	0	1	α	$\alpha+1$
0	0	1	α	$\alpha+1$
1	1	0	$\alpha+1$	α
α	α	$\alpha+1$	0	1
$\alpha+1$	$\alpha+1$	α	1	0

$\cdot$	0	1	α	$\alpha+1$
0	0	0	0	0
1	0	1	α	$\alpha+1$
α	0	α	$\alpha+1$	1
$\alpha+1$	0	$\alpha+1$	1	α

Now the multiplicative inverses of 1, α and $\alpha+1$ are, respectively, 1, $\alpha+1$, and α.

Évariste Galois is one of the founders of modern algebra. He was killed in a duel at the age of 19; already he had worked out, and published, the construction of finite fields (he did much more than we have seen, showing that the number of elements in a finite field must be a power of a prime, and that for each prime power there is a unique finite field of that order). Finite fields are called *Galois*

fields in his honour; the field with q elements is denoted by $\mathrm{GF}(q)$. So the field we constructed above is $\mathrm{GF}(4)$.

But his major work, in which he showed how group theory could be used to decide when a "solution by radicals" for a general polynomial equation could be found, had been lost by referees at the French Academy (who were probably unable to understand it). Its main impact came fifteen years later when it was rediscovered and published.

As well as algebra, Galois was deeply involved in the revolutionary politics of his time. The duel in which he was shot and killed was apparently over a woman; but historians have uncovered evidence that it had been set up, either by the Royalist police, or by the revolutionaries to whom he had offered himself as a sacrifice to spark a general uprising. If the second explanation is true, then he died in vain, as there was no uprising.

7.5 Appendix: Laws for polynomials

In Proposition 7.1, we asserted that, if R satisfies the system (CRI) of laws, then so does $R[x]$, the set of polynomials over R. In this appendix I will say a few words about the proof. First, let us be clear about the definitions.

A polynomial over R is an expression

$$f(x) = a_n x^n + \cdots + a_1 x + a_0 = \sum_{i=0}^{n} a_i x^i.$$

Suppose that $g(x)$ is another polynomial:

$$g(x) = b_m x^m + \cdots + b_0 = \sum_{i=0}^{m} b_i x^i.$$

To add $f(x)$ and $g(x)$, we first assume that $m = n$. (If $m < n$, we add extra terms $0x^i$ for $i = m+1, \ldots, n$ to the polynomial $g(x)$, and similarly if $n < m$ we add zero terms to $f(x)$. Then

$$(f+g)(x) = \sum_{i=0}^{n} (a_i + b_i) x^i.$$

The rule for multiplication is a bit more complicated:

$$(fg)(x) = \sum_{i=0}^{m+n} c_i x^i,$$

where

$$c_i = \sum a_j b_{i-j},$$

the sum being over all values of j for which both a_j and b_{i-j} are defined; that is, we have $0 \le j \le n$ and $0 \le i-j \le m$, so that $i-m \le j \le i$. We can summarise these two sets of conditions by saying

$$\max(0, i-m) \le j \le \min(i, n).$$

Consider, for example, the distributive law

$$(f+g)h = fh+gh.$$

We assume that f and g are as above (with $m = n$) and that

$$h(x) = \sum_{i=0}^{p} d_i x^i.$$

Then the coefficient of x^i in $(f+g)h$ is

$$\sum (a_j + b_j) d_{i-j} = \sum a_j d_{i-j} + b_j d_{i-j},$$

using the distributive law for R; and the coefficient in $fh+gh$ is

$$\sum a_j d_{i-j} + \sum b_j d_{i-j}.$$

Now all the sums are over the same range $\max(0, i-m) \le j \le \min(i, p)$, and rearranging the terms shows that the two expressions are equal.

Exercises

7.1 Suppose that $ax+b$ divides $c_n x^n + \cdots + c_0$ in $\mathbb{Z}[x]$, where $a, b, c_0, \ldots, c_n$ are integers. Show that a divides c_n, and b divides c_0. Hence show that $x^n - 2$ is irreducible over $\mathbb{Z}$ for any positive integer n.

7.2 (a) Show that the polynomial x^2+1 is irreducible over $\mathbb{Z}_3$.

(b) Construct a field with nine elements.

7.3 Verify the associative law for multiplication of polynomials.

Chapter 8

Rings

We have seen many different types of structure (numbers, matrices, polynomials, sets, modular arithmetic) which satisfy very similar laws. Now we take the obvious next step: we consider systems satisfying these laws abstractly, and prove things about them directly from the laws they satisfy. The results will then be true in our systems no matter what they are made up of. This is called the *axiomatic method*.

8.1 Rings

A *ring* is a set R of elements with two operations, *addition* (written $+$) and *multiplication* (written $\cdot$ or just by juxtaposing the factors) which satisfies the following laws. (Most of these we have seen before, but we state them all formally here.)

Additive laws:

(A0) Closure law: For al $a, b \in R$, 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) Zero law: There is an element $0 \in R$ with the property that $a+0 = 0+a = a$ for all $a \in R$.

(A3) Additive 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 al $a, b \in R$, 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 laws: For all $a, b, c \in R$, we have $a(b+c) = ab+ac$ and $(b+c)a = ba+ca$.

Before we go further, a couple of comments:

- The closure laws are new. Strictly speaking, they are not necessary, since to say that $+$ is an operation on R means that the output $a+b$ when we input a and b to the black box belongs to R. We put them in as a reminder that, when we are checking that something is a ring, we have to be sure that this holds.

- We have stated the identity and inverse laws for addition in a more complicated way than necessary. Since we are going on to state the commutative law for addition, we could simply have said that $a+0 = a$ and $a+(-a) = 0$. We'll see the reason soon.

We have already seen that sometimes the multiplication satisfies further laws, which resemble the laws for addition. This won't always be the case, so we give special names to rings in which these laws hold.

Let R be a ring. We say that R is a *ring with identity* if

(M2) Identity law: There is an element $1 \in R$ (with $1 \neq 0$) such that $1a = a1 = a$ for all $a \in R$.

We say that R is a *division ring* if it satisfies (M2) and also

(M3) Multiplicative inverse law: for all $a \in R$, if $a \neq 0$, then there exists $b \in R$ such that $ab = ba = 1$. We write b as a^{-1}.

We say that R is a *commutative ring* if

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

(Note that the word "commutative" here refers to the multiplication; the addition in a ring is always commutative.) Finally, we say that R is a *field* if it satisfies (M2), (M3) and (M4).

The condition (CRI) which we introduced in the last chapter thus stands for "commutative ring with identity".

In a non-commutative ring, we need to assume both parts of the identity and multiplicative inverse laws, since one does not follow from the other. Similarly, we do need both parts of the distributive law.

8.2 Examples of rings

We have a ready-made stock of examples:

- $\mathbb{Z}$ is a commutative ring with identity.
- $\mathbb{Q}$, $\mathbb{R}$ and $\mathbb{C}$ are fields.
- If R is a commutative ring with identity, then so is the *polynomial ring* $R[x]$.
- If R is a ring, then so is the set $M_{n\times n}(R)$ of all $n \times n$ matrices over R (with the usual definitions of matrix addition and multiplication). If R has an identity, then so does $M_{n\times n}(R)$. But this ring is usually not commutative.
- For any m, the set $\mathbb{Z}_m$ of integers mod m is a commutative ring with identity. It is a field if and only if m is prime.
- If R is a field and $g(x)$ a monic polynomial of degree at least 1 over R, then the set of congruence classes of polynomals mod $g(x)$ is a commutative ring with identity. It is a field if (and only if) $g(x)$ is an irreducible polynomial.

Note that the third and fourth of these constructions (polynomials and matrices) are methods of building new rings from old ones. You may guess that the fifth and sixth can also be made into constructions of new rings from old. This is correct, but the construction is beyond the scope of this course. You will meet it next year in Algebraic Structures I.

Some other familiar structures do not form rings. For example, the set of natural numbers $\mathbb{N}$ is not a ring, since the additive inverse law does not hold.

At the end of the last chapter, we constructed a field with four elements.

8.3 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 8.1 *In a ring R,*

there is a unique zero element;

any element has a unique additive inverse.

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

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

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

This justifies us in calling the unique zero element 0.

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

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

Hence

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

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

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

Proposition 8.2 *Let R be a ring.*

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

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

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

The next result is called the *cancellation law*.

Proposition 8.3 *Let R be a ring. If $a+b = a+c$, then $b = c$.*

Proof

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

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

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

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

Proposition 8.4 *Let R be a ring. For any element $a \in R$, we have $0a = a0 = 0$.*

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

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

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

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

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

for all $a, b, c \in X$. This means that we can write $a \circ b \circ c$ 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:

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

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

This result generalises:

Proposition 8.5 *Let $\circ$ be an operation on a set X which satisfies the associative law. Then the value of the expression*

$$a_1 \circ a_2 \circ \cdots \circ a_n$$

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

I won't give the inductive proof here; you are encouraged to try it yourself! You will find the proof in an appendix to the notes.

8.4 Units

Let R be a ring with identity element 1. An element $u \in R$ is called a *unit* if there is an element $v \in R$ such that $uv = vu = 1$. The element v is called the *inverse* of u, written u^{-1}. By Proposition 8.2, a unit has a unique inverse.

Here are some properties of units.

Proposition 8.6 *Let R be a ring with identity.*

(a) 0 *is not a unit.*

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

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

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

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

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

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

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

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

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

Here is how Hermann Weyl explains Proposition refunits(d), the statement that $(uv)^{-1} = v^{-1}u^{-1}$, 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.

Here are some examples of units in familiar rings.

- In a field, every non-zero element is a unit.
- In $\mathbb{Z}$, the only units are 1 and -1.
- Let F be a field. Then a polynomial in the polynomial ring $F[x]$ is a unit if and only if it is a non-zero constant polynomial. For we have

$$\deg(f(x)g(x)) = \deg(f(x)) + \deg(g(x)),$$

  so if $f(x)g(x) = 1$ then $f(x)$ must have degree zero, that is, it is a constant polynomial.
- Let F be a field and n a positive integer. An element A of the ring $M_{n\times n}(F)$ is a unit if and only if the determinant of A is non-zero. In particular, $\begin{pmatrix} a & b \\ c & d \end{pmatrix}$ is a unit in $M_{2\times 2}(\mathbb{R})$ if and only if $ad - bc \neq 0$; if this holds, then its inverse is

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

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

Proposition 8.7 *Suppose that $m > 1$.*

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

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

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

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

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

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

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

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

a contradiction.

Example The table shows, for each non-zero element $[a]_{12}$ of $\mathbb{Z}_{12}$, an element $[b]_{12}$ such that the product is either 1 or 1. To save space we write a instead of $[a]_{12}$.

a	1	2	3	4	5	6	7	8	9	10	11
ab	$1\cdot 1=1$	$2\cdot 6=0$	$3\cdot 4=0$	$4\cdot 3=0$	$5\cdot 5=1$	$6\cdot 2=0$	$7\cdot 7=1$	$8\cdot 3=0$	$9\cdot 4=0$	$10@cdot6=0$	$11\cdot 11=1$
Unit?	$\surd$	$\times$	$\times$	$\times$	$\surd$	$\times$	$\surd$	$\times$	$\times$	$\times$	$\surd$

So the units in $\mathbb{Z}_{12}$ are $[1]_{12}$, $[5]_{12}$, $[7]_{12}$, and $[11]_{12}$.

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

8.5 Appendix: The associative law

In this section we give the proof that, if $\circ$ is an operation on a set X which satisfies the associative law, then the composition of n terms doesn't depend on how we put in the brackets (Proposition 8.5).

The proof is by induction on n. For $n = 2$, there are no brackets in $a_1 \circ a_2$, and nothing to prove. For $n = 3$, there ae two ways to put in the brackets, viz. $a_1 \circ (a_2 \circ a_3)$ and $(a_1 \circ a_2) \circ a_3$; the associative law asserts that they are equal. In the notes we saw that, for $n = 4$, there are five bracketings, and the five expressions are all equal.

So now suppose that the statement is true for expressions with fewer than n terms, and consider any two bracketings of $a_1 \circ \cdots \circ a_n$. Now for any bracketing, when we work it out "from the inside out", in the last step we have just two expressions to be composed; that is, the expression looks like

$$(x_1 \circ \cdots \circ x_k) \circ (x_{k+1} \circ \cdots \circ x_n).$$

There may be further brackets inside the two terms, but (according to the inductive hypothesis) they don't affect the result. We will say that the expression *splits after k terms*.

Suppose that the first expression splits after k terms, and the second splits after l terms.

Case $k = l$ Both expressions now have the form

$$(x_1 \circ \cdots \circ x_k) \circ (x_{k+1} \circ \cdots \circ x_n),$$

and by induction the bracketed terms don't depend on any further brackets. So they are equal.

Case $k < l$ Now the first expression is

$$(x_1 \circ \cdots \circ x_k) \circ (x_{k+1} \circ \cdots \circ x_n)$$

and the second is

$$(x_1 \circ \cdots \circ x_l) \circ (x_{l+1} \circ \cdots \circ x_n).$$

By the induction hypothesis, the value of the term $x_1 \circ \cdots \circ x_k$ doesn't depend on where the brackets are; so we can rearrange the brackets so that this expression splits after k terms, so that the whole expression is

$$((x_1 \circ \cdots \circ x_k) \circ (x_{k+1} \circ \cdots \circ x_l)) \circ (x_{l+1} \circ \cdots \circ x_n).$$

In the same way, we can rearrange the second expression as

$$(x_1 \circ \cdots \circ x_k) \circ ((x_{k+1} \circ \cdots \circ x_l) \circ (x_{l+1} \circ \cdots \circ x_n)).$$

Now the two expressions are of the form $(a \circ b) \circ c$ and $a \circ (b \circ c)$, where

$$\begin{aligned} a &= x_1 \circ \cdots \circ x_k, \\ b &= x_{k+1} \circ \cdots \circ x_l, \\ c &= x_{l+1} \circ \cdots \circ x_n. \end{aligned}$$

The associative law shows that they are equal.

Case $k > l$ This case is almost identical to the preceding one.

Exercises

8.1 Let $n\mathbb{Z}$ be the set of all integers divisible by n. Show that $n\mathbb{Z}$ is a ring (with the usual addition and multiplication). Is it commutative? Does it have an identity?

8.2 Let $\mathscr{P}(S)$ denote the set of all subsets of the set S. For $A, B \in \mathscr{P}(S)$, define $A + B = A \triangle B$ (symmetric difference), and $AB = A \cap B$ (intersection). Show that $\mathscr{P}(S)$ is a ring. Show also that $A^2 = A$ for all $A \in \mathscr{P}(S)$.

8.3 Let R be a ring in which $a^2 = a$ for all $a \in R$. By considering $(a+b)^2$, show

(a) R is commutative;

(b) $a + a = 0$ for all $a \in R$.

(Such a ring is called a *Boolean ring*.)

Chapter 9

Groups

The additive and multiplicative axioms for rings are very similar. This similarity suggests considering a structure with a single operation, called a group. In this section we study groups and their properties.

9.1 Definition

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

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

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

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

(G3) Inverse law: for all $a \in G$, there exists $b \in G$ such that $a \circ b = b \circ a = e$, 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 $a \circ b = b \circ a$

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

9.2 Elementary properties

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

Proposition 9.1 *Let G be a group.*

(a) The composition of n elements has the same value however the brackets are inserted.

(b) The identity of G is unique.

(c) Each element has a unique inverse.

(d) Cancellation law; if $a \circ b = a \circ c$ then $b = c$.

Proof (a) Proved in the appendix to the last section of the notes. (b) If e and e^* are identities then

$$e = e \circ e^* = e^*.$$

(c) If b and b^* are inverses of a then

$$b = b \circ e = b \circ a \circ b^* = e \circ b^* = b^*.$$

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

9.3 Examples of groups

We have some ready-made examples.

- Let R be a ring. Take $G = R$, with operation $+$; the identity is 0 and the inverse of a is $-a$. This group is called the *additive group* of the ring R. It is an abelian group.

- Let R be a ring with identity, and let $U(R)$ denote the set of units of R, with operation multiplication in R. This is a group;

  - the closure, identity and inverse laws follow from Proposition xx in the last part of the notes;
  - the associative law follows from the ring axiom (M1).

  This group is called the *group of units* of R. The next couple of examples are special cases.

- In particular, if F is a field, then the group $U(F)$ of units of F consists of all the non-zero elements of F. This is called the *multiplicative group* of F.

- Let F be a field and n a positive integer. The set $M_{n\times n}(F)$ of all $n\times n$ matrices with elements in F is a ring. We saw that a matrix is a unit in this ring if and only if its determinant is non-zero. The group $U(M_{n\times n}(F))$ is called the *general linear group* of dimension n over F, written $\mathrm{GL}(n,F)$.
- Let V be a vector space. Then, with the operation of vector addition, V is an abelian group; the identity is the zero vector $\mathbf{0}$, and the inverse of $\mathbf{v}$ is $-\mathbf{v}$.

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

Remark on notation I have used here a neutral symbol $\circ$ for the group operation. In books, you will often see the group operation written as multiplication, or (in abelian groups) as addition. Here is a table comparing the different notations.

Notation	Operation	Identity	Inverse
General	$a\circ b$	e	a'
Multiplicative	ab, $a\cdot b$	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 $(G,\circ)$ (or $(G,+)$, or $(G,\cdot)$) be a group". The rest of the notation should then be fixed as in the table.

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

9.4 Cayley tables

If a group is finite, it can be represented by its operation table. In the case of groups, this table is more usually called the *Cayley table*, after Arthur Cayley who pioneered its use. Here, for example, is the Cayley table of the group of units of the ring $\mathbb{Z}_{12}$.

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

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 $a\circ x=b$. This equation has the unique solution $x=a^{-1}\circ b$, where a^{-1} is the inverse of a. A similar argument applies to the columns.

Example Let G be a group with three elements e, a, b, with e the identity. We know part of the Cayley table:

$\circ$	e	a	b
e	e	a	b
a	a		
b	b		

Now consider $a \circ b$, the element in the second row and third column. This cannot be a, since we already have a in the row; and it cannot be b, since we already have b in the column. So $a \circ b = e$. With similar arguments we can find all the other entries.

So there is only one "type" of group with three elements.

We will just stop and look at what this means. Let $(G, \circ)$ and $(H, *)$ be groups. We say that G and H are *isomorphic* if there is a bijective (one-to-one and onto) function $F : G \to H$ such that $F(g_1 \circ g_2) = F(g_1) * F(g_2)$ for all $g_1, g_2 \in G$. In other words, we can match elements of G with elements of H such that the group operation works in the same way on elements of G and the matched elements of H. The function F is called an *isomorphism*.

Thus, the argument we just gave shows that any two groups with three elements are isomorphic.

9.5 Subgroups

Let $(G, \circ)$ be a group, and H a sub*set* of G, that is, a selection of some of the elements of G. For example, let $G = (\mathbb{Z}, +)$ (the additive group of integers), and $H = 4\mathbb{Z}$ (the set of multiples of 4).

We say that H is *subgroup* of G if H, with the same operation (addition in our example) is itself a group.

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

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

(G1) H should satisfy the associative law; that is, $(h_1 \circ h_2) \circ h_3 = h_1 \circ (h_2 \circ h_3$, 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. But, by the uniqueness of the identity, this must be the same as the identity element of G. So this condition requires that H should contain the identity of G.

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

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

Proposition 9.2 *A non-empty subset H of a group $(G, \circ)$ is a subgroup if and only if, for all $h_1, h_2 \in H$, we have $h_1 \circ h_2^{-1} \in H$.*

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

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

In our example, $G = \mathbb{Z}$, $H = 4\mathbb{Z}$, take two elements of H, say $4a$ and $4b$; then since the group operation is $+$, the inverse of $4b$ is $-4b$, and we have to check whether $4a - 4b \in H$. The answer is yes, since $4a - 4b = 4(a-b) \in 4\mathbb{Z}$. So $4\mathbb{Z}$ is a subgroup.

9.6 Cosets and Lagrange's Theorem

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

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

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

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

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

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

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

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

so $g_1 \sim g_3$.

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

What do cosets look like?

For any $g \in G$, let

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

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

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

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

Theorem 9.3 *Let H be a subgroup of G. Then the cosets of H in G are the sets of the form*

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

and they form a partition of G.

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

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

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

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

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

Theorem 9.4 *Let G be a finite group, and H a subgroup of G. Then $|H|$ divides $|G|$. The quotient $|G|/|H|$ is equal to $|G:H|$, the index of H in G.*

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

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

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

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

9.7 Orders of elements

Remember that the order of a group is the number of elements in the group. We will define in this section the order of an element of a group. This is quite different – be careful not to get them confused – but there is a connection, as we will see.

Let g be an element of a group G. We define g^n for every integer n in the following way:

$$\begin{aligned} g^0 &= e, \\ g^n &= g^{n-1} \circ g \text{ for } n > 0, \\ g^{-n} &= (g^n)^{-1} \text{ for } n > 0. \end{aligned}$$

Now it is possible to prove that the *exponent laws* hold:

Proposition 9.5 *For any integers m and n,*

(a) $g^m \circ g^n = g^{m+n}$,

(b) $(g^m)^n = g^{mn}$.

The proof is not difficult but needs a lot of care. It follows from the definition that

$$g^n = \begin{cases} g \circ \cdots \circ g \ (n \text{ factors}) & \text{if } n > 0, \\ g^{-1} \circ \cdots \circ g^{-1} \ (-n \text{ factors}) & \text{if } n < 0. \end{cases}$$

Now consider g^{m+n}. There are four cases.

- If m and n are both positive then

$$g^m \circ g^n = g \circ \cdots \circ g \ (m+n \text{ factors}) = g^{m+n}.$$

- If one of m and n is positive, say $m > 0$, $n < 0$, then
  - If $m+n > 0$, so that $m > -n$, then $-n$ of the factors g cancel all the factors g^{-1}, leaving $m+n$ factors g, so the result is g^{m+n}.
  - If $m+n < 0$, then m of the factors g^{-1} cancel all the factors g, leaving $-m-n$ factors g^{-1}; again we have g^{m+n}.
- Finally, if m and n are both negative, a similar argument to the first case applies.

If one of m and n is zero, say $m = 0$, then the product is $e \circ g^n = g^n$.

The argument for the second exponent law is similar.

It follows from the second exponent law that $(g^n)^{-1} = g^{-n}$. This also follows because $g^n \circ g^{-n} = g^0 = e$.

Now we make two definitions.

- The *order* of the element g is the smallest positive number n for which $g^n = e$, if such a number exists; if no positive power of g is equal to e, we say that g has infinite order.
- The *subgroup generated by g* is the set

  $$\{g^n : n \in \mathbb{Z}\}$$

  of all powers of g. We write it as $\langle g \rangle$.

It is not clear from what has been said so far that "the subgroup generated by g" is actually a subgroup! In fact it is; this and more are contained in the next Proposition. Remember that the word "order" has two different meanings; the first is the number of elements in the subgroup, the second is the number we have just defined.

Proposition 9.6 *For any element g of a group G, the set $\langle g \rangle$ is a subgroup of G, and its order is equal to the order of g.*

Proof To show that $\langle g \rangle$ is a subgroup, we apply the Subgroup Test. Take two elements of this set, say g^m and g^n. Then

$$g^m \circ (g^n)^{-1} = g^m \circ g^{-n} = g^{m-n} \in \langle g \rangle.$$

Next we show that, if g has order n, then

- $g^m = e$ if and only if n divides m;

- $g^k = g^l$ if and only if $k \equiv_n l$.

Suppose that $m = nq$. Then $g^m = (g^n)^q = e^q = e$. Conversely, suppose that $g^m = e$. By the Division Rule, $m = nq + r$, with $0 \leq r \leq n-1$. Now $g^n = g^m = e$, so $g^r = e$. But n is the smallest positive integer such that the nth power of g is e; since $r < n$ we must have $r = 0$, and n divides m.

Now $g^k = g^l$ if and only if $g^{l-k} = e$. By the preceding paragraph, this holds if and only if n divides $l-k$, that is, if and only if $k \equiv_n l$.

We see that if g has order n, then the set $\langle g \rangle$ contains just n elements (one for each congruence class mod n), so it is a subgroup of order n.

Similarly, if g has infinite order, then all the elements of $\langle g \rangle$ are distinct (since if $g^k = g^l$ then $g^{l-k} = e$), so $\langle g \rangle$ is an infinite subgroup.

Corollary 9.7 *Let g be an element in a finite group of order n. Then $g^n = e$.*

Proof The order of g cannot be infinite, since $\langle g \rangle$ is a finite set in this case. Suppose the order of g is m. Then the order of the subgroup $\langle g \rangle$ is m. By Lagrange's Theorem, m divides $n = |G|$.

Now we can revisit Fermat's Little Theorem and prove a stronger version.

Proposition 9.8 *Let n be a positive integer, and a an integer such that* $\gcd(a,n) = 1$. *Then $a^{\phi(n)} \equiv_n 1$, where ϕ is Euler's totient function.*

Proof Let U_n be the group of units of $\mathbb{Z}_n$. Then $|U_n| = \phi(n)$, and $[a]_n \in U_n$. By the preceding corollary, $[a^{\phi(n)}]_n = [a]_n^{\phi(n)} = [1]_n$; in other words, $a^{\phi(n)} \equiv_n 1$.

Example There are four units in $\mathbb{Z}_{12}$, namely $1, 5, 7, 11$. (We write a instead of $[a]_{12}$.) By the Corollary, if a is one of these four numbers, then $a^4 \equiv_{12} 1$. In fact, in this case $a^2 \equiv_{12} 1$ for each of the four numbers.

9.8 Cyclic groups

A group G is a *cyclic group* if $G = \langle g \rangle$ for some element $g \in G$.

The prototypical cyclic group of order n is $(\mathbb{Z}_n, +)$, while the prototypical infinite cyclic group is $(\mathbb{Z}, +)$. In each case, the group is generated by the element 1.

Proposition 9.9 *Any two cyclic groups of the same order are isomorphic.*

Proof We show that a cyclic group of order n is isomorphic to $\mathbb{Z}_n$, while an infinite cyclic group is isomorphic to $\mathbb{Z}$.

Let $G = \langle g \rangle$ be a cyclic group of order n. We saw in the last section that the element g has order n, and that $g^k = g^l$ if and only if $k \equiv_n l$. Now the map $[k]_n \mapsto g^k$ is well-defined and is one-to-one and onto, that is, a bijection, from $\mathbb{Z}_n$ to G; and it is an isomorphism, since

$$g^k \circ g^l = g^m \Leftrightarrow k + l \equiv_n m.$$

The proof for infinite groups is even simpler and is left to you.

Exercises

9.1 Show that, if $b \circ a = c \circ a$, then $b = c$.

9.2 Let G be a group of order n. Show that G is a cyclic group if and only if G contains an element whose order is n. Hence show that any group of prime order is cyclic.

9.3 Let G be a group of order 4; say $G = \{e, a, b, c\}$, where e is the identity. Suppose that G is **not** a cyclic group.

(a) Show that $a^2 = b^2 = c^2 = e$.

(b) Determine the Cayley table of G.

(c) Show that G is abelian.

Chapter 10

Permutations

We have seen rings and groups whose elements are numbers, polynomials, matrices, and sets. In this chapter we meet another type of object: permutations. The operation on permutations is composition, and we construct groups of permutations which play and important role in general group theory.

10.1 Definition and representation

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

In this section we consider only the case when X is a finite set, and we take X to be the set $\{1, 2, \ldots, n\}$ for convenience. As an example of a permutation, we will take $n = 8$ and let f be the function which maps $1 \mapsto 4$, $2 \mapsto 7$, $3 \mapsto 3$, $4 \mapsto 8$, $5 \mapsto 1$, $6 \mapsto 5$, $7 \mapsto 2$, and $8 \mapsto 6$.

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

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

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

of permutations is

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

This number is called $n!$ (read "n factorial" or "factorial n"), the product of the natural numbers from 1 to n. Thus we have proved:

Proposition 10.1 *The number of permutations of the set* $\{1, \ldots, n\}$ *is* $n!$.

10.2 The symmetric group

Let f_1 and f_2 be permutations. We define the *composition* of f_1 and f_2 to be the permutation obtained by applying f_1 and then f_2.

Warning If you write the image of x under the permutation f as $f(x)$, then the composition of f_1 and f_2 maps x to $f_2(f_1(x))$ – note the reversal! In order to make the notation work better, we change the way we write the image of x under f by putting f on the right, as xf (or sometimes up in the air, as x^f). Then we have $x(f_1 \circ f_2) = (xf_1)f_2$, which is easier to remember.

You should be aware, though, that some people choose to resolve the problem the other way, by defining the composition of f_1 and f_2 to be "first f_2, then f_1".

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

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

The easiest way to calculate $f \circ g$ is to take each of the numbers $1, \ldots, 8$, map it by f, map the result by g, and write down the result to get the bottom row of the two-line form for $f \circ g$. Thus, f maps 1 to 4, and g maps 4 to 8; so $f \circ g$ maps 1 to 8; f maps 2 to 7, and g maps 7 to 5, so $f \circ g$ maps 2 to 5; and so on.

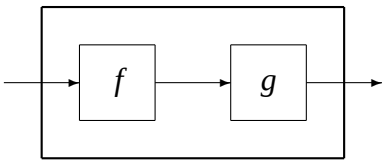
Another way to do it is to re-write the two-line form for g by shuffling the columns around so that the first row agrees with the second row of f. Then the second row will be the second row of $f \circ g$. Thus,

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

so

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

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 the first connected to the input of the second:



Now to calculate the result of applying $f \circ g$ to 1, we feed 1 into the input; the first black box outputs 4, which is input to the second black box, which outputs 8.

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

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

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

Given a permutation f, we define the *inverse permutation* of f to be the permutation which "puts everything back where it came from" – thus, if f maps x to y, then f^{-1} maps y to x. (This is just the inverse function as we defined it before.) It can be calculated directly from this rule. Another method is to take the two-line form for f, shuffle the columns so that the bottom row is $1\,2\,\ldots\,n$, and then interchanging 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 $f \circ f^{-1} = f^{-1} \circ f = e$.

Now you will not be surprised to learn:

Theorem 10.2 *The set of all permutations of $\{1,\ldots,n\}$, with the operation of composition, is a group.*

Proof The composition of two permutations is a permutation. The identity and inverse laws have just been verified above. So all we have to worry about is the associative law. We have

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

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

(Essentially, this last argument shows that the result of applying $f \circ g \circ h$, bracketed in any fashion, is "f, then g, then h".)

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

Proposition 10.3 *S_n is an abelian group if $n \leq 2$, and is non-abelian if $n \geq 3$.*

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

For $n \geq 3$, S_n contains elements f and g, where f interchanges 1 and 2 and fixes $3, \ldots, n$, and g interchanges 2 and 3 and fixed 1, $4 \ldots, n$. Now check that $f \circ g \neq g \circ f$. (For example, $f \circ g$ maps 1 to 3, but $g \circ f$ maps 1 to 2.)

10.3 Cycles

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

Let $a_1, a_2, \ldots, a_k$ be distinct numbers chosen from the set $\{1, 2, \ldots, n\}$. The *cycle* $(a_1, a_2, \ldots, a_k)$ denotes the permutation which maps $a_1 \mapsto a_2$, $a_2 \mapsto a_3$, ..., $a_{k-1} \mapsto a_k$, 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*, and define their *product* to be 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 $\{1, \ldots, n\}$. In a similar way, we define the product of any set of pairwise disjoint cycles.

Theorem 10.4 *Any permutation can be written as a product of disjoint cycles. The representation is unique, up to the facts that the cycles can be written in any order, and each cycle can be started at any point.*

Proof Our proof is an algorithm to find the *cycle decomposition* of a permutation. We will consider first our standard 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 $f : 3 \to 3$, so (3) is a cycle with a single element. Now we have the cycle decomposition:

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

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

The uniqueness is hopefully clear.

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!

We simplify the notation still further. Any element in a cycle of length 1 is fixed by the permutation, and by convention we do not bother writing such cycles. So our example permutation could be written simply as $f = (1,4,8,6,5)(2,7)$. 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 writing one cycle (1) to represent it, or by just calling it e.)

Cycle notation makes it easy to get some information about a permutation:

Proposition 10.5 *The order of a permutation is the least common multiple of the lengths of the cycles in its disjoint cycle representation.*

Proof Recall that the order of f is the smallest positive integer n such that $f^n = e$.

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

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

Now elements in the first cycle return to their starting position after 5 steps, and again after 10, 15, ... steps. So, if $f^n = 1$, 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 $f^n = 1$, then n must also be even. Hence if $f^n = 1$ 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.

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

10.4 Transpositions

A *transposition* is a permutation which swaps two elements i and j and fixes all the other elements of $\{1, \ldots, n\}$. In disjoint cycle form, a transposition looks like (i, j).

Theorem 10.6 *Any permutation in S_n can be written as a product of transpositions. The number of transpositions occurring in a product equal to a given element f is not always the same, but always has the same parity (even or odd) depending on g.*

Proof We begin by observing that

$$(1,2,\ldots,n) = (1,2)(1,3)\cdots(1,n).$$

For, in the product on the right,

- 1 is mapped to 2 by the first factor, and remains there afterwards;
- 2 is mapped to 1 by the first factor, then to 2 by the second, then stays there;

- ...
- $n-1$ is fixed by all factors until the second-last; it is mapped to 1 by the second-last factor and then to n by the last;
- n is fixed by all factors except the last, which takes it to 1.

So the two permutations are equal.

Now in exactly the same way, an arbitrary cycle $(a_1, a_2, \ldots, a_k)$ can be written as a product of transpositions:

$$(a_1, a_2, \ldots, a_k) = (a_1, a_2)(a_1, a_3)\cdots(a_1, a_k).$$

Finally, given an arbitrary permutation, write it in disjoint cycle form, and then write each cycle as a product of transpositions.

The statement about parity is harder to prove, and I have put the proof into an appendix.

Our standard example can be written

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

We call a permutation *even* or *odd* according as it is a product of an even or odd number of transpositions; we call this the *parity* of f. Notice that a cycle of length k is a product of $k-1$ transpositions. So, if the lengths of the cycles of f are $k_1, \ldots, k_r$ (including fixed points), then f is the product of

$$(k_1 - 1) + (k_2 - 1) + \cdots + (k_r - 1) = n - r$$

transpositions (since the cycle lengths add up to n). In other words, if we define $c(f)$ to be the number of cycles in the cycle decomposition of f, then the parity of f is the same as the parity of $n - c(f)$.

Theorem 10.7 *Suppose that $n \geq 2$. Then the set of even permutations in S_n is a subgroup of S_n having order $n!/2$ and index 2.*

Proof Let A_n be the set of even permutations in S_n. If $f_1, f_2 \in A_n$, then f_2^{-1} has the same cycle lengths as f_2 (since we just reverse all the cycles), so it is also in A_n. Thus, f_1 and f_2^{-1} are each products of an even number of transpositions; and then so, obviously, is $f_1 \circ f_2^{-1}$. By the Subgroup Test, A_n is a subgroup.

Let $\sim$ be the equivalence relation defined by this subgroup; that is, $f_1 \sim f_2$ if and only if $f_1 \circ f_2^{-1} \in A_n$. By considering each of f_1 and f_2 as products of transpositions, we see that $f_1 \sim f_2$ if and only if f_1 and f_2 have the same parity. So there are just two cosets of A_n.

By Lagrange's Theorem,

$$|A_n| = |S_n|/2 = n!/2.$$

The subgroup A_n consisting of even permutations is called the *alternating group* of degree n.

Example For $n = 3$, we have $|S_3| = 3! = 6$, so $|A_3| = 3$. The three even permutations are e, $(1,2,3)$ and $(1,3,2)$; the remaining three permutations are the transpositions $(1,2)$, $(1,3)$ and $(2,3)$ form the other coset of A_3 in S_3.

Remark The formula for a 3×3 determinant can be expressed as follows. For each permutation $f \in S_3$, we do the following. Pick the elements in row i and column if of the matrix, and multiply them together. That is, choose one term from each row and column in all possible ways. Now multiply the product by $+1$ if f is an even permutation, and by -1 if f is an odd permutation. Finally, add up these terms for all the permutations.

For example, if

$$A = \begin{pmatrix} a & b & c \\ l & m & n \\ p & q & r \end{pmatrix},$$

the terms are as follows:

Permutation	Product	Sign
e	amr	$+$
$(1,2,3)$	bnp	$+$
$(1,3,2)$	clq	$+$
$(1,2)$	blr	$-$
$(1,3)$	cmp	$-$
$(2,3)$	anq	$-$

So $\det(A) = amr + bnp + clq - blr - cmp - anq$.

Now exactly the same procedure defines the determinant of an $n \times n$ matrix, for any positive integer n. The drawback is that the number of terms needed for an $n \times n$ determinant is $n!$, a rapidly growing function; so the work required becomes unreasonable very quickly. This is not a practical way to compute determinants; but it is as good a definition as any!

10.5 Even and odd permutations

In this Appendix, we prove that the parity (even or odd) of a permutation does not depend on the way we write it as a product of transpositions. We will give two entirely different proofs.

First proof

For this proof, we see what happens when we multiply a permutation by a transposition. We find that the number of cycles changes by 1 (it may increase or decrease). There are two cases, depending on whether the two points transposed lie in different cycles or the same cycle of the permutation. So let f be a permutation and t a transposition.

Case 1: Transposing two points in different cycles. We may suppose that f contains two cycles $(a_1,\ldots,a_k)$ and $(b_1,\ldots,b_l)$, and that $t=(a_1,b_1)$ (this is because we can start each of the cycles at any point). Cycles of f not containing points moved by t will be unaffected. Now we find

$$f\circ t : a_1 \mapsto a_2 \mapsto \cdots \mapsto a_k \mapsto b_1 \mapsto b_2 \mapsto \cdots \mapsto b_l \mapsto a_1,$$

so the two cycles of f are "stitched together" into a single cycle in $f\circ t$, and the number of cycles decreases by 1.

Case 2: Transposing two points in the same cycle. This time let $(a_1,\ldots,a_m,\ldots,a_k)$ be a cycle of f, and assume that $t=(a_1,a_m)$, where $1<m\leq k$. This time

$$\begin{aligned} f\circ t : \quad & a_1 \mapsto a_2 \mapsto \cdots \mapsto a_{m-1} \mapsto a_1 \\ & a_m \mapsto a_{m+1} \mapsto \cdots \mapsto a_k \mapsto a_m \end{aligned}$$

so the single cycle of f is "cut apart" into two cycles.

Now any permutation f can be written as

$$f = t_1\circ t_2\circ\cdots\circ t_s,$$

where $t_1,\ldots,f_s$ are transpositions. Let f_i be the product of the first i of the transpositions, and consider the quantity $n-c(f_i)$, where $c(f)$ denotes the number of cycles of f (including fixed points). We start with $f_0=e$, having n fixed points, so $n-c(f_0)=0$. Now, at each step, we multiply by a transposition, so we change $c(f_i)$ by one, and hence change $n-c(f_i)$ by one. So the final value $n-c(f)$ is even or odd depending on whether the number s of transpositions is even or odd. But $n-c(f)$ is defined just by the cycle decomposition of f, independent of how we express it as a product of transpositions. So in any such expression, the parity of the number of transpositions will be the same.

Second proof

Let $x_1, \ldots, x_n$ be n indeterminates, and consider the function

$$F(x_1, \ldots, x_n) = \prod_{i<j} (x_j - x_i).$$

For example, for $n = 3$, we have

$$F(x_1, x_2, x_3) = (x_2 - x_1)(x_3 - x_1)(x_3 - x_2).$$

Given a permutation f, we define a new function F^f of the same indeterminates by applying the permutation f to their indices:

$$F^f(x_1, \ldots, x_n) = \prod_{i<j} (x_{jf} - x_{if}).$$

For example, if $n = 3$ and $f = (2,3)$, then

$$F^{(1,2)}(x_1, x_2, x_3) = (x_3 - x_1)(x_2 - x_1)(x_2 - x_3) = -F(x_1, x_2, x_3).$$

The result of applying f_1 and then f_2 to F is just the result of applying $f_2 \circ f_1$ to F, as you may check. We show that, for any transposition t, we have

$$F^t(x_1, \ldots, x_n) = -F(x_1, \ldots, x_n).$$

It will follow that, if f is expressed as the product of s transpositions, then

$$F^f(x_1, \ldots, x_n) = (-1)^s F(x_1, \ldots, x_n).$$

Since the value of F^f does not depend on which expression as a product of transpositions we use, we see that $(-1)^s$ must be the same for all such expressions for f, and hence the number of transpositions in the product must always have the same parity, as required.

To prove our claim, take the transposition $t = (k, l)$, where $k < l$, and see what it does to F. We look at the bracketed terms $(x_j - x_i)$ and see what happens to them. There are several cases.

- If $\{k, l\} \cap \{i, j\} = \emptyset$, then the term is unaffected by the permutation t.
- If $i < k$, then the terms $(x_k - x_i)$ and $(x_l - x_i)$ are interchanged, and there is no effect on F.
- If $k < i < l$, then the term $(x_i - x_k)$ goes to $(x_i - x_l) = -(x_l - x_i)$, and the term $(x_l - x_i)$ goes to $(x_k - x_i) = -(x_i - x_k)$; the two sign changes cancel out.

- If $i > l$, then the terms $(x_i - x_k)$ and $(x_i - x_l)$ are interchanged, and there is no effect on F.
- Finally, the term $(x_j - x_i)$ is mapped to $(x_i - x_j) = -(x_j - x_i)$.

So the overall effect of t is to introduce one minus sign, and we conclude that $F^t = -F$, as required.

Exercises

10.1 Let $g = (1,5,4,9,6,3)(7,8)$ and $h = (1,4,3)(6,8,7)(5,9,2)$ be permutations in the symmetric group S_9. Find $g \circ h$, g^2, g^{-1}, and $g^{-1} \circ h \circ g$. Show that h and $g^{-1} \circ h \circ g$ have the same order.

10.2 If g and h are elements of any group, show that

$$(g^{-1} \circ h \circ g)^n = g^{-1} \circ h^n \circ g$$

for any integer n, and deduce that h and $g^{-1} \circ h \circ g$ necessarily have the same order.

10.3 List the elements of S_4, and say whether each is an even or odd permutation.

Index

Abel, Niels, 79
abelian group, 79
addition, 69
additive group, 80
Al-Khwarizmi, iii
alternating group, 96
Argand diagram, 21
associative law, 16, 70, 73, 76, 79

bijective, 41, 42
blackboard bold, 15
Boolean ring, 77
Borel, Émil, 16

cancellation law, 72, 80
Carroll, Lewis, 12
Cartesian product, 36
Cayley table, 81
Cayley, Arthur, 81
closure law, 70, 79
commutative group, 79
commutative law, 16, 17, 70, 79
commutative ring, 70
commutative ring with identity, 61
complement, 33
complex numbers, 19, 65
complex plane, 21
composition
    of permutations, 90
congruence, 55, 63
conjecture, 11
converse, 12
coordinates, 35
corollary, 10
coset, 84
coset representative, 84
counterexample, 7
    minimal, 10
CRI, 61
cycle notation, 92
cyclic group, 87

De Moivre's Theorem, 23
definition, 12
degree
    of polynomial, 31
Descartes, René, 35
determinant, 96
difference
    of sets, 33
distributive law, 16, 17, 68, 70
divides, 48
division algorithm, 48
division ring, 70
division rule, 47, 51
dot diagrams, 17

equivalence relation, 38
Equivalence Relation Theorem, 39
equivalent conditions, 6
Euclid, 5
Euclid's algorithm, 49, 53
Euclid's Theorem, 5
Euler's function, 76
even permutation, 95
exponent laws, 85

Factor Theorem, 63

factorial, 90
Fermat's Last Theorem, 11
Fermat's Little Theorem, 59, 87
Fermat, Pierre, 11
field, 64, 70, 71
    finite, 66
finite field, 66
function, 40
Fundamental Theorem of Algebra, 46

Galois field, 67
Galois, Évariste, 67
gcd, 48, 53
general linear group, 81
Goldbach's conjecture, 11
greatest common divisor, 48, 53
group, 79
    alternating, 96
    cyclic, 87
    symmetric, 92
group of units, 80

identity law, 16, 70, 79
identity permutation, 91
if, 6
if and only if, 2, 6
implied by, 6
implies, 6
index, 84
induction, 7
injective, 41, 42
integers, 17
intersection, 32
inverse function, 42
inverse law, 17, 70, 79
inverse permutation, 91
isomorphic, 82, 87
isomorphism, 82

Johnson, Samuel, iii

Kronecker, Leopold, 16

Lagrange's Theorem, 84
laws, 16, 25
lcm, 49
least common multiple, 49
lemma, 10

matrices, 28
matrix addition, 28
matrix multiplication, 28
matrix ring, 71
minimal counterexample, 10
modular arithmetic, 57
monic, 63
monic polynomial, 52
multiplication, 69
multiplicative group, 80

natural numbers, 1, 16
necessary and sufficient condition, 6
necessary condition, 6

odd permutation, 95
one-to-one, 41, 42
one-to-one correspondence, 41
only if, 6
onto, 41, 42
operation, 43
order
    of element, 86
    of group, 84
    of permutation, 94
ordered pair, 35

parity
    of permutation, 95
partition, 38
permutation, 89
polynomial, 31, 51, 62
    monic, 63
polynomial ring, 71
prime numbers, 5
proof, 11

by contradiction, 3
by induction, 7
proposition, 10
Pythagoras, 3
Pythagoras' Theorem, 3

rational numbers, 3, 18
real numbers, 19
reflexive, 38
relations, 37
Remainder Theorem, 63
ring, 69
Boolean, 77
commutative with identity, 61
ring with identity, 70

scalar product, 27
sets, 32
subgroup, 82
generated by element, 86
subgroup test, 83
sufficient condition, 6
surjective, 41, 42
symmetric, 38
symmetric difference, 33
symmetric group, 92

theorem, 10
totient function, 76, 87
transitive, 38
transposition, 94
trigonometry, 23
two-line notation
for permutation, 89

union, 32
unit, 74
unordered pair, 35

vector, 25, 81
vector addition, 26
vector product, 27
Venn diagrams, 32

Weyl, Hermann, 74
Wiles, Andrew, 11
Wilson's Theorem, 59

zero law, 70