MA151: Algebra 1 2022/23

Contents

Chapter 1 - What is abstract algebra?

Chapter 2 - Sets and Binary operations

Chapter 3 - Groups

Chapter 4 - First theorems and notation

Chapter 5 - The order of an element

Chapter 6 - Subgroups

CHAPTER 17 - FACTORISATION IN RINGS

CHAPTER 18 - COSETS AND QUOTIENT RINGS

CHAPTER 19 - RING HOMOMORPHISMS

Many thanks to Samir Siksek for allowing me to use his notes as the basis for these course notes. Much of what is here was written by Samir. Please e-mail r.m.lissaman@warwick.ac.uk with details of any errors you notice. You are advised to download a new copy of these notes at intervals throughout the course as each version is subject to change, particularly because this is a brand new module this year. This will mainly be correcting errors and typos, but I also like to add detail every now and then depending on how much time we get to discuss things in lectures. Details of any substantial changes will be posted to the Moodle page for MA151.

Richard Lissaman

September 2022

Chapter 1 - What is abstract algebra?

In this course we'll be studying *algebraic structure*. Often a module like this has the words *abstract algebra* in the title. The main examples we'll see are *groups* and *rings*. Let's not worry about what these are just yet.

First of all, let's explore the words *abstract* and *algebra*. When we think of the word algebra we think of symbols replacing specific instances of numbers and manipulationg them to various ends.

1.1 Getting fussy about algebra

Think about some typical algebra that you may have seen at school/college.

$$\begin{aligned} x(x+y) &= yx \\ \Rightarrow x^2+xy-yx &= 0 \\ \Rightarrow x^2 &= 0 \\ \Rightarrow x &= 0 \end{aligned}$$

Let's think very precisely (much more precisely than usual) about which properties of numbers we are using here.

First of all we have used the fact that $x(x+y) = x^2+xy$ for any two numbers.

We've also used that facts that $xy = yx$, that $yx + (-yx) = 0$ and that $x^2 + 0 = x^2$.

Finally we've used the fact that the only number whose square is 0 is 0 itself when we've concluded that $x^2 = 0 \Rightarrow x = 0$.

All of these facts are certainly true if x and y are regular numbers, '+' means regular addition and xy mean $x \times y$ the regular multiplication of two numbers x and y. However some of these facts would not be true in other cases.

For example if x and y are both matrices and xy means matrix multiplication of x and y then we can't say that $xy = yx$.

Here is a case in point

$$\begin{pmatrix} 2 & 1 \\ 3 & 2 \end{pmatrix} \begin{pmatrix} 2 & 0 \\ 3 & 2 \end{pmatrix} \neq \begin{pmatrix} 2 & 0 \\ 3 & 2 \end{pmatrix} \begin{pmatrix} 2 & 1 \\ 3 & 2 \end{pmatrix}.$$

We can't say that if the square of a matrix is the zero matrix then that matrix itself must be the zero matrix since

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

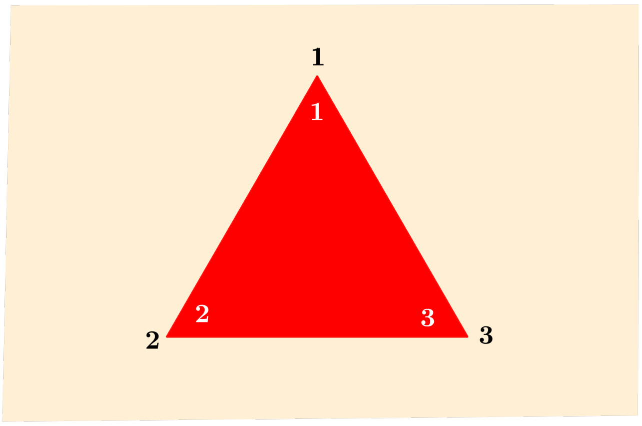
If x and y are vectors and xy means the vector product of x and y then we can't say that $xy = yx$. Don't worry if you have not met the vector product just yet.

In this module we'll be paying a lot of attention to the exact properties of operations like this. It will be important to be able to decide which objects and operations have which properties because we want to explore what happens when we abstract these from any particular instance of it. Probably this last sentence doesn't mean much to you now but it should mean more to you as the module goes on.

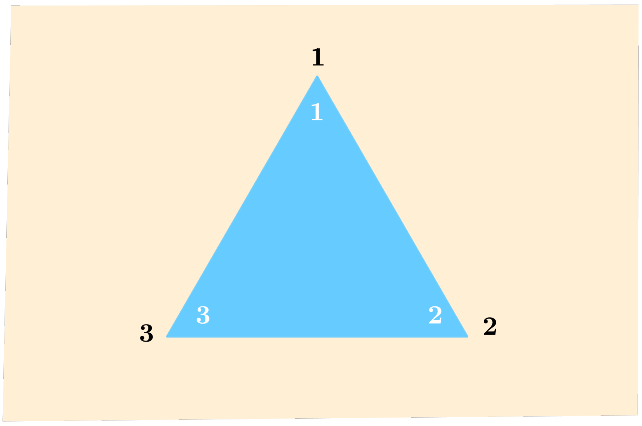
1.2 Algebra and symmetry

Sometimes, we can also think about the algebraic structures we'll be exploring geometrically. This approach reflects more closely the origins of the abstract algebra as a discipline in its own right. Here is an example, to give a flavour of this.

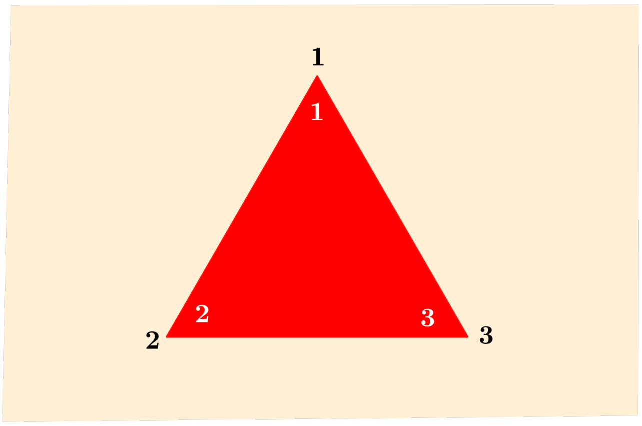
Imagine the following, an equilateral triangle made of red paper fitting exactly into a yellow paper frame.



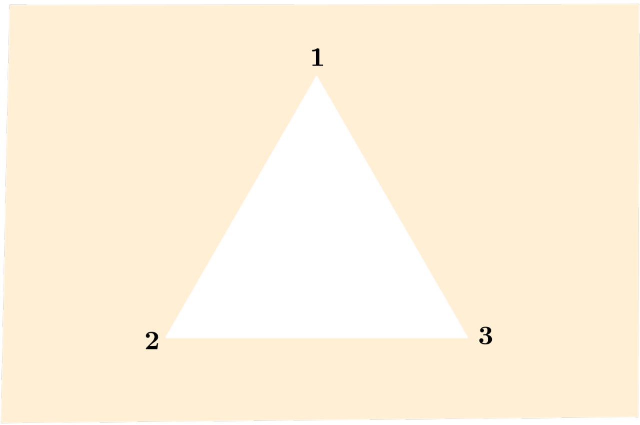
The back of the triangle is blue, from the back it looks like this (notice the 2 and 3 are swapped on this reverse view).



Let's turn it back over again...

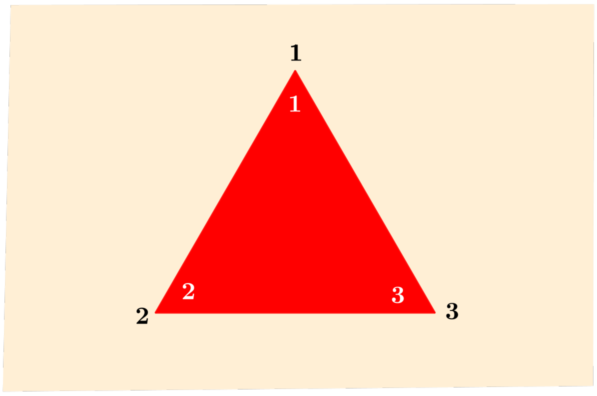


and lift the triangle out of its frame.

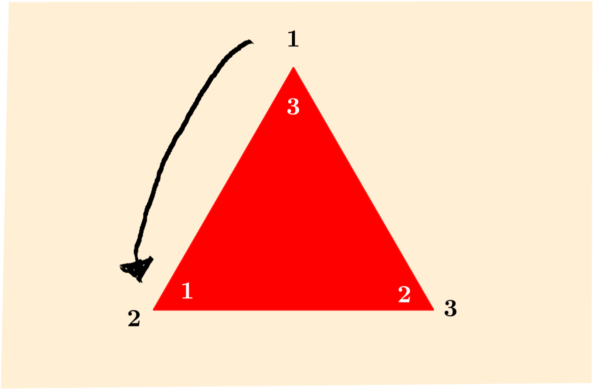


How many ways can we put the triangle back into its frame so that it occupies the frame exactly? Let's count them.

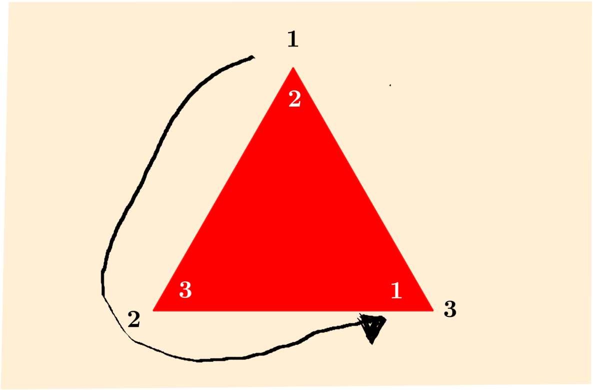
1. We could put it straight back down, exactly in the position it was in when we picked it up:



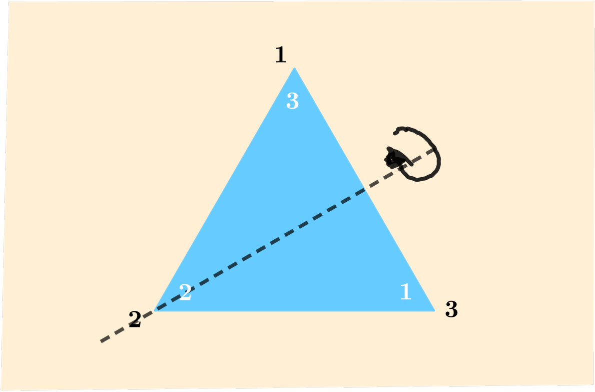
2. We could rotate it by $\frac{1}{3}$ of a turn anticlockwise.



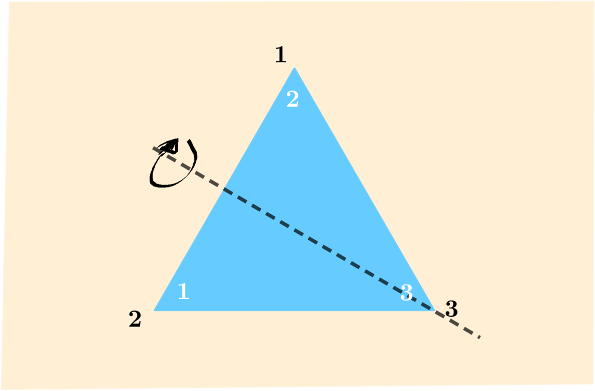
3. We could rotate it by $\frac{2}{3}$ of a turn anti-clockwise.



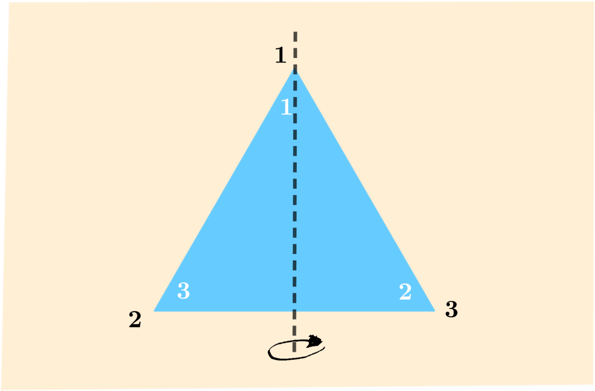
4. We could reflect it in this line (you might prefer to think of this as a 180 degree rotation about this axis given by the line). This means the triangle will be the other way up afterwards and we'll see the blue side.



5. or about this line/axis



6. or about this one



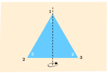
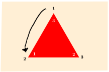
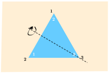
These six are all the possibilities. Here is an argument as to why there can be no more than 6.

For any 'way' the corner labelled 1 can end in one of three 'frame corners'. After that the corner labelled 2 can end in either of the two remaining frame corners. After that the corner labelled 3 will end in the other frame corner.

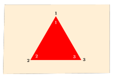
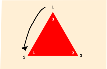
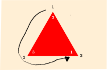
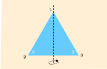
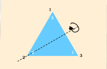
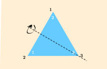
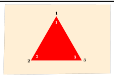
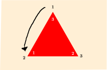
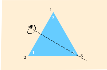
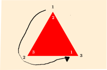
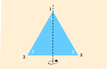
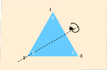
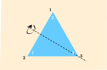
This means there are at most $3 \times 2 \times 1 = 6$ possible 'ways'. Since we have already found six different 'ways' these must be all six of them.

At this point it's probably worth pointing out the the 'ways' are called 'symmetries'. More precisely these are the symmetries of an equilateral triangle.

Finally, and very importantly, think about what would happen if you pick the triangle up from its orginal position, move it according to on of the 'ways', then move the new triangle according to another (possibly the same) 'way'. The triangle is occupying its frame after these two moves so what you have done must be the same as one of the six 'ways'.

Convince yourself that doing this to the triangle  and then doing this to the new triangle  is the same as just doing this to the original triangle .

Considering this for all possible choices of 'do thing one' then 'do thing two' we can start to fill in this table.

		Do this first					
	∘						
Do this second							
							
							
							
							
							

The complete table looks like this.

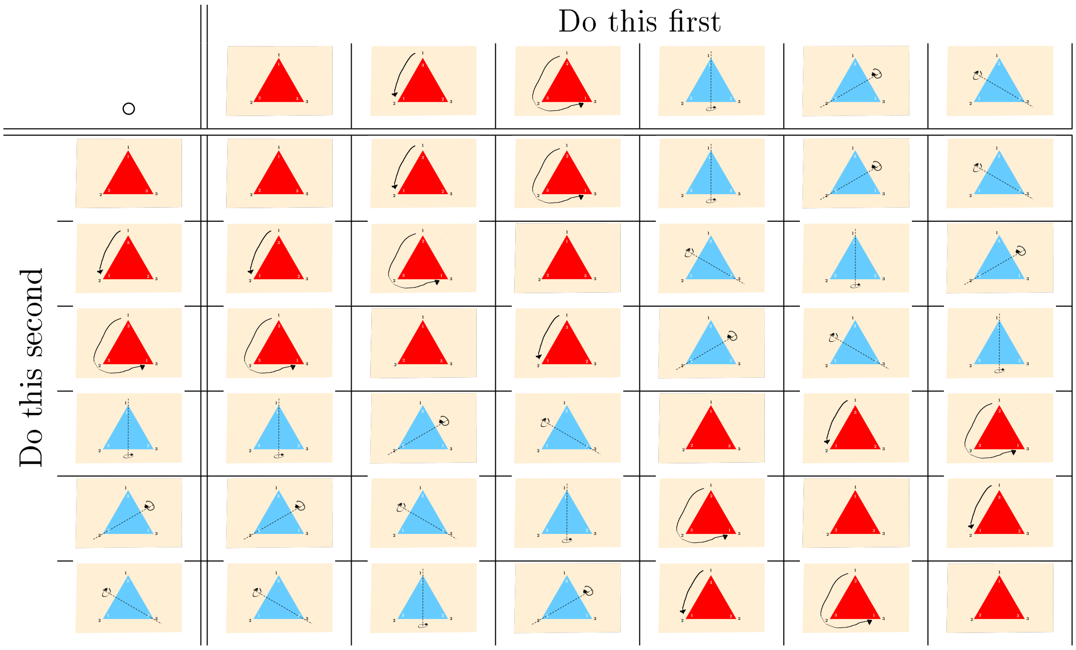


Figure 1: The six symmetries of an equilateral triangle and how any two of them 'combine'.

What did we do?

We took a mathematical object (an equilateral triangle) and we looked at transformations of it that preserve some property (sitting in the frame) and which were 'undo-able' (i.e. we could apply some other transformation to get the object back to its orginal state).

This is actually our first example of a group! This one is called D_6, the *dihedral group* of symmetries of an equilateral triangle.

Later we'll see that groups provide a way to study situations like this in an abstract sense.

Chapter 2 - Sets and Binary operations

2.1 What is a set?

2.1.1 Definition A *set* is simply a collection of objects.

We use curly brackets to denote sets. For example, if we write

$$A = \{2, 5, 13\},$$

then we're saying that the set A consists of the elements 2, 5, 13. This is one way of specifying a set; we simply list all its elements between curly brackets. The notation $x \in S$ means *x is a member of the set S* and the notation $x \notin S$ means *x is not a member of the set S*. For the set A above, we know $13 \in A$ but $11 \notin A$. We can also specify some infinite sets in this fashion; for example, the set of all integers

$$\mathbb{Z} = \{\dots, -3, -2, -1, 0, 1, 2, 3, \dots\}.$$

This is absolutely standard notation: when you see $\mathbb{Z}$, you're expected to know that it's the set of integers. The set of natural numbers is

$$\mathbb{N} = \{0, 1, 2, 3, 4 \dots\}.$$

Again this is standard notation (but not all mathematicians include 0 in the natural numbers).i

Here is an example of another way of specifying a set:

$$B = \{x \in \mathbb{Z} \mid x^2 = 16\}.$$

This is saying that B is the set of all integers x satisfying the equation $x^2 = 16$. Of course, another way of specifying the same set would be to write $B = \{-4, 4\}$.

If we write

$$C = \{x \in \mathbb{N} \mid x^2 = 16\},$$

then $C = \{4\}$.

To get more practice with this notation, observe that another way of specifying the natural numbers is to write

$$\mathbb{N} = \{x \in \mathbb{Z} \ \mid \ x \geq 0\}.$$

Yet another correct—although admittedly silly way—is to write

$$\mathbb{N} = \{x \in \mathbb{Z} \ \mid \ x \geq -0.5\}.$$

2.2 The empty set

2.2.1 Definition The *empty set* is the set containing no objects. It is denoted by $\emptyset$.

If we write

$$D = \{u \in \mathbb{Z} \mid u^3 = 2\},$$

then D is the set of integers u satisfying $u^3 = 2$. There are no integers satisfying this equation, so D is the empty set. We denote the empty set by $\emptyset$, so we can write $D = \emptyset$.

Here are a couple more examples of empty sets:

$$\{w \in \mathbb{N} \mid w \leq -1\} = \emptyset, \qquad \{v \in \mathbb{Z} \mid 3.01 \leq v \leq 3.99\} = \emptyset.$$

2.3 More sets (and more notation)

Here are some other sets you need to know:

1. $\mathbb{Q}$ is the set of *rational numbers.* We can write this as

   $$\mathbb{Q} = \left\{ \frac{a}{b} \mid a, b \in \mathbb{Z}, \quad b \neq 0 \right\}.$$

   Examples of elements of $\mathbb{Q}$ are 0, 5, $-7/11$, $3/2$, $6/4$ (the last two being the same element). From *Foundations* you should/will know that $\sqrt{2}$ is irrational. You can write this statement in set notation: $\sqrt{2} \notin \mathbb{Q}$. Other examples of irrational numbers are e and π.

2. $\mathbb{R}$ is the set of *real numbers.* It isn't possible to write $\mathbb{R}$ in straightforward way as for the sets above, but you can think of the elements of $\mathbb{R}$ as points on the real line. Examples of elements of $\mathbb{R}$ are -7, $3/5$, 3.85, $\sqrt{7}$, $(\pi + 1)/2$, $\sin 5$.

3. $\mathbb{C}$ is the set of *complex numbers.* You have seen complex numbers in your *Further Mathematics (or equivalent)* A-Level.

   Recall that i is a symbol that satisfies $i^2 = -1$. We can write the set of complex numbers as

   $$\mathbb{C} = \{ \ a + bi \mid a, b \in \mathbb{R} \ \}.$$

4. We define the set of n-dimensional points/vectors with real coordinates/components as

$$\mathbb{R}^n = \{(x_1, x_2, \dots, x_n) \mid x_1, x_2, \dots, x_n \in \mathbb{R}\}.$$

Thus $\mathbb{R}^2$ is the set of points/vectors in the plane, and $\mathbb{R}^3$ is the set of points/vectors in 3- dimensional space.

Notice that in the above notation vectors/points are written as rows, for example $(1, 2) \in \mathbb{R}^2$.

You might be more used to writing vectors as columns, so you might write

$$\begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}$$

instead of $(x_1, x_2, \dots, x_n)$. For example,

$$\begin{pmatrix} 1 \\ 2 \end{pmatrix}$$

instead of $(1, 2)$.

Just like $\mathbb{Q}$, $\mathbb{R}$ and $\mathbb{C}$ where we have the four basic arithmetic operations, $\mathbb{R}^n$ has some additional structure defined on it. *Vector addition* is defined by

$$(x_1, x_2, \dots, x_n) + (y_1, y_2, \dots, y_n) = (x_1 + y_1, x_2, +y_2, \dots, x_n + y_n)$$

or, in column notation

$$\begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} + \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} x_1 + y_1 \\ x_2 + y_2 \\ \vdots \\ x_n + y_n \end{pmatrix}$$

For example, in $\mathbb{R}^3$,

$$(2, 3, -4) + (2, 1, 0) = (4, 4, -4)$$

or, in column notation,

$$\begin{pmatrix} 2 \\ 3 \\ -4 \end{pmatrix} + \begin{pmatrix} 2 \\ 1 \\ 0 \end{pmatrix} = \begin{pmatrix} 4 \\ 4 \\ -4 \end{pmatrix}$$

Scalar multiplication is defined as follows. If λ is a scalar (i.e. $\lambda \in \mathbb{R}$) and $\mathbf{x} = (x_1, x_2, \ldots, x_n) \in \mathbb{R}^n$ is a vector, we define

$$\lambda\mathbf{x} = \lambda(x_1, x_2, \ldots, x_n) = (\lambda x_1, \lambda x_2, \ldots, \lambda x_n).$$

For example, in $\mathbb{R}^3$, if $\lambda = 3$ and $\mathbf{x} = (1, 2, 3)$ then

$$3(1, 2, 3) = (3, 6, 9).$$

You learn in *Algebra II* about how these two operations give $\mathbb{R}^n$ the structure of a *vector space* over $\mathbb{R}$ ('over $\mathbb{R}$' just means that, in this case, the scalars are real numbers).

2.4 What is a binary operation?

2.4.1 Definition Let S be a set. A *binary operation* on S is a rule by which any two elements of S can be combined to give another element of S. $\Diamond$

We are going to use the symbol $\star$ for binary operations. It's use will mostly be reserved for when we are talking about a general, non-specified, binary operation.

So given $s_1, s_2 \in S$ we have a further element $s_1 \star s_2 \in S$.

2.4.2 Examples

1. Addition is a binary operation on $\mathbb{R}$, because given any two real numbers, their sum is a real number. One way mathematicians like to say this is, "$\mathbb{R}$ *is closed under addition*". All that means is that the sum of two real numbers is a real number. $\Diamond$

2. Addition is also a binary operation on $\mathbb{C}$, $\mathbb{Q}$, $\mathbb{Z}$ and $\mathbb{N}$. Likewise, multiplication is a binary operation on $\mathbb{N}$, $\mathbb{Z}$, $\mathbb{Q}$, $\mathbb{R}$, $\mathbb{C}$. $\Diamond$

3. Is subtraction a binary operation? This question does not make sense because we haven't specified the set. Subtraction is a binary operation on $\mathbb{Z}$, $\mathbb{Q}$, $\mathbb{R}$, $\mathbb{C}$. Subtraction is not a binary operation on $\mathbb{N}$; for example $1,\ 2 \in \mathbb{N}$ but $1 - 2 = -1 \notin \mathbb{N}$. Thus $\mathbb{N}$ *is not closed under subtraction.* $\Diamond$

4. Is division a binary operation on $\mathbb{R}$? No, because 1, 0 are real numbers but $1/0$ is not defined. Thus $\mathbb{R}$ *is not closed under division.* $\Diamond$

5. Let us define $\mathbb{R}^*$ to be the set of non-zero real numbers:

$$\mathbb{R}^* = \{\ x \in \mathbb{R} \mid x \neq 0\ \}.$$

Now division is a binary operation on $\mathbb{R}^*$. But notice that addition is no longer a binary operation on $\mathbb{R}^*$; for example $5,\ -5 \in \mathbb{R}^*$ but $5 + (-5) = 0 \notin \mathbb{R}^*$. $\Diamond$

2.5 Operations on vectors

Recall

$$\mathbb{R}^n = \{(x_1, x_2, \ldots, x_n) \mid x_1, x_2, \ldots, x_n \in \mathbb{R}\}.$$

and that in the above notation vectors are written as rows, for example $(1, 2) \in \mathbb{R}^2$ but they can equally be written as columns:

$$\begin{pmatrix} 1 \\ 2 \end{pmatrix}.$$

Vector addition defined by

$$(x_1, x_2, \ldots, x_n) + (y_1, y_2, \ldots, y_n) = (x_1 + y_1, x_2, +y_2, \ldots, x_n + y_n)$$

is then a binary operation on $\mathbb{R}^n$, the result of adding two vectors in $\mathbb{R}^n$ is another vector in $\mathbb{R}^n$.

Vector subtraction, given by

$$(x_1, x_2, \ldots, x_n) - (y_1, y_2, \ldots, y_n) = (x_1 - y_1, x_2 - y_2, \ldots, x_n - y_n),$$

is another binary operation.

What about multiplication by a scalar? Recall that if λ is a scalar (i.e. $\lambda \in \mathbb{R}$) and $\mathbf{x} = (x_1, x_2, \ldots, x_n) \in \mathbb{R}^n$ is a vector, we define

$$\lambda\mathbf{x} = \lambda(x_1, x_2, \ldots, x_n) = (\lambda x_1, \lambda x_2, \ldots, \lambda x_n).$$

Notice that the result is in $\mathbb{R}^n$, but still multiplication by a scalar is *not* a binary operation on $\mathbb{R}^n$, because we're not 'combining' two elements of $\mathbb{R}^n$, but one element of $\mathbb{R}$ which is λ, and one element of $\mathbb{R}^n$ which is $\mathbf{x}$.

What about the dot product? The dot product is defined on $\mathbb{R}^n$ for all n. If $\mathbf{x} = (x_1, \ldots, x_n)$ and $\mathbf{y} = (y_1, \ldots, y_n)$ we define their dot product to be

$$\mathbf{x} \cdot \mathbf{y} = x_1y_1 + x_2y_2 + \cdots + x_ny_n.$$

Notice that the result is in $\mathbb{R}$, not $\mathbb{R}^n$, so the dot product is not a binary operation.

What about the cross product (also known as the vector product)? It is defined on $\mathbb{R}^3$ only as follows. Mixing row and column notation (!), if $\mathbf{x} = (x_1, x_2, x_3)$ and $\mathbf{y} = (y_1, y_2, y_3) \in \mathbb{R}^3$ then

$$\mathbf{x} \times \mathbf{y} = \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} \times \begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix} = \begin{pmatrix} x_2y_3 - y_2x_3 \\ -(x_1y_3 - y_1x_3) \\ x_1y_2 - y1x_2 \end{pmatrix}$$

So, if $\mathbf{x}$, $\mathbf{y} \in \mathbb{R}^3$ then $\mathbf{x} \times \mathbf{y}$ is again in $\mathbb{R}^3$. This means that the cross product is a binary operation on $\mathbb{R}^3$.

2.6 Operations on matrices

2.6.1 Definition $M_{m\times n}(\mathbb{R})$ is the set of $m \times n$ matrices with entries in $\mathbb{R}$. We similarly define $M_{m\times n}(\mathbb{C})$, $M_{m\times n}(\mathbb{Q})$, $M_{m\times n}(\mathbb{Z})$, etc. ♢

2.6.2 Example

$$M_{2\times 2}(\mathbb{R}) = \left\{ \begin{pmatrix} a & b \\ c & d \end{pmatrix} \mid a,\, b,\, c,\, d \in \mathbb{R} \right\}.$$

♢

Matrix addition and subtraction are binary operations on the set $M_{m\times n}(\mathbb{R})$.

Matrix multiplication is a binary operation on the set $M_{n\times n}(\mathbb{R})$, note that square matrices are needed here so that they can be multiplied together with the result being the same size square matrix (see the *Notes on Matrices* document on the module Moodle page for more details).

2.7 Operations on polynomials

2.7.1 Definition $\mathbb{R}[x]$ is the set of polynomials in x with real coefficients. Elements of $\mathbb{R}[x]$ are polynomials and so have the form

$$a_n x^n + a_{n_1} x^{n-1} + a_1 x + a_0$$

where $a_0, a_1, \ldots a_n$ are real numbers.

We can vary the set that we take the coefficients coefficients from to get other sets of polynomials.

$\mathbb{C}[x]$ is the set of polynomials in x with complex coefficients, $\mathbb{Q}[x]$ is the set of polynomials in x with rational coefficients, and $\mathbb{Z}[x]$ is the set of polynomials in x with integer coefficients. $\Diamond$

Notice that $Z[x] \subseteq \mathbb{Q}[x] \subseteq \mathbb{R}[x] \subseteq \mathbb{C}[x]$.

Addition, subtraction and multiplication of polynomials are defined as you would expect. For example in $\mathbb{Z}[x]$ we have

$$(4x^3 - 2x^2 + 3x + 6) + (x^2 - 10x + 3) = 4x^3 - x^2 - 7x + 9$$

$$(4x^3 - 2x^2 + 3x + 6) - (x^2 - 10x + 3) = 4x^3 - 3x^2 + 13x + 3$$

and

$$(4x^3 - 2x^2 + 3x + 6) \times (x^2 - 10x + 3) = 4\,x^5 - 42\,x^4 + 35\,x^3 - 30\,x^2 - 51\,x + 18$$

This addition, multiplication and subtraction each give a binary operation on each of $Z[z], \mathbb{Q}[x], \mathbb{R}[x], \mathbb{C}[x]$.

You'll be aware that we can create functions by dividing a polynomial by another, so called *rational functions.* For example $f : \mathbb{R} \to \mathbb{R}$ and $g : \mathbb{R} \setminus \{1\} \to \mathbb{R}$ given by

$$f(x) = \frac{x^3 + 3x + 1}{x^2 + 1}, \; g(x) = \frac{x + 2}{x - 1}.$$

However $\dfrac{x^3+3x+1}{x^2+1}$ cannot be written as a polynomial. So polynomial division is not a binary operation on any of $Z[z], \mathbb{Q}[x], \mathbb{R}[x], \mathbb{C}[x]$.

2.8 Multiplication tables

Recall our definition of a binary operation on a set S: it is simply a rule which for any pair of elements of S produces a third "output" element. This binary operation does not have to be 'natural', whatever that means. It does not have to be something we met before, like addition, multiplication etc. We can simply invent a set S and binary operation on it. If the S is finite, this is easy by means of a *multiplication table* which tells us for any pair of elements of S what the output element is.

Let $S = \{a, b, c\}$. Let $\star$ be the binary operation on S with the following mutiplication table:

$\star$	a	b	c
a	b	c	a
b	a	c	a
c	b	b	c

The result of the multiplication $a \star b$, is found at the intersection of the row headed by a with the column headed by b. In other words, for multiplication tables, the first element determines the row and the second determines the column. Thus for the multiplication table above,

$$a \star b = c, \qquad b \star a = a, \qquad c \star b = b, \qquad a \star a = b, \ldots .$$

You might think that this example is somewhat contrived, and you'd be right. But later on we'll meet more natural multiplication tables that arise from studying groups, permutations, etc. $\Diamond$

2.9 Commutativity and associativity

2.9.1 Definitions Let S be a set and $\star$ a binary operation.

We say that the binary operation $\star$ is *commutative on* S if $a \star b = b \star a$ for all a, $b \in S$.

We say that the binary operation $\star$ is *associative on* S if $(a \star b) \star c = a \star (b \star c)$ for all a, b, $c \in S$. $\Diamond$

Let's consider this for some of the binary operations we've met already:

2.9.2 Examples

1. Addition and multiplication on $\mathbb{R}$ (or $\mathbb{C}$ or $\mathbb{R}[x]$ or $\dots$) are both commutative and associative. When operations are commutative and associative, order and bracketing do not matter (though it's suprisingly tricky to give a formal proof of this, we'll not give one here):

   $$e+((c+b)+(d+a))=a+b+c+d+e, \qquad e\cdot((c\cdot b)\cdot(d\cdot a))=a\cdot b\cdot c\cdot d\cdot e.$$

   Of course subtraction is neither commutative nor associative (write some examples). $\Diamond$

2. Addition is commutative and associative on $\mathbb{R}^n$. The cross product is not commutative on $\mathbb{R}^3$. If you have met the cross product before you will probably know that if $\mathbf{x}$, $\mathbf{y} \in \mathbb{R}^3$ then

   $$\mathbf{y}\times\mathbf{x}=-\mathbf{x}\times\mathbf{y}.$$

   We say that the cross product is *anti-commutative.* $\Diamond$

3. Matrix addition is both commutative and associative on $M_{m\times n}(\mathbb{R})$.

   Matrix multiplication is an associative but not commutative on the set $M_{n\times n}(\mathbb{R})$ (see the *Notes on Matrices* document on the module Moodle page for more details).

4. Let $S=\{a,b,c\}$ and let $\star$ be the binary operation given by the composition table we looked at in section 2.8. Then $\star$ is not commutative; for example

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

   It is also not associative; for example

   $$(a\star b)\star c=c\star c=c, \qquad a\star(b\star c)=a\star a=b.$$

   $\Diamond$

5. When a binary operation is associative bracketing doesn't matter. For example,

   $$(a\star b)\star((c\star d)\star e)=(a\star(b\star c))\star(d\star e).$$

As long as we keep a, b, c, d, e in the same order from left to right, then the order in which we do the multiplications does not matter. Thus there would be no ambiguity in writing

$$(a \star b) \star ((c \star d) \star e) = a \star b \star c \star d \star e.$$

This fact that bracketing doesn't matter as long as we keep the same order is called the general associativity theorem. For a proper formulation and proof see

`https://proofwiki.org/wiki/General_Associativity_Theorem/Formulation_2/Proof_1` $\Diamond$

6. Are there binary operations that are commutative but not associative? Yes but it isn't easy to come up with 'natural' examples. However it is easy to invent a finite set and a composition table that is commutative but not associative. Let $S = \{a, b, c\}$. Let $\star$ be the binary operation on S with the following composition table:

$\star$	a	b	c
a	b	c	a
b	c	c	a
c	a	a	c

Note that $\star$ is commutative; you can see this by noting that the table is symmetric about the diagonal from the top left corner to the bottom right corner. But it isn't associative. For example,

$$(b \star c) \star a = a \star a = b, \qquad b \star (c \star a) = b \star a = c.$$

$\Diamond$

2.9.3 Exercise In the following, is $\circ$ a binary operation on A? If so, is it commutative? Is it associative? In each case justify your answer.

(a) $A = \mathbb{R}$ is the set of real numbers and $a \star b = a/b$.

(b) $A = \{1, 2, 3, 4, \dots\}$ is the set of positive integers and $a \star b = a^b$.

(c) $A = \{\dots, 1/8, 1/4, 1/2, 1, 2, 4, 8, \dots\}$ is the set of powers of 2 and $a \circ b = ab$.

(d) $A = \mathbb{C}$ is the set of complex numbers and $a \star b = |a - b|$.

Chapter 3 - Groups

3.1 The definition of a group

A *group* is a pair $(G, \star)$ where G is a set and $\star$ is a binary operation on G, such that the following four properties hold:

(i) (closure) for all a, $b \in G$, $a \star b \in G$;

(ii) (associativity) for all a, b, $c \in G$,

$$a \star (b \star c) = (a \star b) \star c;$$

(iii) (existence of the identity element) there is an element $e \in G$ such that for all $a \in G$,

$$a \star e = e \star a = a;$$

(iv) (existence of inverses) for every $a \in G$, there is an element $b \in G$ (called the inverse of a) such that

$$a \star b = b \star a = e.$$

◊

If $\star$ is a binary operation then (i) automatically holds. So why did is it listed in the definition? It's there for good measure! When you suspect an operation gives you a group the first thing you should check is that the operation is really a binary operation.

99% of mathematicians call (i)–(iv) the "group axioms"even through they are the "defining properties of a group" . This could be considered to be a bit odd, the word axiom is usually reserved for statements of 'universal truth'.

3.2 First examples (and non-examples)

3.2.1 Example $(\mathbb{R}, +)$ is a group. We know already that addition is a binary operation on $\mathbb{R}$, so 'closure' holds. We know addition of real numbers is associative. What is the identity element? We want an element $e \in \mathbb{R}$ so that $a+e = e+a = a$ for all $a \in \mathbb{R}$. It is clear that $e = 0$ works and is the only possible choice. Moreover, the (additive) inverse of a is $-a$: $a + (-a) = (-a) + a = 0$. ◊

3.2.2 Example Recall our definition of the natural numbers:

$$\mathbb{N} = \{0, 1, 2, \dots\}.$$

Is $(\mathbb{N}, +)$ a group? Conditions (i), (ii) are satisfied. For condition (iii) we can take the identity element to be 0 (again the only possible choice). But (iv) does not hold. For example, if we take $a = 1$, there is no $b \in \mathbb{N}$ such that $a + b = b + a = 0$. Thus $(\mathbb{N}, +)$ is not a group. $\Diamond$

3.2.3 Example $(\mathbb{Z}, +)$, $(\mathbb{Q}, +)$ and $(\mathbb{C}, +)$ are groups. $\Diamond$

3.2.4 Example Recall we defined

$$\mathbb{R}^* = \{\alpha \in \mathbb{R} : \alpha \neq 0\}.$$

Then $(\mathbb{R}^*, \cdot)$ is a group, where $\cdot$ means multiplication. Again closure and associativity are obvious. If e is the identity element then it has to satisfy $\alpha \cdot e = e \cdot \alpha = \alpha$ for all $\alpha \in \mathbb{R}$. Thus $e = 1$ and this is the only choice possible. Then the inverse of α is α^{-1}.

We can define $\mathbb{C}^*$ and $\mathbb{Q}^*$ in the same way and obtain groups $(\mathbb{C}^*, \cdot)$ and $(\mathbb{Q}^*, \cdot)$.

Can we obtain from $\mathbb{Z}$ a group with respect to multiplication? In view of the above, the obvious candidate is

$$U = \{\alpha \in \mathbb{Z} : \alpha \neq 0\}.$$

But $(U, \cdot)$ is not a group. It is true that (i), (ii) and (iii) hold with 1 being the identity element. But, for example, $2 \in U$ does not have an inverse: there is no $b \in U$ such that $b \cdot 2 = 2 \cdot b = 1$. So $(U, \cdot)$ is not a group. But the answer is not no; all we've done is shown that the obvious choice for a group $(\mathbb{Z}^*, \cdot)$ made up of integers does not work. We'll return to this question and answer it fully later. $\Diamond$

3.2.5 Example $(\mathbb{R}^2, +)$ is a group. Let's prove this. We're allowed to assume the usual properties of the real numbers (see Section 2.9.2). The elements of $\mathbb{R}^2$ are pairs (a_1, a_2) where a_1, a_2 are real numbers. Addition is defined by

$$(a_1, a_2) + (b_1, b_2) = (a_1 + b_1, a_2 + b_2).$$

Note that the entries $a_1 + b_1$ and $a_2 + b_2$ are real numbers, and so $(a_1 + b_1, a_2 + b_2)$ is a pair of real numbers. Hence $(a_1 + b_1, a_2 + b_2)$ is in $\mathbb{R}^2$. In other words, $\mathbb{R}^2$ is closed under addition, which shows that $(\mathbb{R}^2, +)$ satisfies condition (i). Next we want to prove associativity of addition. Consider $\mathbf{a}$, $\mathbf{b}$, $\mathbf{c}$ in $\mathbb{R}^2$. We can write

$$\mathbf{a} = (a_1, a_2), \qquad \mathbf{b} = (b_1, b_2), \qquad \mathbf{c} = (c_1, c_2).$$

Here a_1, a_2, b_1, b_2 and c_1, c_2 are real numbers. Note that

$$(\mathbf{a} + \mathbf{b}) + \mathbf{c} = \left((a_1 + b_1) + c_1, (a_2 + b_2) + c_2\right).$$

Likewise,

$$\mathbf{a} + (\mathbf{b} + \mathbf{c}) = \left(a_1 + (b_1 + c_1), a_2 + (b_2 + c_2)\right).$$

Because addition of real numbers is associative, we know that

$$(a_1 + b_1) + c_1 = a_1 + (b_1 + c_1), \qquad (a_2 + b_2) + c_2 = a_2 + (b_2 + c_2).$$

Hence

$$(\mathbf{a} + \mathbf{b}) + \mathbf{c} = \mathbf{a} + (\mathbf{b} + \mathbf{c}).$$

This shows that $(\mathbb{R}^2, +)$ satisfies (ii).

Next we need an identity element, and the obvious candidate is $\mathbf{0} = (0, 0)$. Then

$$(a_1, a_2) + (0, 0) = (a_1 + 0, a_2 + 0) = (a_1, a_2),$$

and

$$(0, 0) + (a_1, a_2) = (0 + a_1, 0 + a_2) = (a_1, a_2).$$

Thus (iii) is satisfied.

Finally we want an inverse. If $\mathbf{a} = (a_1, a_2)$ is in $\mathbb{R}^2$ then the inverse we choose (there's no other choice) is $\mathbf{b} = (-a_1, -a_2)$. This is in $\mathbb{R}^2$ and satisfies

$$\mathbf{a} + \mathbf{b} = \mathbf{b} + \mathbf{a} = (0, 0).$$

Hence (iv) is satisfied and so $(\mathbb{R}^2, +)$ is a group. $\Diamond$

What matters is that you realize that the properties of addition in $\mathbb{R}^2$ simply follow from the definition of addition in $\mathbb{R}$ and corresponding properties of the real numbers.

The proofs for the following examples 3.2.6, 3.2.7, and 3.2.8 are similar.

3.2.6 $(\mathbb{R}^n, +)$ is a group for any $n \geq 2$. $\Diamond$

3.2.7 $(\mathbb{R}[x], +)$ is a group. $\Diamond$

3.2.8 $(M_{m\times n}(K), +)$ are groups for $K = \mathbb{C}$, $\mathbb{R}$, $\mathbb{Q}$, $\mathbb{Z}$, with $0_{m\times n}$, the $m \times n$ matrix all of whose entries are 0, the identity element. $\Diamond$

3.2.9 All the groups we have met so far are infinite. Here is an example of a finite group. Let $A = \{+1, -1\}$. Then $(A, \cdot)$ is a group (where of course $\cdot$ is multiplication). $\Diamond$

3.2.10 Let $B = \{1, i, -1, -i\}$, where $i = \sqrt{-1}$. Then $(B, \cdot)$ is another example of a finite group. $\Diamond$

3.2.11 Let $C = \{1, i\}$. Then $(C, \cdot)$ is not a group since it isn't closed; for example $i \cdot i = -1 \notin C$. $\Diamond$

3.3 Abelian groups

3.3.1 Definition We say that a group $(G, \star)$ is *abelian* if (in addition to properties (i)–(iv) in definition 3.1) it also satisfies

(v) (commutativity) for all a, $b \in G$,

$$a \star b = b \star a.$$

$\Diamond$

All the groups we have seen above are actually abelian: $(\mathbb{R}, +)$, $(\mathbb{C}, +)$, $(\mathbb{Q}, +)$, $(\mathbb{R}[x], +)$, $(\mathbb{R}^n, +)$, $(\mathbb{R}^*, \cdot)$, $(\mathbb{C}^*, \cdot)$, $(M_{m\times n}(\mathbb{R}), +)$, ...

Are there any non-abelian groups? Our first example uses matrix multiplication as the binary operation.

3.4 A matrix group

We saw that $(M_{2\times 2}(\mathbb{R}), +)$ is a group. This in fact is **not** a particularly interesting group because addition of matrices is not a very interesting operation. Multiplication of matrices is a far more interesting and natural operation; as we saw, if A, B represent certain geometric operations (e.g. scaling, reflection, rotation, etc.) then BA is the operation that one obtains from doing A first then B; if this doesn't sound familiar look again at the *Notes on Matrices* on the module Moodle page. Can we obtain a group out of (say) 2×2 matrices under multiplication?

To answer, let's look back to Example 3.2.4. There we obtained a multiplicative group from the real numbers by removing 0. We had to remove 0 because it doesn't have a multiplicative inverse.

It will not be enough for us to exclude the zero matrix, because there are non-zero matrices that do not have an inverse (again see the *Notes on Matrices* on the module Moodle page for an example). What if we exclude all non-invertible matrices; do we get a group under multiplication?

3.4.1 Definition Define

$$\mathrm{GL}_2(\mathbb{R}) = \left\{ \begin{pmatrix} a & b \\ c & d \end{pmatrix} | a,\, b,\, c,\, d \in \mathbb{R} \text{ and } ad - bc \neq 0 \right\}.$$

$\diamondsuit$

Recall that $ad - bc$ is the determinant of the 2×2-matrix $\left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right)$, and the matrix is invertible if and only if this determinant is non-zero. So $\mathrm{GL}_2(\mathbb{R})$ contains all the invertible 2×2 matrices (with real entries) and none of the non-invertible ones.

3.4.2 Theorem $\mathrm{GL}_2(\mathbb{R})$ is group under multiplication of matrices. We call $\mathrm{GL}_2(\mathbb{R})$ the *general linear group.*

Proof. The first thing to check is that $\mathrm{GL}_2(\mathbb{R})$ is closed under multiplication. If A and B are in $\mathrm{GL}_2(\mathbb{R})$ then AB is a 2×2 matrix with real entries. Also, we know that $\det(AB) = \det(A)\det(B)$. Because A and B have non-zero determinants, so does AB. So AB is in $\mathrm{GL}_2(\mathbb{R})$.

We already know that matrix multiplication is associative.

The identity matrix

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

is in $\mathrm{GL}_2(\mathbb{R})$ (because it has non-zero determinant) and is the multiplicative identity element; it satisfies $AI_2 = I_2A = A$ for any 2×2 matrix A.

Finally, we should ask if every matrix in $\mathrm{GL}_2(\mathbb{R})$ has an inverse. We defined $\mathrm{GL}_2(\mathbb{R})$ so every element is invertible, but we need to make sure that the inverse is also in $\mathrm{GL}_2(\mathbb{R})$. If $A \in \mathrm{GL}_2(\mathbb{R})$ then $\det(A) \neq 0$. We know, since

$\det(A)\det(A^{-1}) = 1$ that $\det(A^{-1}) \neq 0$ and indeed

$$\det(A^{-1}) = \frac{1}{\det(A)}.$$

Moreover, A^{-1} is a 2×2 matrix with real entries. Hence $A^{-1} \in \mathrm{GL}_2(\mathbb{R})$. ◊

We can define $\mathrm{GL}_2(\mathbb{Q})$ and $\mathrm{GL}_2(\mathbb{C})$ in a similar way and show that they are groups. However, as this very important exercise shows we can't do this with the integers.

3.4.3 Exercise Show that

$$\left\{ \begin{pmatrix} a & b \\ c & d \end{pmatrix} | a,\, b,\, c,\, d \in \mathbb{Z} \text{ and } ad - bc \neq 0 \right\}$$

is **not** a group with respect to multiplication. ◊

You may remember D_6, the symmetries of an equliaterial triangle, from chapter 1. That is a non-abelian group.

In the next section we give another similar example of a non-abelian group, but this time, instead of an equilateral triangle, we'll look at the symmetries of a square.

3.5 D_8 the symmetries of a square

In many ways the examples above are misleading for three reasons:

- Most of the examples of groups we have met above have additional structure. For example, in $\mathbb{R}$ we can add, but we can also multiply and we can divide by non-zero numbers.

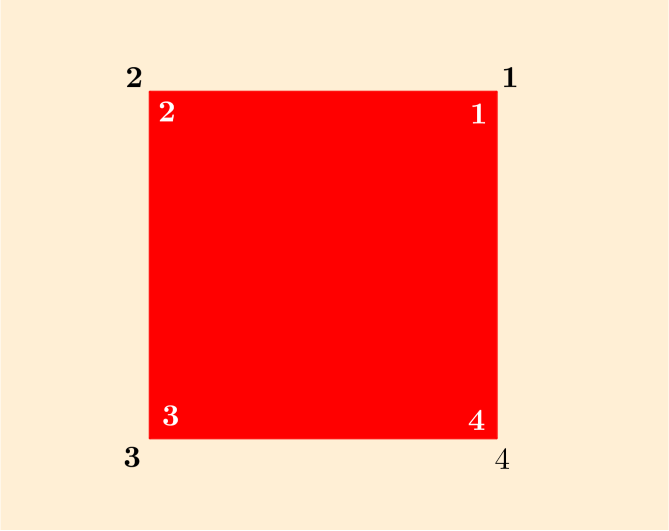
  In fact $\mathbb{R}$ is an example of a *field*. Like in $\mathbb{R}^2$ we have addition and scalar multiplication, so $\mathbb{R}^2$ is an example of a *vector space*. This doesn't stop $(\mathbb{R}, +)$ and $(\mathbb{R}^2, +)$ from being groups, but if you want to test your own ideas in group theory, it is best to also look at examples where there aren't any of these additional structures.

- Most of groups you've met so far are abelian, the exception is $\mathrm{GL}_2(\mathbb{R})$. The theory of abelian groups is rather close in flavour to linear algebra. Many of the most interesting groups that you'll come across during your degree will be non-abelian.

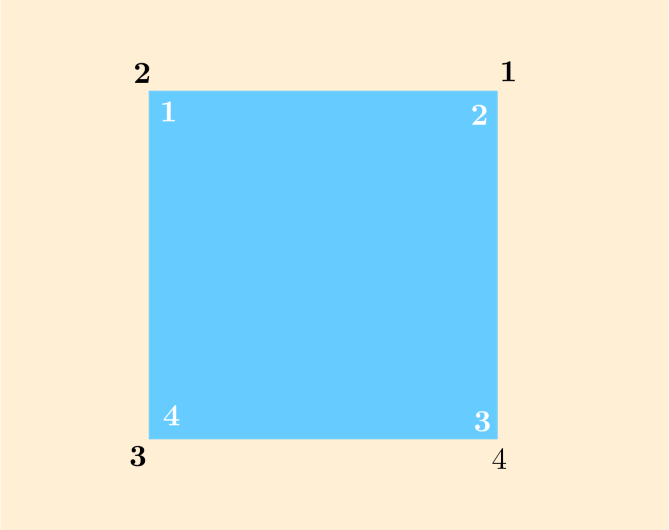
- All the groups above, except for Example 3.2.9 and Example 3.2.10, are infinite. Although infinite groups are important and interesting, some theorems we will do in this course will apply only to finite groups. Thus it is essential to become familiar with examples of finite groups.

Here is a great example of a group!

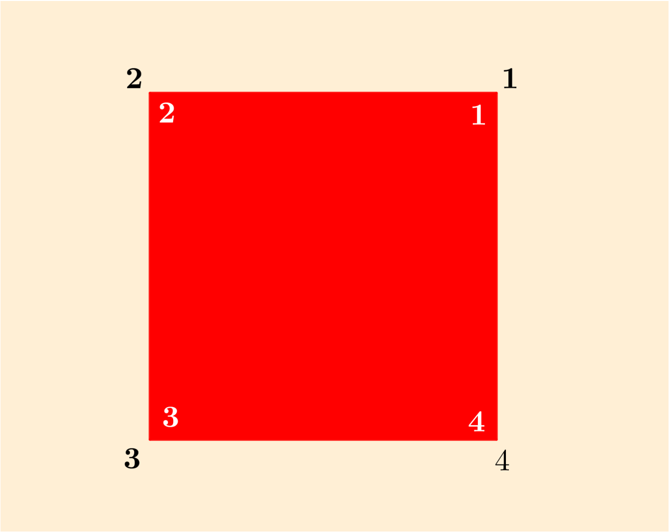
Imagine the following (this might ring a bell); a square made of red paper fitting exactly into a yellow paper frame.



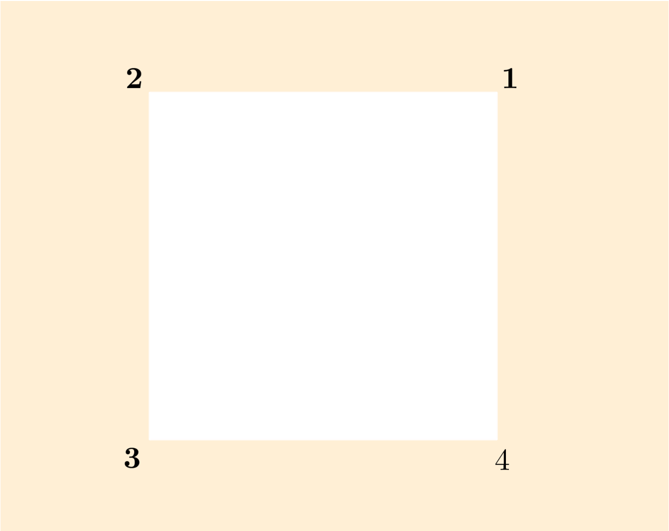
On the back it looks like this.



Let's turn it back over again.

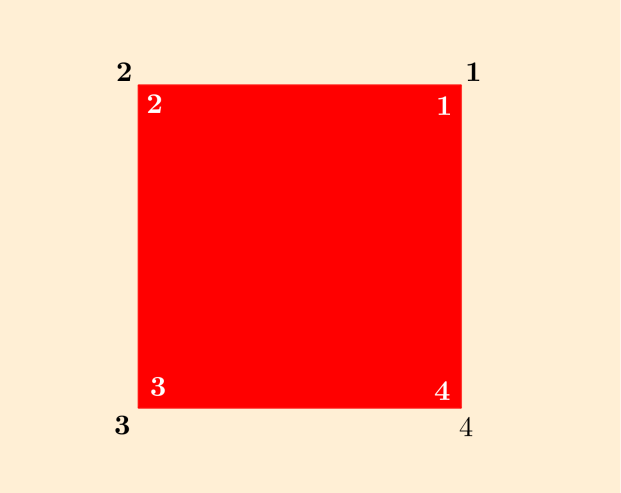


and lift the square out of its frame.

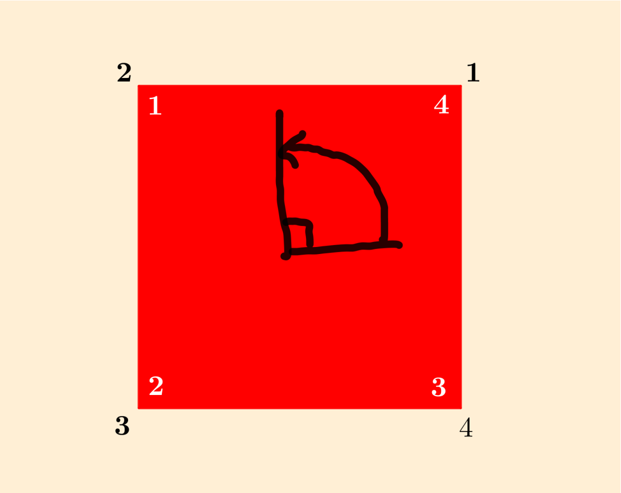


How many ways can we put the square back into its frame so that it occupies the frame exactly? Let's count them.

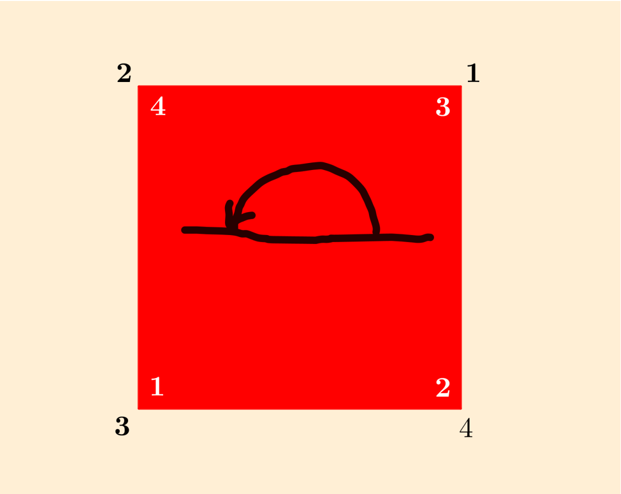
1. We could put it straight back down, exactly in the position it was in when we picked it up:



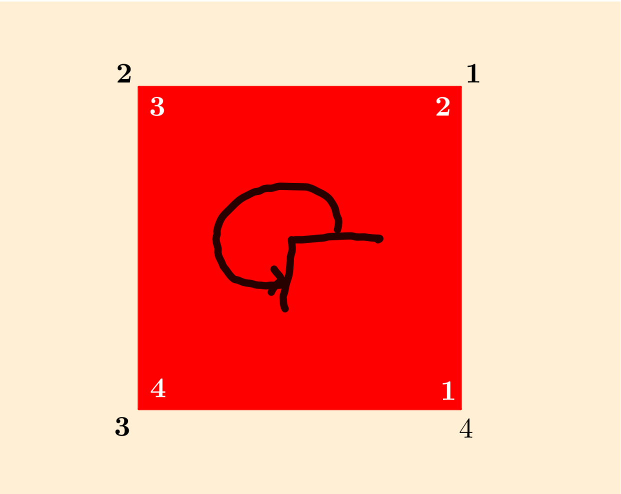
2. We could rotate it by $\frac{1}{4}$ of a turn anticlockwise.



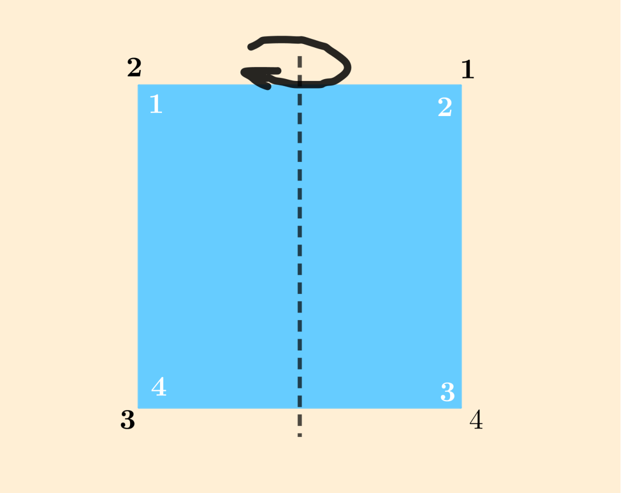
3. We could rotate it by $\frac{1}{2}$ of a turn anitclockwise.



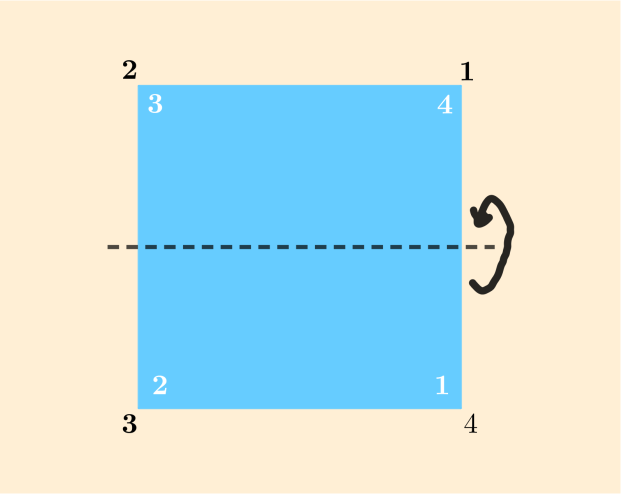
4. We could rotate it by $\frac{3}{4}$ of a turn anitclockwise.



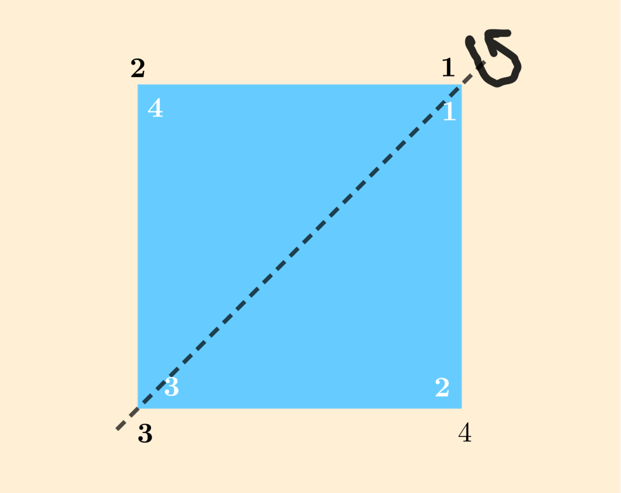
5. We could reflect it in this line (you might prefer to think of this as a 180 degree rotation about this axis given by the line). This means the square will be the other way up afterwards and we'll see the blue side.



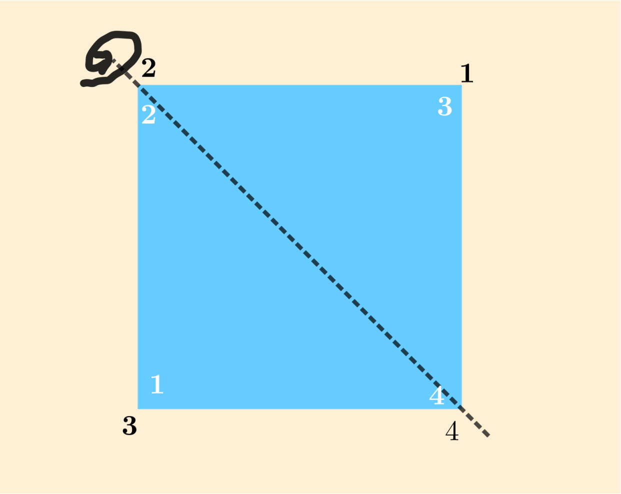
6. or about this line/axis



7. or about this one



8. or this one

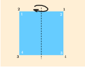
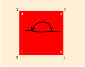
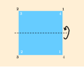


These eight are all the possibilities.

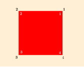
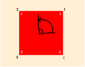
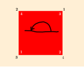
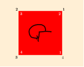
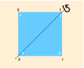
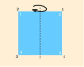
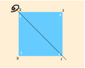
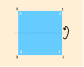
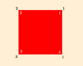
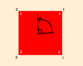
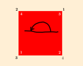
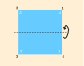
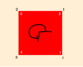
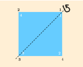
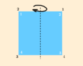
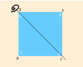
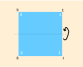
Here is an argument as to why there can be no more than 8 'ways'. For any 'way' the corner labelled 1 can end in one of four 'frame corners'. After that the corner labelled 2 can end in either of the two corners adjacent to the one the the corner labelled 1 landed in.After that the new positions of corner 3 and corner 4 are completely determined.

This means there are at most $4 \times 2 = 8$ possible 'ways'. Since we have already found eight different 'ways' these must be all eight of them.

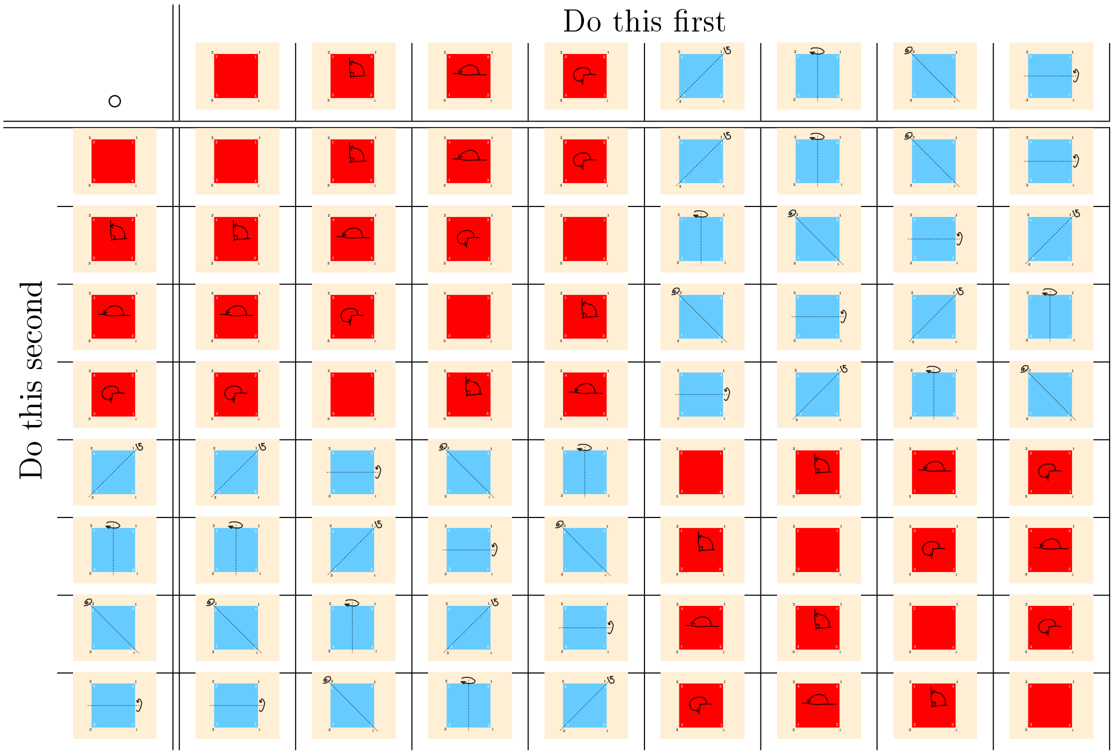
Finally, and very importantly, think about what would happen if you pick the square up from its orginal position, move it according to on of the 'ways', then move the newly positioned square according to another (possibly the same) 'way'. The square is occupying its frame after these two moves so what you have done must be the same as one of the eight 'ways'.

Convince yourself that doing this to the triangle  and then doing this to the new triangle  is the same as just doing this to the original triangle .

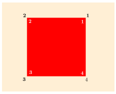
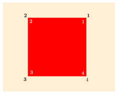
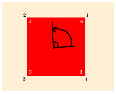
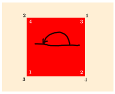
Considering this for all possible choices of 'do thing one' then 'do thing two' we can start to fill in this table.

		Do this first							
	○								
Do this second									
									
									
									
									
									
									
									

The complete table looks like this.



You may be thinking that at this point we really need some notation to save drawing the pictures representing the ’moves’ every time!

Let ρ_0, ρ_1, ρ_2, ρ_3 be anticlockwise rotations of the square about O by 0°, 90°, 180° and 270°. So these are , ,  and  respectively.

Let’s give the reflections names too. As in Figure 2:

- σ_0 the reflection about the diagonal joining the top-right vertex to the bottom-left vertex;
- σ_1 the reflection about the line joining the midpoint of top side and the midpoint of bottom side;
- σ_2 the reflection about the diagonal joining top-left vertex and the bottom-right vertex;
- σ_3 the reflection about the line joining the midpoint of the left side and the midpoint of the right side.

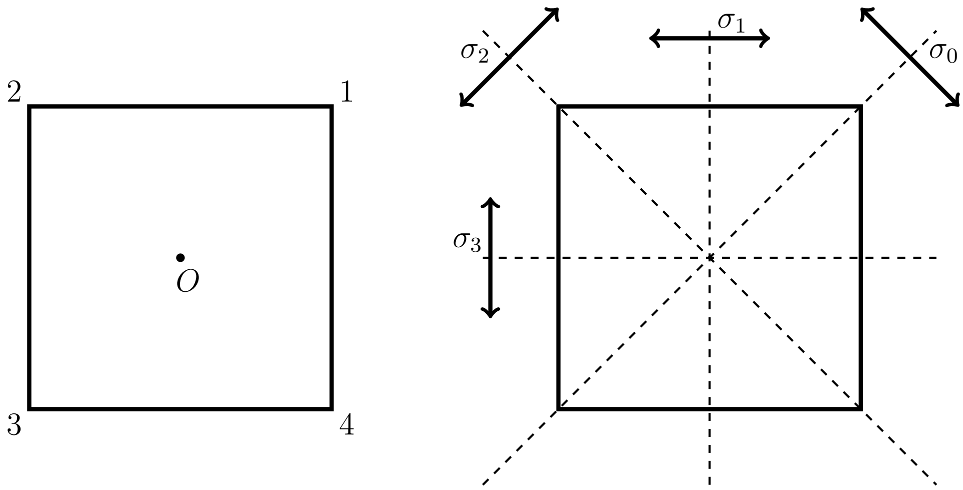


Figure 2: Left: the square with vertices labelled 1, 2, 3, 4. Right: the reflections σ_0, σ_1, σ_2, σ_3.

These are the ’symmetries of a square’ and put into a set they look like this.

$$D_8 = \{\rho_0, \rho_1, \rho_2, \rho_3, \sigma_0, \sigma_1, \sigma_2, \sigma_3\}.$$

We talked about a group of symmetries, so it is not enough to just list the symmetries, but we have to specify a binary operation.

We’ve discussed that the compostion/multiplication is ’do the first move, then do the second move’. Notice that if we think of the moves as functions then this is just compostion of functions. (In fact you could put the square in the x, y plane with its centre at $(0, 0)$ and write down the matrix which corresponds to each of the moves.)

We need to be clear about how the notation works here. If $\alpha, \beta \in D_8$ then $\alpha \circ \beta$ means the symmetry which is "apply β first then α" (not the other way round). This might feel a bit strange. The reason for this choice is that we want to sometimes think of the elements of D_8 as functions, and when we do that we want composition in D_8 to agree with the usual composition of functions. Recall that $f \circ g$ means apply g first then f.

Now we can write out a composition/multiplication table (remember: $\alpha \circ \beta$ means you take α from the left column of ’row headings’ and β from the top row of ’column headings’):

$\circ$	ρ_0	ρ_1	ρ_2	ρ_3	σ_0	σ_1	σ_2	σ_3
ρ_0	ρ_0	ρ_1	ρ_2	ρ_3	σ_0	σ_1	σ_2	σ_3
ρ_1	ρ_1	ρ_2	ρ_3	ρ_0	σ_1	σ_2	σ_3	σ_0
ρ_2	ρ_2	ρ_3	ρ_0	ρ_1	σ_2	σ_3	σ_0	σ_1
ρ_3	ρ_3	ρ_0	ρ_1	ρ_2	σ_3	σ_0	σ_1	σ_2
σ_0	σ_0	σ_3	σ_2	σ_1	ρ_0	ρ_3	ρ_2	ρ_1
σ_1	σ_1	σ_0	σ_3	σ_2	ρ_1	ρ_0	ρ_3	ρ_2
σ_2	σ_2	σ_1	σ_0	σ_3	ρ_2	ρ_1	ρ_0	ρ_3
σ_3	σ_3	σ_2	σ_1	σ_0	ρ_3	ρ_2	ρ_1	ρ_0

It is not worth your while to check every entry in the table, but make sure you check four or five entries at random to get an idea of how to compose symmetries, and let me know if there are any mistakes!

Let's convince ourselves that $(D_8, \circ)$ is a group.

The first thing we should ask about is closure. This is clear from the table (either the picture one or the symbol one); when you compose two elements of D_8 you get an element of D_8.

It is clear that ρ_0 (=do nothing) is an identity element.

It is also (geometrically) clear that every element has an inverse which does belong to D_8. If you reflect twice in the same line you end up where you started, so $\sigma_i \circ \sigma_i = \rho_0$; in other words, σ_i is its own inverse for $i = 0, 1, 2, 3$. The inverse of an anticlockwise rotation around O by 90° is an anticlockwise rotation around O by 270°. We find that the inverses of ρ_0, ρ_1, ρ_2 and ρ_3 respectively are ρ_0, ρ_3, ρ_2 and ρ_1.

What's left is to prove associativity. Composing symmetries is associative for exactly the same reason as composing functions is associative.

Doing "symmetry A followed by symmetry B" to the square and then doing symmetry C to what you got is the same as doing symmetry A to the square and then doing "symmetry B followed by symmetry C" to what you got (they both are doing A then B then C to the square).

If you think the square as being centred at the origin and the matrices for each symmetry then you really can think of the elements as functions.

D_8 non-abelian group is our first example of a non-abelian group. To check that it isn't abelian all we have to do is give a pair of symmetries that don't commute.

For example,

$$\sigma_0 \circ \rho_1 = \sigma_3, \qquad \rho_1 \circ \sigma_0 = \sigma_1.$$

3.5.1 Exercise In this exercise you will write out the composition/multiplication table for the group D_6 which is the group of symmetries of an equilateral triangle. Sketch an equilateral triangle and label the vertices 1, 2, 3 in anticlockwise order (see Chapter for a reminder about this). Label the centre of the triangle with O. Let ρ_0, ρ_1, ρ_2 denote anticlockwise rotations about O through angles 0, $2\pi/3$ and $4\pi/3$. Let σ_1, σ_2, σ_3 denote reflections about the lines respectively joining vertices 1, 2, 3 to O. Let

$$D_6 = \{\rho_0, \rho_1, \rho_2, \sigma_1, \sigma_2, \sigma_3\}.$$

Write down a composition/multiplication table for D_3 and explain why it is a group [1]. You can see from the table that D_6 is not abelian.

3.5.2 Exercise Write down the symmetries of a triangle that is isoceles but not equilateral and a composition table for them. Do they form a group?

[1] More generally, D_n denotes the group of symmetries of a regular polygon with n sides. These are called the *dihedral groups*. Some mathematicians denote D_{2n} by D_n because it is the symmetries of a regular n-gon.

Chapter 4 - First theorems and notation

Our first two theorems deal with subconscious assumptions. One of the defining properties of a group is the ‘existence of the identity element’ (property (iii)). The word ‘the’ contains a hidden assumption; how do we know there is only one identity element? Shouldn’t we be talking about the ‘existence of an identity element’?

4.1 Uniqueness of the identity element and inverses

4.1.1 Theorem Let $(G, \star)$ be a group. Then $(G, \star)$ has a unique identity element.

Proof. Suppose that e and e' are identity elements. Thus, for all $a \in G$ we have

$$a \star e = e \star a = a, \tag{1}$$

and

$$a \star e' = e' \star a = a. \tag{2}$$

Now let us try evaluating $e \star e'$. If we let $a = e$ and use (2) we find

$$e \star e' = e.$$

But if we let $a = e'$ and use (1) we find

$$e \star e' = e'.$$

Thus $e = e'$. In other words, the identity element is unique. ◊

4.1.2 Theorem Let $(G, \star)$ be a group and let a be an element of G. Then a has a unique inverse.

Proof. Our proof follows the same pattern as the proof of Theorem 4.1.1, and you’ll see this pattern again and again during your undergraduate career. Almost all uniqueness proofs follow the same pattern: suppose that there are two of the thing that we want to prove unique; show that these two must be equal; therefore it is unique.

For our proof we suppose that b and c are both inverses of a. We want to show that $b = c$. By definition of inverse (property (iv) in the definition of a group) we know that

$$a \star b = b \star a = e, \qquad a \star c = c \star a = e,$$

where e is of course the identity element of the group. Thus

$$\begin{aligned} b &= b \star e \qquad \text{by (iii) in the definition of a group} \\ &= b \star (a \star c) \qquad \text{from the above } a \star c = e \\ &= (b \star a) \star c \qquad \text{by (ii) in the definition of a group} \\ &= e \star c \qquad \text{from the above } b \star a = e \\ &= c \qquad \text{by (iii) again.} \end{aligned}$$

Thus $b = c$. Since any two inverses of a must be equal, we see that the inverse of a is unique. $\Diamond$

4.2 Getting relaxed about notation

It is quite tedious to keep writing $\star$ for the group operation. If $(G, \star)$ is a group and a, $b \in G$, we shall write ab for $a \star b$, unless there is reason for possible confusion.

For example if $(G, \star) = (\mathbb{R}, +)$ then it is silly to write ab for $a + b$ because the usual meaning for ab is "$a \times b$". But it is OK most of the time, and when it is OK we will do it. Moreover, we shall often say "let G be a group", without giving an explicit name to the binary operation. When we talk of the groups $\mathbb{R}$, $\mathbb{R}^2$, $\mathbb{R}[x]$, $\mathbb{R}^*$, etc. we shall mean the groups $(\mathbb{R}, +)$, $(\mathbb{R}^2, +)$, $(\mathbb{R}[x], +)$, $(\mathbb{R}^*, \cdot)$, etc. This may feel a bit odd at first but it's something you will get used to.

If G is a group, and we're writing ab for $a \star b$, then it makes sense to use 1 to denote the identity element instead of e. We write a^{-1} for the (unique) inverse of a. So

$$a \star b = b \star a = e,$$

where b is the inverse of a, becomes

$$aa^{-1} = a^{-1}a = 1,$$

which looks familiar.

Here are a couple of crucial results that you should get used to.

4.2.1 Theorem Let G be a group and $a \in G$. Then

$$(a^{-1})^{-1} = a.$$

Proof. We're being asked to prove that a is the inverse of a^{-1}. Thinking carefully about what this would mean, we want to show that

$$a^{-1}a = 1 = aa^{-1}$$

But this is clearly true because a^{-1} is the inverse of a. ◊

The above proof is a good exercise in getting your head around definitions. Make sure you understand how we are proving that a is acting as the inverse element to a^{-1} in the above (usually we think of this the other way round).

4.2.2 Theorem Let G be a group and a, $b \in G$. Then

$$(ab)^{-1} = b^{-1}a^{-1}.$$

Notice that we reverse the order when taking inverse. You have probably seen this before when you did matrices at school/college

Proof. We're being asked to prove that $b^{-1}a^{-1}$ is the inverse of ab. So we want to show that

$$(b^{-1}a^{-1})(ab) = 1 = (ab)(b^{-1}a^{-1}).$$

Now

$$\begin{aligned}(b^{-1}a^{-1})(ab) &= b^{-1}(a^{-1}a)b \qquad \text{by associativity}\\ &= b^{-1}1b\\ &= 1,\end{aligned}$$

and similarly $(ab)(b^{-1}a^{-1}) = 1$. ◊

Note that you shouldn't write $\frac{a}{b}$ unless the group is abelian. This notation is ambiguous; does $\frac{a}{b}$ mean $b^{-1}a$ or ab^{-1}? The two aren't the same in a non-abelian group.

4.2.3 Exercise Use D_8 to give counterexamples to the following:

- $b^{-1}a = ab^{-1}$,
- $(ab)^{-1} = a^{-1}b^{-1}$,
- $a^{-1}ba = b$.

4.2.4 Exercise Let G be a group satisfying $a^2 = 1$ for all a in G. Show that G is abelian.

Going back to our discussion of notation, if n is a positive integer we shall define

$$a^n = \underbrace{aa\cdots a}_{n \text{ times}}.$$

We define $a^0 = 1$. If n is a negative integer we define $a^n = (a^{-n})^{-1}$. Again we should reflect a little to make sure we're not being reckless. Does a^3 mean $(a\star a)\star a$ or $a \star (a \star a)$? It doesn't matter because of the associativity property of a group.

4.2.5 Example Let $\star$ be the binary operation on $S = \{a, b, c\}$ in Example 2.9.2 3. Note that $(S, \star)$ is definitely not a group, as $\star$ is not associative. Now you can check that

$$(a \star a) \star a = a, \qquad a \star (a \star a) = c.$$

Thus writing a^3 in this context does not make any sense. $\Diamond$

The following theorem deals with some consequences of this notation, which should look reasonably familiar to you.

4.2.6 Theorem Let G be a group, and let $a \in G$. Then

1. $a^n \in G$ for all $n \in \mathbb{Z}$.

2. If $n \in \mathbb{Z}$ then $(a^{-1})^n = (a^n)^{-1} = a^{-n}$.

3. Moreover, if m, n are integers then

$$(a^m)^n = a^{mn}, \qquad a^m a^n = a^{m+n}.$$

4. Further, if the group G is abelian, a, $b \in G$ and n an integer then

$$(ab)^n = a^n b^n.$$

Proof.

1. Let's deal with the case of $n = 0$ separately. $a^0 = 1$ and certainly $1 \in G$ so the statement is true then.

Now let's prove it for positive integers by induction on n. If $n = 1$ then $a^n = a^1 = a$ since $a \in G$ so the statement is true when $n = 1$.

Now suppose we know that $a^k \in G$ for some positive integer k. Then $a^{k+1} = a^k a \in G$ since both a^k and a are in G which is closed under the binary operation. It follows by induction that $a^n \in G$ for any positive integer n.

Now let's prove that $a^n \in G$ when n is a negative integer.

Then $a^n = (a^{-n})^{-1}$. But we know that $a^{-n} \in G$ since $-n$ is positive by earlier in the proof. So $(a^{-n})^{-1}$ is also in G by the definition of a group. Therefore $a^n \in G$.

2. Clearly this is true if $n = 0$ since all three of the expressions are equal to 1.

   If n is a positive integer then $a^n(a^{-1})^n = \underbrace{aa\cdots a}_{n \text{ times}}\underbrace{a^{-1}a^{-1}\cdots a^{-1}}_{n \text{ times}}$.

   Each aa^{-1} in the centre collapses to a 1 and eventually we see that this expression is 1. Therefore it's true to say that $(a^{-1})^n = (a^n)^{-1}$. Also $(a^n)^{-1} = a^{-n}$ just by notational definition.

   If n is a negative integer then

$$a^n(a^{-1})^n = (a^{-n})^{-1}((a^{-1})^{-n})^{-1} = (a^{-1})^{-n}((a^{-1})^{-1})^{-n} = (a^{-1})^{-n}a^{-n} = 1$$

   where the expressions either side of the second and third equals sign are the same by what we have already proved for positive integers. The last equals sign follows in a similar way to above (remembering that $-n$ is a positive integer).

   Again we can conclude that $(a^{-1})^n = (a^n)^{-1}$. Also $(a^n)^{-1} = ((a^{-n})^{-1})^{-1} = a^{-n}$.

3. Here we need to first prove that for any integers m and n,

$$(a^m)^n = a^{mn}, \qquad a^m a^n = a^{m+n}.$$

If m and n are positive integers then

$$(a^m)^n = \underbrace{\underbrace{aa\cdots a}_{m \text{ times}}\underbrace{aa\cdots a}_{m \text{ times}}\cdots\underbrace{aa\cdots a}_{m \text{ times}}}_{n \text{ times}}.$$

From this it can be seen that $(a^m)^n = a^{mn}$.

If m is positive and n is negative then we have

$$(a^m)^n = ((a^m)^{-n})^{-1} = (a^{-mn})^{-1} = ((a^{mn})^{-1})^{-1} = a^{mn}$$

using what we have just proved for positive integers and part ii).

The other cases (m negative and n positive and m and n both negative are similar).

We also need to show that $a^m a^n = a^{m+n}$.

If m and n are positive integers then

$$a^m a^n = \underbrace{aa\cdots a}_{m \text{ times}}\underbrace{aa\cdots a}_{n \text{ times}} = a^{m+n}.$$

If m is positive and n is negative then, since $n = -k$ where $k = -n$ is a postive integer we have

$$a^m a^n = a^m a^{-k} = a^m(a^{-1})^k\,\underbrace{aa\cdots a}_{m \text{ times}}\underbrace{a^{-1}a^{-1}\cdots a^{-1}}_{k \text{ times}} = a^{m-k} = a^{m+n}.$$

Again, the other cases are similar.

4. Suppose n is positive and G is abelian.

   Then
   $$(ab)^n = \underbrace{abab\cdots ab}_{n \text{ times}} = \underbrace{aa\cdots a}_{n \text{ times}}\underbrace{bb\cdots b}_{n \text{ times}} = a^n b^n.$$

   Note that the expressions either side of the second equals sign are the same only because G is abelian.

   Then, if n is negative, we have

$$(ab)^n = ((ab)^{-n})^{-1} = ((ab)^{-1})^{-n} = (b^{-1}a^{-1})^{-n} = (b^{-1})^{-n}(a^{-1})^{-n} = b^n a^n = a^n b^n.$$

The equalities in the above rely on what we have already proved in this theorem and results from earlier in this chapter (it's a useful exercise to make sure you can see which ones). $\Diamond$

The next example shows that we must have an ablelian group for 4. in the theorem 4.2.6 above to hold.

4.2.7 Example In D_8 you can check that

$$\rho_1^2\sigma_0^2 = \rho_2, \qquad (\rho_1\sigma_0)^2 = \rho_0,$$

and so $\rho_1^2\sigma_0^2 \neq (\rho_1\sigma_0)^2$. $\Diamond$

4.3 Additive notation

For some groups the binary operation is 'addition' (whatever that means). These include $(\mathbb{R}, +)$, $(\mathbb{Z}, +)$, $(\mathbb{R}[x], +)$, $(\mathbb{R}^2, +)$ etc. An important convention is that additive notation is only ever used for abelian groups. A multiplicative group can be abelian, such as $(\mathbb{R}^*, \cdot)$, and can be non-abelian, such as $(D_8, \circ)$.

You need to rephrase statements appropriately when using additive notation. For example, instead of speaking of

$$a^n = \underbrace{aa\cdots a}_{n \text{ times}},$$

you need to talk about

$$na = \underbrace{a + a + \cdots + a}_{n \text{ times}}.$$

Instead of b^{-1} write $-b$. We will mostly state and prove theorems in multiplicative notation, but it's up to you to translate these into additive notation for groups where the binary operation is addition. Let's do this for Theorem 4.2.6. Here is the translation.

4.3.1 Theorem Let G be an (abelian) group with addition as the binary operation, and let $a \in G$. Then

1. $na \in G$ for all $n \in \mathbb{Z}$.
2. $n(-a) = -(na) = (-n)a$ for any $n \in \mathbb{Z}$
3. Moreover, if m, n are integers then

$$m(na) = (mn)a, \qquad ma + na = (m + n)a.$$

4. Further, if a, $b \in G$ and n an integer then

$$n(a+b) = na + nb.$$

Chapter 5 - The order of an element

We return to using multiplicative notation. In Theorem 4.2.6 we observed that if G is a group containing an element a, then a^n is also in G for all integers n. It seems at first sight that this makes every group infinite: just pick an element a and you have an infinite list of elements

$$\ldots, a^{-4}, a^{-3}, a^{-2}, a^{-1}, 1, a, a^2, a^3, a^4, a^5, \ldots.$$

The group D_8 is finite, so what goes wrong? Take $a = \rho_1 \in D_8$ which represents anti-clockwise rotation by 90°. Then $a^4 = 1$. Thus the seemingly infinite list above simply becomes

$$\ldots, 1, a, a^2, a^3, 1, a, a^2, a^3, 1, \ldots.$$

In reality the list consists of exactly four elements $1, a, a^2, a^3$.

5.1 The Order of an Element

The above discussion leads us to the following definition.

5.1.1 Definition The **order** of an element a in a group G is the smallest positive integer n such that $a^n = 1$. If there is no such positive integer n, we say a has **infinite order**. $\Diamond$

In a finite group (a group with a finite number of elements) every element must have finite order.

5.1.2 Lemma Let G be a finite group and g be an element of G. The g has finite order.

Proof. Suppose $g \in G$ has infinite order. Then $g^0, g^1, g^2, g^3, \ldots$ are elements of G which are distinct from one another, since if $g^m = g^n$ for some natural numbers m,n with $m < n$ then $e = g^0 = g^{m-m} = g^{n-m}$ and g has finite order.

This cannot happen in a finite group and so g must have finite order. $\Diamond$

5.1.3 Example The order of ρ_1 is D_8 is 4. The order of ρ_2 is 2. The order of ρ_0 is 1. What are the orders of the other elements? $\Diamond$

5.1.4 Example In $(\mathbb{R}^*, \cdot)$, the element 1 has order 1 and the element -1 has order 2. What is the order of 7? Is there a *positive integer* n such that $7^n = 1$? No. Thus 7 has infinite order.

What are the elements of finite order in $\mathbb{R}^*$. These are the non-zero real numbers a such that $a^n = 1$ for some positive integer n. You should know that the only such real numbers are 1 and -1. So the only elements of finite order in $\mathbb{R}^*$ are 1 and -1 and all the other elements have infinite order. $\Diamond$

5.1.5 Example When you saw the equation $a^n = 1$ in the above example, you may have thought of the n-th roots of unity. The n-th roots of unity don't all live in $\mathbb{R}$; they live in $\mathbb{C}$. In fact, they live in $\mathbb{C}^*$.

For concreteness we take $n = 3$. You will know from *Foundations* that there are three cube roots of unity. These are $1, \zeta, \zeta^2$, where $\zeta = e^{2\pi i/3}$. See Figure 3. Let us think of these inside the group $\mathbb{C}^*$. Then ζ and ζ^2 have order 3. Let's check this for ζ^2. We note

$$(\zeta^2)^1 = \zeta^2, \qquad (\zeta^2)^2 = \zeta^4 = \zeta \cdot \zeta^3 = \zeta, \qquad (\zeta^2)^3 = (\zeta^3)^2 = 1^2 = 1.$$

So the least positive integer n such that $(\zeta^2)^n = 1$ is $n = 3$, so ζ^2 has order 3. Don't forget that 1 has order 1. So there are three cube roots of unity. Two have order 3 and one has order 1.

Now let us think briefly about the fourth roots of unity. These are $1, i, i^2, i^3$. Again see Figure 3. Note that $i^2 = -1$ and $i^3 = -i$. Of the four, only two have order 4 and these are i and i^3 (check). Of course, -1 has order 2 and 1 has order 1. $\Diamond$

5.1.6 Exercise Write down and sketch the sixth roots of unity. What are their orders? Repeat with the eighth roots of unity.

5.1.7 Exercise $\mathbb{C}^*$ has lots of elements of infinite order. Find a few.

5.1.8 Exercise Let $G = \mathrm{GL}_2(\mathbb{R})$. Show that

$$A = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}, \qquad B = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}$$

belong to G. Determine their orders.

Whilst reading the above examples and working out your own, you may have noticed the following:

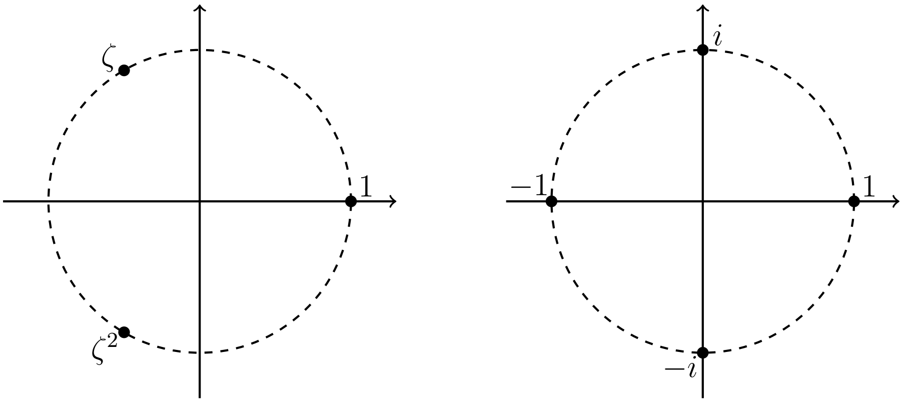


Figure 3: On the left, the three cube roots of unity: here $\zeta = e^{2\pi i/3}$. On the right, the four fourth roots of unity. Note that $e^{2\pi i/4} = e^{\pi i/2} = i$, so the fourth roots of unity are 1, i, $i^2 = -1$, and $i^3 = -i$.

5.1.9 Lemma Let G be a group and g be an element of G.

(i) g has order 1 if and only if g is the identity element.

(ii) Let m be a **non-zero** integer. Then $g^m = 1$ if and only if g has finite order d with $d \mid m$.

Proof. Let G be a group. Suppose g has order 1. By definition of order, $g^1 = 1$. Thus $g = 1$ which is the identity element of G. Conversely, the identity element clearly has order 1. This proves (i).

Part (ii) is an 'if and only if' statement. Suppose that g has order d and $d \mid m$. Then $g^d = 1$ and $m = qd$ where q is an integer. So $g^m = (g^d)^q = 1$. Let us prove the converse. Suppose $g^m = 1$ where m is a non-zero integer. Then $g^{|m|} = 1$, and $|m|$ is a positive integer. Thus g has finite order, which we denote by d. By the *division algorithm* which you met in *Foundations* we may write

$$m = qd + r, \qquad q,\, r \in \mathbb{Z} \text{ and } 0 \le r < d.$$

Now $g^d = 1$ by definition of order, so $1 = g^m = (g^d)^q \cdot g^r = g^r$. But $0 \le r < d$. As d is the order, it is the **least positive** integer such that $g^d = 1$. So $g^r = 1$ is possible with $0 \le r < d$ if and only if $r = 0$. This happens if and only if $m = qd$ which is the same as $d \mid m$. ♢

5.1.10 Exercise Let G be an abelian group. Suppose a, b are elements of orders m and n. Let $d = \operatorname{lcm}(m, n)$. Show that $(ab)^d = 1$, ensuring that you point out where you have used the fact the G is abelian. Give a counterexample to show that this does not have to be true if G is non-abelian. **Hint:** Look at D_6.$\Diamond$

Now we return to our examples. We've looked at various multiplicative groups, but what about additive groups? If $(G, +)$ is a group where the binary operation is addition, what is the order of an element a? Of course, it is the smallest positive integer n such that $na = 0$. If there is no such positive integer that a has infinite order.

5.1.11 Example In $(\mathbb{R}, +)$, $(\mathbb{Z}, +)$, $(\mathbb{R}[x], +)$, $(\mathbb{C}, +)$, the only element of finite order is 0, which has order 1. All other elements have infinite order.

How do we know this. Look at the equation $na = 0$ with a in the group and n a positive integer. We can divide both sides by n and obtain $a = 0$. $\Diamond$

5.2 Preview of Lagrange's Theorem

Lagrange's theorem is a beautiful result which says something about how only the number of elements in a finite group determines some of its internal structure. We won't state of prove Lagrange's theorm in this module. We'll prove a result which could be deduced from Lagrange's theorem but without using Lagrange's theorem!

5.2.1 Definition Let G be a group. The *order* of G is the number of elements that G has. We denote the order of G by $|G|$ or $\#G$. $\Diamond$

We will prove that in a finite abelian group the order of each element is a divisor of the order of the group. We'll need a two lemmas first.

5.2.2 Lemma Let G be a group and $1 \neq g \in G$. Then g has order 2 if and only if $g = g^{-1}$.

Proof. By definition, if g has order 2 then $g^2 = 1$. Multiplying both sides of this on the left by g^{-1} gives $g = g^{-1}g^2 = g^-1$.

For the converse suppose that $g = g^{-1}$. Since $g \neq 1$, the order of g is not 1. Mutiplying both sides of $g = g^{-1}$ on the left by g gives $g^2 = gg^{-1} = 1$. Therefore the order of g is 2. $\Diamond$

5.2.3 Lemma Let G be a group and $1 \neq g \in G$ and let n be a positive integer. Then g has order n if and only if g^{-1} has order n.

Proof. Suppose g has order n. Then $g^n = 1$ and so, multiplying both sides by $(g^n)^{-1}$ gives $1 = (g^n)^{-1} = (g^{-1})^n$. This means that the order of g is n or a positive integer less than n. But if $(g^{-1})^k = 1$ for some $0 < k < n$ then $g^k = 1$ and this contradicts n being the order of g.

For the converse suppose that g^{-1} has order n. Then $(g^{-1})^n = 1$ and so $(g^n)^{-1} = 1$. But this means that $g^n = 1$ so the order of g is n or a positive integer less than n. But if $g^j = 1$ for some $0 < j < n$ then $(g^{-1})^j = 1$ which contradicts n being the order of g^{-1}. ♢

5.2.4 Lemma Let G be a finite abelian group with elements $\{1 = g_1, g_2, ..., g_n\}$. Then $g_1^2 g_2^2 \dots g_n^2 = 1$.

Proof. First rename the elements, keeping $1 = g_1$ so that $g_2, g_3, \dots, g_m$ are all the elements of order 2 and $g_{m+1}, g_{m+2}, \dots, g_n$ are all the elements with order greater than 2.

By Lemma 5.2.2, none of $g_{m+1}, g_{m+2}, \dots, g_n$ are their own inverse and clearly none of them have inverse $1 = g_1$. Therefore, for all i with $m + 1 \leq i \leq n$ there exists a unique j with $m + 1 \leq j \leq n$ and with $i \neq j$ such that $g_j = g_i^{-1}$. (Note that this also follows from Lemma 5.2.3.)

Now renumber $g_{m+1}, g_{m+2}, \dots g_n$ so that in each set

$$\{g_{m+1}, g_{m+2}\}, \{g_{m+3}, g_{m+4}\}, \dots \{g_{n-1}, g_n\}$$

the two elements are inverse to one another. Notice that any two distinct sets (i.e. sets which are not equal) in this list are disjoint (i.e. have no elements in common) as follows:

Suppose $\{g_i, g_{i+1}\}$ and $\{g_j, g_{j+1}\}$ are two such (distinct) sets and suppose $\{g_i, g_{i+1}\} \cap \{g_j, g_{j+1}\} \neq \emptyset$ then at least one of the following must be true

1. $g_i = g_j$, in which case $g_{i+1} = g_{j+1}$ as they are both the inverse of g_i
2. $g_i = g_{j+1}$ in which case $g_{i+1} = g_j$ as there are both the inverse of g_i
3. $g_{i+1} = g_j$ in which case $g_i = g_{j+1}$ as they are both the inverse of g_{i+1} or

4. $g_{i+1} = g_{j+1}$ in which case $g_i = g_j$ as they are both the inverse of g_{i+1}.

In all four cases $\{g_i, g_{i+1}\} = \{g_j, g_{j+1}\}$. This contradicts the two sets being distinct. (Notice that it follows from this that there must be an even number of elements which have order greater than 2.)

Now consider $g_1^2g_2^2\dots g_n^2$. Since $g_1^2 = g_2^2 = \cdots = g_m^2 = 1$ this is equal to $g_{m+1}^2g_{m+2}^2\cdots g_n^2$. But, since G is abelian, this is the same as

$$(g_{m+1}g_{m+2})^2(g_{m+3}g_{m+4})^2\dots(g_{n-1}g_n)^2 = 1$$

because g_{m+2} is the inverse of g_{m+1},g_{m+4} is the inverse of g_{m+3} and so on.

Therefore $g_1^2g_2^2\dots g_n^2 = 1$. $\Diamond$

5.2.5 Theorem Let G be a finite abelian group. Let $x \in G$. Then the order of x divides the order of G.

Proof. Note first that x has finite order by Lemma 5.1.2. Suppose $G = \{g_1, g_2, \dots g_n\}$ where $n = |G|$. Note that this means that $x = g_i$ for some i.

Consider the set $\{xg_1, xg_2, \dots xg_n\}$. We will show that $\{xg_1, xg_2, \dots xg_n\} = G$. Clearly $\{xg_1, xg_2, \dots xg_n\} \subseteq G$. But if $g \in G$ then $g = x(x^{-1}g) \in \{xg_1, xg_2, \dots xg_n\}$. Therefore $G \subseteq \{xg_1, xg_2, \dots xg_n\}$ and so $\{xg_1, xg_2, \dots xg_n\} = G$.

Now consider

$$g_1(xg_1)g_2(xg_2)\dots g_n(xg_n).$$

Since G is abelian, is the product of the squares of all the elements and so by Lemma 5.2.4 it is equal to 1.

On the other hand, since G is abelian

$$g_1xg_1g_2xg_2\dots g_nxg_n = x^ng_1^2g_2^2\dots g : X \to \mathbb{R}^2_n.$$

But $g_1^2g_2^2\dots g : X \to \mathbb{R}^2_n = 1$, again by Lemma 5.2.4. Therefore

$$1 = g_1xg_1g_2xg_2\dots g_nxg_n = x^ng_1^2g_2^2\dots g : X \to \mathbb{R}^2_n = x^n.$$

The result now follows by Lemma 5.1.9. $\Diamond$

5.2.6 Example The set $\{1, i, -1, -i\}$ forms a group of order 4 under multiplication (convince yourself that this is true). Then 1 has order 1; -1 has order 2; i and $-i$ have order 4. This is all consistent with the Theorem 5.2.5. $\Diamond$

Chapter 6 - Subgroups

6.1 Informal discussion of subgroups of D_8

The group D_8 contains rotations and reflections. Here is its multiplication table (look back at chapter 3 for a reminder about the notation).

$\circ$	ρ_0	ρ_1	ρ_2	ρ_3	σ_0	σ_1	σ_2	σ_3
ρ_0	ρ_0	ρ_1	ρ_2	ρ_3	σ_0	σ_1	σ_2	σ_3
ρ_1	ρ_1	ρ_2	ρ_3	ρ_0	σ_1	σ_2	σ_3	σ_0
ρ_2	ρ_2	ρ_3	ρ_0	ρ_1	σ_2	σ_3	σ_0	σ_1
ρ_3	ρ_3	ρ_0	ρ_1	ρ_2	σ_3	σ_0	σ_1	σ_2
σ_0	σ_0	σ_3	σ_2	σ_1	ρ_0	ρ_3	ρ_2	ρ_1
σ_1	σ_1	σ_0	σ_3	σ_2	ρ_1	ρ_0	ρ_3	ρ_2
σ_2	σ_2	σ_1	σ_0	σ_3	ρ_2	ρ_1	ρ_0	ρ_3
σ_3	σ_3	σ_2	σ_1	σ_0	ρ_3	ρ_2	ρ_1	ρ_0

Let us now look at the rotations on their own and the reflections on their own, R is the set of rotations and S is the set of reflections below:

$$R = \{\rho_0, \rho_1, \rho_2, \rho_3\}, \qquad S = \{\sigma_0, \sigma_1, \sigma_2, \sigma_3\}.$$

For now let us look at the part of the composition table that involves only rotations:

$\circ$	ρ_0	ρ_1	ρ_2	ρ_3
ρ_0	ρ_0	ρ_1	ρ_2	ρ_3
ρ_1	ρ_1	ρ_2	ρ_3	ρ_0
ρ_2	ρ_2	ρ_3	ρ_0	ρ_1
ρ_3	ρ_3	ρ_0	ρ_1	ρ_2

Notice from the table that if we compose two rotations we obtain a rotation. We didn't really need the table for this; we can see it from the geometry. Thus $\circ$ is a binary operation on R (as well as being a binary operation on D_8).

We can ask whether $(R, \circ)$ is a group, and it is easy to see that the answer is yes (with the same reasoning as before). We have an interesting phenomenon, which is a group $(R, \circ)$ contained in another group $(D_8, \circ)$. We say that $(R, \circ)$ is a *subgroup* of $(D_8, \circ)$.

We will discuss subgroups at length later. It is also interesting to note that $(R, \circ)$ is abelian. An algebraic way of seeing the $(R, \circ)$ is abelian is to note that its composition table is symmetric about the leading diagonal. But you should also see geometrically that if you compose two rotations (centred at the same point) then the order does not matter. So $(R, \circ)$ is an abelian subgroup of the non-abelian group $(D_8, \circ)$.

What about $(S, \circ)$? Do the reflections of the square form a group? By looking at the composition table the first thing we notice is that S is **not** closed under composition. So $(S, \circ)$ is not a group. Are there any other subgroups inside $(D_8, \circ)$ besides $(R, \circ)$? Yes. See Figure 4 for a complete list.

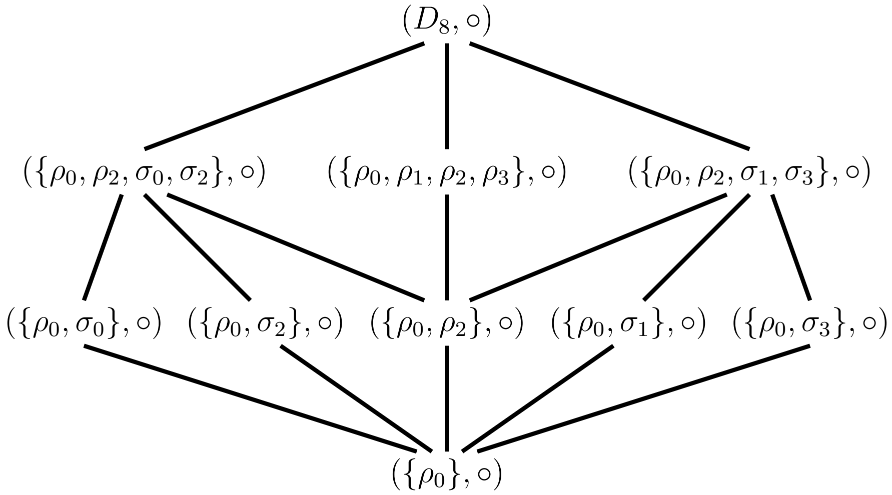

$(D_8, \circ)$
$(\{\rho_0, \rho_2, \sigma_0, \sigma_2\}, \circ)$
$(\{\rho_0, \rho_1, \rho_2, \rho_3\}, \circ)$
$(\{\rho_0, \rho_2, \sigma_1, \sigma_3\}, \circ)$
$(\{\rho_0, \sigma_0\}, \circ)$
$(\{\rho_0, \sigma_2\}, \circ)$
$(\{\rho_0, \rho_2\}, \circ)$
$(\{\rho_0, \sigma_1\}, \circ)$
$(\{\rho_0, \sigma_3\}, \circ)$
$(\{\rho_0\}, \circ)$

Figure 4: The figure shows the subgroups of $(D_8, \circ)$ and how they fit inside each other.

Again, check that a couple of these are subgroups. Don't waste time checking there aren't other subgroups of $(D_8, \circ)$; when you know a lot more about groups and subgroups you can come back to this question.

Now let us write down the formal definition of a subgroup and give some examples.

6.2 Definition and examples

6.2.1 Definition Let $(G, \star)$ be a group. Let H be a subset of G and suppose that $(H, \star)$ is also a group. Then we say that H is a subgroup of G (or more formally $(H, \star)$ is a subgroup of $(G, \star)$). ◊

For H to be a subgroup of G, we want H to a group with respect to *the same binary operation* that makes G a group.

6.2.2 Example $\mathbb{R}^*$ is a subset of $\mathbb{R}$ and both are groups. But $\mathbb{R}^*$ is **not** a subgroup of $\mathbb{R}$, since the operation that makes $\mathbb{R}^*$ a group is multiplication and the operation that makes $\mathbb{R}$ a group is addition. $\diamondsuit$

6.2.3 Example $\mathbb{Z}$ is a subgroup of $\mathbb{R}$ (or more formally, $(\mathbb{Z}, +)$ is a subgroup of $(\mathbb{R}, +)$); because $\mathbb{Z}$ is a subset of $\mathbb{R}$ and both are groups with respect to the same binary operation which is addition. $\diamondsuit$

6.2.4 Example $\mathbb{R}$ is a subgroup of $\mathbb{R}[x]$ since any real number can be viewed as a polynomial of degree 0. $\diamondsuit$

6.2.5 Example $(\emptyset, +)$ is **not** a subgroup of $(\mathbb{R}, +)$, simply because $(\emptyset, +)$ is not a group; a group has to be non-empty since it has to contain an identity element. $\diamondsuit$

6.3 Criterion for a subgroup

6.3.1 Theorem Let G be a group. A subset H of G is a subgroup if and only if it satisfies the following three conditions

(a) $1 \in H$,

(b) if $a, b \in H$ then $ab \in H$,

(c) if $a \in H$ then $a^{-1} \in H$.

Proof. The theorem has an "if and only if" statement.

It's usual when proving an "if and only if" statement to break it up into an "if" part, and an "only if" part, and prove each part separately. This is what we will do here. The "if" part says: "if H is a subset of G that satisfies (a),(b),(c) then it is a subgroup of G". The "only if" part says: "if H is a subgroup of G then H satisfies (a), (b), (c)".

Let us do the "if" part of the proof first. We have a group G and a subset H of G. All we have been told is that H satisfies conditions (a), (b), (c) in the statement of the theorem. We want to show that H is a group, where the binary operation on H is the same as the binary operation on G. This means that we have to show that H satisfies properties (i), (ii), (iii), (iv) in the definition of a group.

Property (i) is ‘closure’: we want that if a, $b \in H$ then $ab \in H$. But this is what (b) is saying. So (i) is satisfied.

Property (ii) is associativity. We want to show that for all a, b, $c \in H$, we have $(ab)c = a(bc)$. But if a, b, c are elements of H then they are also elements of G. We know that associativity holds in G: $(ab)c = a(bc)$. So (ii) holds [2].

Property (iii) is the existence of the identity element in H. But (a) tells us that $1 \in H$. This 1 is the identity element of G and so satisfies $a1 = 1a = a$ for all a in G. Since every a in H is also in G we have that $a1 = 1a = a$ for all a in H so 1 is the identity element of H, and so (iii) holds.

Finally, property (iv) asserts the existence of an inverse for every $a \in H$. This follows from (c). Hence H is a group contained in G and so a subgroup. We have now finished the proof of the “if” part.

Next we do the “only if”part of the proof. Here we assume that H is a subgroup of $|G|$ as in Definition 6.2.1 and need to show that it then satisfies conditions (a), (b) and (c) in the statement of this theorem.

In some ways proving (a) is the most tricky. Remember the 1 in (a) is the identity from G. Let’s call that element 1_G in this part of the proof to remind us of that. All we know is that H is a group and so it will contain an identity element. Let’s call this 1_H. So $1_H \in G$ and if we can show that $1_H = 1_G$ then it will follow that $1 \in H$.

We have $1_H = 1_H 1_H$. Now $1_H \in G$ and so it has an inverse in G, 1_H^{-1} with the property that $1_H 1_H^{-1} = 1_G$ (think carefully about this and make sure you are happy with why 1_G is on the right hand side as opposed to 1_H).

Multiplying both sides of $1_H = 1_H 1_H$ on the left by 1_H^{-1} gives $1_H^{-1} 1_H = 1_H^{-1} 1_H 1_H$ or $1_G = 1_G 1_H = 1_H$.

So the identity element of G equals the identity element of H and in particular (returning to the usual notation $1_G = 1$) $1 \in H$.

(b) follows because, as a group in its own right under the same binary operation as G, H is closed under that binary operation.

[2] There is a subtle point here that is camouflaged by our notation, and that is that the binary operation we’re using on H is precisely the same one as the binary operation we’re using on G. If it was different we would have no right to say: because associativity holds in G it holds in H.

To show (c) we have to show that if $a \in H$ then the inverse of a in G is in H. Because H is a subgroup it will have an inverse in H, call this a^{-1}. Then $1_H = aa^{-1} = a^{-1}a$. But, since $1 = 1_H$, this means that $1 = aa^{-1} = a^{-1}a$ and a^{-1} is the inverse of a in G (by the uniqueness of inverses, Theorem 4.1.2). ♢

Now let's try out the theorem.

6.3.2 Example Let's take $G = \mathbb{R}^*$ and H the subset of positive real numbers:

$$H = \{a \in \mathbb{R}^* : a > 0\}.$$

Let's show that H is a subgroup of G. First, 1 is positive, so $1 \in H$. Hence condition (a) is satisfied.

To check (b), suppose that a, b are in H. Thus a and b are positive, and so their product ab is also positive. Hence $ab \in H$ and we know that (b) is satisfied.

Finally, we want to check condition (c). Suppose a is an element of H. Then a is positive, and so a^{-1} is positive. Hence a^{-1} is also an element of H. It follows that condition (c) is satisfied.

By Theorem 6.3.1, H is a subgroup of $\mathbb{R}^*$. ♢

6.3.3 Example Let

$$2\mathbb{Z} = \{2a : a \in \mathbb{Z}\} = \{\dots, -6, -4, -2, 0, 2, 4, 6, \dots\}.$$

In other words, $2\mathbb{Z}$ is the set of even integers. Now $2\mathbb{Z}$ is a subset of $\mathbb{Z}$, but is it a subgroup of $\mathbb{Z}$? We should check the three conditions in the theorem, where $G = \mathbb{Z}$ and $H = 2\mathbb{Z}$. Condition (a) is "$1 \in H$". What does that mean in our context? 1 is not the number 1. The 1 in the theorem is the identity element for the group operation on $\mathbb{Z}$. The group operation on $\mathbb{Z}$ is addition. The identity element is 0. As 0 is an even number (after all $0 = 2 \times 0$) we have $0 \in 2\mathbb{Z}$. Thus condition (a) is satisfied.

Let's move on to condition (b). This says "if a, $b \in H$ then $ab \in H$". Again ab doesn't always mean the product of a and b; it is shorthand for $a \star b$ where $\star$ is the binary operation on G. Here $G = \mathbb{Z}$ and the binary operation on $\mathbb{Z}$ is $+$. So to check (b) what we must check is the following "if a, $b \in 2\mathbb{Z}$ then $a + b \in 2\mathbb{Z}$". In words this just says "the sum of two even integers is even", which is true so (b) holds.

Finally we have to interpret (c) in our context. Here a^{-1} is the inverse of a with respect to addition, so it just means $-a$. Thus to check (c) we want to check that "if a is an even integer then $-a$ is also even". Again this is true, so (c) holds.

It follows from Theorem 6.3.1 that $2\mathbb{Z}$ is a subgroup of $\mathbb{Z}$.

By contrast, the set of odd integers

$$\{\dots, -5, -3, -1, 1, 3, 5, \dots\}$$

is not a subgroup of $\mathbb{Z}$. For example, it does not contain the identity element 0, so does not satisfy (a). $\Diamond$

6.3.4 Example In Subsection 6.1, we listed the ten subgroups of D_8. Go back to that list, and use Theorem 6.3.1 to verify that a couple of them are indeed subgroups. $\Diamond$

6.3.5 Example Let

$$V = \{(a, a) : a \in \mathbb{R}\}.$$

In other words V is the subset of $\mathbb{R}^2$ where the x-coordinate equals the y-coordinate. Thus V is the line $y = x$ in $\mathbb{R}^2$. It is geometrically obvious that V contains the origin, which is the identity element of $\mathbb{R}^2$; that if we add two vectors belonging to it the result also belongs to it; and that if we multiply any vector belonging to this diagonal by -1 the result also belongs to V. Figure 5 will help you visualise this. But you also need to be able to write a proof in symbols. Let us do that:

First note that $\mathbf{0} = (0, 0) \in V$. Secondly, suppose $\mathbf{u} \in V$ and $\mathbf{v} \in V$. By definition of V, $\mathbf{u} = (a, a)$ and $\mathbf{v} = (b, b)$ for some a, $b \in \mathbb{R}$. Thus $\mathbf{u} + \mathbf{v} = (a+b, a+b)$ which again belongs to V. Finally, suppose that $\mathbf{v} \in V$. By definition of V, $\mathbf{v} = (a, a)$ for some $a \in \mathbb{R}$. So $-\mathbf{v} = (-a, -a)$ which is in V. This shows that V is a subgroup of $\mathbb{R}^2$. $\Diamond$

6.3.6 Example This time we take $W = \{(a, a) : a \in \mathbb{R},\ a \geq 0\}$. The set W is not all the line $y = x$ but a 'ray' as in Figure 6. Note that W does satisfy the first two conditions (a), (b) for being a subgroup. However, it does not satisfy condition (c); for example, $\mathbf{v} = (1, 1)$ belongs to W but $-\mathbf{v} = (-1, -1)$ does not. Hence W is not subgroup of $\mathbb{R}^2$.

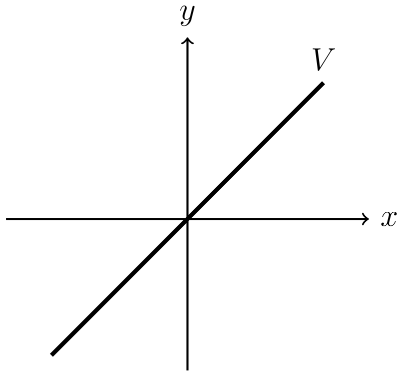


Figure 5: The set $V = \{(a,a) : a \in \mathbb{R}\}$ is the line $y = x$. It contains the identity element $(0,0)$, is closed under addition and negation. Therefore it is a subgroup of $\mathbb{R}^2$.

To show that W is not a subgroup, we gave a **counterexample**. This means that we gave an example to show that at least one of the requirements in Theorem 6.3.1 is not satisfied. ♢

6.3.7 Example Let

$$V = \{(a,a) : a \in \mathbb{R}\}, \qquad V' = \{(-a,a) : a \in \mathbb{R}\}.$$

You know from Example 6.3.5 that V is a subgroup of $\mathbb{R}^2$ (and is the line $y = x$). You can show, in a similar way, that V' (which happens to be the line $y = -x$) is also a subgroup of $\mathbb{R}^2$. What about their union $U = V \cup V'$? You can check that U satisfies conditions (a) and (c) of Theorem 6.3.1. However, $(1,1)$ and $(-1,1)$ are in U but their sum $(0,2)$ is not in U. So U does not satisfy (b), and is therefore not a subgroup of $\mathbb{R}^2$. See Figure 7.

On the other hand, the intersection $V \cap V' = \{(0,0)\}$ is a subgroup of $\mathbb{R}^2$. ♢

6.3.8 Exercise Let G be a group and let H_1, H_2 be subgroups. Show that $H_1 \cap H_2$ is also a subgroup of G.

6.3.9 Example Let's take

$$C = \{(a, a^3) : a \in \mathbb{R}\}.$$

Clearly C is a subset of $\mathbb{R}^2$; in fact it is the graph $y = x^3$ (see Figure 8). But is it a subgroup? It contains the identity element $(0,0)$. Moreover, $-(a,a^3) =$

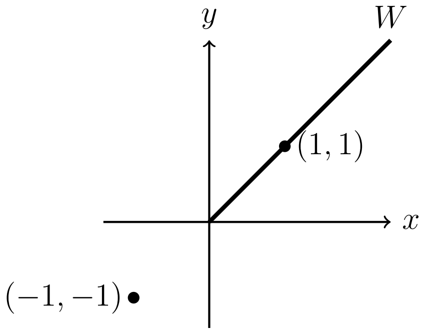


Figure 6: The ray $W = \{(a, a) : a \in \mathbb{R},\ a \geq 0\}$ is not a subgroup of $\mathbb{R}^2$. It contains the identity element $(0, 0)$ and is closed under addition. The problem is with the existence of additive inverses; e.g. $(1, 1)$ is in W but its inverse $(-1, -1)$ isn't in W.

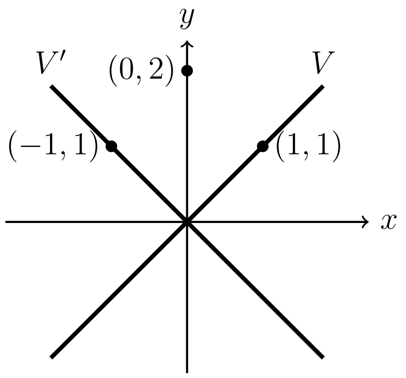


Figure 7: The lines $y = x$ and $y = -x$ are subgroups of $\mathbb{R}^2$. Their union is not.

$(-a, (-a)^3)$. So C satisfies condition (c) for subgroups. But it doesn't satisfy condition (b). To show this we give a counterexample. Note that $(1, 1)$ is in C but $(1, 1) + (1, 1) = (2, 2)$ is not in C. ♢

6.3.10 Example $\mathbb{Z}^2$ is a subgroup of $\mathbb{R}^2$. ♢

6.3.11 Exercise Which lines in $\mathbb{R}^2$ define a subgroup? Justify your answer.

6.3.12 Example Recall that

$$\mathbb{C}^* = \{\alpha \in \mathbb{C} : \alpha \neq 0\}.$$

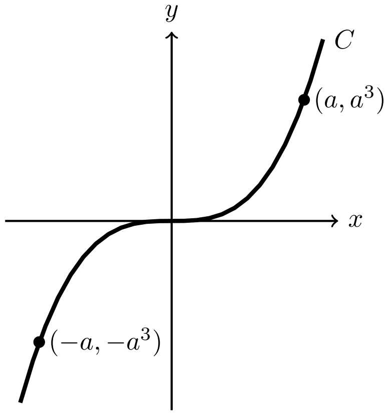


Figure 8: The set $C = \{(a, a^3) : a \in \mathbb{R}\}$ is the graph $y = x^3$. It satisfies conditions (a) and (c) for subgroups but not condition (b).

Geometrically, $\mathbb{C}^*$ is the whole complex plane minus the origin. We have observed before that $\mathbb{C}^*$ is a group (where the binary operation is multiplication of complex numbers). Let

$$\mathbb{S} = \{\alpha \in \mathbb{C} : |\alpha| = 1\}.$$

The set $\mathbb{S}$ is the set of all points in the complex plane with distance 1 from the origin. Of course this is just the unit circle (the circle centred at the origin with radius 1) as in Figure 9. Let us check that $\mathbb{S}$ is a subgroup of $\mathbb{C}^*$; it is clearly a subset. Of course the identity element of $\mathbb{C}^*$ is 1 and $|1| = 1$ so $1 \in \mathbb{S}$, which proves (a). Suppose α, $\beta \in \mathbb{S}$. Then $|\alpha| = 1$ and $|\beta| = 1$. From the properties of

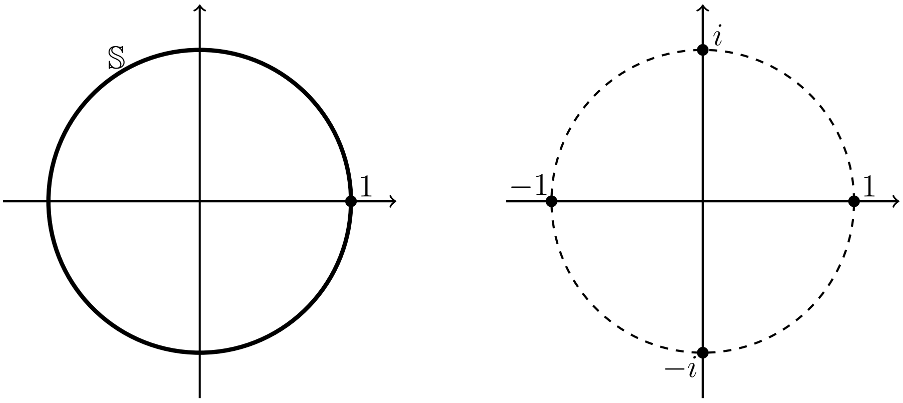


Figure 9: On the left, the group $\mathbb{S}$ which is just the unit circle. On the right, the subgroup of the fourth roots of unity.

the absolute value [3] we have

$$|\alpha\beta| = |\alpha||\beta| = 1.$$

Thus $\alpha\beta \in \mathbb{S}$. This proves (b).

To check (c), suppose $\alpha \in \mathbb{S}$, so that $|\alpha| = 1$. Then, again from the properties of the absolute value,

$$|\alpha^{-1}| = \frac{1}{|\alpha|} = 1,$$

so $\alpha^{-1} \in \mathbb{S}$. By Theorem 6.3.1, $\mathbb{S}$ is indeed a subgroup of $\mathbb{C}^*$.

We shall call $\mathbb{S}$ the *circle group*. Notice that $\mathbb{S}$ is an infinite subgroup of $\mathbb{C}^*$. But $\mathbb{C}^*$ has plenty of finite subgroups too. An example is $\{1, i, -1, -i\}$. This is the set of solutions to the equation $x^4 = 1$ (check). The solutions to $x^4 = 1$ are called the fourth roots of unity. Check for yourself that $\{1, i, -1, -i\}$ is a subgroup of $\mathbb{C}^*$ (and in fact a subgroup of $\mathbb{S}$). Can you find a finite subgroup of $\mathbb{C}^*$ that isn't a subgroup of $\mathbb{S}$? We'll return to roots of unity later. $\diamondsuit$

6.3.13 Exercise In the following, is H a subgroup of the group G? Give full justification.

Before you start answering: You might be wondering why the binary operation on G isn't specified. Mathematicians generally don't; you're expected to figure it out from the context [4].

(i) $G = \mathbb{R}$, $H = \mathbb{R}^*$.

(ii) $G = \mathbb{R}^*$, $H = \{1, -1\}$.

(iii) $G = \mathbb{C}$, $H = 2\mathbb{Z}$.

(iv) $G = \mathbb{C}$, $H = \{a + ai : a \in \mathbb{R}\}$.

(v) $G = \mathbb{C}^*$, $H = \{\alpha \in \mathbb{C}^* : \alpha^3 = 1\}$.

[3] At school/college you might have called $|\alpha|$ the *modulus* of α. Most mathematicians call $|\alpha|$ the *absolute value* of α.

[4] We know that addition makes $\mathbb{R}$ into a group, and multiplication doesn't. But are there really no other binary operations on $\mathbb{R}$ that make it into a group?

Yes, there are binary operations other than addition that make the set of real numbers into a group. But if this was anything other than the usual or obvious operation you'd have been told so.

(vii) $G = \mathbb{R}[x]$, $H = \mathbb{Z}[x]$.

(viii) $G = \mathbb{R}[x]$, $H = \{f \in \mathbb{R}[x] : f(0) = 0\}$.

(ix) $G = \mathbb{R}[x]$, $H = \{f \in \mathbb{R}[x] : f(0) = 1\}$.

6.3.14 Exercise Let

$$D = \{\alpha \in \mathbb{C}^* : |\alpha| \leq 1\}.$$

Sketch D. Show that D is not a subgroup of $\mathbb{C}^*$.

6.3.15 Exercise Let r be a positive real number. Let

$$\mathbb{S}_r = \{\alpha \in \mathbb{C}^* : |\alpha| = r\}.$$

What does $\mathbb{S}_r$ represent geometrically? For what values of r will $\mathbb{S}_r$ be a subgroup of $\mathbb{C}^*$?

6.4 Roots of unity

Let n be a positive integer. Let $\zeta = e^{2\pi i/n}$. The n-th roots of unity are the solutions in $\mathbb{C}$ to the equation $x^n = 1$. Recall that there are exactly n of them:

$$1, \zeta, \zeta^2, \ldots, \zeta^{n-1}.$$

See Figure 3 for the roots of unity when $n = 3$ and $n = 4$ and note how they're distributed on the unit circle. Write

$$U_n = \{1, \zeta, \zeta^2, \ldots, \zeta^{n-1}\}.$$

That is, U_n is the set of n-th roots of unity.

6.4.1 Lemma U_n is a subgroup of $\mathbb{C}^*$ of order n.

Proof. Clearly U_n is a subset of $\mathbb{C}^*$ containing 1. Suppose a, $b \in U_n$. We want to check that $ab \in U_n$. But since $a^n = b^n = 1$ we know that $(ab)^n = a^n b^n = 1$. So ab is also an n-th root of unity and so $ab \in U_n$. Likewise, $(a^{-1})^n = (a^n)^{-1} = 1$. So a^{-1} is an n-th root of unity and so $a^{-1} \in U_n$. Thus U_n is indeed a subgroup of $\mathbb{C}^*$. Since it has n elements, it has order n.

Notation Warning. The notation U_n is not standard. Why do I point this out? You must always be careful with notation: do other people understand you? If you write $\mathbb{C}^*$ then this is standard notation and every mathematician will know what you mean. If you write U_n, others (e.g. your tutor and supervisor) will not know what you mean. They will of course know that the n-th roots of unity are a subgroup of $\mathbb{C}^*$, but they will not know that you're denoting this subgroup by U_n. If you write U_n, even in your homework, then you have to say what it is.

6.4.2 Exercise Is $U_2 \cup U_3$ a subgroup of $\mathbb{C}^*$?

6.5 Non-trivial and proper subgroups

It's very easy for you to prove the following proposition.

6.5.1 Proposition Let G be a group. Then G and $\{1\}$ are subgroups.

Here, of course, $\{1\}$ is the subset containing the identity element of G. We call $\{1\}$ the *trivial* subgroup of G; any other subgroup is called *non-trivial*. A subgroup of G that is not equal to G is called *proper*. The subgroups $\{1\}$ and G aren't particularly interesting, since they're always there. The interesting subgroups are the proper non-trivial subgroups.

6.5.2 Example The trivial subgroup of $\mathbb{Z}$ is $\{0\}$. Examples of a non-trivial subgroups are $\mathbb{Z}$ and $2\mathbb{Z}$. The subgroup $2\mathbb{Z}$ is proper and non-trivial. $\Diamond$

6.5.3 Example Consider the group U_4 which is the group of fourth roots of unity. Thus $U_4 = \{1, i, -1, -i\}$; of course the binary operation is multiplication. The trivial subgroup is $\{1\}$. We note that $U_2 = \{1, -1\}$ is a non-trivial proper subgroup. Are there any others?

Suppose H is *another* non-trivial proper subgroup of U_4. Then $1 \in H$, as subgroups always contain the identity element. Since H is non-trivial, and $H \neq \{1, -1\}$, it must contain either i or $-i$. Suppose H contains i. Then H contains $i^2 = -1$ and $i^3 = -i$. Therefore $H = U_4$, which contradicts the assumption that H is proper. Similarly if H contains $-i$ then $H = U_4$ (check). Therefore the only non-trivial proper subgroup of U_4 is $U_2 = \{1, -1\}$. $\Diamond$

Chapter 7 - Cyclic groups

Cyclic groups are the simplest groups to understand. These are those groups where the elements are powers of one particular element. We say they are *generated* by that element.

7.1 Cyclic subgroups

7.1.1 Theorem Let G be a group, and let g be an element of G. Write $\langle g\rangle$ for the set

$$\langle g\rangle = \{g^n \mid n \in \mathbb{Z}\} = \{\dots, g^{-2}, g^{-1}, 1, g, g^2, g^3, \dots\}.$$

Then $\langle g\rangle$ is a subgroup of G.

Proof. We'll prove this using Theorem 6.3.1. So we need to show that the subset $\langle g\rangle$ of G satisfies conditions (a), (b) and (c) in that theorem.

Let's start with (a). We have that $g^0 = 1$ and, since $0 \in \mathbb{Z}$ this means that the identity from G, $1 \in \langle g\rangle$.

Now for (b). We need to show that if $a, b \in \langle g\rangle$ then $ab \in \langle g\rangle$. We do this as follows.

Let $a, b \in \langle g\rangle$. Then $a = g^m$ and $b = g^n$ for some $m, n \in \mathbb{Z}$. Then $ab = g^m g^n = g^{m+n}$ by Theorem 4.2.6. Since $m + n \in \mathbb{Z}$ this means that $ab \in \langle g\rangle$.

Finally (c). We need to show that if $a \in \langle g\rangle$ then $a^{-1} \in \langle g\rangle$:

Let $a \in \langle g\rangle$. The $a = g^m$ for some $m \in Z$. Then $a^{-1} = (g^m)^{-1} = g^{-m}$ by Theorem 4.2.6. Since $-m \in \mathbb{Z}$ this means that $a^{-1} \in \langle g\rangle$. ♢

7.1.2 Definition Let G be a group and let $g \in G$. We call $\langle g\rangle$ the *cyclic subgroup of* G generated by g. If $G = \langle g\rangle$ for some $g \in G$ then we call G a *cyclic group*, and we say that g is a *generator* of G.

7.1.3 Example As roots of unity are fresh in your mind, let's start with them. The group of n-th roots of unity U_n is cyclic, since every element is a power of $\zeta = e^{2\pi i/n}$; indeed the elements of U_n are precisely

$$\zeta^0 = 1, \zeta, \zeta^2, \dots, \zeta^{n-1}.$$

Thus $U_n = \langle \zeta \rangle$ and ζ is a generator.

Let's consider U_6, and calculate the cyclic subgroup generated by each element. Write $\zeta = e^{2\pi i/6}$. Note that $\zeta^6 = 1$. Consider for example $h = \zeta^2$. The powers of h are $1, h, h^2$. Indeed, note that $h^3 = \zeta^6 = 1$. Thus

$$h^4 = h, \; h^5 = h^2, \; h^6 = 1, \; h^7 = h, \ldots.$$

What about h^{-1}. We know that $h^3 = 1$; multiplying both sides by h^{-1} we deduce that $h^{-1} = h^2$. Thus

$$h^{-2} = h, \; h^{-3} = 1, \; h^{-4} = h^2, \; h^{-5} = h, \ldots.$$

Thus the distinct powers of h are $1, h, h^2$, which are $1, \zeta^2, \zeta^4$. We can't write all the elements of U_6 as powers of h; therefore h is not a generator of U_6.

However, let us consider $g = \zeta^5$. We can write the powers of g and simplify them using the fact that $\zeta^6 = 1$. For example,

$$g^2 = \zeta^{10} = \zeta^6 \zeta^4 = \zeta^4.$$

We find that $1, g, g^2, g^3, g^4, g^5$ are respectively, $1, \zeta^5, \zeta^4, \zeta^3, \zeta^2, \zeta$. Since every element of U_6 is a power of $g = \zeta^5$, we see that g is also a generator of U_6. Table 1 lists the elements of U_6 and the subgroups they generate.

g	$\langle g \rangle$
1	$\{1\}$
ζ	$\{1, \zeta, \zeta^2, \zeta^3, \zeta^4, \zeta^5\}$
ζ^2	$\{1, \zeta^2, \zeta^4\}$
ζ^3	$\{1, \zeta^3\}$
ζ^4	$\{1, \zeta^2, \zeta^4\}$
ζ^5	$\{1, \zeta, \zeta^2, \zeta^3, \zeta^4, \zeta^5\}$

Table 1: The six elements of U_6 and the cyclic subgroups they generate.

7.1.4 Example Recall the group D_8 of the symmetries of the square. It has 8 elements. It's easy to write down the subgroup generated by each element (see Section 3.5 to remind yourself of the notation):

g	$\langle g \rangle$
1	$\{1\}$
ρ_1	$\{1, \rho_1, \rho_2, \rho_3\}$
ρ_2	$\{1, \rho_2\}$
ρ_3	$\{1, \rho_1, \rho_2, \rho_3\}$
σ_0	$\{1, \sigma_0\}$
σ_1	$\{1, \sigma_1\}$
σ_2	$\{1, \sigma_2\}$
σ_3	$\{1, \sigma_3\}$

None of the elements of D_8 generates it. We see that D_8 is not a cyclic group. $\Diamond$

7.1.5 Theorem Cyclic groups are abelian.

Proof. Let G be a cyclic group generated by g. Let a, b be elements of G. We want to show that $ab = ba$. Now, $a = g^m$ and $b = g^n$ for some integers m and n. So, $ab = g^m g^n = g^{m+n}$ and $ba = g^n g^m = g^{n+m}$. But $m + n = n + m$ (addition of integers is commutative). So $ab = ba$.

Whilst working through the above examples, you will have noticed a pattern about $\langle g \rangle$, which we state in the following theorem.

7.1.6 Theorem Let G be a group and let g be an element of finite order n. Then

$$\langle g \rangle = \{1, g, g^2, \ldots, g^{n-1}\}.$$

In particular, the order of the subgroup $\langle g \rangle$ is equal to the order of g.

Proof. Observe that $\langle g \rangle$ is a set, and $\{1, g, \ldots, g^{n-1}\}$ is a set. We want to show that these sets are the same.

Whenever you have two sets, A and B, and you want to prove that they're equal, one way to do this is to show that every element of A belongs to B and every element of B belongs to A. You will see this principle again and again throughout your undergraduate career.

Let's apply this principle in our situation. By definition,

$$\langle g \rangle = \{g^n : n \in \mathbb{Z}\} = \{\ldots, g^{-2}, g^{-1}, 1, g, g^2, g^3, \ldots\}.$$

That is $\langle g \rangle$ is the set of all powers of g. It is obvious that every element of $\{1, g, \ldots, g^{n-1}\}$ belongs to $\langle g \rangle$. What about the other way round. Suppose that h

is an element of $\langle g \rangle$. We want to show that h is an element of $\{1, g, \dots, g^{n-1}\}$. We can write $h = g^m$ where m is an integer (positive or negative). We want to show that $h = g^r$ where r is one of $0, 1, 2, \dots, n-1$. For this we will use the *division algorithm* which you met in *Foundations*. We can write

$$m = qn + r, \qquad q, r \in \mathbb{Z}, \quad 0 \le r < n.$$

Here we simply divided m by n; the integers q, r are respectively the quotient and the remainder. Thus

$$h = g^m = g^{qn+r} = (g^n)^q \cdot g^r.$$

However, $g^n = 1$ since g has order n. So $h = g^r$. Since $0 \le r < n$, we see that r is one of $0, 1, \dots, n-1$. Therefore h is in $\{1, g, \dots, g^{n-1}\}$. By our principle, we see that $\langle g \rangle = \{1, g, \dots, g^{n-1}\}$.

7.1.7 Exercise In each of the following groups G, write down the cyclic subgroup generated by g.

(a) $G = \mathbb{S}$, $g = \exp(2\pi i/7)$.

(b) $G = \mathrm{GL}_2(\mathbb{R})$, $g = \left(\begin{smallmatrix} 0 & 1 \\ -1 & 0 \end{smallmatrix}\right)$.

7.1.8 Exercise Which of the following groups G are cyclic? Justify your answer for each, and if G is cyclic then write down a generator.

(a) $G = k\mathbb{Z}$ (where k is a non-zero integer).

(b) D_3.

7.1.9 Exercise In this exercise, you will show using contradiction that $\mathbb{R}^*$ is not cyclic. Suppose that it is cyclic and let $g \in \mathbb{R}^*$ be a generator. Then $\mathbb{R}^* = \langle g \rangle$. In particular, $|g|^{1/2} \in \mathbb{R}^*$ and so $|g|^{1/2} = g^m$ for some integer m. Show that the only solutions to this equation are $g = \pm 1$. Where's the contradiction?

7.1.10 Exercise In this exercise you'll show that $\mathbb{Q}$ is not cyclic. Let a, b be integers with $b \neq 0$. Let p be a prime that does not divide b. Show that $1/p$ cannot be written in the form na/b with n an integer. Deduce that $\mathbb{Q}$ is not cyclic.

7.1.11 Exercise Show that $\mathbb{S}$ is not cyclic.

7.2 Congruence classes modulo n under addition

In *Foundations* you have defined an equivalence relation on $\mathbb{Z}$ for a fixed postive integer n as follows.

> 'Where a, b are integers, a is said to be *congruent modulo* n to b if $a - b$ is a multiple of n, i.e. if there exists an integer k such that $a = b + kn$.'

In *Foundations* you go on to say that the equivalence classes for this equivalence relation are called the congruence classes modulo n and you write $[a]_n$ for the congruence class modulo n containing a. This means that $[a]_n = \{a + kn \mid k \in \mathbb{Z}\}$. You call the set of equivalence classes $\mathbb{Z}/n\mathbb{Z}$.

You then define the following two operations on $\mathbb{Z}/n\mathbb{Z}$, denoted $+_n$ and $\times_n$, as follows:

- $[a]_n +_n [b]_n = [a + b]_n$
- $[a]_n \times_n [b]_n = [a \times b]_n$.

In *Foundation* you make sure that these are well-defined because they are defined in terms of representatives of equivalence classes. Notice that these are both binary operations on $\mathbb{Z}/n\mathbb{Z}$ (since both $[a + b]_n \in \mathbb{Z}/n\mathbb{Z}$ and $[a \times b]_n \in \mathbb{Z}/n\mathbb{Z}$.

Given this, in this module, we might ask the following questions.

i. Is $(\mathbb{Z}/n\mathbb{Z}, +_n)$ a group?

ii. Is $(\mathbb{Z}/n\mathbb{Z}, \times_n)$ a group?

Let's look at the 'addition table' for $(\mathbb{Z}/6\mathbb{Z}, +_6)$ (case i. above with $n = 6$).

$+_6$	$[0]_6$	$[1]_6$	$[2]_6$	$[3]_6$	$[4]_6$	$[5]_6$
$[0]_6$	$[0]_6$	$[1]_6$	$[2]_6$	$[3]_6$	$[4]_6$	$[5]_6$
$[1]_6$	$[1]_6$	$[2]_6$	$[3]_6$	$[4]_6$	$[5]_6$	$[0]_6$
$[2]_6$	$[2]_6$	$[3]_6$	$[4]_6$	$[5]_6$	$[0]_6$	$[1]_6$
$[3]_6$	$[3]_6$	$[4]_6$	$[5]_6$	$[0]_6$	$[1]_6$	$[2]_6$
$[4]_6$	$[4]_6$	$[5]_6$	$[0]_6$	$[1]_6$	$[2]_6$	$[3]_6$
$[5]_6$	$[5]_6$	$[0]_6$	$[1]_6$	$[2]_6$	$[3]_6$	$[4]_6$

We can see that this is a group in which $[0]_6$ is the identity element, the inverse of $[m]_6$ is $[6 - m]_6 = [-m]_6$ (associativity follows from the associativity of regular addition of integers). This is proved formally below in Theorem 7.2.2.

There is nothing special about $n = 6$ here, the same would be true for any positive integer n.

Now let's look at the multiplication table for $(\mathbb{Z}/6\mathbb{Z}, \times_6)$ (case ii. above with $n = 6$).

$\times_6$	$[0]_6$	$[1]_6$	$[2]_6$	$[3]_6$	$[4]_6$	$[5]_6$
$[0]_6$	$[0]_6$	$[0]_6$	$[0]_6$	$[0]_6$	$[0]_6$	$[0]_6$
$[1]_6$	$[0]_6$	$[1]_6$	$[2]_6$	$[3]_6$	$[4]_6$	$[5]_6$
$[2]_6$	$[0]_6$	$[2]_6$	$[4]_6$	$[0]_6$	$[2]_6$	$[4]_6$
$[3]_6$	$[0]_6$	$[3]_6$	$[0]_6$	$[3]_6$	$[0]_6$	$[3]_6$
$[4]_6$	$[0]_6$	$[4]_6$	$[2]_6$	$[0]_6$	$[4]_6$	$[2]_6$
$[5]_6$	$[0]_6$	$[5]_6$	$[4]_6$	$[3]_6$	$[2]_6$	$[1]_6$

We can see that this is not a group. You could pick out many reasons for this but one of them is that the only candidate for the identity is $[1]_6$ and then $[0]_6$ has no inverse element.

Maybe the problem is just with the 'zero element', like we have seen before with the need to remove 0 to get $\mathbb{R}^*$ as a group under multiplication.

If we remove $[0]_6$ from the table row and column headings above we get

$\times_6$	$[1]_6$	$[2]_6$	$[3]_6$	$[4]_6$	$[5]_6$
$[1]_6$	$[1]_6$	$[2]_6$	$[3]_6$	$[4]_6$	$[5]_6$
$[2]_6$	$[2]_6$	$[4]_6$	$[0]_6$	$[2]_6$	$[4]_6$
$[3]_6$	$[3]_6$	$[0]_6$	$[3]_6$	$[0]_6$	$[3]_6$
$[4]_6$	$[4]_6$	$[2]_6$	$[0]_6$	$[4]_6$	$[2]_6$
$[5]_6$	$[5]_6$	$[4]_6$	$[3]_6$	$[4]_6$	$[2]_6$

The appearance of $[0]_6$ within this table shows that this is no longer even a binary operation on $\{[1]_6, [2]_6, [3]_6, [4]_6, [5]_6\}$ so it certainly doesn't give us a group.

It's worth noting here that had we done the above but with $n = 5$ rather than $n = 6$ we would still have seen that $(\mathbb{Z}/5\mathbb{Z}, \times_5)$ is not a group but removing $[0]_5$ then gives

$\times_5$	$[1]_5$	$[2]_5$	$[3]_5$	$[4]_5$
$[1]_5$	$[1]_5$	$[2]_5$	$[3]_5$	$[4]_5$
$[2]_5$	$[2]_5$	$[4]_5$	$[1]_5$	$[3]_5$
$[3]_5$	$[3]_5$	$[1]_5$	$[4]_5$	$[2]_5$
$[4]_5$	$[4]_5$	$[3]_5$	$[2]_5$	$[1]_5$

which is a group (convince yourself it has the required properties by looking at the table above). Keeping to our convention of * meaning 'remove the zero element', we would say that $(\mathbb{Z}/6\mathbb{Z})^*$ with binary operation $\times_6$ is not a group but $((\mathbb{Z}/5\mathbb{Z})^*, \times_5)$ is a group.

We'll explore this more when we come to the rings part of the module. At this point we'll also explain footnote 9, about the notation $\mathbb{Z}/n\mathbb{Z}$, at the bottom of page 29 in the *Foundations* notes.

7.2.1 Exercise Write down the addition and multiplication tables for $\mathbb{Z}/4\mathbb{Z}$ and $\mathbb{Z}/5\mathbb{Z}$.

Here is the proof that, for any positive integer n, $(\mathbb{Z}/n\mathbb{Z}, +_n)$ is a group.

7.2.2 Theorem Let n be an integer satisfying $n \geq 1$. Then $(\mathbb{Z}/n\mathbb{Z}, +)$ is an abelian group.

Proof. To show that $\mathbb{Z}/n\mathbb{Z}$ a group, we want to check that $\mathbb{Z}/n\mathbb{Z}$ is closed under addition, that addition is associative, that there is an identity element, and that every element has an additive inverse.

We define, as in *Foundations*, $\mathbb{Z}/n\mathbb{Z}$ to be the set of congruence classes modulo m. There, the sum of classes $[a]_n$ and $[b]_n$ is defined to be $[a+b]_n$ which is a congruence class modulo m (and it is checked that this is well-defined). So $\mathbb{Z}/n\mathbb{Z}$ is closed under addition. Let's prove associativity. Note

$$\begin{aligned}
([a]_n + [b]_n) + [c]_n &= [a+b]_n + [c]_n \\
&= [(a+b)+c]_n \\
&= [a+(b+c)]_n \qquad \text{addition in } \mathbb{Z} \text{ is associative} \\
&= [a]_n + [b+c]_n \\
&= [a]_n + ([b]_n + [c]_n).
\end{aligned}$$

Thus addition in $\mathbb{Z}/n\mathbb{Z}$ is associative. Obviously $[0]_n$ is the additive identity. What about the additive inverse? Note that $[a]_n + [-a]_n = [0]_n$ so every class has an additive inverse [5].

Thus $(\mathbb{Z}/n\mathbb{Z}, +)$ is a group. The proof that it is abelian is left as an easy exercise. $\Diamond$

Note that we can safely refer to 'the group $\mathbb{Z}/n\mathbb{Z}$' without specifying the binary operation as $+_n$ because $\mathbb{Z}/n\mathbb{Z}$ is never a group under $\times_n$.

7.2.3 Example For each element of the group $\mathbb{Z}/n\mathbb{Z}$, we can write down the cyclic group it generates. Note that since $\mathbb{Z}/n\mathbb{Z}$ is an additive group, the subgroup generated by g is $\langle g \rangle = \{mg \mid m \in \mathbb{Z}\}$. That is, it is the set of multiples of g rather than the set of powers of g. This is done below in the case $n = 6$. See Table 2.

$[a]_6$	$\langle [a]_6 \rangle$
$[0]_6$	$\{[0]_6\}$
$[1]_6$	$\{[0]_6, [1]_6, [2]_6, [3]_6, [4]_6, [5]_6\}$
$[2]_6$	$\{[0]_6, [2]_6, [4]_6\}$
$[3]_6$	$\{[0]_6, [3]_6\}$
$[4]_6$	$\{[0]_6, [2]_6, [4]_6\}$
$[5]_6$	$\{[0]_6, [1]_6, [2]_6, [3]_6, [4]_6, [5]_6\}$

Table 2: The six elements of $\mathbb{Z}/6\mathbb{Z}$ and the cyclic subgroups that they generate.

You can see from the above example how $\mathbb{Z}/n\mathbb{Z} = \langle [1]_n \rangle$ for any positive integer n, so that $\mathbb{Z}/n\mathbb{Z}$ is cyclic for any positive integer n.

7.2.4 Exercise In each of the following groups G, write down the both the order of the element g and the cyclic subgroup generated by g.

(a) $G = \mathbb{Z}/5\mathbb{Z}$, $g = [4]_5$.

(b) $G = \mathbb{Z}/8\mathbb{Z}$, $g = [2]_8$.

(c) $G = \mathbb{Z}/12\mathbb{Z}$, $g = [8]_{12}$.

[5] Perhaps you prefer the inverse of $[a]_n$ where $0 \leq a < n$ to be of the form $[b]_n$ where b also satisfies $0 \leq b < n$. In this case, if $0 < a < m$, then observe that $0 < n - a < n$, and $[a]_n + [n - a]_n = [0]_n$, since $a + (n - a) \equiv 0 \pmod{n}$. Moreover $-0 \equiv 0 \pmod{n}$, thus $-[0]_n = [0]_n$.

Chapter 8 - Composition as a binary operation

Virtually all of this chapter is revision from *Foundations*. It's mainly included for completeness and so mostly only definitions and statements of results are included. Here we define a function, define what it means to say that two function are equal, define the terms injective, surjective, bijective and inverse function and state related results.

There is one result which warrants particular attention, this is the associativity of composition of functions, covered in Theorem 8.2.3. In the two chapters which follow this one we'll be looking at some groups which have functions as their elements and which use compostion of functions as their binary operation and so this is essential for that.

8.1 Definitions

8.1.1 Definition Given two sets X and Y, a *function*, f, is a rule which associates a unique element $y \in Y$ to each element $x \in X$.

We write $f : X \to Y$. If $y \in Y$ is the unique element of Y associated with $x \in X$ we write $y = f(x)$.

X is called the *domain* of the function and Y is called the *codomain*. The subset $\{f(x) \mid x \in X\}$ of Y is called the *image* of f.

8.1.2 Definition Let X, Y, A and B be sets and let $f : X \to Y$ and $g : A \to B$ be functions.

We say that $f = g$ if all of the following hold

1. $X = A$
2. $Y = B$
3. $f(x) = g(x)$ for all $x \in X$. ♢

8.2 Composition, injective and surjective functions, inverses

8.2.1 Definition Let S_1, S_2 and S_3 be sets and f, g be functions

$$f : S_1 \to S_2, \qquad g : S_2 \to S_3.$$

We can define the *composition* $g \circ f : S_1 \to S_3$ by the rule: $(g \circ f)(x) = g(f(x))$, i.e. $g \circ f$ is the function obtained by substituting f into g.

8.2.2 Example Here is an example of composition of functions. Let

$$f : \mathbb{R} \to \mathbb{R}, \qquad f(x) = x^2 - 5$$

and

$$g : \mathbb{R} \to \mathbb{R}, \qquad g(x) = 3x + 2.$$

Notice that because the codomain and the domain are both $\mathbb{R}$ for both f and g we can compose these functions, and we can do that both 'ways round':

$$\begin{aligned}(f \circ g)(x) &= f(g(x)) = f(3x + 2) = (3x + 2)^2 - 5 = 9x^2 + 12x - 1,\\ (g \circ f)(x) &= g(f(x)) = g(x^2 - 5) = 3(x^2 - 5) + 2 = 3x^2 - 13.\end{aligned}$$

The order matters here: $f \circ g$ is the result of substituting g into f, and $g \circ f$ is the result of substituting f into g.

The following lemma might look quite basic, but it one of the most important results we shall meet in this module, and we shall use it again and again. it tells us that compostion of functions is associative.

8.2.3 Lemma Let S_1, S_2, S_3, S_4 be sets and let f, g, h be functions

$$h : S_1 \to S_2, \qquad g : S_2 \to S_3, \qquad f : S_3 \to S_4.$$

Then $f \circ (g \circ h) = (f \circ g) \circ h$.

Proof
Think about what these two functions do to an element $x \in S_1$.

Let's start with $f \circ (g \circ h)$. Here we 'do' $g \circ h$ to x first, this will give us $g(h(x))$. Then we 'do' f to this. So we will end up with $f(g(h(x)))$.

Now think about the effect of $(f \circ g) \circ h)$ on x Here we 'do' h to x first, this will give us $h(x)$. Then we 'do' $f \circ g$ to this, this means doing g to it, then doing f to the result. So we will again end up with $f(g(h(x)))$.

Therefore $[f \circ (g \circ h)](x) = f(g(h(x))) = [(f \circ g) \circ h](x)$ for all $x \in S_1$ and this is what it means to say that $f \circ (g \circ h)$ and $(f \circ g) \circ h$ are equal functions. $\Diamond$

8.2.4 Definition Let X and Y be sets. Let $f : X \to Y$ be a function. Suppose there exists a function $g : Y \to X$ such that $(g \circ f)(x) = x$ for all $x \in X$ and $(f \circ g)(y) = y$ for all $y \in Y$. Then f is said to be *invertible* and g is said to be the *inverse function* to f.

Note that then g is also invertible and f is the inverse function to g by the symmetry in the definition. Also it is usual to then write g as f^{-1}.

8.2.5 Definitions Let X and Y be sets. The function $f : X \to Y$ is said to be *injective* or *one-to-one* if whenever $x_1, x_2 \in X$ with $x_1 \neq x_2$ then $f(x_1) \neq f(x_2)$. Note that this is equivalent to saying that if $x_1, x_2, \in X$ and $f(x_1) = f(x_2)$ then $x_1 = x_2$.

The function $f : X \to Y$ is said to be *surjective* or *onto* if for all $y \in Y$ there exists $x \in X$ such that $f(x) = y$.

The function $f : X \to Y$ is said to be *a bijection* it is both injective and surjective. ♢

The next theorem, from *Foundations*, is going to be really important to us when we look at groups whose elements are functions. If we are to have a group, we will need these functions to be invertible.

8.2.6 Theorem Let X and Y be sets. The function $f : X \to Y$ is bijective if and only if it is invertible.

8.3 Composition of functions as a binary operation

In example 8.2.2 we started with functions $\mathbb{R} \to \mathbb{R}$ (i.e. with domain and codomain which are both the set of real numbers) and composed them to obtain functions $\mathbb{R} \to \mathbb{R}$. Likewise, in definition 8.2.1, if $S_1 = S_2 = S_3 = S$ say, so that f and g are functions $S \to S$ then $g \circ f$ is a function $S \to S$. In this case (i.e. when the domains and codomains are equal) $\circ$ is a binary operation.

It's easy to get confused about this. Although the set S is involved in this, this is not a binary operation on S, because it doesn't take two elements of S and give us another element. It is a binary operation on the set of functions from S to itself.

Composition of functions from a set S to itself is associative but not commutative. We know that it is associative from Lemma 8.2.3. We know that it isn't commutative by Example 8.2.2.

This sets the scene for the next two chapters where we will see some examples of groups whose elements are functions from a set to itself.

Chapter 9 - Groups of isometries of the plane

9.1 Isometries of the plane that fix the origin

Here we will think about functions $f : \mathbb{R}^2 \to \mathbb{R}^2$ which 'preserve distance'. These are called *isometries.* Two groups will arise from this but first let's make these ideas precise with some definitions.

9.1.1 Definition Let $\mathbf{a}, \mathbf{b} \in \mathbb{R}^2$ so that $\mathbf{a} = (a_1, a_2)$ and $\mathbf{b} = (b_1, b_2)$ where $a_1, a_2, b_1, b_2 \in \mathbb{R}$. Then the (Euclidean) distance between $\mathbf{a}$ and $\mathbf{b}$ is given by

$$|\mathbf{a} - \mathbf{b}| = \sqrt{(a_1 - b_1)^2 + (a_2 - b_2)^2}.$$

9.1.2 Definition Let $f : \mathbb{R}^2 \to \mathbb{R}^2$ be a function such that for any $\mathbf{a}, \mathbf{b} \in \mathbb{R}^2$, $|f(\mathbf{a}) - f(\mathbf{b})| = |\mathbf{a} - \mathbf{b}|$. Then f is called an *isometry* of $\mathbb{R}^2$. $\Diamond$

This just means that, in an isometry, the distance between any two given points and the distance between their image points is the same. You may hear such maps referred to as being *distance preserving.*

Note that such functions in this chapter are isometries of $\mathbb{R}^2$ so we'll sometimes just call them isometries (dropping the 'of $\mathbb{R}^2$').

9.1.3 Exercise Write down some functions which are isometries of $\mathbb{R}^2$. $\Diamond$

The following lemma is key to how we can see isometries as the elements of a group. It says that if we compose two isometries then the new thing we get is still an isometry. This means that composition is a binary operation on the set of isometries.

9.1.4 Lemma Let $f : \mathbb{R}^2 \to \mathbb{R}^2$ and $g : \mathbb{R}^2 \to \mathbb{R}^2$ be isometries of $\mathbb{R}^2$. Then $g \circ f : \mathbb{R}^2 \to \mathbb{R}^2$ is an isometry of $\mathbb{R}^2$.

Proof. Let $\mathbf{a}, \mathbf{b} \in \mathbb{R}^2$. Since f is an isometry $|f(\mathbf{a}) - f(\mathbf{b})| = |\mathbf{a} - \mathbf{b}|$. Since g is an isometry, $|g(f(\mathbf{a})) - g(f(\mathbf{b}))| = |f(\mathbf{a}) - f(\mathbf{b})|$. But then

$$|(g \circ f)(\mathbf{a}) - (g \circ f)(\mathbf{b})| = |g(f(\mathbf{a})) - g(f(\mathbf{b}))| = |f(\mathbf{a}) - f(\mathbf{b})| = |\mathbf{a} - \mathbf{b}|.$$

This means that $g \circ f$ is an isometry. $\Diamond$

Being an isometry is a very restrictive condition on a function and it turns out that we can specify that an isometry must take one of a very small number of forms. We start by finding these forms for isometries which fix the origin. In the following lemma we consider isometries which fix the orgin and the point $(1, 0)$.

9.1.5 Lemma Let $f : \mathbb{R}^2 \to \mathbb{R}^2$ be a function such that

1. $f((0, 0)) = (0, 0)$ and $f((1, 0)) = (1, 0)$
2. f is an isometry of $\mathbb{R}^2$.

Then f is either the identity map (i.e. $f((x, y)) = (x, y)$ for all $x, y \in \mathbb{R}$ or it is a reflection in the x-axis (i.e $f((x, y)) = (x, -y)$ for all $x, y \in \mathbb{R}$).

Proof. First consider the image of the point $(0, 1)$ under f. Let's call this (s, t), i.e. $f((0, 1)) = (s, t)$. (s, t) must be on the unit circle because it remains the same distance from the origin as $(0, 1)$ is. It must also be the same distance from $(1, 0)$ as $(0, 1)$ is. So it is on a circle centred at $(1, 0)$ with radius $\sqrt{2}$. These two circles are shown in Figure 10.

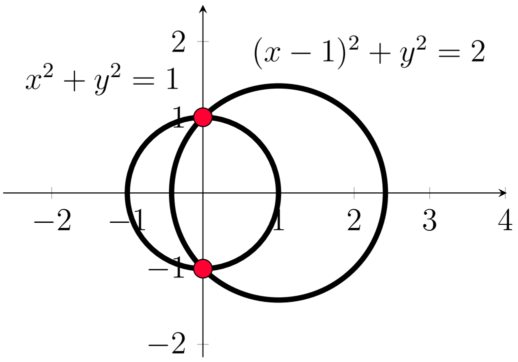


Figure 10: The two possibilities for the image of $(0, 1)$ are shown in red.

So (s, t) is on both of these circles which means that

$$s^2 + t^2 = 1 \text{ and } (s - 1)^2 + t^2 = 2.$$

Solving these simultaneously gives $s = 0$ and $t = \pm 1$. So there are only two possibilites for the image of $(0, 1)$, one is $(0, 1)$ and the other is $(0, -1)$. This gives us two cases in the proof.

Case 1 When $f((0,1)) = (0,1)$.

Let $\mathbf{p} = (x,y) \in \mathbb{R}^2$. Suppose $f(p) = \mathbf{q} = (u,v) \in \mathbb{R}^2$. The distance of $\mathbf{p}$ from $(0,0)$ is the same as the distance of $\mathbf{q}$ from $f((0,0)) = (0,0)$. Therefore

$$u^2 + v^2 = x^2 + y^2.$$

The distance of $\mathbf{p}$ from $(1,0)$ is the same as the distance of $\mathbf{q}$ from $f((1,0)) = (1,0)$. Therefore

$$(u-1)^2 + v^2 = (x-1)^2 + y^2.$$

By subracting one of theses equations from the other (the RHSs and the LHSs) we get $2u = 2x$. Therefore $u = x$.

We also have that the distance of $\mathbf{p}$ from $(0,1)$ is the same as the distance of $\mathbf{q}$ from $f((0,1)) = (0,1)$. Therefore

$$u^2 + (v-1)^2 = x^2 + (y-1)^2.$$

Recall that $u^2 + v^2 = x^2 + y^2$. Substracting one of these equations from the other, we have $2v = 2y$ and so $v = y$. So, in this case, $x = u$ and $y = v$ and so $\mathbf{p} = \mathbf{q}$ and f is the identity map.

Case 2 When $f((0,1)) = (0,-1)$.

Again let $\mathbf{p} = (x,y) \in \mathbb{R}^2$ and suppose $f(\mathbf{p}) = \mathbf{q} = (u,v) \in \mathbb{R}^2$. As in case 1. we can deduce that $u = x$.

The distance of $\mathbf{p}$ from $(0,1)$ is the same as the distance of $\mathbf{q}$ from $f((0,1)) = (0,-1)$. Therefore

$$u^2 + (v+1)^2 = x^2 + (y-1)^2.$$

Again $u^2 + v^2 = x^2 + y^2$. Substracting one of these equations from the other, we have $2v = -2y$ and so $v = -y$. Therefore, in this case, $u = x$ and $v = -y$ and so $f((x,y)) = (x,-y)$ and f is the reflection in the x-axis. $\Diamond$

9.1.6 Theorem Let $f : \mathbb{R}^2 \to \mathbb{R}^2$ be a function such that

1. $f((0,0)) = (0,0)$
2. f is an isometry of $\mathbb{R}^2$.

Then f is either a rotation about the origin or it is a reflection in the x-axis followed by a rotation about the origin.

Proof. Let $f((1,0)) = (u,v)$. Let θ be the angle between the line joining (u,v) to the origin and the line joining $(1,0)$ to the origin, measured anti-clockwise from the latter to the former. Let $g_\theta : \mathbb{R}^2 \to \mathbb{R}^2$ be the function which is an anti-clockwise rotation through angle θ about the origin. This clearly has an inverse function (or, equivalently, is a bijection). Consider the function $(g_\theta)^{-1} \circ f$. We have that

$$(g_\theta^{-1} \circ f)(1,0) = g_\theta^{-1}((u,v)) = (1,0).$$

Also $(g_\theta^{-1} \circ f)(0,0) = g_\theta^{-1}(f(0,0)) = g_\theta^{-1}(0,0) = (0,0)$ and $g_\theta^{-1} \circ f$ an isometry (this is by Lemma 9.1.4 since both f and $(g_\theta)^{-1}$ are isometries). Therefore by Lemma 9.1.5, $(g_\theta)^{-1} \circ f$ is either the identity map or a reflection in the x-axis.

If $(g_\theta)^{-1} \circ f$ is the identity map, then $f = (g_\theta \circ (g_\theta)^{-1}) \circ f = g_\theta \circ ((g_\theta)^{-1} \circ f) = g_\theta$.

If $(g_\theta)^{-1} \circ f$ is reflection in the x axis, which we'll call r_x, then $f = (g_\theta \circ (g_\theta)^{-1}) \circ f = g_\theta \circ ((g_\theta)^{-1} \circ f) = g_\theta \circ r_x$.

This means that f is either a rotation about the origin or it is a reflection in the x-axis followed by a rotation about the origin. $\Diamond$

Now we make a useful switch to the notation of complex numbers. If we think of $\mathbb{R}^2$ as being the complex plane (Argand diagram), we can describe describe a rotation by multiplication by a complex number with modulus 1 (which will therefore take the form $e^{i\theta}$) and the reflection in the x-axis by conjugation. So we'll now think of the point (x,y) as being, or being identified with if you prefer, the complex number $x + iy$ in the Argand diagram.

In these terms Theorem 9.1.6 says the followng. Let $f : \mathbb{C} \to \mathbb{C}$ be a function such that

1. $f(0) = 0$
2. f preserves distance, i.e. for any $w, z \in \mathbb{C}$, $|f(w) - f(z)| = |w - z|$.

Then f either has form $f(z) = e^{i\theta}z$ for some $\theta \in \mathbb{R}$ or form $f(z) = e^{i\theta}\overline{z}$.

The set of all such functions is actually a group (under composition of functions)!

9.1.7 Theorem The set of isometries of $\mathbb{R}^2$, here identified with $\mathbb{C}$, which fix the origin is a group under composition of funtions. We call this group $O_2(\mathbb{R})$, the orthogonal group on $\mathbb{R}^2$.

Proof. We need to check (i) - (iv) in the definition of a group.

(i) Closure. We could deduce this using Lemma 9.1.4 and noting that the composition of two fiunctions which fix the orgin will fix the origin. However it's useful to look at the way multiplication (i.e. composition) really works in this group. By Theorem 9.1.6 every element of $O_2(\mathbb{R})$ is a function f which either has form form $f(z) = e^{i\theta}z$ or form $f(z) = e^{i\theta}\overline{z}$ for some $\theta \in \mathbb{R}$. 'Multiplication' (i.e. composition) in this group works like this:

		Type of first isometry	
	$\circ$	$e^{i\theta}z$	$e^{i\theta}\overline{z}$
Type of second isometry	$e^{i\omega}z$	$e^{i(\omega+\theta)}z$	$e^{i(\omega+\theta)}\overline{z}$
	$e^{i\omega}\overline{z}$	$e^{i(\omega-\theta)}\overline{z}$	$e^{i(\omega-\theta)}z$

Figure 11: The types of isometry of the plane which fix the origin and how they combine.

From this we can see that the binary operation is closed because if we combine any of the two types in any order the result is still an isometry.

(ii) Associativity. This is a group where the elements are functions from a set to itself and the binary operation is composition. Such a binary operation is always associative. See Lemma 8.2.3.

(iii) Identity. Clearly the identity map is an isometry of $\mathbb{R}^2$ which fixes the origin. As one of the two standard forms it is the map $f(z) = e^{0\times i}z = 1 \times z = z$ .This is the identity element in the group.

(iii) Existence of inverses. If f has form $f(z) = e^{i\theta}z$ then $f^{-1}(z) = e^{-i\theta}z$ (which we can see is of the first form) and if f has form $f(z) = e^{i\theta}\overline{z}$ then $f^{-1}(z) = \overline{e^{-i\theta}z} = e^{i\theta}\overline{z}$ (which we can see is of the second form) so inverses exist. $\Diamond$

It's worth noting that we could have expressed all of the above in matrix notation, thinking of z as $x + iy$:

1. $f(z) = e^{i\theta}z$ is equivalent to $f\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} \cos(\theta) & -\sin(\theta) \\ \sin(\theta) & \cos(\theta) \end{pmatrix}\begin{pmatrix} x \\ y \end{pmatrix}$.

2. $f(z) = e^{i\theta}\overline{z}$ is equivalent to $f\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} \cos(\theta) & -\sin(\theta) \\ \sin(\theta) & \cos(\theta) \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix}$.

We'll make this idea of being able to think of a group in more than one way a bit more precise in Chapter 12. We'll generally stick with complex numbers notation for this as it's a bit more elegant and certainly more compact.

9.1.8 Exercise Multiply out $\begin{pmatrix} \cos(\theta) & -\sin(\theta) \\ \sin(\theta) & \cos(\theta) \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}$. What sort of transformation does this represent?

9.1.9 Exercise Show that the inverse matrix to both $\begin{pmatrix} \cos(\theta) & -\sin(\theta) \\ \sin(\theta) & \cos(\theta) \end{pmatrix}$ and $\begin{pmatrix} \cos(\theta) & -\sin(\theta) \\ \sin(\theta) & \cos(\theta) \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}$ (once multiplied out) is its transpose. Matrices like this are called *orthogonal.*

9.1.10 Theorem The set of isometries, $f : \mathbb{R}^2 \to \mathbb{R}^2$ of the form $f(z) = e^{i\theta}z$, where $\mathbb{R}^2$ has been identified with $\mathbb{C}$, is a subgroup of $O_2(\mathbb{R})$. This is the subgroup of all the rotations about the orgin. It is called $SO_2(\mathbb{R})$, the special orthogonal group on $\mathbb{R}^2$. $\Diamond$

Proof. We need to check conditions (a), (b) and (c) in Theorem 6.3.1.

(a) The identity in $O_2(\mathbb{R})$ is $f(z) = z = e^{i\times 0}z$. This is $f(z) = e^{i\theta}z$ with $\theta = 0$ and so it is in $SO_2(\mathbb{R})$.

(b) Suppose $f, g \in SO_2(\mathbb{R})$. Then $f(z) = e^{i\theta_1}z$ and $g(z) = e^{i\theta_2}z$ for some $\theta_1, \theta_2 \in \mathbb{R}$. Then $(g \circ f)(z) = g(f(z)) = g(e^{i\theta_1 z} = e^{i\theta_1}e^{i\theta_2}z = e^{i(\theta_1+\theta_2)}z$. This is in $SO_2(\mathbb{R})$.

(c) Suppose $f \in SO_2(\mathbb{R})$. Then $f^{-1}(z) = e^{-i\theta}z$ which can be seen to be in $SO_2(\mathbb{R})$. $\Diamond$

9.2 General isometries of the plane

Now lets consider general isometries of the plane, i.e. those which don't necessarily fix the origin. It turns out we have already done much of the work to understand these. We'll stick with the idea of having identified $\mathbb{R}^2$ with $\mathbb{C}$ here.

9.2.1 Theorem Let f be an isometry of $\mathbb{R}^2$, identified with $\mathbb{C}$ as above. Then f has one of the following two forms.

1. $f(z) = e^{i\theta}z + w$ where $\theta \in \mathbb{R}$ and $w \in \mathbb{C}$
2. $f(z) = e^{i\theta}\overline{z} + w$, where $\theta \in \mathbb{R}$ and $w \in \mathbb{C}$.

Proof. Let $w = f(0)$. Let h_{-w} be the translation through $-w$, i.e. $h_{-w}(z) = z - w$. Clearly h_{-w} is an isometry. By Lemma 9.1.4 $h_{-w} \circ f$ is an isometry.

$(h_{-w} \circ f)(0) = h_{-w}(f(0)) = h_{-w}(w) = w - w = 0$ so $h_{-w} \circ f$ is an isometry which fixes the origin. By Theorem 9.1.6 either

1. there is a real number θ such that $h_{-w} \circ f(z) = e^{i\theta}z$ for all $z \in \mathbb{C}$. This means that, for all $z \in \mathbb{C}$, $f(z) - w = e^{i\theta}z$, or $f(z) = e^{i\theta}z + w$, or
2. there is a real number θ such that $h_{-w} \circ f(z) = e^{i\theta}\overline{z}$ for all $z \in \mathbb{C}$. This means that, for all $z \in \mathbb{C}$, $f(z) - w = e^{i\theta}\overline{z}$, or $f(z) = e^{i\theta}\overline{z} + w$. $\Diamond$

The set of all isometries of $\mathbb{R}^2$ is again a group. This group is called $\mathrm{Eucl}(\mathbb{R}^2)$. Here is a sketch of the details.

The inverses of the two fypes are $f^{-1}(z) = e^{-i\theta}(z - w)$ and $f^{-1}(z) = \overline{e^{-i\theta}(z - w)} = e^{i\theta}\overline{z} - e^{i\theta}\overline{\,}(w)$ respectively.

Here is how these isometries combine:

		Type of first isometry	
	$\circ$	$e^{i\theta}z + u$	$e^{i\theta}\overline{z} + u$
Type of second isometry	$e^{i\omega}z + v$	$e^{i(\omega+\theta)}z + e^{i\omega}u + w$	$e^{i(\omega+\theta)}\overline{z} + e^{i\omega}u + w$
	$e^{i\omega}\overline{z} + v$	$e^{i(\omega-\theta)}\overline{z} + e^{i\omega}\overline{u} + w$	$e^{i(\omega-\theta)}z + e^{i\omega}\overline{u} + w$

Figure 12: The types of isometry of the plane and how they combine.

It's clear that the identity map is an isometry of $\mathbb{R}^2$, so we again have a group, under composition of functions.

Again, it's worth noting that we could have expressed all of the above in matrix notation, thinking of z as $x + iy$ and w as $a + bi$:

1. $f(z) = e^{i\theta}z + w$ is equivalent to

$$f\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} \cos(\theta) & -\sin(\theta) \\ \sin(\theta) & \cos(\theta) \end{pmatrix}\begin{pmatrix} x \\ y \end{pmatrix} + \begin{pmatrix} a \\ b \end{pmatrix}.$$

2. $f(z) = e^{i\theta}\overline{z} + w$ is equivalent to

$$f\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} \cos(\theta) & -\sin(\theta) \\ \sin(\theta) & \cos(\theta) \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} + \begin{pmatrix} a \\ b \end{pmatrix}.$$

9.2.2 Exercise Prove that $O_2(\mathbb{R})$ is a subgroup of $\mathrm{Eucl}(\mathbb{R}^2)$.

9.2.3 Exercise Show that set of elements of $\mathrm{Eucl}(\mathbb{R}^2)$ which have the form $f(z) = z + w$ for some $w \in \mathbb{C}$ is a subgroup of $\mathrm{Eucl}(\mathbb{R}^2)$.

Chapter 10 - Symmetric Groups

Probably the most interesting groups have elements that are functions. In the last chapter we saw some examples and the groups we are about to meet in this chapter, the symmetric groups, are other examples. It turns out that every finite group can be thought of as a subgroup of one of the symmetric groups. This is called Cayley's Theorem, where this statement is made precise. You won't meet this until later courses in group theory.

10.1 Motivation

Let A be a set, and let f, g be functions from A to itself. We know that we can compose f, g to obtain $f \circ g$ which is also a function from A to itself. We shall write $\mathrm{Map}(A)$ for the set of functions from A to itself. Then $\circ$ is a binary operation on $\mathrm{Map}(A)$. And it's natural to ask if this makes $\mathrm{Map}(A)$ into a group. After all, we know by Lemma 8.2.3 that composition of functions is associative. The following example will help clarify these ideas.

10.1.1 Example Let $A = \{1, 2\}$. You will quickly convince yourself that there are only four functions from A to itself, which are given in Figure 13.

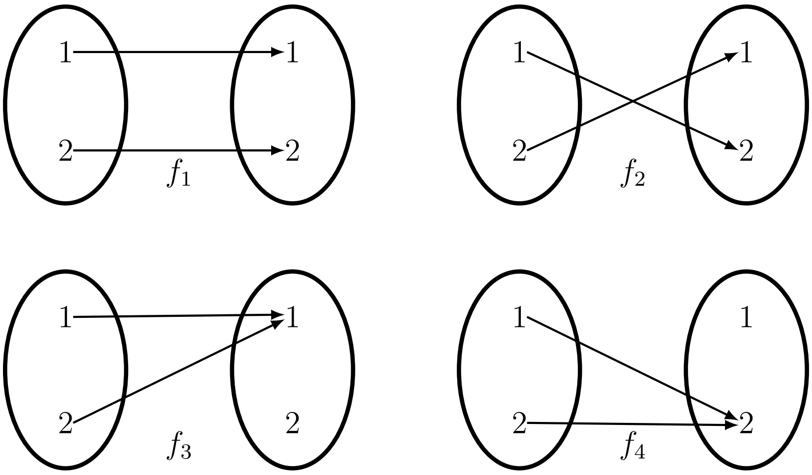


Figure 13: f_1, f_2, f_3 and f_4 are the four functions from $\{1, 2\}$ to itself.

Thus $\mathrm{Map}(A) = \{f_1, f_2, f_3, f_4\}$. Is $\mathrm{Map}(A)$ a group with respect to composition of functions? Here is the composition table for $\mathrm{Map}(A)$:

$\circ$	f_1	f_2	f_3	f_4
f_1	f_1	f_2	f_3	f_4
f_2	f_2	f_1	f_4	f_3
f_3	f_3	f_3	f_3	f_3
f_4	f_4	f_4	f_4	f_4

Make sure you understand the table. The entry for $f_i \circ f_j$ is at the intersection of the i-th row and j-th column. As always, $f_i \circ f_j$ means apply f_j first then f_i. We know that composition of functions is associative by Lemma 8.2.3. Moreover, it is clear from the table that f_1 is the identity element. But f_3 and f_4 don't have inverses; we can't combine either of them with any of the four functions to obtain the identity f_1.

But if you look carefully at the table, you will see a group with respect to composition. It is the subset: $\{f_1, f_2\}$. We already know why f_1, f_2 have inverses (which in this case happen to be f_1 and f_2 respectively), and f_3, f_4 don't: the functions f_1, f_2 are bijections and f_3 and f_4 are not. This was Theorem 8.2.6 which says that a function f from a set X to a set Y has an inverse if and only if it is a bijection. In the example we have been looking at $X = Y = \{1, 2\}$.

$\diamondsuit$

10.2 The symmetric group on a general set A

Let A be a set. We shall denote the set of bijections from A to itself by $\mathrm{Sym}(A)$.

10.2.1 Example In Example 10.1.1 we wrote down all the functions from $A = \{1, 2\}$ to itself and found that exactly two of these are bijections. These were called f_1 and f_2 in Figure 13. Hence $\mathrm{Sym}(A) = \{f_1, f_2\}$. Note that $f_1 = \mathrm{id}_A$, the function which sends every element of A to itself. In that example, we noted that $\{f_1, f_2\}$ is a group under composition with f_1 being the identity element. Check this again, and note that the group is abelian. $\diamondsuit$

10.2.2 Theorem Let A be a set. Then $(\mathrm{Sym}(A), \circ)$ is a group with id_A as the identity element.

We call $\mathrm{Sym}(A)$ the *symmetric group* on A.

Proof.

We need to do this usual checks: closure; associativity; identity element; existence of inverses.

Closure in this group means showing that if $f : A \to A$ and $g : A \to A$ are both bijections then $(g \circ f) : A \to A$ is also a bijection. We do this as follows.

Let $c \in A$. Since g is surjective there exists $b \in A$ such that $g(b) = c$. Since f is surjective there exists $a \in A$ such that $f(a) = b$. Then $(g \circ f)(a) = g(f(a)) = g(b) = c$ and so $g \circ f$ is surjective.

Let $a_1, a_2 \in A$ be such that $(g \circ f)(a_1) = (g \circ f)(a_2)$. Then $g(f(a_1)) = g(f(a_2))$ and by the injectivitiy of g, $f(a_1) = f(a_2)$. Then, by the injectivitiy of f, $a_1 = a_2$ and so $g \circ f$ is injective.

Therefore $g \circ f$ is a bijection.

Composition of functions is associative by Lemma 8.2.3.

Clearly id_A is a bijection and so is in $\mathrm{Sym}(A)$. We want to check that id_A is the identity for composition, which means that for any $f \in \mathrm{Sym}(A)$ we want $f \circ \mathrm{id}_A = \mathrm{id}_A \circ f = f$. Note

$$(f \circ \mathrm{id}_A)(x) = f(\mathrm{id}_A(x)) = f(x), \qquad (\mathrm{id}_A \circ f)(x) = \mathrm{id}_A(f(x)) = f(x).$$

Thus $f \circ \mathrm{id}_A = \mathrm{id}_A \circ f = f$ holds.

Finally we want every element of $\mathrm{Sym}(A)$ to have an inverse in $\mathrm{Sym}(A)$. This is true by Theorem 8.2.6.

10.2.3 Exercise Let $f : \mathbb{Z} \to \mathbb{Z}$ and $g : \mathbb{R} \to \mathbb{R}$ be given by $x \mapsto 2x$. Show that $f \notin \mathrm{Sym}(\mathbb{Z})$ but $g \in \mathrm{Sym}(\mathbb{R})$. Write down g^n for integers n.

10.2.4 Exercise Let $f : \mathbb{C} \to \mathbb{C}$, $g : \mathbb{C} \to \mathbb{C}$, $h : \mathbb{C} \to \mathbb{C}$ be given by $f(z) = z + 1$, $g(z) = z + i$, $h(z) = iz$. Describe f, g, h geometrically. Show that f, g, h are in $\mathrm{Sym}(\mathbb{C})$. Show that f and g commute. What about f and h or g and h? What are the orders of f, g and h?

10.3 S_n

We define S_n to be the group $\mathrm{Sym}(\{1, 2, \dots, n\})$. We call S_n the *n-th symmetric group*. In Example 10.2.1 we found that S_2 is a group of order 2.

10.3.1 Theorem S_n has order $n!$.

Proof. S_n is the set of bijections from $\{1, 2, \ldots, n\}$ to itself. So we want to count these bijections. It's clear that any injective function from $\{1, 2, \ldots, n\}$ to itself will be surjective (because if distinct elements get sent to distinct elements then the number of elements that get 'hit' must be n, i.e. all elements of $\{1, 2, \ldots, n\}$ are 'hit' and the function is surjective).

So let's count the injections. Let f be an injection from $\{1, 2, \ldots, n\}$ to itself. Then $f(1)$ can be any of $1, 2, \ldots, n$; that is, there are n choices for $f(1)$. If we fix $f(1)$ then $f(2) \neq f(1)$. So there are $n-1$ choices for $f(2)$ once we've chosen $f(1)$. Likewise there are $n-2$ choices for $f(3)$ once we've chosen $f(1)$ and $f(2)$. It is now clear that the number of injections is

$$n \times (n-1) \times \cdots \times 1 = n!.$$

The elements of S_n are called *permutations*. One way of representing permutations is to use diagrams such as those for f_1, $f_2 \in S_2$ in Figure 13. The following is a more economical way. Let $a_1, a_2, \ldots, a_n$ be the numbers $1, 2, \ldots, n$ in some order. Then

$$\begin{pmatrix} 1 & 2 & \cdots & n \\ a_1 & a_2 & \cdots & a_n \end{pmatrix}$$

represents the unique permutation in S_n that sends 1 to a_1, 2 to a_2, …, and n to a_n.

10.3.2 Example S_2 has two elements:

$$\begin{pmatrix} 1 & 2 \\ 1 & 2 \end{pmatrix}, \qquad \begin{pmatrix} 1 & 2 \\ 2 & 1 \end{pmatrix}.$$

These are respectively the same as f_1, f_2 in Figure 13. The first of these is the identity element. We noted in Example 10.2.1 that $S_2 = \mathrm{Sym}(\{1, 2\})$ is abelian. $\Diamond$

10.3.3 Example We know from Theorem 10.3.1 that S_3 has 6 elements. These are

$$\begin{pmatrix} 1 & 2 & 3 \\ 1 & 2 & 3 \end{pmatrix}, \qquad \begin{pmatrix} 1 & 2 & 3 \\ 2 & 3 & 1 \end{pmatrix}, \qquad \begin{pmatrix} 1 & 2 & 3 \\ 3 & 1 & 2 \end{pmatrix}$$
$$\begin{pmatrix} 1 & 2 & 3 \\ 1 & 3 & 2 \end{pmatrix}, \qquad \begin{pmatrix} 1 & 2 & 3 \\ 2 & 1 & 3 \end{pmatrix}, \qquad \begin{pmatrix} 1 & 2 & 3 \\ 3 & 2 & 1 \end{pmatrix}.$$

Again, the first of these is the identity element. It is important that you know what the notation means and how to multiply two permutations written in this notation, so let's have some practice. Let

$$\rho = \begin{pmatrix} 1 & 2 & 3 \\ 3 & 1 & 2 \end{pmatrix}, \qquad \mu = \begin{pmatrix} 1 & 2 & 3 \\ 1 & 3 & 2 \end{pmatrix}.$$

Never forget that these are bijections from $\{1, 2, 3\}$ to itself. To find out what ρ does, look at the columns. ρ is the function that sends 1 to 3, 2 to 1 and 3 to 2. Thus

$$\rho(1) = 3, \qquad \rho(2) = 1, \qquad \rho(3) = 2. \tag{3}$$

Likewise,

$$\mu(1) = 1, \qquad \mu(2) = 3, \qquad \mu(3) = 2.$$

Now let us compute $\rho\mu$. As always, this means apply μ first then ρ. So

$$\begin{aligned} (\rho\mu)(1) &= \rho(\mu(1)) = \rho(1) = 3; \\ (\rho\mu)(2) &= \rho(\mu(2)) = \rho(3) = 2; \\ (\rho\mu)(3) &= \rho(\mu(3)) = \rho(2) = 1. \end{aligned}$$

Thus

$$\rho\mu = \begin{pmatrix} 1 & 2 & 3 \\ 3 & 2 & 1 \end{pmatrix}.$$

Similarly,

$$\mu\rho = \begin{pmatrix} 1 & 2 & 3 \\ 2 & 1 & 3 \end{pmatrix}.$$

Note that $\mu\rho \neq \rho\mu$, so S_3 is non-abelian. How do we compute ρ^{-1}? From (3) we find

$$1 = \rho^{-1}(3), \qquad 2 = \rho^{-1}(1), \qquad 3 = \rho^{-1}(2).$$

We rearrange this:

$$\rho^{-1}(1) = 2, \qquad \rho^{-1}(2) = 3, \qquad \rho^{-1}(3) = 1.$$

Hence

$$\rho^{-1} = \begin{pmatrix} 1 & 2 & 3 \\ 2 & 3 & 1 \end{pmatrix}.$$

◊

10.3.4 Exercise Let ρ and τ be the following permutations:

$$\rho = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 \\ 2 & 3 & 5 & 1 & 4 \end{pmatrix}, \qquad \tau = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 \\ 3 & 1 & 2 & 5 & 4 \end{pmatrix}.$$

Compute ρ^{-1}, $\rho\tau$, τ^2.

10.3.5 Exercise Show that S_n is non-abelian for $n \geq 3$.

10.4 Cycle notation

Let $a_1, a_2, \dots, a_m$ be distinct elements of the set $\{1, 2, \dots, n\}$. By the notation

$$(a_1, a_2, \dots, a_m) \tag{4}$$

we mean the element of S_n that takes a_1 to a_2, a_2 to a_3, $\dots$, a_{m-1} to a_m and a_m back to a_1, and fixes all other elements of $\{1, 2, \dots, n\}$. The permutation (4) is called a *cycle of length m*. A cycle of length 2 is called a *transposition*.

10.4.1 Example Let $\mu = (1, 4, 5) \in S_5$. The cycle μ is of length 3 and is illustrated in Figure 14.

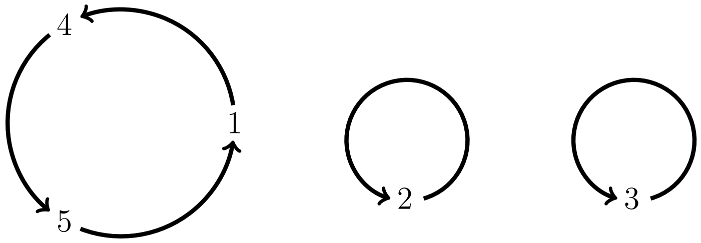


Figure 14: The cycle $(1, 4, 5) \in S_5$.

We can write $(1, 4, 5)$ using our old notation:

$$(1, 4, 5) = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 \\ 4 & 2 & 3 & 5 & 1 \end{pmatrix}.$$

Notice that $(1, 4, 5) = (4, 5, 1) = (5, 1, 4)$. However, $(1, 4, 5) \neq (1, 5, 4)$.

The transposition $(1, 5) \in S_5$ is given in Figure 15.

In our old notation, the transposition $(1, 5)$ is written as follows:

$$(1, 5) = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 \\ 5 & 2 & 3 & 4 & 1 \end{pmatrix}.$$

Note that $(1, 5) = (5, 1)$.

Finally (1) is the cycle that takes 1 to itself and fixes all the other elements. Clearly $(1) = (2) = (3) = (4) = (5) = \mathrm{id}$ is nothing other than the identity permutation. $\diamondsuit$

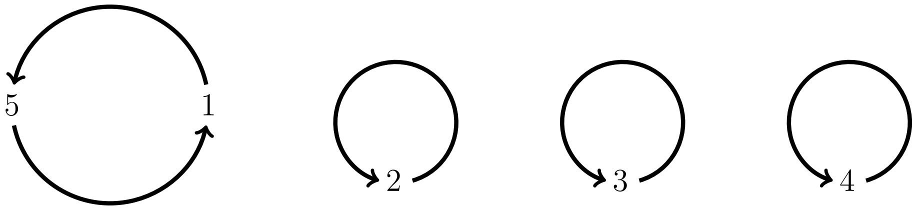


Figure 15: The transposition $(1, 5) \in S_5$. This merely swaps 1 and 5, and fixes all other elements.

I hope that the above example has convinced you that cycle notation is simultaneously more concise and and more transparent than the old notation. If so, the following theorem, where we show that every permutation can be written as a product of disjoint cycles will come as a pleasant surprise!

Before we state it, what does *disjoint* mean? Two cycles $(a_1, a_2, \ldots, a_n)$ and $(b_1, b_2, \ldots, b_m)$ are said to be disjoint if $a_i \neq b_j$ for all integers i, j with $1 \leq i \leq n$ and $1 \leq j \leq m$. What does *product* mean? The product of two permutations is their composition as functions.

10.4.2 Theorem Every permutation can be written as a product of disjoint cycles.

Proof. Let ρ be an element of S_n. Consider the sequence

$$1, \quad \rho 1, \quad \rho^2 1, \quad \rho^3 1, \ldots$$

Every term in this infinite sequence is contained in the finite set $\{1, 2, \ldots, n\}$. Thus the sequence must contain repetition. Let $\rho^u 1$ be the first term in the sequence that has already appeared. Thus $\rho^u 1 = \rho^v 1$ for some $0 \leq v < u$. Apply ρ^{-v} to both sides. We obtain $\rho^{u-v} 1 = 1$. Note that $0 < u - v \leq u$. If $u - v < u$, then $\rho^{u-v} 1$ is in fact the first term in the sequence that has already appeared, which contradicts our assumption. Therefore, $u - v = u$ and so $v = 0$. Hence $\rho^u 1 = 1$, and $1, \rho 1, \ldots, \rho^{u-1} 1$ are distinct.

Let μ_1 be the cycle of length u

$$\mu_1 = \left(1, \rho 1, \rho^2 1, \ldots, \rho^{u-1} 1\right).$$

It is clear that μ_1 has the same effect as ρ on the elements $1, \rho 1, \ldots, \rho^{u-1} 1$.

Now let a be the first element of the set $\{1, 2, \ldots, n\}$ not appearing in the list $1, \rho 1, \ldots, \rho^{u-1} 1$. Repeat the above argument with the sequence

$$a, \rho a, \rho^2 a, \rho^3 a, \ldots.$$

We deduce the existence of a cycle

$$\mu_2 = \big(a, \rho a, \ldots, \rho^{v-1} a\big)$$

such that μ_2 and ρ have the same effect on the elements $a, \rho a, \ldots, \rho^{v-1} a$. Let us show that μ_1 and μ_2 are disjoint. Suppose otherwise. Then $\rho^i 1 = \rho^j a$ for some $0 \leq i < u$ and $0 \leq j < v$. Now apply ρ^{v-j} to both sides to obtain $\rho^k 1 = a$ where $k = i + v - j$. This contradicts our assumption that a does not appear in the list $1, \rho 1, \ldots, \rho^{u-1} 1$. Hence the cycles μ_1 and μ_2 are disjoint. Now the product $\mu_1 \mu_2$ has the same effect as ρ on the elements $1, \rho 1, \ldots, \rho^{u-1} 1, a, \rho a, \ldots, \rho^{v-1} a$.

We repeat the process, starting with the first element of $\{1, 2, \ldots, n\}$ not appearing in either cycle μ_1, μ_2 to construct a μ_3 that is disjoint from both μ_1 and μ_2, etc. As the set $\{1, 2, \ldots, n\}$ is finite, this process must terminate eventually with some μ_r. The product of disjoint cycles $\mu_1 \mu_2 \ldots \mu_r$ has the same effect on $\{1, \ldots, n\}$ as ρ. Therefore

$$\rho = \mu_1 \mu_2 \cdots \mu_r.$$

$\diamondsuit$

Let's see an example where we write down a permutation as a product of cycles.

10.4.3 Example Let

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

Write ρ as a product of disjoint cycles.

Answer: We start with 1 are repeatedly apply ρ to it:

$$1 \mapsto 5 \mapsto 8 \mapsto 3 \mapsto 1.$$

Therefore ρ contains the cycle $(1, 5, 8, 3)$. Now we start with an element of the set $\{1, 2, \ldots, 8\}$ that is not contained in the cycle $(1, 5, 8, 3)$. For example start with 2 and repeatedly apply ρ to it:

$$2 \mapsto 7 \mapsto 6 \mapsto 2.$$

So ρ also contains the cycle $(2,7,6)$. Note that the cycles $(1,5,8,3)$ and $(2,7,6)$ are disjoint, and ρ contains the product (or composition) $(1,5,8,3)(2,7,6)$. There still remains one element of the set $\{1,2,\dots,8\}$ that does not appear as either of the two cycles $(1,5,8,3)$ and $(2,7,6)$ and this is 4. Applying ρ to 4 we find:

$$4 \mapsto 4.$$

So

$$\rho = (1,5,8,3)(2,7,6)(4)$$

as a product of disjoint cycles. Recall that (4) is just the identity, so it is usual to omit it and write,

$$\rho = (1,5,8,3)(2,7,6).$$

You might be wondering why we wrote ρ as above and not $\rho = (2,7,6)(1,5,8,3)$. This does not matter since **disjoint cycles commute**; more on this below. $\diamondsuit$

10.4.4 Example Let

$$\sigma = (1,3,10,9)(2,5,6), \qquad \tau = (4,3,10)(1,5,8).$$

Express $\sigma\tau$ and σ^{-1} as a product of disjoint cycles.

Answer: We start with 1 and follow the same procedure as the above example. Note that $\sigma\tau 1$ means apply τ first to 1 and then apply σ to the result. Now $\tau 1 = 5$ and $\sigma 5 = 6$. So $\sigma\tau 1 = 6$. Next we apply $\sigma\tau$ to 6. The permutation τ does not have 6 in its cycle decomposition, so $\tau 6 = 6$. So $\sigma\tau 6 = \sigma 6 = 2$. We keep applying $\sigma\tau$ until we return to 1:

$$1 \mapsto 6 \mapsto 2 \mapsto 5 \mapsto 8 \mapsto 3 \mapsto 9 \mapsto 1.$$

Thus $\sigma\tau$ has the cycle $(1,6,2,5,8,3,9)$ in its decomposition as a product of disjoint cycles. We note that this cycle has no 4 in it. So we apply $\sigma\tau$ repeatedly starting with 4:

$$4 \mapsto 10 \mapsto 4.$$

Hence $\sigma\tau$ has the product $(1,6,2,5,8,3,9)(4,10)$ in its decomposition as a product of disjoint cycles. Finally, note that of the elements of the set $\{1,2,\dots,10\}$, the only one not appearing in the product $(1,6,2,5,8,3,9)(4,10)$ is 7. However $\sigma\tau 7 = 7$. So

$$\sigma\tau = (1,6,2,5,8,3,9)(4,10)$$

as a product of disjoint cycles.

You may have noticed that we were tacitly assuming that σ and τ are elements of S_{10} and computed the product under that assumption. In fact, we would have obtained the same result had σ and τ been elements of $S_{11}, S_{12}, \ldots$. Indeed viewed as elements of S_{11}, the permutations σ and τ, and the cycles $(1,6,2,5,8,3,9)$ and $(4,10)$ all fix 11.

To compute σ^{-1} we start with $\sigma = (1,3,10,9)(2,5,6)$ and reverse the arrows: Therefore $\sigma^{-1} = (1,9,10,3)(2,6,5)$. Check for yourself that $\sigma\sigma^{-1}$ is indeed the identity permutation. ♢

10.4.5 Exercise Let ρ and τ be as given in Exercise 10.3.4. Write ρ and τ as products of disjoint cycles.

10.4.6 Exercise Which of the following pairs of permutations are equal elements of S_6?

(i) $(1,2,3)(4,6)$ and $(6,4)(2,3,1)(5)$.

(ii) $(4,5,6)(1,2,3)$ and $(3,1,2)(5,4,6)$.

10.4.7 Exercise Let $\rho = (1,2,3)(4,5)$ and $\tau = (1,2,3,4)$. Write the following in cycle notation (i.e. as a product of disjoint cycles): ρ^{-1}, τ^{-1}, $\rho\tau$, $\tau\rho^2$.

10.4.8 Lemma Disjoint cycles commute.

Proof. Let σ and τ be disjoint cycle in S_n and write

$$\sigma = (a_1, a_2, \ldots, a_k), \qquad \tau = (b_1, b_2, \ldots, b_\ell).$$

Since σ and τ are disjoint $a_i \neq b_j$ for $i = 1, \ldots, k$ and $j = 1, \ldots, \ell$.

We want to show that $\sigma\tau = \tau\sigma$. This means that $\sigma\tau x = \tau\sigma x$ for all $x \in \{1, 2, \ldots, n\}$. We subdivide into three cases:

Case 1: x does not equal any of the a_i or b_j. Then $\tau x = x$ and $\sigma x = x$. Therefore

$$\sigma\tau x = \sigma x = x = \tau x = \tau\sigma x.$$

Case 2: $x = a_i$ for some $i = 1, \ldots, k$. Thus x does not equal any of the b_j, and so $\tau x = x$. Hence $\sigma\tau x = \sigma x = \sigma a_i = a_{i+1}$; here a_{k+1} is interpreted as being a_1. Let's compute $\tau\sigma x$. This is $\tau\sigma a_i = \tau a_{i+1} = a_{i+1}$ since a_{i+1} does not equal any of the b_j.

Hence $\sigma\tau x = \tau\sigma x$.

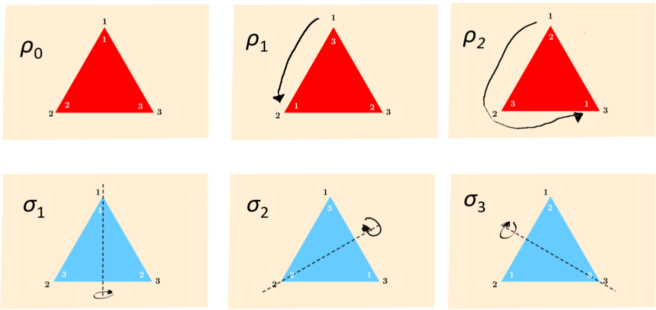
Case 3: $x = b_j$ for some $j = 1, \dots, \ell$. This is similar to Case 2.

We conclude that $\sigma\tau = \tau\sigma$ as required.

10.5 D_{2n} and S_n

Cast your mind back to Chapter 1 when we met the six symmetries of an equilateral triangle.

We gave each of the symmetries a name:



And this is the the composition/table using the binary operation 'followed by'(essentially composition of symmetries):

$\circ$	ρ_0	ρ_1	ρ_2	σ_1	σ_2	σ_3
ρ_0	ρ_0	ρ_1	ρ_2	σ_1	σ_2	σ_3
ρ_1	ρ_1	ρ_2	ρ_0	σ_3	σ_1	σ_2
ρ_2	ρ_2	ρ_0	ρ_1	σ_2	σ_3	σ_1
σ_1	σ_1	σ_2	σ_3	ρ_0	ρ_1	ρ_2
σ_2	σ_2	σ_3	σ_1	ρ_2	ρ_0	ρ_1
σ_3	σ_3	σ_1	σ_2	ρ_1	ρ_2	ρ_0

This is the group D_6.

Remember that we labelled the vertices of the equilateral triangle with the numbers 1, 2, 3. Notice how there is a natural correspondence between the elements of D_6 and elements of S_3. For example σ_2 swaps vertices 1 and 3 and so it corresponds to $(1,3) \in S_3$.

The elements of S_3 which correspond to

$$\rho_0, \rho_1, \rho_2, \rho_3, \sigma_1, \sigma_2, \sigma_3 \in D_6$$

written in disjoint cycle notation are, respectively:

$$(1); (1,2,3); (1,3,2); (2,3); (1,3); (1,2).$$

Notice that the list above is the whole of S_3. Here is the composition/multiplication table for S_3.

$\circ$	(1)	$(1,2,3)$	$(1,3,2)$	$(2,3)$	$(1,3)$	$(1,2)$
(1)	(1)	$(1,2,3)$	$(1,3,2)$	$(2,3)$	$(1,3$	$(1,2)$
$(1,2,3)$	$(1,2,3$	$(1,3,2)$	(1)	$(1,2)$	$(2,3)$	$(1,3)$
$(1,3,2)$	$(1,3,2)$	(1)	$(1,2,3)$	$(1,3)$	$(1,2)$	$(2,3)$
$(2,3)$	$(2,3)$	$(1,3)$	$(1,2))$	(1)	$(1,2,3)$	$(1,3,2)$
$(1,3)$	$(1,3)$	$(1,2)$	$(2,3)$	$(1,3,2)$	(1)	$(1,2,3)$
$(1,2)$	$(1,2)$	$(2,3)$	$(1,3)$	$(1,2,3)$	$(1,3,2)$	(1)

This is identical to the D_6 table if you carefully swap all the entries according to the correspondence above! From an asbtract algebra point of view, when we are really only interested in the elements and how they combine, is there really any distinction between D_6 and S_3? Chapter 12 contains more on this.

Now cast your mind back to Chapter 3 when we met D_8 and labelled the vertices of the square with 1, 2, 3, 4 as in Figure 2. Go back there and remind yourself of the notation for the elements of D_8.

In the notation established there, the elements of D_8 are

$$\rho_0, \rho_1, \rho_2, \rho_3, \sigma_0, \sigma_1, \sigma_2, \sigma_3.$$

The elements of S_4 which correspond to these by considering where the four vertices end up, written in disjoint cycle notation are, respectively:

$$(1); (1,2,3,4); (1,3)(2,4); (1,4,3,2); (2,4); (12)(34); (13); (14)(23).$$

Notice that this is not the whole of S_4 because S_4 has 24 elements. Why is it that the permutation $(1,2,3) \in S_4$, for example, does not correspond to a symmetry of the square?

Here is the composition/multiplication table for D_8

$\circ$	ρ_0	ρ_1	ρ_2	ρ_3	σ_0	σ_1	σ_2	σ_3
ρ_0	ρ_0	ρ_1	ρ_2	ρ_3	σ_0	σ_1	σ_2	σ_3
ρ_1	ρ_1	ρ_2	ρ_3	ρ_0	σ_1	σ_2	σ_3	σ_0
ρ_2	ρ_2	ρ_3	ρ_0	ρ_1	σ_2	σ_3	σ_0	σ_1
ρ_3	ρ_3	ρ_0	ρ_1	ρ_2	σ_3	σ_0	σ_1	σ_2
σ_0	σ_0	σ_3	σ_2	σ_1	ρ_0	ρ_3	ρ_2	ρ_1
σ_1	σ_1	σ_0	σ_3	σ_2	ρ_1	ρ_0	ρ_3	ρ_2
σ_2	σ_2	σ_1	σ_0	σ_3	ρ_2	ρ_1	ρ_0	ρ_3
σ_3	σ_3	σ_2	σ_1	σ_0	ρ_3	ρ_2	ρ_1	ρ_0

Here is the composition/multiplication table for the corresponding 8 elements of S_4.

$\circ$	(1)	$(1,2,3,4)$	$(1,3)(2,4)$	$(1,4,3,2)$	$(2,4)$	$(1,2)(3,4)$	$(1,3)$	$(1,4)(2,3)$
(1)	(1)	$(1,2,3,4)$	$(1,3)(2,4)$	$(1,4,3,2)$	$(1,3)$	$(1,2)(3,4)$	$(1,3)$	$(1,4)(2,3)$
$(1,2,3,4)$	$(1,2,3,4)$	$(1,3)(2,4)$	$(1,4,3,2)$	(1)	$(1,2)(3,4)$	$(1,3)$	$(1,4)(2,3)$	$(1,3)$
$(1,3)(2,4)$	$(1,3)(2,4)$	$(1,4,3,2)$	(1)	$(1,2,3,4)$	$(1,3)$	$(1,4)(2,3)$	$(2,4)$	$(1,2)(3,4)$
$(1,4,3,2)$	$(1,4,3,2)$	(1)	$(1,2,3,4)$	$(1,3)(2,4)$	$(1,4)(2,3)$	$(2,4)$	$(1,2)(3,4)$	$(1,3)$
$(2,4)$	$(2,4)$	$(1,4)(2,3)$	$(1,3)$	$(1,2)(3,4)$	(1)	$(1,4,3,2)$	$(1,3)(2,4)$	$(1,2,3,4)$
$(1,2)(3,4)$	$(1,2)(3,4)$	$(2,4)$	$(1,4)(2,3)$	$(1,3)$	$(1,2,3,4)$	(1)	$(1,4,3,2)$	$(1,3)(2,4)$
$(1,3)$	$(1,3)$	$(1,2)(3,4)$	$(2,4)$	$(1,4)(2,3)$	$(1,3)(2,4)$	$(1,2,3,4)$	(1)	$(1,4,3,2)$
$(1,4)(2,3)$	$(1,4)(2,3)$	$(1,3)$	$(1,2)(3,4)$	$(2,4)$	$(1,4,3,2)$	$(1,3)(2,4)$	$(1,2,3,4)$	(1)

Notice that the above table shows that

$$H = \{(1); (1,2,3,4); (1,3)(2,4); (1,4,3,2); (2,4); (1,2)(3,4); (1,3); (1,4)(2,3)\}$$

is a subgroup of S_4. Again, is this subgroup really any different to D_8 itself?

This means there is a subgroup which is a essentially a 'copy' of D_8 inside S_4. We'll make this idea more precise in chapter 12 and you will eventually be comfortable enough to say that there is no distinction between D_8 and this copy of D_8 and say that D_8 is a subgroup of S_4.

10.5.1 Exercise Think about the group D_{10}, the symmetries of a regular pentagon. How many elements does it have? How many elements does S_5 have? How is there a subgroups of S_5 which is a 'copy' of D_{10}? Think about how this generalises to any integer $n > 2$.

Chapter 11 - The Alternating Group

Here we'll meet the *alternating group* A_n which is a subgroup of S_n. This group has had many applications in mathematics. Notably the fact that there is no formula for the roots of a quintic from its coefficients based on the four arithmetic operations and taking n^{th} roots can be proved as a consquence of the internal subgroup structure of this group.

11.1 Permutations and transpositions

11.1.1 Lemma Every permutation can be written as a product of transpositions.

Note the absence of the word 'disjoint'.

Proof. We know that every permutation can be written a product of cycles. So it is enough to show that a cycle can be written as a product of transpositions. Check for yourself that

$$(a_1, a_2, \ldots, a_m) = (a_1, a_m) \cdots (a_1, a_3)(a_1, a_2). \tag{5}$$

11.1.2 Example Equation (5) gives a recipe for writing any cycle as a product of transpositions. For example,

$$(1, 5, 9) = (1, 9)(1, 5).$$

Note that these transpositions are not disjoint and so they don't have to commute. Check that

$$(1, 9)(1, 5) \neq (1, 5)(1, 9).$$

One thing to be careful about is that decomposition of a permutation as a product of transpositions is not in any way unique. For example, using (5) we have

$$(1, 2, 3, 4) = (1, 4)(1, 3)(1, 2).$$

However, you can also check that

$$(1, 2, 3, 4) = (2, 3)(1, 3)(3, 5)(3, 4)(4, 5).$$

So we can write $(1, 2, 3, 4)$ as a product of 3 transpositions and as a product of 5 transpositions. Can we write it as a product of 4 transpositions? Spend no more and no less than five minutes thinking about this. ◊

11.2 Even and odd Permutations

Let $n \geq 2$ be an integer. Let $x_1, x_2, \ldots, x_n$ be variables, and let P_n be the polynomial

$$P_n = \prod_{1 \leq i < j \leq n} (x_i - x_j).$$

The polynomial P_n is called the *n-th alternating polynomial.* It will help us to discover an important subgroup of S_n called the *alternating group* and denoted by A_n. Let us write down the first three alternating polynomials:

$$P_2 = x_1 - x_2, \qquad P_3 = (x_1 - x_2)(x_1 - x_3)(x_2 - x_3),$$
$$P_4 = (x_1 - x_2)(x_1 - x_3)(x_1 - x_4)(x_2 - x_3)(x_2 - x_4)(x_3 - x_4).$$

If $\sigma \in S_n$ then define

$$\sigma(P_n) = \prod_{1 \leq i < j \leq n} (x_{\sigma(i)} - x_{\sigma(j)}).$$

11.2.1 Example Let $\sigma = (1, 2) \in S_3$. Then

$$\begin{aligned} \sigma(P_3) &= (x_{\sigma 1} - x_{\sigma 2})(x_{\sigma 1} - x_{\sigma 3})(x_{\sigma 2} - x_{\sigma 3}) \\ &= (x_2 - x_1)(x_2 - x_3)(x_1 - x_3) \\ &= -P_3. \end{aligned}$$

We obtain the equality in the final step of the calculation by comparing the factors of P_3 with the factors of $\sigma(P_3)$, and **not** by expanding! Note that the first factor of P_3 changed sign and the last two factors are swapped. So $\sigma(P_3) = -P_3$.

Now let $\tau = (1, 2, 3) \in S_3$. Then

$$\begin{aligned} \tau(P_3) &= (x_{\tau(1)} - x_{\tau(2)})(x_{\tau(1)} - x_{\tau(3)})(x_{\tau(2)} - x_{\tau(3)}) \\ &= (x_2 - x_3)(x_2 - x_1)(x_3 - x_1) \\ &= P_3. \end{aligned}$$

Again we obtain equality by comparing factors. Write down $\rho(P_3)$ for the other four elements $\rho \in S_3$. ◊

This example involving P_3 for an element of S_3 is quite straightforward. Let's look at a more difficult example.

11.2.2 Example Let $\tau = (2, 4) \in S_5$. We want to check that $\tau(P_5) = -P_5$. Some factors of P_5 are unaffected. For example, $\tau(x_1 - x_3) = x_{\tau(1)} - x_{\tau(3)} = x_1 - x_3$. The ones that aren't affected are the ones that don't contain either of x_2 or x_4. These are,

$$x_1 - x_3, \qquad x_1 - x_5, \qquad x_3 - x_5.$$

We will split the other factors of P_5 into four groups [6]:

$$\begin{array}{lll}
\text{(I)} & x_1 - x_2, & x_1 - x_4, \\
\text{(II)} & x_2 - x_3, & x_3 - x_4, \\
\text{(III)} & x_2 - x_5, & x_4 - x_5, \\
\text{(IV)} & x_2 - x_4. &
\end{array}$$

Let's study what τ does to each group. Note that

$$\tau(x_1 - x_2) = x_1 - x_4, \qquad \tau(x_1 - x_4) = x_1 - x_2.$$

Thus τ swaps the factors in group (I) whilst *keeping their signs the same.* But

$$\tau(x_2 - x_3) = x_4 - x_3 = -(x_3 - x_4), \qquad \tau(x_3 - x_4) = x_3 - x_2 = -(x_2 - x_3).$$

Thus τ swaps the factors in group (II) and *changes the sign of each.* Moreover,

$$\tau(x_2 - x_5) = x_4 - x_5, \qquad \tau(x_4 - x_5) = x_2 - x_5.$$

So τ swaps the factors in group (III) whilst *keeping their signs the same.* Finally,

$$\tau(x_2 - x_4) = x_{\tau 2} - x_{\tau 4} = x_4 - x_2 = -(x_2 - x_4).$$

So the one factor in group (IV) simply *changes sign.* We see that $\tau(P_5)$ has the same factors as P_5 with three sign changes: $\tau(P_5) = (-1)^3 P_5 = -P_5$. $\diamondsuit$

11.2.3 Lemma Let $\tau \in S_n$ be a transposition. Then $\tau(P_n) = -P_n$.

Proof. Let $\tau = (\ell, m)$. The transposition (ℓ, m) swaps ℓ and m, and keeps everything else fixed. In particular $(\ell, m) = (m, \ell)$. So we can suppose that $\ell < m$. Any factor $x_i - x_j$ where neither i nor j is equal to ℓ nor m, is unaffected by τ.

[6] The word "groups" here is used in its English language sense, not in its mathematical sense.

We pair off the other factors as follows:

$$
\begin{aligned}
&\text{(I)} \quad \begin{cases} x_1 - x_\ell, & x_1 - x_m, \\ x_2 - x_\ell, & x_2 - x_m, \\ \vdots & \vdots \\ x_{\ell-1} - x_\ell, & x_{\ell-1} - x_m, \end{cases} \\
&\text{(II)} \quad \begin{cases} x_\ell - x_{\ell+1}, & x_{\ell+1} - x_m, \\ x_\ell - x_{\ell+2}, & x_{\ell+2} - x_m, \\ \vdots & \vdots \\ x_\ell - x_{m-1}, & x_{m-1} - x_m, \end{cases} \\
&\text{(III)} \quad \begin{cases} x_\ell - x_{m+1}, & x_m - x_{m+1}, \\ x_\ell - x_{m+2}, & x_m - x_{m+2}, \\ \vdots & \vdots \\ x_\ell - x_n, & x_m - x_n, \end{cases} \\
&\text{(IV)} \quad \begin{cases} x_\ell - x_m. \end{cases}
\end{aligned}
$$

Now τ swaps each pair in (I), keeping the signs the same; it swaps each pair in (II) and changes the sign of each; it swaps each pair in (III), keeping the signs the same; it changes the sign of $x_\ell - x_m$. So $\tau(P_n)$ has exactly the same factors as P_n, up to a certain number of sign changes. How many sign changes? The number of sign changes is:

$$2(m - \ell - 1) + 1.$$

The 1 is for changing the sign of $x_\ell - x_m$. There are 2 sign changes coming from each pair in (II). The number of such pairs is $m - \ell - 1$. Since the number of sign changes is odd, we see that $\tau(P_n) = -P_n$.

11.2.4 Lemma If $\sigma \in S_n$ then $\sigma(P_n) = \pm P_n$. More precisely, if σ is a product of an even number of transpositions then $\sigma(P_n) = P_n$ and if σ is a product of an odd number of transpositions then $\sigma(P_n) = -P_n$.

Proof. Recall, by Lemma 11.1.1, that we can write every permutation as a product of transpositions. Every transposition changes the sign of P_n. The lemma follows.

11.2.5 Example We have noted in Example 11.1.2 that the way we express a permutation as a product of transpositions is not unique. Indeed we saw that

$$(1, 2, 3, 4) = (1, 4)(1, 3)(1, 2), \qquad (1, 2, 3, 4) = (2, 3)(1, 3)(3, 5)(3, 4)(4, 5).$$

So we can write $(1,2,3,4)$ as a product of 3 transpositions and as a product of 5 transpositions. We asked the question of whether $(1,2,3,4)$ can be written as a product of 4 transpositions? Write $\sigma = (1,2,3,4)$. From the above lemma, we see that $\sigma(P_n) = -P_n$. If we're able to write σ as a product of an even number of transpositions then $\sigma(P_n) = P_n$. We would then have $P_n = -P_n$ which is a contradiction. Therefore we cannot write σ as a product of 4 transpositions. $\Diamond$

This example tells us how to prove the following theorem.

11.2.6 Theorem Every permutation in S_n can be written as a product of either an even number of transpositions, or an odd number of transpositions but ***not both***.

Proof. Let $\sigma \in S_n$. Then, by Lemma 11.1.1, S_n can be written as a product of transpositions. Suppose σ can be written both as a product of an even number of transpositions and a a product of an odd number of transpositions. Then by Lemma 11.2.4 we have $\sigma(P_n) = P_n$ and $\sigma(P_n) = -P_n$. This implies that $P_n = -P_n$, a contradiction. The conclusion follows. $\Diamond$

11.2.7 Definition We shall call a permutation *even* if we can write it as a product of an even number of transpositions, and we shall call it *odd* if we can write it as a product of an odd number of transpositions.

11.2.8 Example $(1,2,3,4)$ is an odd permutation because we can write it as the product of 3 transpositions:

$$(1,2,3,4) = (1,4)(1,3)(1,2).$$

Indeed, a cycle of length n can be written as product of $n-1$ transpositions by (5). So we need to be careful because a cycle of length n is even if n is odd, and it is odd if n is even!

The permutation $(1,2,3)(4,5)$ is the product of an even permutation which is $(1,2,3)$ and an odd permutation which is the transposition $(4,5)$. Thus $(1,2,3)(4,5)$ is an odd permutation.

What about the identity element id? Note that $\mathrm{id}(P_n) = P_n$, so id must be even. We must be able to write it as a product of an even number of transpositions. A mathematician would say that the identity element is the product of zero transpositions, so it is even. If you don't feel comfortable with that kind of reasoning, instead, note that

$$\mathrm{id} = (1,2)(1,2),$$

which does allow us to check that id is indeed even.

We now come to define a very important group. Let $n \geq 2$. We define the *n-th alternating group* to be

$$A_n = \{\sigma \in S_n \mid \sigma \text{ is even}\}.$$

As usual, all we've done is specify a subset of S_n which we've denoted by A_n and we must indeed show that A_n is a group.

11.2.9 Theorem A_n is a subgroup of S_n.

Proof. We've already seen that the identity element id is even, so $\mathrm{id} \in A_n$. If σ, $\rho \in A_n$ then we can write each as an even number of transpositions. Therefore the product $\sigma\rho$ can be written as an even number of transpositions (even+even=even). Hence $\sigma\rho \in A_n$.

Finally we must show that the inverse of an even permutation is even. Suppose σ is even. We can write

$$\sigma = \tau_1\tau_2\ldots\tau_m$$

where the τ_i are transpositions, and m is even. Now

$$\begin{aligned}\sigma^{-1} &= (\tau_1\tau_2\cdots\tau_m)^{-1}\\ &= \tau_m^{-1}\tau_{m-1}^{-1}\cdots\tau_1^{-1}\\ &= \tau_m\tau_{m-1}\cdots\tau_1.\end{aligned}$$

Here you should convince yourself that $\tau^{-1} = \tau$ for any transposition τ. Since m is even, we find that σ^{-1} is even and so $\sigma^{-1} \in A_n$.

Hence A_n is a subgroup of S_n. ♢

11.2.10 Example Recall that $S_2 = \{\mathrm{id}, (1,2)\}$. We see that $A_2 = \{\mathrm{id}\}$ is the trivial subgroup. ♢

11.2.11 Example Recall that S_3 has $3! = 6$ elements:

$$S_3 = \{\text{id}, (1,2), (1,3), (2,3), (1,2,3), (1,3,2)\}.$$

Then

$$A_3 = \{\text{id}, (1,2,3), (1,3,2)\}.$$

Note that S_3 is non-abelian, but you can check that A_3 is abelian. $\diamondsuit$

In the above examples we saw that A_n has half the number of elements of S_n for $n = 2$, 3. In fact, this pattern continues.

11.2.12 Theorem Let $n \geq 2$. Then A_n has order

$$|A_n| = \frac{1}{2}|S_n| = \frac{n!}{2}.$$

Proof. We will show there are the same number of even permutations as odd permutations in S_n.

Suppose $\sigma \in S_n$ is an even permutation. Let $\tau = (1,2)$, an odd permutation. Then $\tau\sigma$ is an odd permutation. Moreover if σ_1 and σ_2 are distinct even permutations then $\tau\sigma_1$ and $\tau\sigma_2$ are distinct odd permutations (think about this, if not you could multiply them both on the left by τ again and get that $\sigma_1 = \sigma_2$). So for every even permutation we can create a distinct odd permutation in this way. Therefore the number of odd permutations is greater than or equal to the number of even permutations.

Now suppose $\sigma \in S_n$ is an odd permutation. Let $\tau = (1,2)$, an odd permutation. Then $\tau\sigma$ is an even permutation. Moreover if σ_1 and σ_2 are distinct odd permutations then $\tau\sigma_1$ and $\tau\sigma_2$ are distinct even permutations. So for every odd permutation we can create a distinct even permutation in this way. Therefore the number of even permutations is greater than or equal to the number of odd permutations.

Combining these statements we must have that the number of even permutations is the same as the number of odd permuations and so $|A_n| = \frac{1}{2}|S_n| = \frac{n!}{2}$. $\diamondsuit$

11.2.13 Exercise Let ρ and τ be as given in Exercise 10.3.4. Write ρ and τ as products of transpositions and state if they're even or odd.

11.2.14 Exercise Write down the elements of A_3 and check that it is cyclic (and hence abelian). Show that A_n is non-abelian for $n \geq 4$.

11.2.15 Exercise Let ρ and $\tau \in S_n$. Show that τ is even if and only if $\rho^{-1}\tau\rho$ is even. (**Hint: It will help to show that if $\rho = c_1c_2\cdots c_m$ as a product of transpositions, then $\rho^{-1} = c_mc_{m-1}\ldots c_1$**).

Chapter 12 - Isomorphisms of Groups

At the end of the Chapter 10 we saw that D_6 and S_3 have arisen in two different way but have 'the same' multiplication table

Here is D_6 in the usual notation.

$\circ$	ρ_0	ρ_1	ρ_2	σ_1	σ_2	σ_3
ρ_0	ρ_0	ρ_1	ρ_2	σ_1	σ_2	σ_3
ρ_1	ρ_1	ρ_2	ρ_0	σ_3	σ_1	σ_2
ρ_2	ρ_2	ρ_0	ρ_1	σ_2	σ_3	σ_1
σ_1	σ_1	σ_2	σ_3	ρ_0	ρ_1	ρ_2
σ_2	σ_2	σ_3	σ_1	ρ_2	ρ_0	ρ_1
σ_3	σ_3	σ_1	σ_2	ρ_1	ρ_2	ρ_0

Remember that we labelled the vertices of the equilateral triangle with the numbers 1, 2, 3. Notice how there is a natural correspondence between the elements of D_6 and elements of S_3. For example σ_2 swaps vertices 1 and 3 and so it corresponds to $(1,3) \in S_3$.

The elements of S_3 which correspond to

$$\rho_0, \rho_1, \rho_2, \rho_3, \sigma_1, \sigma_2, \sigma_3 \in D_6$$

written in disjoint cycle notation are, respectively:

$$(1); (1,2,3); (1,3,2); (2,3); (1,3); (1,2).$$

Notice that the list above is the whole of S_3. If you carefully swap each element in the D_6 table for its corresponding element in the S_3 table you get the composition/multiplication table for S_3.

$\circ$	(1)	$(1,2,3)$	$(1,3,2)$	$(2,3)$	$(1,3)$	$(1,2)$
(1)	(1)	$(1,2,3)$	$(1,3,2)$	$(2,3)$	$(1,3)$	$(1,2)$
$(1,2,3)$	$(1,2,3)$	$(1,3,2)$	(1)	$(1,2)$	$(2,3)$	$(1,3)$
$(1,3,2)$	$(1,3,2)$	(1)	$(1,2,3)$	$(1,3)$	$(1,2)$	$(2,3)$
$(2,3)$	$(2,3)$	$(1,3)$	$(1,2))$	(1)	$(1,2,3)$	$(1,3,2)$
$(1,3)$	$(1,3)$	$(1,2)$	$(2,3)$	$(1,3,2)$	(1)	$(1,2,3)$
$(1,2)$	$(1,2)$	$(2,3)$	$(1,3)$	$(1,2,3)$	$(1,3,2)$	(1)

From this abstract point of view they *actually are* the same.

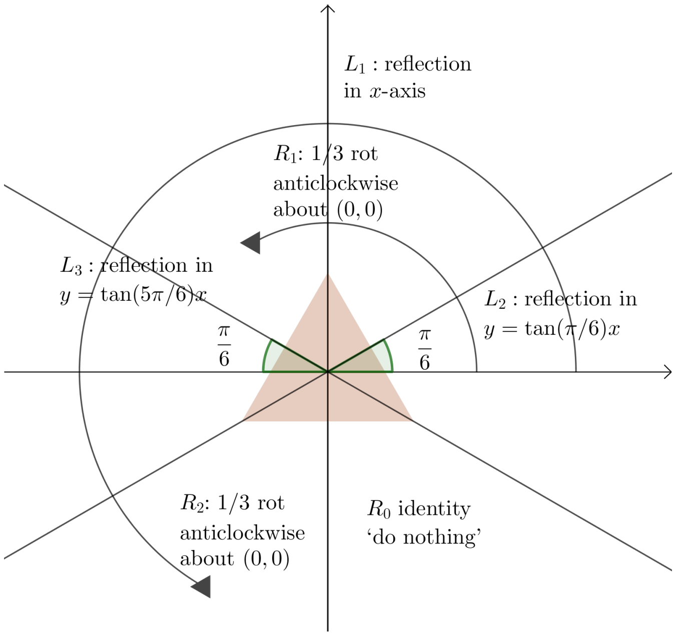
There is another group that has 'the same' multiplication table as the two above. It is a group consisting of six matrices representing rotations about the origin and reflections in lines through the origin:

$$R_0 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, R_1 = \begin{pmatrix} -\frac{1}{2} & -\frac{\sqrt{3}}{2} \\ \frac{\sqrt{3}}{2} & -\frac{1}{2} \end{pmatrix}, R_2 = \begin{pmatrix} -\frac{1}{2} & \frac{\sqrt{3}}{2} \\ -\frac{\sqrt{3}}{2} & -\frac{1}{2} \end{pmatrix},$$

$$L_1 = \begin{pmatrix} -1 & 0 \\ 0 & 1 \end{pmatrix}, L_2 = \begin{pmatrix} \frac{1}{2} & \frac{\sqrt{3}}{2} \\ \frac{\sqrt{3}}{2} & -\frac{1}{2} \end{pmatrix}, L_3 = \begin{pmatrix} \frac{1}{2} & -\frac{\sqrt{3}}{2} \\ -\frac{\sqrt{3}}{2} & -\frac{1}{2} \end{pmatrix}.$$

1. R_0 is the identity map;
2. R_1 is the rotation anticlockwise about the origin throuh $2\pi/3$ radians;
3. R_2 is the rotation anticlockwise about the origin throuh $4\pi/3$ radians;
4. L_1 is the reflection in the $y - axis$;
5. L_2 is the reflection in the line $y = \tan(\pi/6)x$;
6. L_3 is the reflection in the line $y = \tan(5\pi/6)x$.

There are shown in the diagram below. Notice an equaliteral triangle centred at the origin has been added to the diagram below. You should be able to convince yourself that the triangle would be mapped to itself by each of these matrices and, thinking of them like that, the correspondence with D_6 feels very natural.


L_1 : reflection in x-axis
R_1: 1/3 rot anticlockwise about $(0,0)$
L_3 : reflection in $y = \tan(5\pi/6)x$
L_2 : reflection in $y = \tan(\pi/6)x$
$\frac{\pi}{6}$
$\frac{\pi}{6}$
R_2: 1/3 rot anticlockwise about $(0,0)$
R_0 identity 'do nothing'

Here is the multiplication table.

$\circ$	R_0	R_1	R_2	L_1	L_1	L_3
R_0	R_0	R_1	R_2	L_1	L_2	L_3
R_1	R_1	R_2	R_0	L_3	L_1	L_1
R_2	R_2	R_0	R_1	L_1	L_3	L_1
L_1	L_1	L_1	L_3	R_0	R_1	R_2
L_1	L_1	L_3	L_1	R_2	R_0	R_1
L_3	L_3	L_1	L_1	R_1	R_2	R_0

Note that these six matrices under matrix multiplication are a subgroup of $\mathrm{GL}_2(\mathbb{R})$ and indeed a subgroup of $\mathrm{O}_2(\mathbb{R})$, at least when thought of in complex number form, but not a subgroup of $\mathrm{SO}_2(\mathbb{R})$. Convince yourself about these statements.

Here is another example of the same thing.

U_4 has multiplication table as below

$\times$	1	i	-1	$-i$
1	1	i	-1	$-i$
i	i	-1	$-i$	1
-1	-1	$-i$	1	i
$-i$	$-i$	1	i	-1

If you replace each 1 with $[0]_4$, each i with $[1]_4$, each -1 with $[2]_4$ and each $-i$ with $[3]_4$, then switch $\times$ for $+_4$ you get the table for $(\mathbb{Z}/4\mathbb{Z}, +_4)$ as below.

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

If you now replace each $[0]_4$ with ρ_0, each $[1]_4$ with ρ_1, each $[2]_4$ with ρ_2 and each $[3]_4$ with ρ_3 and then switch $+_4$ for 'followed by' you get the table for the subgroup $\{\rho_0, \rho_1, \rho_2, \rho_3\}$ of D_8 as below

followed by	ρ_0	ρ_1	ρ_2	ρ_3
ρ_0	ρ_0	ρ_1	ρ_2	ρ_3
ρ_1	ρ_1	ρ_2	ρ_3	ρ_0
ρ_2	ρ_2	ρ_3	ρ_0	ρ_1
ρ_3	ρ_3	ρ_0	ρ_1	ρ_2

Here is another(slightly silly) example of two groups which are essentially the same. The set $\{e, a, b\}$ with binary operation $*$ given in the table below is a group.

$\star$	e	a	b
e	e	a	b
a	a	b	e
b	b	e	a

So is the set $\{\text{apple}, \text{strawberry}, \text{lemon}\}$ with binary operation $\diamond$ given in the table below is a group.

$\diamond$	apple	strawberry	lemon
apple	apple	strawberry	lemon
strawberry	strawberry	lemon	apple
lemon	lemon	apple	strawberry

The point is that the names are not important, it's the interactions between the elements (i.e. how they combine according to the binary operation) which characterise a group.

Here is one more example, this time involving infinite groups, which you may need to think a bit more about to take in.

In $\mathbb{S}$ we can think of each element as $e^{i\theta}$ for some real number θ. Then such elements combine as $e^{i\theta} \times e^{i\omega} = e^{i(\theta+\omega)}$.

In $\mathrm{SO}_2(\mathbb{R})$ each element f is an isometry of the form $f(z) = e^{i\theta}z$ for some real number θ. Composing two elements $f, g \in \mathrm{SO}_2(\mathbb{R})$, where $g(z) = e^{i\omega}z$ gives

$$(f \circ g)(z) = f(g(z)) = f(e^{\omega i}z) = e^{i\theta}e^{i\omega}z = e^{i(\theta+\omega)}z.$$

This one is a little more tricky to see but in both cases the operation is essentially about adding the values θ and ω and we are only interested in that value up to multiples of 2π (because changing the value by a multiple of 2π has no effect on either the element of $\mathbb{S}$ or the element of $\mathrm{SO}_2(\mathbb{R})$. Again we could ask whether these two groups really any different?

We'll see here that these are all examples of *isomorphic* groups. What does this mean?

12.1 What is an isomorphism?

12.1.1 Definition Let $(G, \diamond)$ and $(H, *)$ be groups. We say that the function $\phi : G \to H$ is an *isomorphism* if it is a bijection and it satisfies

$$\phi(g_1 \diamond g_2) = \phi(g_1) * \phi(g_2)$$

for all g_1, g_2 in G. In this case we say that $(G, \diamond)$ and $(H, *)$ are *isomorphic.*

Isomorphic groups may look different, but in essence they are the same. An isomorphism is a way of relabeling the elements of one group to obtain another group, as the following examples will make clear.

12.1.2 Example Define $\phi : \mathbb{Z}/m\mathbb{Z} \to U_m$ by the rule

$$\phi([a]_m) = \zeta^a, \qquad a = 0, 1, \ldots, m-1.$$

Then ϕ is a bijection and satisfies the property

$$\phi([a]_m +_m [b]_m) = \phi([a+b]_m) = \zeta^{a+b} = \zeta^a \cdot \zeta^b = \phi([a]_m)\phi([b]_m).$$

So ϕ is an isomorphism. $\Diamond$

12.1.3 Example Recall that, for a real number θ, the map $f_\theta : \mathbb{C} \to \mathbb{C}$ given by $f_\theta(z) = e^{i\theta}z$, the anti-clockwise rotation through and angle θ about 0, is an isometry of $\mathbb{C}$ (notice how we can use the value θ as a subscript in the function name to indicate the size of the rotation). As an aside, it's also worth noting that, in the language of matrices and $\mathbb{R}^2$, this would correspond to the transformation given by this rotation matrix:

$$R_\theta = \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix}.$$

It represents anticlockwise rotation around the origin through an angle θ.

Returning to complex numbers notation, if $\theta, \phi \in \mathbb{R}$ then

$$(f_\theta \circ f_\phi)(z) = f_\theta(f_\phi(z)) = f_\theta(e^{i\phi}z) = e^{i\theta}e^{i\phi}z = e^{i(\theta+\phi)}z = f_{\theta+\phi}(z).$$

This gives the identity

$$f_{\theta+\phi} = f_\theta \circ f_\phi$$

(or in matrix language

$$R_{\theta+\phi} = R_\theta R_\phi)$$

which 'turns addition into multiplication', indeed it comes from the identity

$$e^{i(\theta+\phi)} = e^{i\theta}e^{i\phi}.$$

So f_θ and $e^{i\theta}$ are analogues and we might expect that the groups $\mathrm{SO}_2(\mathbb{R})$ and $\mathbb{S}$ are isomorphic. Recall that $\mathrm{SO}_2(\mathbb{R})$ is the special orthogonal group (defined in Theorem 9.1.10) given by

$$\mathrm{SO}_2(\mathbb{R}) = \{f_\theta \mid \theta \in \mathbb{R}\},$$

and $\mathbb{S}$ is the circle group (page 62) given by

$$\mathbb{S} = \{\alpha \in \mathbb{C} \mid |\alpha| = 1\} = \{e^{i\theta} \mid \theta \in \mathbb{R}\}.$$

You can satisfy yourself that the map

$$\phi : \mathrm{SO}_2(\mathbb{R}) \to \mathbb{S}, \qquad \phi(f_\theta) = e^{i\theta}$$

is an isomorphism.

By this correspondence, we see that there are isometry analogues of the n-th roots of unity (which you can translate to matrix analogues if you prefer). If we let

$$\mathcal{Z} = f_{2\pi/n} \in \mathrm{SO}_2(\mathbb{R})$$

(in other words the isometry which is the anticlockwise rotation about the origin through angle $2\pi/n$) then $\mathrm{id}_\mathbb{C} = \mathcal{Z}^0, \mathcal{Z}, \ldots, \mathcal{Z}^{n-1}$ all satisfy the relationship $A^n = \mathrm{id}_\mathbb{C}$m their n^{th} power is the identity. ◊

12.1.4 Exercise Let $\mathcal{Z} = f_{2\pi/6}$. Show that $\{1, \mathcal{Z}, \ldots, \mathcal{Z}^5\}$ is a subgroup of $\mathrm{SO}_2(\mathbb{R})$. Write down the orders of its elements.

12.1.5 Exercise Write down an isomporphism ϕ from D_6 to S_3. To do this you need to specify with of $\mathrm{id}, (1\ 2), (2\ 3), (1\ 3), (1\ 2\ 3), (1\ 3\ 2) \in S_3$ is for each element in $D_6 = \{\rho_0, \rho_1, \rho_2, \sigma_1, \sigma_2, \sigma_3\}$. You don't need to check that your ϕ is an isomorphism.

12.1.6 Exercise Write down an isomporphism ϕ from D_8 to a subgroup of S_3. You don't need to check that your ϕ is an isomorphism.

12.1.7 Exercise Suppose groups G and H are isomorphic. Show that G and H have the same order. Show that G is abelian if and only if H is abelian. Show that G is cylic if and only if H is cyclic.

12.2 Direct products of groups

We'll now look at a way to form new groups from existing ones. Given groups $(G, \star)$, $(H, \diamond)$ we can define a group structure on the set of ordered pairs $G \times H = \{(g, h) \mid g \in G, h \in H\}$.

12.2.1 Definition Given groups $(G, \star)$, $(H, \diamond)$ define a binary operation . on $G \times H = \{(g, h) \mid g \in G, h \in H\}$ as follows. If $(g_1, h_1), (g_2, h_2) \in G \times H$ then

$$(g_1, h_1).(g_2, h_2) = (g_1 \star g_2, h_1 \diamond h_2).$$

Then $(G \times H, .)$ is a group called the *the direct product of* G *and* H. $\diamondsuit$

12.2.2 Exercise Prove that $(G \times H, .)$ as above is a group by showing

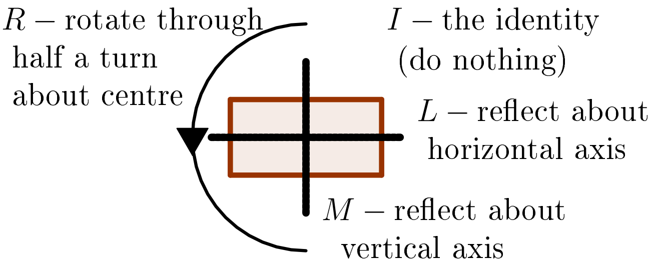
1. that . is an associative binary operation on $G \times H$,
2. that $(1_G, 1_H)$ is the identity element where 1_G is the identity in G and 1_H is the identity in H
3. that (g^{-1}, h^{-1}) is the inverse element to (g, h) where g^{-1} is the inverse to $g \in G$ and h^{-1} is the inverse to $h \in H$.

12.2.3 Example Let $G = \mathbb{Z}/2\mathbb{Z}$ and $H = \mathbb{Z}/2\mathbb{Z}$ in the definition above. The elements of $G \times H$ are then $([0]_2, [0]_2)$, $([1]_2, [0]_2)$,$([0]_2, [1]_2)$ and $([1]_2, [1]_2)$ and its table is

$+$	$([0]_2, [0]_2)$	$([1]_2, [0]_2)$	$([0]_2, [1]_2)$	$([1]_2, [1]_2)$
$([0]_2, [0]_2)$	$([0]_2, [0]_2)$	$([1]_2, [0]_2)$	$([0]_2, [1]_2)$	$([1]_2, [1]_2)$
$([1]_2, [0]_2)$	$([1]_2, [0]_2)$	$([0]_2, [0]_2)$	$([1]_2, [1]_2)$	$([0]_2, [1]_2)$
$([0]_2, [1]_2)$	$([0]_2, [1]_2)$	$([1]_2, [1]_2)$	$([0]_2, [0]_2)$	$([1]_2, [0]_2)$
$([1]_2, [1]_2)$	$([1]_2, [1]_2)$	$([0]_2, [1]_2)$	$([1]_2, [0]_2)$	$([0]_2, [0]_2)$

Exceptionally, we have used $+$ for the binary operation in the direct product here because the two underlying groups are both additive groups.

12.2.4 Example You've actually seen the example above before, on Assignment 1. Here are the elements of the group of isometries of a rectangle which isn't a square, there are four of them.



And here is its group table.

$\circ$	I	R	L	M
I	I	R	L	M
R	R	I	M	L
L	L	M	I	R
M	M	L	R	I

Can you find a one-to-one correspondence between the four elements of $\mathbb{Z}/2\mathbb{Z} \times \mathbb{Z}/2\mathbb{Z}$ and the four elements of the group above which maps the table for $\mathbb{Z}/2\mathbb{Z} \times \mathbb{Z}/2\mathbb{Z}$ (in other words can you find an isomorphism between the two groups)? There are actually several isomorphisms.

This group actually has a special name, it's called the Klein 4-group and it's denoted K_4. Here is the generic multiplication table for K_4 where we just call the elements $1, a, b$ and c.

$\circ$	1	a	b	c
1	1	a	b	c
a	a	1	c	b
b	b	c	1	a
c	c	b	a	1

Notice

- the group is abelian
- each element other than the identity has order 2

- the product of any two distinct elements which are both not the identity is the other element which is not the identity.

These three properties characterise the entire table (i.e. you can write down the whole table from just knowing these three properties).

12.3 Classifying groups up to isomorphism

This is really what abstract group theory is all about. We want to answer questions like

- What is a complete list, up to isomorphism, of all the groups with 3 elements?
- What is a complete list, up to isomorphism, of all the groups with 4 elements?

And so on....

The following theorem, which generalises Theorem 5.2.5 to non-Abelian groups will help us to answer the two questions above at least.

12.3.1 Theorem Let G be a finite group. Let $g \in G$. Then the order of g divides the order of G.

Proof. By Theorem 5.1.2, g has finite order n. Consider the subgroup generated by g i.e.

$$\langle g \rangle = \{1, g, g^2, \dots, g^{n-1}\}.$$

This contains n elements. If $G = \langle g \rangle$ then we are done because then $|G| = n$ and the order of g is n so clearly the order of g then divides the order of G.

If not we can find an element $x \in G$ such that $x \notin \langle g \rangle$. For such an element x consider the set

$$X = \{x, xg, xg^2, \dots, xg^{n-1}\}.$$

We will show that X contains n elements by showing that all the listed elements of X are distinct. We wil also show that $X \cap \langle g \rangle = \emptyset$.

Suppose that $xg^i = xg^j$ for integers i, j with $0 \leq i < j \leq n-1$. Then, multiplying both sides of this on the right by g^{-i} gives $x = g^{j-i}$. But this implies that $x \in \langle g \rangle$ which contradicts the way that x was chosen.

Now suppose that $X \cap \langle g \rangle \neq \emptyset$. Then there are integers i, j between 0 and $n-1$ inclusive such that $g^i = xg^j$. Multiplying both sides of this on the right by

g^{-j} gives $g^{i-j} = x$ which again implies that $x \in \langle g \rangle$, a contradiction.

Suppose $G = \langle g \rangle \cup X$. Then G has $2n$ elements and the order of g, n, divides the order of G.

If not we can choose an element $y \in G$ such that $y \notin \langle g \rangle \cup X$. In a similar way to before with X and x let

$$Y = \{y, yg, yg^2, \ldots, yg^{n-1}\}.$$

As before all the listed elements of Y are distinct. As before $Y \cap \langle g \rangle = \emptyset$. Also $Y \cap X = \emptyset$ as follows. If not there are integers i, j between 0 and $n-1$ inclusive such that $yg^i = xg^j$. Multiplying both sides of this on the right by g^{-i} gives $y = xg^{j-i}$ which implies that $y \in X$, contradicting the way that y has been chosen.

Suppose $G = \langle g \rangle \cup X \cup Y$. Then G has $3n$ elements and the order of g, n, divides the order of G.

If not choose $z \in G$ such that $z \notin \langle g \rangle \cup X \cup Y$ and let $Z =$...(argument continues as before)...

This process must end at some point because G is a finite group. At the point it does we have deduced that other order of g divides the order of G. ♢

12.3.2 Example We'll now use this to show that there are actually only two groups of order 4, up to isomorphism.

Suppose we have a group G with four elements. Let's call them $1, a, b$ and c, they are all distinct from one another.

Then either

1. G has an element of order 4
2. G has no elements of order 4.

Case 1. Let's suppose, renaming if necessary and so without any loss of generality, that a is an element of order 4. Then the whole group is $\{1, a, a^2, a^3\}$ and so, again renaming if necessary, without any loss of generality we can assume that $b = a^2$ and $c = a^3$.

Then the group table looks like this

$\circ$	1	a	$b = a^2$	$c = a^3$
1	1	a	b	c
a	a	b	c	1
$b = a^2$	b	c	1	a
$c = a^3$	c	1	a	b

So it's a cyclic group of order 4. U_4, $\mathbb{Z}/4\mathbb{Z}$ and the subgroup $\{\rho_0, \rho_1, \rho_2, \rho_3\}$ of D_8 are all examples of this type, i.e. they are all isomorphic to one another and to the group shown above.

Case 2. In the second case we have no elements of order 4. By Theorem 12.3.1 all the elements other than 1 must have order 2. In other words $a^2 = b^2 = c^2 = 1$. Combining this with what we know about the identity element, a lot of the multiplcation table is already filled in

$\circ$	1	a	b	c
1	1	a	b	c
a	a	1		
b	b		1	
c	c			1

Now consider ab. It must be one of 1, a, b or c.

If $ab = 1$ then, multiplying on the left by a, gives $a^2b = a$. But $a^2 = 1$ so this gives $b = a$. This is not possible a and b are distinct elements in the group.

If $ab = a$ then, multiplying on the left by a, gives $a^2b = a^2$. But $a^2 = 1$ so this gives $b = 1$. This is not possible 1 and b are distinct elements in the group.

If $ab = b$ then, multiplying on the right by b, gives $ab^2 = b^2$. But $b^2 = 1$ so this gives $a = 1$. This is not possible 1 and a are distinct elements in the group.

The only possibility left is that $ab = c$!

You can argue similarly that $ba = c$, $ac = b$, $ca = b$, $bc = a$ and $cb = a$. This means the complete table must be

$\circ$	1	a	b	c
1	1	a	b	c
a	a	1	c	b
b	b	c	1	a
c	c	b	a	1

and the group is K_4. $\Diamond$

12.3.3 Exercise Show that there is only one group with order 3 (up to isomorphism). Show that there is only one group with order 5 (up to isomorphim). What is the smallest non-abelian group? $\Diamond$

12.4 What's so special about S_n?

We started lecture 10 by looking at symmetry groups of arbitrary sets A. Then we restricted ourself to $S_n = \text{Sym}(\{1, 2, \ldots, n\})$. This is not as big a restriction as it looks. Suppose the set A is finite, and let $n = |A|$, the number of elements of A. Then $\text{Sym}(A)$ is isomorphic to S_n. One way of seeing this is convince ourselves that every permutation of $\{1, 2, \ldots, n\}$ gives us a permutation of A. For example, suppose $A = \{a_1, a_2, a_3\}$. Then the permutation $\{1, 2, 3\}$ given by

$$\begin{pmatrix} 1 & 2 & 3 \\ 2 & 3 & 1 \end{pmatrix}$$

corresponds to the permutation of A given by

$$\begin{pmatrix} a_1 & a_2 & a_3 \\ a_2 & a_3 & a_1 \end{pmatrix}.$$

Understanding $\text{Sym}(A)$ with $|A| = n$ is the same as understanding S_n.

Chapter 13 - Group Homomorphisms

13.1 What is a group homomorphism?

A homomorphism is like an isomorphism but we take out the requirement to be a bijection. As we noted, when groups are isomorphic they are essentially the same. When there is a homomorphism between groups which isn't a bijection they have a coarser similarity. You will explore this much more in future modules about group theory, this chapter just serves as an introduction.

13.1.1 Definition Let $(G, \diamond)$ and $(H, \star)$ be groups. We say that the function $\phi : G \to H$ is a *homomorphism* if it satisfies

$$\phi(g_1 \diamond g_2) = \phi(g_1) \star \phi(g_2)$$

for all g_1, g_2 in G.

This means that all isomorphims are homomorphisms and so you have examples of homomorphisms already, all the isomorphisms you saw in the last chapter.

We also define some associated objects.

13.1.2 Definitions Let G and H be groups and let $\phi : G \to H$ be a homomorphism. Then

1. $\text{Ker}\,\phi = \{g \in G \mid \phi(g) = 1_H\}$ is called the *kernel* of ϕ.
2. $\text{Im}\,\phi = \{h \in H \mid \text{there exists } g \in G \text{ with } \phi(g) = h\} = \{\phi(g) \mid g \in G\}$ is called the *image* of ϕ

Let's look at some examples of homomorphisms that are not isomorphisms.

13.1.3 Example Let $G = \mathbb{Z}$ and let $H = \{1, -1\}$ (under multiplication). Define a function $\phi : G \to H$ as follows:

$$\phi(m) = \begin{cases} 1 & \text{if } m \text{ is even} \\ -1 & \text{if } m \text{ is odd} \end{cases}$$

To check that this is a homomorphism we need compare $\phi(m)\phi(n)$ with $\phi(m+n)$ and show they are equal for all possible combinations of m, n being odd or even.

If m and n are both even then $\phi(m) = \phi(n) = 1$ and so $\phi(m)\phi(n) = 1$. Also $m + n$ is even and so $\phi(m + n) = 1$ and we have $\phi(m + n) = \phi(m)\phi(n)$.

If m and n are both odd then $\phi(m) = \phi(n) = -1$ and so $\phi(m)\phi(n) = 1$. Also $m + n$ is even and so $\phi(m + n) = 1$ and we have $\phi(m + n) = \phi(m)\phi(n)$.

If m is odd and n is even then $\phi(m) = -1$ and $\phi(n) = 1$ and so $\phi(m)\phi(n) = -1$. Also $m+n$ is odd and so $\phi(m+n) = -1$ and we have $\phi(m+n) = \phi(m)\phi(n)$. The case of m even and n odd is similar.

So ϕ is a homomorphism. It's clear that Im $\phi = \{-1, 1\}$ and that Ker $\phi = 2\mathbb{Z}$.

13.1.4 Example Let $G = \mathbb{Z}$ and let $H = \mathbb{Z}/n\mathbb{Z}$. Define a function $\phi : G \to H$ as follows:

$$\phi(m) = [m]_n$$

To check that this is a homomorphism we need to show that $\phi(k) + \phi(l) = \phi(k + l)$ for all integers k, l.

We have

$$\phi(k) + \phi(l) = [k]_n + [l] + n = [k + l]_n = \phi(k + l).$$

So ϕ is a homomorphism. Ker $\phi = \{k \in \ Z \mid [k]_n = [0]_n\} = n\mathbb{Z}$, the multiples of n. Convince yourself that Im $\phi = \mathbb{Z}/n\mathbb{Z}$.

13.1.5 Example Let $G = \mathrm{Eucl}(\mathbb{R}^2)$ and $H = \mathrm{O}_2(\mathbb{R})$. Recall from chapter 9 that there are two possible forms for an isometry in G so we define a function $\phi : G \to H$ by describing what happens to each of the two forms, as follows:

If $f \in G$ is given by $f(z) = e^{i\theta}z + w$ where $\theta \in \mathbb{R}$ and $w \in \mathbb{C}$ then $\phi(f)$ is the element of H given by $(\phi(f))(z) = e^{i\theta}z$.

If $f \in G$ is given by $f(z) = e^{i\theta}\overline{z} + w$ where $\theta \in \mathbb{R}$ and $w \in \mathbb{C}$ then $\phi(f)$ is the element of H given by $(\phi(f))(z) = e^{i\theta}\overline{z}$.

In both cases we are just dropping the translation part of the map, the '$+w$' which comes at the end.

To prove that this is a homomorphism we need to show that $\phi(f \circ g) = \phi(f) \circ \phi(g)$ for all possible combinations of isometry types for f and g.

For example, if $f(z) = e^{i\theta}z + w$ and $g(z) = e^{i\mu}z + v$ then $(\phi(f))(z) = e^{i\theta}z$ and $(\phi(g))(z) = e^{i\mu}z$. This means that $(\phi(f) \circ \phi(g))(z) = e^{i(\theta+\mu}z$. Also

$$(f \circ g)(z) = e^{i(\theta+\mu)} + e^{i\theta}v + w$$

and so $\phi(f \circ g) = e^{i(\theta+\mu)}z$. Therefore $\phi(f \circ g) = \phi(f) \circ \phi(g)$.

If $f(z) = e^{i\theta}\overline{z} + w$ and $g(z) = e^{i\mu}z + v$ then $(\phi(f))(z) = e^{i\theta}\overline{z}$ and $(\phi(g))(z) = e^{i\mu}z$. This means that $(\phi(f) \circ \phi(g))(z) = e^{i(\theta-\mu}z$. Also

$$(f \circ g)(z) = e^{i(\theta-\mu)} + e^{i\theta}v + w$$

and so $\phi(f \circ g) = e^{i(\theta-\mu)}z$. Therefore $\phi(f \circ g) = \phi(f) \circ \phi(g)$.

The other two cases are similar.

Now we might ask what is Ker ϕ? Giving this some thought, we would need f to have form $f(z) = z + w$ for it to become the identity function upon dropping the translation part, the '$+w$'. So the kernel of ϕ is the set of isometries which are pure translation, those with form $f(z) = z + w$ for some $w \in \mathbb{C}$.

Convince yourself that Im $\phi = \mathrm{O}_2(\mathbb{R})$ (ϕ is surjective).

13.1.6 Exercise Let $G = S_n$ and let $H = \{1, -1\}$ (under multiplication). Define a function $\phi : G \to H$ as follows:

$$\phi(\sigma) = \begin{cases} 1 & \text{if } \sigma \text{ is even} \\ -1 & \text{if } \sigma \text{ is odd} \end{cases}$$

Check that ϕ is a homomorphism and write down Ker ϕ and Im ϕ.

13.1.7 Exercise Let $G = \mathbb{R}^2$ and let $H = \mathbb{R}^2$. Define a function $\phi : G \to H$ as follows:

$$\phi(x, y) = (x, 0).$$

Show that ϕ is a homomorphism and calculate Ker ϕ and Im ϕ.

13.2 The kernel and the image of a homomorphism

You will learn much more about the kernel and the image of homomorphisms in future group theory modules. They are fundamental in understanding exactly how groups are related and the structure of general groups.

For now we will just prove that the kernel and the image of a homomorphism are both subgroups of the groups they lie in. We need a lemma for this.

13.2.1 Lemma Let G and H be groups and let $\phi : G \to H$ be a homomorphism. Then $\phi(1_G) = 1_H$ (here we are using 1_G and 1_H for the respective identity elements rather than the usual 1 for emphasis).

Proof. We have $1_G 1_G = 1_G$. Therefore, since ϕ is a homomorphism, we have

$$\phi(1_G) = \phi(1_G 1_G) = \phi(1_G)\phi(1_G) \text{ (*)}.$$

But $\phi(1_G)$ is just an element of H and as such has an inverse in H, $\phi(1_G)^{-1}$. Multiplying both sides of (*) on the left by $\phi(1_G)^{-1}$ gives

$$1_H = \phi(1_G)^{-1}\phi(1_G) = \phi(1_G)^{-1}\phi(1_G)\phi(1_G) = 1_H\phi(1_G) = \phi(1_G).$$

$\Diamond$

13.2.2 Theorem Let G and H be groups and let $\phi : G \to H$ be a homomorphism. Then

1. Ker $\phi = \{g \in G \mid \phi(g) = 1\}$ is a subgroup of G
2. Im $\phi = \{\theta(g) \mid g \in G$ is a subgroup of H.

Proof.

In both cases we'll use the (a), (b), (c) test for a subgroup from Theorem 6.3.1.

1. (a) Identity. By Lemma 13.2.1, $1 \in$ Ker $\phi = \{g \in G \mid \phi(g) = 1\}$.

   (b) Closure. Suppose $g_1, g_2 \in$ Ker ϕ. Then $\phi(g_1) = \phi(g_2) = 1$. Therefore $\phi(g_1 g_2) = \phi(g1)\phi(g_2) = 1$ and $g_1 g_2 \in$ Ker ϕ.

   (c) Inverses. Suppose $g \in$ Ker ϕ. Then $\phi(g) = 1$. We need to show that $\phi(g^{-1}) = 1$ so that $g^{-1} \in$ Ker ϕ. We do this as follows:

$$\phi(g^{-1}) = \phi(g^{-1}1 = \phi(g^{-1})\phi(g) = \phi(gg^{-1}) = \phi(1) = 1.$$

2. Im $\phi = \{\theta(g) \mid g \in G$ is a subgroup of H. (a) Identity. By Lemma 13.2.1, $1 \in$ Im $\phi = \{\theta(g) \mid g \in G$.

   (b) Closure. Suppose $h_1, h_2 \in$ Im ϕ. Then there exist $g_1, g_2 \in G$ such that $\phi(g_1) = h_1$ and $\phi(g_2) = h_2$. Then $\phi(g_1g_2) = \phi(g_1)\phi(g_2) = h_1h_2$ and so $h_1h_2 \in$ Im ϕ.

   (c) Inverses. Suppose $h \in$ Im ϕ. We need to show that h^{-1} (which exists in H) is also in Im ϕ.

   Since $h \in$ Im ϕ there exists $g \in G$ such that $\phi(g) = h$. Then

$$\phi(g^{-1}\phi(g) = \phi(g^{-1}g) = \phi(1) = 1 \text{ (**)}.$$

   Multiplying both sides of (**) on the right by $\phi(g)^{-1}$ gives

$$\phi(g^{-1}) = \phi(g^{-1}\phi(g)\phi(g)^{-1} = \phi(g)^{-1} = h^{-1}.$$

   This means that $h^{-1} \in$ Im ϕ, as required. ◊

13.2.3 Exercise Let G and H be groups and let $\phi : G \to H$ be a homomorphism. Show that ϕ is injective if and only if Ker $\phi = \{1\}$.

13.2.4 Exercise Let G and H be groups and let $\phi : G \to H$ be an injective homomorphism. Show that H contains a subgroup with is isomorphic to G.

Chapter 14 - Rings

The remaining chapters are about another structure in abstract algebra. These are called rings. Rings have two binary operations, an 'addition' and a 'multiplication'. The set of integers $\mathbb{Z}$ with its usual addition and multiplication is in many ways the prototype ring.

In the integers we have the likes of 'division with remainder'; 'unique factorisation into products of primes' and 'fractions built from integers (the rationals)'. Some of ring theory is concerned with the extent to which other rings have properties like this. We'll take particular interest in rings of polynomials in this module.

If you start to miss groups don't worry, we will see them again in the remaining chapters!

14.1 Definition

A *ring* is a triple $(R, +, \cdot)$, where R is a set and $+$, $\cdot$ are binary operations on R such that the following properties hold

(i) (closure) for all a, $b \in R$, $a + b \in R$ and $a \cdot b \in R$;

(ii) (associativity of addition) for all a, b, $c \in R$

$$(a + b) + c = a + (b + c);$$

(iii) (existence of an additive identity element) there is an element $0 \in R$ such that for all $a \in R$,

$$a + 0 = 0 + a = a.$$

(iv) (existence of additive inverses) for all $a \in R$, there an element, denoted by $-a$, such that

$$a + (-a) = (-a) + a = 0;$$

(v) (commutativity of addition) for all a, $b \in R$,

$$a + b = b + a;$$

(vi) (associativity of multiplication) for all a, b, $c \in R$,

$$a \cdot (b \cdot c) = (a \cdot b) \cdot c;$$

(vii) (distributivity) for all a, b, $c \in R$,

$$a \cdot (b + c) = a \cdot b + a \cdot c; \qquad (b + c) \cdot a = b \cdot a + c \cdot a;$$

(viii) (existence of a multiplicative identity) there is an element $1 \in R$ so that for all $a \in R$,

$$1 \cdot a = a \cdot 1 = a.$$

Moreover, a ring $(R, +, \cdot)$ is said to be *commutative*, if it satisfies the following additional property:

(ix) (commutativity of multiplication) for all a, $b \in R$,

$$a \cdot b = b \cdot a.$$

Note that the word ‘commutative’ in the phrase ‘commutative ring’ refers to multiplication. Commutativity of addition is part of the definition of ring. Some textbooks omit property (viii) from the definition of a ring. Those textbooks call a ring satisfying (viii) a *ring with unity*. We shall always assume that our rings satisfy (viii).

Observe, from properties (i)–(v), if $(R, +, \cdot)$ is a ring, then $(R, +)$ is an abelian group.

14.2 Examples

14.2.1 Example You know lots of examples of rings: $\mathbb{Z}$, $\mathbb{Q}$, $\mathbb{R}$, $\mathbb{C}$, $\mathbb{R}[x]$, etc. All these examples are commutative rings.

14.2.2 Example Let

$$M_{2\times 2}(\mathbb{R}) = \left\{ \begin{pmatrix} a & b \\ c & d \end{pmatrix} \mid a, b, c, d \in \mathbb{R} \right\}.$$

This is the set of 2×2 matrices with real entries. From the properties of matrices it is easy to see that $M_{2\times 2}(\mathbb{R})$ is a ring with the usual addition and multiplication of matrices. The additive identity is the zero matrix, and the multiplicative identity is I_2. The ring $M_{2\times 2}(\mathbb{R})$ is an example of a non-commutative ring, as matrix multiplication is non-commutative.

Similarly we define $M_{2\times 2}(\mathbb{C})$, $M_{2\times 2}(\mathbb{Z})$, $M_{2\times 2}(\mathbb{Q})$. These are all non-commutative rings. $\Diamond$

14.2.3 Theorem Let m be an integer satisfying $m \geq 2$. Then $\mathbb{Z}/m\mathbb{Z}$ is a ring.

Proof. We really mean that $(\mathbb{Z}/m\mathbb{Z}, +_m, \times_m)$ is a commutative ring. We've already seen that that $(\mathbb{Z}/m\mathbb{Z}, +_m)$ is an abelian group. That leaves associativity of multiplication, distributivity and the existence of a multiplicative identity to check. These all follow from the corresponding properties in $\mathbb{Z}$ as follows. Given $a, b, c \in \mathbb{Z}$ we have:

associativity of multiplication: $[a]_m \times_m ([b]_m \times_m [c]_m) = [a]_m \times_m [bc]_m = [a(bc)]_m = [(ab)c]_m = [ab]_m \times_m [c]_m = ([a]_m \times_m [b]_m) \times_m [c]_m$.

distributivity: $[a]_m \times_m ([b]_m +_m [c]_m) = [a]_m.[b+c]_m = [a(b+c)]_m = [ab+ac]_m = [ab]_m +_m [ac]_m = [a]_m \times_m [b]_m +_m [a]_m \times_m [c]_m$.

existence of a multiplicative identity: $[1]_m \times_m [a]_m = [1.a]_m = [a]_m = [a.1]_m = [a]_m \times_m [1]_m$. ♢

14.2.4 Example You're familiar with the following two binary operations on $\mathbb{R}^3$: addition and the cross product (also known as the vector product). Is $(\mathbb{R}^3, +, \times)$ a ring? No. First the cross product is not associative. For example,

$$\mathbf{i} \times (\mathbf{j} \times \mathbf{j}) = 0, \qquad (\mathbf{i} \times \mathbf{j}) \times \mathbf{j} = -\mathbf{i}.$$

We only need one of the properties (i)–(viii) to fail for us to conclude that $(\mathbb{R}^3, +, \times)$ is not a ring. We know that (vi) fails. It is interesting to note that (viii) fails too, as we now show. Indeed,

$$\mathbf{a} \times \mathbf{b} = -\mathbf{b} \times \mathbf{a}. \tag{6}$$

Suppose $\mathbf{1}$ is a vector in $\mathbb{R}^3$ that satisfies

$$\mathbf{a} \times \mathbf{1} = \mathbf{1} \times \mathbf{a} = \mathbf{a}$$

for all $\mathbf{a} \in \mathbb{R}^3$. From (6) we see that $\mathbf{a} = -\mathbf{a}$ for all $\mathbf{a} \in \mathbb{R}^3$. This gives a contradiction. Therefore (viii) fails too. ♢

14.2.5 Example Consider $(\mathbb{R}[x], +, \circ)$, where $\circ$ is composition of polynomials. Is this a ring? No. It is easy to see that all the required properties hold except for distributivity (the "multiplicative identity" is the polynomial $f(x) = x$). Let us give a counterexample to show that distributivity fails. Let

$$f(x) = x^2, \qquad g(x) = x, \qquad h(x) = x.$$

Then

$$f \circ (g + h) = f(2x) = 4x^2; \qquad f \circ g + f \circ h = x^2 + x^2 = 2x^2.$$

♢

14.2.6 Example The **zero ring** is the ring with just one element $\{0\}$. In this ring $1 = 0$, and there is only one possible definition of addition and multiplication: $0 + 0 = 0$, $0 \cdot 0 = 0$. The zero ring is not interesting.

14.2.7 Example Let's step back a little and think about $\mathbb{R}^2$. We know that $(\mathbb{R}^2, +)$ is an abelian group. Is there a way of defining multiplication on $\mathbb{R}^2$ so that we obtain a ring? We will define two different multiplications that make $\mathbb{R}^2$ into a ring. The first is more obvious: we define

$$(a_1, a_2) \times (b_1, b_2) = (a_1 b_1, a_2 b_2).$$

With this definition, you can check that $(\mathbb{R}^2, +, \times)$ is a ring, where the multiplicative identity is $\mathbf{1} = (1, 1)$.

The other way is more subtle: we define

$$(a_1, a_2) \times (b_1, b_2) = (a_1 b_1 - a_2 b_2, a_1 b_2 + a_2 b_1). \tag{7}$$

Where does this definition come from? Recall that $\mathbb{R}^2$ is represented geometrically by the plane, and $\mathbb{C}$ is represented geometrically by the plane. If we're thinking of points in the plane as elements of $\mathbb{R}^2$ then we write them as ordered pairs of real numbers: (a, b). If we're thinking of points in the plane as elements of $\mathbb{C}$ then we write them in the form $a + ib$ where again a, b are real numbers. We multiply in $\mathbb{C}$ using the rule

$$(a_1 + ia_2) \times (b_1 + ib_2) = (a_1 b_1 - a_2 b_2) + i(a_1 b_2 + a_2 b_1). \tag{8}$$

Notice that definitions (7), (8) are exactly the same at the level of points on the plane. We've used the multiplicative structure of $\mathbb{C}$ to define multiplication on $\mathbb{R}^2$. With this definition, $(\mathbb{R}^2, +, \times)$ is a ring. What is the multiplicative identity? It's not $(1, 1)$. For example $(1, 1) \times (1, 1) = (0, 2)$. Think about the multiplicative identity in $\mathbb{C}$. This is simply $1 = 1 + 0i$. So the multiplicative identity in $(\mathbb{R}^2, +, \times)$ (with multiplication defined as in (7)) is $(1, 0)$. Check for yourself that

$$(a_1, a_2) \times (1, 0) = (1, 0) \times (a_1, a_2) = (a_1, a_2).$$

$\Diamond$

Here are couple of quick lemmas about rings.

14.2.8 Lemma Let R be a ring and $a \in R$. Then $0.a = 0 = a.0$.

Proof. We have (make sure you can see why each of the two equalities in the below is true):

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

Adding the additive inverse of $0.a$, namely $-(0.a)$ to both sides of this equation gives $0 = 0.a - 0.a = 0.a + 0.a - 0.a = 0.a$.

That $0 = a.0$ holds similarly. ◊

There is a consequence of the above for a ring R in which $1 = 0$. Then $a = a \cdot 1 = a \cdot 0 = 0$ for all $a \in R$ and so R is the zero ring. To summarise a ring is the zero ring if and only if $1 = 0$.

14.2.9 Lemma Let R be a ring and $a, b \in R$. Then $-(a.b) = (-a).b = a.(-b)$.

Proof. We have that $ab + (-a).b = (a + (-a)).b = 0.b = 0$ by Lemma 14.2.8. By the uniqueness of additive inverses (see 4.1.2, noting that this is written in mutiplicative notation) it follows that $(-a).b = -(ab)$. To get the other result consider $ab + a.(-b)$ in a similar way. ◊

Chapter 15 - Subrings and ideals

15.1 Subrings

Just as we have subgroups, so we have subrings.

15.1.1 Definition Let $(R, +, \cdot)$ be a ring. Let S be a subset of R and suppose that $(S, +, \cdot)$ is also a ring with the same multiplicative identity. Then we say that S is a subring of R (or more formally $(S, +, \cdot)$ is a subring of $(R, +, \cdot)$). $\Diamond$

For S to be a subring of R, we want S to a ring with respect to *the same two binary operations* that makes R a ring, and $1_R \in S$ where 1_R is the multiplicative identity of R.

15.1.2 Example $\mathbb{Z}$ is a subring of $\mathbb{R}$; $\mathbb{Q}$ is a subring of $\mathbb{R}$; $\mathbb{Z}$ is a subring of $\mathbb{Q}$; $\mathbb{R}$ is a subring of $\mathbb{R}[x]$. $\Diamond$

Theorem 6.3.1 gave a criterion for a subset of a group to be a subgroup. As you'd expect we have a similar criterion for a subset of a ring to be a subring.

15.1.3 Theorem Let R be a ring. A subset S of R is a subring if and only if it satisfies the following conditions

(a) $0, 1 \in S$ (that is S contains the additive and multiplicative identity elements of R);

(b) if $a, b \in S$ then $a + b \in S$;

(c) if $a \in S$ then $-a \in S$;

(d) if $a, b \in S$ then $ab \in S$.

Proof. Let's prove this from 'left to right' first. So suppose that the subset S is a subring of R.

By (i),(ii),(iii) and (iv) in the definition of a ring $(S, +)$ is a subgroup of $(R, +)$.

By theorem 6.3.1, since $(S, +)$ is a subgroup of $(R, +)$, $0 \in S$ and both (b) and (c) above are true.

We know from the definition of a subring that the multiplicative identity from R is in S, so we now know that (a) above is true. Also from the definition of a subring, if $a, b \in S$ then $ab \in S$, so (d) above is true.

Now for ‘right to left’ . Suppose S is a subset of R and $(a), (b), (c), (d)$ above are true.

Since $0 \in S$ and since (b) and (c) are true, by 6.3.1 $(S, +)$ is a subgroup of $(R, +)$.

(a) above ensures the existence of a multiplicative identity in S and (d) tells us that S is closed under multiplication.

All that remains to check is the commutativity of the addition in S, the associativity of the multiplication in S and the distributive rules in S. But these all follow immediately because they hold in R and any elements of S are also elements of R. ♢

15.1.4 Example In Example 6.3.3, we saw that the set of even integers $2\mathbb{Z}$ is a subgroup of $\mathbb{Z}$. Strictly speaking, $(2\mathbb{Z}, +)$ is a subgroup of $(\mathbb{Z}, +)$. Now we know that $(\mathbb{Z}, +, \cdot)$ is a ring. Is $(2\mathbb{Z}, +, \cdot)$ a subring? From Theorem 15.1.3 we see that it isn't because $1 \notin 2\mathbb{Z}$. ♢

15.1.5 Example In view of the previous example, let's try to discover if $\mathbb{Z}$ has any subrings other than itself. Let S be a subring of $\mathbb{Z}$. We know that $0,\ 1 \in S$. Also, by (b) we know that $2 = 1 + 1 \in S$. Repeating the argument, $3 = 2 + 1 \in S$ and so on. By induction we know that $0, 1, 2, \ldots$ are all in S. But by (c), if $a \in S$ then $-a \in S$. So $\ldots, -3, -2, -1$ are also in S. Hence $\mathbb{Z}$ is contained in S. But S is a subset of $\mathbb{Z}$. So they must be equal: $S = \mathbb{Z}$.

Therefore, the only subring of $\mathbb{Z}$ is $\mathbb{Z}$ itself.

♢

15.1.6 Exercise As a diversion, contrast the above to the situation with subgroups of the group $(\mathbb{Z}, +)$.

1. Show that $n\mathbb{Z} = \{nm \mid m \in \mathbb{Z}\}$ is a subgroup of $(\mathbb{Z}, +)$ for any integer n. Deduce that there are infinitely many proper subgroups of $(\mathbb{Z}, +)$.

2. Let H be a non-trivial, proper subgroup of $\mathbb{Z}$. Let n be the smallest positive integer in H. Using, division with remainder, show that $H = n\mathbb{Z}$. ♢

Back to rings!

15.1.7 Exercise Let m be an integer satisfying $m \geq 2$. Show that the only subring of $\mathbb{Z}/m\mathbb{Z}$ is $\mathbb{Z}/m\mathbb{Z}$ itself. $\Diamond$

The easiest way to show that a set is a ring is to show that it is a subring of a known ring. If you do this, you only have four properties to check (a),(b),(c),(d). If you don't do this, you'll have eight properties to check (i)–(viii). The following two examples will help you appreciate this principle.

15.1.8 Example Let

$$S = \left\{ \frac{a}{2^r} \mid a, r \in \mathbb{Z}, r \geq 0 \right\}.$$

Then S is a ring.

First think of a ring that contains S. The elements of S are rational numbers whose denominator is a power of 2; for example

$$7 = \frac{7}{2^0}, \qquad \frac{-1}{2}, \qquad \frac{15}{8} = \frac{15}{2^3}$$

are elements of S. An obvious choice of a ring that contains S is $\mathbb{Q}$, the ring of rational numbers. So let's show that S is a subring of $\mathbb{Q}$. Clearly $0 = 0/2^0$ and $1 = 1/2^0$ are in S. Suppose α, β are elements of S. We can write

$$\alpha = \frac{a}{2^r}, \qquad \beta = \frac{b}{2^s},$$

where a, b, r, $s \in \mathbb{Z}$ and r, $s \geq 0$. We want to check that $\alpha + \beta$, $-\alpha$ and $\alpha\beta$ are in S. Note that

$$-\alpha = \frac{-a}{2^r}, \qquad \alpha\beta = \frac{ab}{2^{r+s}}.$$

Clearly $-\alpha$, $\alpha\beta$ are in S, since $-a$, $a + b$, r, $r + s$ are integers and r, $r + s \geq 0$. Now for the sum, we'll assume without loss of generality that $r \geq s$. Then

$$\alpha + \beta = \frac{a + 2^{r-s}b}{2^r}.$$

Now since a, b, r, s are integers and $r \geq s$, we have $a + 2^{r-s}b$ is also an integer. Clearly, $\alpha + \beta$ is in S. By Theorem 15.1.3, S is a subring and therefore a ring. $\Diamond$

15.1.9 Exercise Which of the following are subrings of $M_{2\times 2}(\mathbb{R})$? If so, are they commutative?

(i) $\left\{\begin{pmatrix} a & b \\ 0 & c \end{pmatrix} \mid a, b, c \in \mathbb{R}\right\}$.

(ii) $\left\{\begin{pmatrix} a & b \\ 0 & 0 \end{pmatrix} \mid a, b \in \mathbb{R}\right\}$.

(iii) $\left\{\begin{pmatrix} a & b \\ 0 & 1 \end{pmatrix} \mid a, b \in \mathbb{R}\right\}$.

(iv) $\left\{\begin{pmatrix} a & 0 \\ 0 & b \end{pmatrix} \mid a \in \mathbb{R}, b \in \mathbb{Z}\right\}$.

(v) $\left\{\begin{pmatrix} a & b \\ -b & a \end{pmatrix} \mid a, b \in \mathbb{R}\right\}$.

(vi) $\{A \in M_{2\times 2}(\mathbb{R}) \mid \det(A) = 1\}$.

15.2 Ideals

We can get some understanding of a ring by considering its subring but generally the study of rings depends on other subsets of the ring called *ideals*. We'll use these most in chapter 18 but they will also crop up in chapter 17.

15.2.1 Definition Let R be a ring. Let I be a subset of R. Then I is said to be a (two-sided) ideal of R if

1. $(I, +)$ is a subgroup of $(R, +)$
2. For every $x \in I$ and $r \in R$, both $xr \in I$ and $rx \in I$.

15.2.2 Examples

1. In any ring R, $\{0\}$ and R itself are ideals.
2. $n\mathbb{Z}$ is an ideal of $\mathbb{Z}$ for any $n \in \mathbb{Z}$.
3. The set of polynomials of the form $a_n x^n + a_{n-1}x^{n-1} + \cdots + a_2x^2 + a_1x$ (i.e. those with no constant term), where $a_i \in \mathbb{R}$ is an ideal of $\mathbb{R}[x]$.
4. For any commutative ring R and any $r \in R$, $rR = \{rs \mid s \in R\}$ is an ideal of $\mathbb{R}$.
5. $M_n(m\mathbb{Z})$, i.e the set of matrices 2×2 matrices whose entries are all multiples of m, is an ideal of $M_n(\mathbb{Z})$ for any integer m.

15.2.3 Lemma Let R be a non-zero ring and let I be an ideal of R. Then $I = R$ if and only if $1 \in I$.

Proof. Suppose $I = R$. Then, since $1 \in R$, $1 \in I$.

For the converse suppose $1 \in I$. Since I is an ideal we have $1.r \in I$ for all $r \in R$. But $1.r = r$ and so this means $r \in I$ for all $r \in R$. Therefore $R = I$. $\Diamond$

We can generalise the result above a little. Suppose that I is an ideal of $\mathbb{Z}$ and $-1 \in I$. Then, for all $n \in \mathbb{Z}$, we would have $n = (-1) \times (-n) \in I$ and therefore $I = \mathbb{Z}$.

-1 is the only element in $\mathbb{Z}$ other than 1 itself to have a multiplicative inverse, namely -1, and you can see how this was used in the argument that is $-1 \in I$, an ideal of $\mathbb{Z}$ then $I = \mathbb{Z}$ just given. As an element in a ring which has a multiplicative inverse is called a *unit* of that ring. Here is the definition.

15.2.4 Definition Let R be a ring. An element u is called a *unit* if there is some element v in R such that $uv = vu = 1$. In other words, an element u of R is a unit if it has a multiplicative inverse that belongs to R.

15.2.5 Example In any non-zero ring, 0 is not a unit and 1 is a unit. $\Diamond$

15.2.6 Example In $\mathbb{R}$, $\mathbb{Q}$, $\mathbb{C}$, every non-zero element has a multiplicative inverse. So the units are the non-zero elements. $\Diamond$

15.2.7 Example What are the units in $\mathbb{Z}$? Suppose u is a unit in $\mathbb{Z}$. Then there is some $v \in \mathbb{Z}$ such that $uv = vu = 1$. This means that $1/u$ is an integer. The only integers u such that $1/u$ is also an integer are ± 1. So the units in $\mathbb{Z}$ are ± 1. $\Diamond$

15.2.8 Example Recall that $\mathbb{R}[x]$ is the ring of polynomials with real coefficients. Then x is not a unit, since there is no polynomial that we can multiply x by to get 1 (think of the degree of resultant polynomial, more on this in chapter 17). However, 2 is a unit, since $1/2$ is a polynomial in $\mathbb{R}[x]$ with real coefficients:

$$\frac{1}{2} = \frac{1}{2} + 0x.$$

Convince yourself that the units in $\mathbb{R}[x]$ are precisely the non-zero constant polynomials. $\Diamond$

So now we will generalise Lemma 15.2.3 slightly, in the form of a corollary.

15.2.9 Corollary Let R be a non-zero ring and let I be an ideal of R. Then $I = R$ if and only if I contains a unit in R.

Proof. Suppose $I = R$. Then, since $1 \in R$, $1 \in I$ and is a unit.

For the converse suppose $u \in I$ and u is a unit. This means that there exists $v \in R$ such that $uv = vu = 1$. But, since I is an ideal $1 = uv \in I$ and $I = R$ by Lemma 15.2.3. $\Diamond$

We can work out exactly what the ideals of $\mathbb{Z}$ look like.

15.2.10 Theorem Let I be an ideal of $\mathbb{Z}$. Then $I = n\mathbb{Z} = \{nm \mid m \in \mathbb{Z}\}$ for some $n \in \mathbb{Z}$.

Proof. If $I = \{0\}$ then let $n = 0$ and we are done.

So assume $I \neq 0$. If $x \in I$ then $-x \in I$ and so I will contain a positive integer. Let n be the smallest positive integer contained in I. Since I is an ideal and $n \in I$, $nm \in I$ for all integers m. This means that $n\mathbb{Z} \subseteq I$.

Now let $x \in I$. By 'division with remainder' there exist $q, r \in \mathbb{Z}$ with $0 \leq r \leq n-1$ such that $x = nq + r$. Since $n \in I$, $nq \in I$ and $-nq \in I$. Also $x \in I$ and therefore $r = x - nq \in I$. By the minimality of n this means $r = 0$ and $x = nq \in n\mathbb{Z}$. Therefore $I \subseteq nZ$ and $I = n\mathbb{Z}$. $\Diamond$

15.2.11 Exercise Let R be a commutative ring and let $x \in R$. Prove that $xR = \{xr \mid r \in R\}$ is an ideal of R. Show by example that this may be false if R is not commutative.

15.2.12 Exercise Let I, J be ideals of a ring R.

1. Prove that $I + J = \{x + y \mid x \in I, y \in J\}$ is an ideal of R.
2. Prove that $I \cap J$ is an ideal of R. Show by example that $I \cup J$ is not necessarily an ideal of R.

15.2.13 Exercise Express each of the following ideals of $\mathbb{Z}$ in the form $m\mathbb{Z}$ where $m \in \mathbb{Z}$.

1. $5\mathbb{Z} \cap 7\mathbb{Z}$

2. $10\mathbb{Z} \cap 15\mathbb{Z}$

3. $3\mathbb{Z} \cap 5\mathbb{Z} \cap 7\mathbb{Z}$.

4. $6\mathbb{Z} + 8\mathbb{Z}$.

5. $6\mathbb{Z} + 5\mathbb{Z}$.

Chapter 16 (non-examinable) - The unit group of a ring

16.1 The unit group of a ring

Recall that we defined $\mathbb{R}^*$, $\mathbb{Q}^*$, $\mathbb{C}^*$ be removing from $\mathbb{R}$, $\mathbb{Q}$, $\mathbb{C}$ the zero element; e.g.

$$\mathbb{R}^* = \{a \in \mathbb{R} \mid a \neq 0\}.$$

We found that $\mathbb{R}^*$ is group with respect to multiplication. In Example 3.2.4 we tried to do the same with $\mathbb{Z}$ and failed to obtain a group. Note that $\mathbb{R}$, $\mathbb{Q}$, $\mathbb{C}$ are rings and so is $\mathbb{Z}$. Given a ring, is there a naturally defined subset that is a group with respect to multiplication? It turns out that the answer is yes, and that for $\mathbb{R}$, $\mathbb{Q}$ and $\mathbb{C}$ we obtain $\mathbb{R}^*$, $\mathbb{Q}^*$, $\mathbb{C}^*$ as we'd expect.

Let's think about what this natural subset might be. It seems sensible to include 1 in the subset because that will certainly act as an identity element in the group we seek. Then, if we include any other element it would have to be a unit from the ring in order that it has a multiplicative inverse. This line of thinking leads us to wonder whether the set of *all* the units in a ring is a group under multiplication. It turns out that it is:

16.1.1 Definition Let R be a ring. We define the *unit group of* R to be the set [7]

$$R^* = \{a \in R \mid a \text{ is a unit in } R\}. \tag{9}$$

Just because we've called R^* the *unit group of* R doesn't get us out of checking that it is really a group!

16.1.2 Lemma Let $(R, +, \cdot)$ be a ring and let R^* be the subset defined in (9). Then $(R^*, \cdot)$ is a group.

Proof. We must first show that R^* is closed under multiplication. Suppose u_1, $u_2 \in R^*$. Thus u_1, u_2 are units of R, and so there are v_1, $v_2 \in R$ such that

$$u_1 v_1 = v_1 u_1 = 1, \qquad u_2 v_2 = v_2 u_2 = 1. \tag{10}$$

We want to show that $u_1 u_2$ is a unit. Note that $v_2 v_1 \in R$ since R is closed under multiplication (it's a ring after all). Moreover,

$$\begin{aligned}(u_1u_2)(v_2v_1) &= u_1(u_2v_2)v_1 \qquad \text{associativity of multiplication}\\ &= u_1 \cdot 1 \cdot v_1 \qquad \text{since } u_2v_2 = 1\\ &= 1 \qquad \text{since } u_1v_1 = 1.\end{aligned}$$

[7] Some mathematicians write $R^\times$ instead of R^*.

Similarly $(v_2v_1)(u_1u_2) = 1$. Thus u_1u_2 is a unit [8] in R, and so $u_1u_2 \in R^*$. We've proved that R^* is closed under multiplication.

We want to show that multiplication is associative in R^*. But multiplication is associative in R since R is a ring. Therefore it is associative in R^*.

Since $1 \cdot 1 = 1$, 1 is a unit and so $1 \in R^*$.

Finally we want to show that every element in R^* has a multiplicative inverse that belongs to R^*. Suppose $u \in R^*$. Then $uv = vu = 1$ for some $v \in R$. Note that this makes v also a unit, and so $v \in R^*$. Thus u has a multiplicative inverse in R^*. This completes the proof that R^* is a group.

16.1.3 Example Note that $\mathbb{R}^*$, $\mathbb{C}^*$, $\mathbb{Q}^*$ have exactly the same meaning as before. ♢

16.1.4 Example We showed that the units of $\mathbb{Z}$ are ±1. Therefore the unit group of $\mathbb{Z}$ is

$$\mathbb{Z}^* = \{1, -1\}.$$

♢

16.1.5 Example Recall that $M_{2\times2}(\mathbb{R})$ is the ring of 2×2 matrices with real entries. It is clear from the definition of a unit, that the units of $M_{2\times2}(\mathbb{R})$ are the invertible matrices. In other words, they are the ones having non-zero determinant. Thus

$$(M_{2\times2}(\mathbb{R}))^* = \mathrm{GL}_2(\mathbb{R}).$$

Similarly,

$$(M_{2\times2}(\mathbb{Q}))^* = \mathrm{GL}_2(\mathbb{Q}), \qquad (M_{2\times2}(\mathbb{C}))^* = \mathrm{GL}_2(\mathbb{C}).$$

What about the unit group of $M_{2\times2}(\mathbb{Z})$? This is more complicated. For example, consider the matrix $A = \left(\begin{smallmatrix} 3 & 1 \\ 1 & 1 \end{smallmatrix}\right)$. The matrix A is invertible, and $A^{-1} = \left(\begin{smallmatrix} 1/2 & -1/2 \\ -1/2 & 3/2 \end{smallmatrix}\right)$. Although A is in $M_{2\times2}(\mathbb{Z})$, its inverse is not in $M_{2\times2}(\mathbb{Z})$, but it is in $M_{2\times2}(\mathbb{Q})$ and $M_{2\times2}(\mathbb{R})$. Thus A is a unit in $M_{2\times2}(\mathbb{Q})$, and $M_{2\times2}(\mathbb{R})$ but not in

[8] Start again. We have u_1, u_2 are units and so satisfy (10) for some v_1, v_2 in R. We want to show that u_1u_2 is a unit. **What is wrong with the following argument?**

$$(u_1u_2)(v_1v_2) = (u_1v_1)(u_2v_2) = 1 \cdot 1 = 1.$$

Similarly $(v_1v_2)(u_1u_2) = 1$. Thus u_1u_2 is a unit.

$M_{2\times 2}(\mathbb{Z})$. The problem is clear: when calculating the inverse of a matrix, we must divide by its determinant, and the result does not have to be an integer.

Let's go back to the definition of a unit. Suppose $A \in M_{2\times 2}(\mathbb{Z})$ is a unit. Then there is a matrix $B \in M_{2\times 2}(\mathbb{Z})$ such that

$$AB = BA = I_2.$$

Taking determinants, are recalling that $\det(AB) = \det(A)\det(B)$ we find that

$$\det(A)\det(B) = 1.$$

Now $\det(A)$ and $\det(B)$ are integers because A and B have integer entries. Thus

$$\det(A) = \det(B) = 1, \qquad \text{or} \qquad \det(A) = \det(B) = -1.$$

Conversely if $A \in M_{2\times 2}(\mathbb{Z})$ has determinant ± 1, then its inverse will have integer entries and so A is a unit. We deduce that

$$(M_{2\times 2}(\mathbb{Z}))^* = \{A \in M_{2\times 2}(\mathbb{Z}) \mid \det(A) = \pm 1\}.$$

We define the group $\mathrm{GL}_2(\mathbb{Z})$ by

$$\mathrm{GL}_2(\mathbb{Z}) = \{A \in M_{2\times 2}(\mathbb{Z}) \mid \det(A) = \pm 1\};$$

then $(M_{2\times 2}(\mathbb{Z}))^* = \mathrm{GL}_2(\mathbb{Z})$. In fact, for a *commutative* ring R we define

$$\mathrm{GL}_2(R) = \{A \in M_{2\times 2}(R) \mid \det(A) \in R^*\}.$$

You will easily see that this is consistent with the earlier definitions of $\mathrm{GL}_2(\mathbb{R})$, $\mathrm{GL}_2(\mathbb{C})$, $\mathrm{GL}_2(\mathbb{Q})$ and $\mathrm{GL}_2(\mathbb{Z})$, and that moreover, $(M_{2\times 2}(R))^* = \mathrm{GL}_2(R)$.

16.1.6 Example Let

$$S = \left\{ \begin{pmatrix} a & b \\ 0 & c \end{pmatrix} \mid a, b, c \in \mathbb{Z} \right\}.$$

We'll show that that S is a ring under the usual addition and multiplication of matrices and then find S^*.

To show that S is a ring it is enough to show that it is a subring of $M_{2\times 2}(\mathbb{Z})$. We leave that as an exercise.

Let us compute the unit group. Suppose $A = \left(\begin{smallmatrix} a & b \\ 0 & c \end{smallmatrix}\right)$ is in S. To be unit it is not enough for this matrix to be invertible, we also want the inverse to belong to S. So we require the determinant ac to be non-zero and we want

$$A^{-1} = \frac{1}{ac}\begin{pmatrix} c & -b \\ 0 & a \end{pmatrix} = \begin{pmatrix} 1/a & -b/ac \\ 0 & 1/c \end{pmatrix}$$

to belong to S. Thus we want the integers a, b, c to satisfy

$$ac \neq 0, \qquad \frac{1}{a}, \frac{1}{c}, -\frac{b}{ac} \in \mathbb{Z}.$$

This happens precisely when $a = \pm 1$ and $c = \pm 1$. Thus

$$S^* = \left\{ \begin{pmatrix} \pm 1 & b \\ 0 & \pm 1 \end{pmatrix} \mid b \in \mathbb{Z} \right\}.$$

◊

16.1.7 Exercise In Example 15.1.8, we showed that

$$S = \left\{ \frac{a}{2^r} \mid a, \, r \in \mathbb{Z}, \, r \geq 0 \right\}$$

is a ring. Find its unit group.

16.1.8 Exercise Let $\mathbb{Z}[\sqrt{2}] = \{a + b\sqrt{2} \mid a, b \in \mathbb{Z}\}$. Show that $\mathbb{Z}[\sqrt{2}]$ is a ring and that $1 + \sqrt{2}$ is a unit. What is its order as an element of the group $\mathbb{Z}[\sqrt{2}]^*$?

16.1.9 Exercise Let $\zeta = e^{2\pi i/3}$ (this is a cube root of unity). Check that $\overline{\zeta} = \zeta^2$. Let $\mathbb{Z}[\zeta] = \{a + b\zeta : a, b \in \mathbb{Z}\}$.

(i) Show that $\zeta^2 \in \mathbb{Z}[\zeta]$ (**Hint:** the sum of the cube roots of unity is ...).

(ii) Show that $\mathbb{Z}[\zeta]$ is a ring.

(iii) Show that ± 1, $\pm\zeta$ and $\pm\zeta^2$ are units in $\mathbb{Z}[\zeta]$.

(iv) (Harder) Show that $\mathbb{Z}[\zeta]^* = \{\pm 1, \pm\zeta, \pm\zeta^2\}$.

(v) Show that this group is cyclic.

16.2 Fields

A *field* $(F, +, \cdot)$ is a commutative ring which is not the zero ring such that every non-zero element is a unit. Thus a commutative ring F is a field if and only if its unit group is

$$F^* = \{a \in F \mid a \neq 0\}.$$

16.2.1 Example $\mathbb{R}$, $\mathbb{C}$, $\mathbb{Q}$ are fields. ◊

16.2.2 Example $\mathbb{Z}$ is not a field, since for example $2 \in \mathbb{Z}$ is non-zero but not a unit. ◊

16.2.3 Example $\mathbb{R}[x]$ is not a field, since for example $x \in \mathbb{R}[x]$ is non-zero but not a unit. ◊

16.2.4 Example

$$\mathbb{Q}[i] = \{a + bi \mid a, b \in \mathbb{Q}\}$$

is a field as follows.

First we have to show that $\mathbb{Q}[i]$ is a commutative ring. For this it is enough to show that $\mathbb{Q}[i]$ is a subring of $\mathbb{C}$. It is clearly a subset of $\mathbb{C}$ that contains 0 and 1. Suppose α, $\beta \in \mathbb{Q}[i]$. We want to show that $\alpha + \beta$, $\alpha\beta$, $-\alpha$ are all in $\mathbb{Q}[i]$. Write

$$\alpha = a + bi, \qquad \beta = c + di$$

where a, b, c, $d \in \mathbb{Q}$. Then

$$\alpha + \beta = (a + c) + (b + d)i.$$

Since $\mathbb{Q}$ is closed under addition, $a + c$ and $b + d \in \mathbb{Q}$. So $\alpha + \beta \in \mathbb{Q}[i]$. Similarly, check for yourself that $\alpha\beta$ and $-\alpha$ are in $\mathbb{Q}[i]$. Thus $\mathbb{Q}[i]$ is a subring of $\mathbb{C}$ and so a ring [9].

Finally we have to show that every non-zero element of $\mathbb{Q}[i]$ is a unit. Suppose α is a non-zero element of $\mathbb{Q}[i]$. We can write $\alpha = a + bi$ where a, $b \in \mathbb{Q}$,

[9] Arguably, we could've made the proof more transparent by writing

$$\alpha = \frac{r}{s} + \frac{u}{v}i, \qquad \beta = \frac{k}{\ell} + \frac{m}{n}i,$$

where r, s, u, v, k, ℓ, m, n are integers and s, v, ℓ, n are non-zero. This would've worked, but it's probably better to get used to thinking of rational numbers as numbers in their own right.

and not both zero. We want to show that existence of some $\beta \in \mathbb{Q}[i]$ such that $\alpha\beta = \beta\alpha = 1$. In other words, we want to show that $1/\alpha$ is in $\mathbb{Q}[i]$. But we know how to compute $1/\alpha$. Recall that to divide complex numbers we multiply the numerator and denominator by the conjugate of the denominator:

$$\begin{aligned}\frac{1}{\alpha} &= \frac{1}{a+bi} \\ &= \frac{1}{a+bi} \cdot \frac{a-bi}{a-bi} \\ &= \frac{a-bi}{a^2+b^2} \\ &= \frac{a}{a^2+b^2} - \frac{b}{a^2+b^2}i.\end{aligned}$$

As a, b are rationals, so are $a/(a^2+b^2)$ and $b/(a^2+b^2)$. So $1/\alpha$ is in $\mathbb{Q}[i]$. Therefore $\mathbb{Q}[i]$ is a field. $\Diamond$

16.2.5 Exercise Let $\mathbb{Q}[\sqrt{2}] = \{a + b\sqrt{2} : a, b \in \mathbb{Q}\}$. Show that $\mathbb{Q}[\sqrt{2}]$ is a field.

16.2.6 Exercise Let

$$F = \left\{ \left(\begin{smallmatrix} a & b \\ -b & a \end{smallmatrix} \right) \mid a, b \in \mathbb{R} \right\}.$$

(a) Show that F is a field (under the usual addition and multiplication of matrices). (**Hint:** Begin by showing that F is a subring of $M_{2\times 2}(\mathbb{R})$. You need to also show that F is commutative and that every non-zero element has an inverse in F.)

(b) Let $\phi : F \to \mathbb{C}$ be given by $\phi \left(\begin{smallmatrix} a & b \\ -b & a \end{smallmatrix} \right) = a + bi$. Show that ϕ is a bijection that satisfies $\phi(A + B) = \phi(A) + \phi(B)$ and $\phi(AB) = \phi(A)\phi(B)$. Note that, although we are not defining it in this module, you have just shown that F and C are isomorphic (as rings/fields).

(c) Show that

$$F' = \left\{ \left(\begin{smallmatrix} a & b \\ -b & a \end{smallmatrix} \right) \mid a, b \in \mathbb{C} \right\}$$

is not a field.

16.3 Units in $\mathbb{Z}/m\mathbb{Z}$

16.3.1 Example What are the unit groups of $\mathbb{Z}/m\mathbb{Z}$ for $m = 2, 3, 4, 5, 6$. To work this out just look at the multiplication table for $\mathbb{Z}/6\mathbb{Z}$

$\times_6$	$[0]_6$	$[1]_6$	$[2]_6$	$[3]_6$	$[4]_6$	$[5]_6$
$[0]_6$	$[0]_6$	$[0]_6$	$[0]_6$	$[0]_6$	$[0]_6$	$[0]_6$
$[1]_6$	$[0]_6$	$[1]_6$	$[2]_6$	$[3]_6$	$[4]_6$	$[5]_6$
$[2]_6$	$[0]_6$	$[2]_6$	$[4]_6$	$[0]_6$	$[2]_6$	$[4]_6$
$[3]_6$	$[0]_6$	$[3]_6$	$[0]_6$	$[3]_6$	$[0]_6$	$[3]_6$
$[4]_6$	$[0]_6$	$[4]_6$	$[2]_6$	$[0]_6$	$[4]_6$	$[2]_6$
$[5]_6$	$[0]_6$	$[5]_6$	$[4]_6$	$[3]_6$	$[2]_6$	$[1]_6$

and scan for the elements which have multiplicative inverses. You'll see that.

$$(\mathbb{Z}/6\mathbb{Z})^* = \{[1]_6, [5]_6\}\,.$$

In the same way you'll find that

$$\begin{aligned} (\mathbb{Z}/2\mathbb{Z})^* &= \{[1]_2\}\,, & (\mathbb{Z}/3\mathbb{Z})^* &= \{[1]_3, [2]_3\}\,, \\ (\mathbb{Z}/4\mathbb{Z})^* &= \{[1]_4, [3]_4\}\,, & (\mathbb{Z}/5\mathbb{Z})^* &= \{[1]_5, [2]_5, [3]_5, [4]_5\}\,. \end{aligned}$$

For example here is the multiplication table for $\mathbb{Z}/5\mathbb{Z}$ from which you can see that every non-zero element is a unit.

$\times_5$	$[0]_5$	$[1]_5$	$[2]_5$	$[3]_5$	$[4]_5$
$[0]_5$	$[0]_5$	$[0]_5$	$[0]_5$	$[0]_5$	$[0]_5$
$[1]_5$	$[0]_5$	$[1]_5$	$[2]_5$	$[3]_5$	$[4]_5$
$[2]_5$	$[0]_5$	$[2]_5$	$[4]_5$	$[1]_5$	$[3]_5$
$[3]_5$	$[0]_5$	$[3]_5$	$[1]_5$	$[4]_5$	$[2]_5$
$[4]_5$	$[0]_5$	$[4]_5$	$[3]_5$	$[2]_5$	$[1]_5$

In particular, $\mathbb{Z}/2\mathbb{Z}$, $\mathbb{Z}/3\mathbb{Z}$ and $\mathbb{Z}/5\mathbb{Z}$ are fields and $\mathbb{Z}/4\mathbb{Z}$, $\mathbb{Z}/6\mathbb{Z}$ are not fields. Can you make a general guess as to which $\mathbb{Z}/m\mathbb{Z}$ are fields and which aren't? Can you prove your guess? $\Diamond$

16.3.2 Theorem Let $[a]_m \in \mathbb{Z}/m\mathbb{Z}$. Then $[a]_m$ is a unit in $\mathbb{Z}/m\mathbb{Z}$ if and only if $\gcd(a, m) = 1$. Thus

$$(\mathbb{Z}/m\mathbb{Z})^* = \{[a]_m \mid 0 \le a \le m-1 \ \text{ and } \ \gcd(a, m) = 1\}\,.$$

Proof. Suppose $[a]_m$ is a unit in $\mathbb{Z}/m\mathbb{Z}$. Then there is some $[b]_m$ in $\mathbb{Z}/m\mathbb{Z}$ so that $ab \equiv 1 \pmod{m}$. Thus, there is some $k \in \mathbb{Z}$ such that $ab - 1 = km$. Write $g = \gcd(a, m)$. Then $g \mid a$ and $g \mid m$. So $g \mid (ab - km) = 1$. But this means that $g = 1$.

Conversely, suppose $\gcd(a, m) = 1$. By Bezout's Lemma (see *Foundations*) we know that we can write $1 = ba + cm$ for some integers b, $c \in \mathbb{Z}$. Thus $ab \equiv 1 \pmod{m}$. Hence $[a]$ is a unit, with multiplicative inverse $[b]_m$. $\Diamond$

16.3.3 Exercise Redo Example 16.3.1 using Theorem 16.3.2.

16.3.4 Example By Theorem 16.3.2, we know that $\overline{19}$ is invertible in $\mathbb{Z}/256\mathbb{Z}$. But the statement of the theorem does not tell us how to find the inverse. It would take us a very long to run through the elements $\overline{u} \in \mathbb{Z}/256\mathbb{Z}$ and check to see if $19u \equiv 1 \pmod{256}$. However, **the proof of the theorem does give us a recipe for finding the inverse.** We know by factoring that $\gcd(19, 256) = 1$, but let's use Euclid's Algorithm [10] to write 1 as a linear combination of 19 and 256:

$$\begin{aligned} \mathbf{256} &= 13 \times \mathbf{19} + \mathbf{9} \\ \mathbf{19} &= 2 \times \mathbf{9} + \mathbf{1}. \end{aligned}$$

Thus

$$\mathbf{1} = \mathbf{19} - 2 \times \mathbf{9} = \mathbf{19} - 2 \times (\mathbf{256} - 13 \times \mathbf{19}) = (1 - 2 \times -13) \times \mathbf{19} - 2 \times \mathbf{256},$$

so

$$\mathbf{1} = 27 \times \mathbf{19} - 2 \times \mathbf{256}.$$

Hence $27 \times 19 \equiv 1 \pmod{256}$, so $[27]_256$ is the inverse of $[19]_256$ in $\mathbb{Z}/256\mathbb{Z}$. $\Diamond$

16.4 Fermat's Little Theorem & Euler's Theorem via group theory

Through the computations you've done so far, you've probably conjectured the following.

16.4.1 Theorem Let p be a prime. Then $\mathbb{Z}/p\mathbb{Z}$ is a field. Therefore,

$$(\mathbb{Z}/p\mathbb{Z})^* = \{[1]_p, [2]_p, \dots, [p-1]_p\}.$$

Proof. We already know that $\mathbb{Z}/m\mathbb{Z}$ is a commutative ring for any integer $m \geq 2$. Now to show that $\mathbb{Z}/p\mathbb{Z}$ is a field, we must show that any non-zero $[a]_p \in \mathbb{Z}/p\mathbb{Z}$ is invertible. But if $[a]_p \in \mathbb{Z}/p\mathbb{Z}$ is non-zero, then a is one of $1, 2, \dots, p-1$. Clearly a is not divisible by p. Since p is prime, $\gcd(a, p) = 1$. Hence by Theorem 16.3.2, $[a]_p$ is invertible in $\mathbb{Z}/p\mathbb{Z}$. This shows that $\mathbb{Z}/p\mathbb{Z}$ is a field.

[10] It is easy to get muddled in the substitutions involved in Euclid's Algorithm. One way to reduce the muddle is to somehow distinguish the numbers you started with, here 256 and 19, and the remainders from the quotients. I did the distinguishing by writing the numbers we started with and the remainders in boldtype. In your calculations, you can underline them.

16.4.2 Exercise Prove the converse of Theorem 16.4.1: if $\mathbb{Z}/m\mathbb{Z}$ is a field then m is prime.

16.4.3 Theorem (Fermat's Little Theorem) Let p be a prime and a an integer such that $p \nmid a$. Then

$$a^{p-1} \equiv 1 \pmod{p}. \tag{11}$$

Proof. We know that $a \equiv b \pmod{p}$ where b is one of $0, 1, 2, \ldots, p-1$. Now as $p \nmid a$, we see that $b \neq 0$. By Theorem 16.4.1, $[b]_p$ is in the unit group of $\mathbb{Z}/p\mathbb{Z}$ which is

$$(\mathbb{Z}/p\mathbb{Z})^* = \{[1]_p, [2]_p, \ldots, [p-1]_p\}.$$

The order of the group $(\mathbb{Z}/p\mathbb{Z})^*$ is clearly $p-1$. By Theorem 12.3.1,

$$[b]^{p-1} = 1.$$

Thus $b^{p-1} \equiv 1 \pmod{p}$. Since $a \equiv b \pmod{p}$, we obtain (11). ♢

Here's a fun application of Fermat's Little Theorem.

16.4.4 Example Here we'll compute $2^{1000} \pmod{13}$.

Since 13 is prime and $13 \nmid 2$, we know by Fermat's Little Theorem that $2^{12} \equiv 1 \pmod{13}$. Now by the Division Algorithm,

$$1000 = 83 \times 12 + 4.$$

Therefore,

$$2^{1000} = 2^{83 \times 12 + 4} = (2^{12})^{83} \times 2^4 \equiv 1^{83} \times 16 \equiv 3 \pmod{13}.$$

♢

Now for Euler's theorem...

16.4.5 Definition Let $m \geq 1$. We denote the order of the group $(\mathbb{Z}/m\mathbb{Z})^*$ by $\varphi(m)$. The function φ is called *Euler's φ-function.*

16.4.6 Example We know that if p is a prime, then $(\mathbb{Z}/p\mathbb{Z})^* = \{[1]_p, [2]_p, \ldots, [p-1]_p\}$., and so $\varphi(p) = p-1$. ♢

16.4.7 Example We know that

$$(\mathbb{Z}/6\mathbb{Z})^* = \{[1]_6, [5]_6\},$$

and so $\varphi(6) = 2$. ♢

16.4.8 Example Let $n \geq 1$. Then $(\mathbb{Z}/2^n\mathbb{Z})^*$ consists of $[a]_{2^n}$ with a in the range $0 \leq a \leq 2^n - 1$ that are coprime to 2^n. These are the odd integers a in the range $0 \leq a \leq 2^n - 1$. Thus

$$(\mathbb{Z}/2^n\mathbb{Z})^* = \{[1]_{2^n}, [3]_{2^n}, \dots, [2^n - 1]_{2^n}\}.$$

Hence $\varphi(2^n) = 2^{n-1}$. ◊

16.4.9 Theorem (Euler's Theorem) Let m be an integer satisfying $m \geq 2$. Let a be an integer such that $\gcd(a, m) = 1$. Then

$$a^{\varphi(m)} \equiv 1 \pmod{m}.$$

Proof. This has almost the same proof as Fermat's Little Theorem. I'll leave the necessary modifications as an exercise.

You're probably wondering if there is a formula for $\varphi(m)$, and in fact there is.

16.4.10 Proposition Write

$$m = p_1^{r_1} \cdots p_k^{r_k}$$

where $p_1, \dots, p_k$ are distinct primes and $r_1, \dots, r_k$ are positive integers. Then

$$\varphi(m) = (p_1^{r_1} - p_1^{r_1-1}) \cdots (p_k^{r_k} - p_k^{r_k-1}).$$

The proof of Proposition 16.4.10 is a little long and we'll not include it in these notes. I'll post a video of this but it's non-examinable.

16.4.11 Exercise Use Euler's Theorem to compute $2^{1000} \pmod{63}$.

16.4.12 Exercise It is known that $(\mathbb{Z}/m\mathbb{Z})^*$ is cyclic if $m = 2$, 4, p^a or $2p^a$ where p is an odd prime. For all other $m \geq 2$, the unit group $(\mathbb{Z}/m\mathbb{Z})^*$ is not cyclic. For more on this, do *Number Theory* in term 3. For now, check that $(\mathbb{Z}/7\mathbb{Z})^*$ is cyclic, but $(\mathbb{Z}/8\mathbb{Z})^*$ is not cyclic.

16.4.13 Exercise Use Theorem 12.3.1 to show that $\varphi(m)$ is even for $m \geq 3$.

Chapter 17 - Factorisation in rings

17.1 Factorisation in the integers

In *Foundations* you have proved the existence and uniqueness of the factorisation of an integer into products of prime numbers. This is Theorem 20.1 in the *Foundations* notes which says:

‘Every $n \in N$ can be written uniquely as a product of primes.’

Some of the key steps in being able to prove this were as follows

- ‘Division with remainder’ in the integers, i.e. for all $a \in \mathbb{Z}$ and $0 \neq b \in \mathbb{N}$ there exist unique $r, q \in \mathbb{Z}$ such that $a = bq + r$ where $0 \leq r < b$. This is Theorem 18.7 in the *Foundations* notes.
- ‘Euclid’s algorithm’ from which you can deduce Bezout’s Lemma. Bezout’s lemma implies that, given $a, b \in \mathbb{N}$ with $a > b > 0$ there exist $x, y \in \mathbb{Z}$ such that $\text{hcf}(a, b) = xa + yb$. These are Theorems 19.1 and 19.3 in the *Foundations* notes.
- The property of a prime number p that, for any $a, b \in \mathbb{Z}$, if $p|ab$ then $p|a$ or $p|b$. This is Theorem 19.5 in the *Foundations* notes.

We will show that there are analogues of all the concepts above in the polynomial ring $F[x]$ where F is a field and from these we will be able to deduce a form of unique factorisation in $F[x]$.

A field was defined in chapter 16 as follows. A *field* $(F, +, \cdot)$ is a commutative ring which is not the zero ring such that every non-zero element is a unit.

$\mathbb{Q}$, $\mathbb{R}$, $\mathbb{C}$ are fields. $\mathbb{Z}$ is not a field (for example 2 does not have a multiplicative inverse). Also we’ll need the fact that $\mathbb{Z}/p\mathbb{Z}$ is a field if p is prime (as proved in chapter 16).

17.2 Factorisation in the ring $F[x]$ where F is a field

Note that in the below we will be restricting to a polynomial ring of the form $F[x]$ where F is a field even though some of the ideas can be applied to more general polynomial rings (i.e. to some polynomial rings where the coefficients aren’t necessarily from a field). You will see more of this in future algebra modules which cover rings.

First we define the degree of a polynomial in $F[x]$ and look at some of its properties.

17.2.1 Definition Let F be a field. Let $f(x) \in F[x]$. Suppose

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

where $a_n \neq 0$. Then we define the degree of $f(x)$ to be n. We write $\deg(f(x)) = n$. If $f(x) = 0$, the zero polynomial, we define $\deg(f(x)) = -\infty$. $\Diamond$

This is just the regular notion of order which you will probably be used to. The order of a polynomial is given by the highest power of x that occurs in it. So the order of $x^2 + 3x + 2$ is 2, the order of $x^1 0 + 1$ is 10, the order of $3x + 2$ is 1 and the order of the constant polynomial 4 is 0. Be careful about the zero polynomial, in the above it is defined to have order $-\infty$ which distinguishes it from the non-zero constant polynomials which have degree 0.

17.2.2 Proposition Let F be a field. Then

1. if $f(x), g(x) \in F[x]$ then $\deg(f(x)g(x)) = \deg(f(x)) + \deg(g(x))$ (here we apply the conventions that $-\infty + n = -\infty = n + (-\infty)$ for any $n \in \mathbb{N}$ and that $-\infty + (-\infty) = -\infty$).
2. $f(x) \in F[x]$ is a unit in $F[x]$ if and only if $\deg(f(x)) = 0$.
3. if $f(x), g(x) \in F[x]$ are such that $f(x)g(x) = 0$ then either $f(x) = 0$ or $g(x) = 0$.
4. if $f(x), g(x) \in F[x]$ with $0 \leq \deg(g(x)) \leq \deg(f(x))$ then there exist $q(x), r(x) \in F[x]$ with $\deg(r(x)) < \deg(g(x))$ and $f(x) = g(x)q(x) + r(x)$.

Proof.

1. First suppose that $f(x) = 0$. Then $\deg(f(x)) = -\infty$. We then have $f(x)g(x) = 0$ and so $\deg(f(x)g(x)) = -\infty = \deg(f(x)) + \deg(g(x))$ by the conventions we are adopting about addition involving $-\infty$.

   The result is true similarly if $g(x) = 0$.

   So assume that $f(x) \neq 0$ and $g(x) \neq 0$. Then $m = \deg(f(x)) \geq 0$ and $n = \deg(g(x)) \geq 0$. Suppose that $f(x) = a_m x^m + a_{m-1}x^{m-1} + ... + a_1 x + a_0$ and $g(x) = b_n x^n + b_{n-1}x^{n-1} + \cdots + b_1 x + b_0$ where $a_i, b_j \in F$ with $a_m \neq 0$

and $b_n \neq 0$. We can see that $a_m b_n \neq 0$ by contradiction as follows.

Suppose $a_m b_n = 0$. Then, since $a_m \neq 0$ and F is a field there exists $a_m^{-1} \in F$ such that $a_m^{-1} a_m = 1$. Mutliplying both sides of $a_m b_n = 0$ by a_m^{-1} gives $b_n = 1 b_n = a_m^{-1} a_m b_n = a_m^{-1} 0 = 0$, contradicting $b_n \neq 0$.

We have

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

and, since $a_m b_n \neq 0$, $\deg(f(x)g(x)) = m + n = \deg(f(x)) + \deg(g(x))$.

2. Suppose $f(x) \in F[x]$ is a unit. Then there exists $g(x) \in F[x]$ such that $f(x)g(x) = 1$ so that $\deg(f(x)g(x)) = 0$. By 1. above this means that $\deg(f(x)) = 0$ (and $\deg(g(x)) = 0$).

   Conversely, suppose $\deg(f(x)) = 0$. Then $f(x) = a_0$ where $0 \neq a_0 \in F$. Since F is a field there exists $a_0^{-1} \in F$ such that $a_0^{-1} a_0 = 1$. This means that $f(x)$ is a unit with inverse $g(x) = a_0^{-1}$.

3. If both $f(x) \neq 0$ and $g(x) \neq 0$ then $\deg(f(x)) \geq 0$ and $\deg(g(x)) \geq 0$. By 1. above this would give $\deg(f(x)g(x)) \geq 0$ and, in particular $f(x)g(x) \neq 0$. Therefore if $f(x)g(x) = 0$ we must have either $f(x) = 0$ or $g(x) = 0$ (or both).

4. Notice that this is nothing more than the long division of polynomials you might have seen in school/college.

   The proof will be by induction on the degree of $f(x)$. If $\deg(f(x)) = 0$ then $f(x) = k_1$ for some $0 \neq k_1 \in F$. Since $0 \leq \deg g(x) \leq \deg(f(x)) = 0$, $g(x) = k_2$ for some $0 \neq k_2 \in F$. But then $f(x) = k_1 = k_1 k_2^{-1} k_2 + 0$ so the statement is true with $q(x) = k_1 k_2^{-1}$ and $r(x) = 0$.

   Now suppose that $f(x) = a_m x^m + a_{m-1} x^{m-1} + ... + a_1 x + a_0$ where $m \geq 1$ and $g(x) = b_n x^n + b_{n-1} x^{n-1} + \cdots + b_1 x + b_0$ where $a_i, b_j \in F$ with $a_m \neq 0$ and $b_n \neq 0$. We have $m = \deg(f(x)) \geq \deg(g(x)) = n$. Let

$$f_1(x) = f(x) - (a_m b_n^{-1}) x^{m-n} g(x).$$

   Observing that the coefficient of x^m in $f_1(x)$ is zero we see that $\deg(f_1) \leq m - 1$. By strong induction on the degree of f there exists $q_1(x), r(x) \in F[x]$ with $\deg(r(x)) < \deg(g(x))$ such that $f_1(x) = q_1(x)g(x) + r_1(x)$. This gives

$$q_1(x)g(x) + r(x) = f(x) - (a_m b_n^{-1}) x^{m-n} g(x).$$

This rearranges to $f(x) = ((a_m b_n^{-1})x^{m-n} + q_1(x))g(x) + r(x)$. If we put $q(x) = (a_m b_n^{-1})x^{m-n} + q_1(x)$ then $f(x) = g(x)q(x) + r(x)$ where $\deg(r(x)) < \deg(g(x))$, as required. $\Diamond$

The proposition above just states properties of polynomials which you are almost certainly already familiar with. For example, 1. in the above is just expressing what we see in the following example

$$(x^3 + 2x + 1)(2x^2 + x + 1) = 2x^5 + 5x^4 + 4x^4 + 4x^2 + 3x + 1$$

a degree 3 polynomial multiplied by a degree 2 polynomial gives a degree 5 polynomial.

17.2.3 Examples

1. In $\mathbb{Q}[x]$, $x + 1 = \dfrac{1}{2}(2x + 1) + \dfrac{1}{2}$. Notice that this cannot be done in $\mathbb{Z}[x]$. If $x + 1 = (2x + 1)q(x) + r(x)$ where $\deg(r(x)) < 1$ then $\deg(q(x)) = 0$ and so $q(x) = k$ for some integer k and $r(x)$ would have to be a constant polynomial (possibly 0). But then, by equating coefficients of x on both sides $2k = 1$ which is not possible for $k \in \mathbb{Z}$.

2. Let $R = \mathbb{Z}/4\mathbb{Z}$. Note that R is not a field because here $[2]_4 \times_4 [2]_4 = [0]_4$ and so $[2]_4$ can have no multiplicative inverse.

   Now consider $f(x) = x^2 + 1, g(x) = 2x + 1 \in R[x]$. Here you need to think of the coefficients 'modulo 4', we are supressing the usual $[]_4$ notation. There cannot exist $q(x), r(x) \in R[x]$ with $\deg(r(x)) < \deg(g(x))$ such that $f(x) = g(x)q(x) + r(x)$. $r(x)$ would have to be a constant polynomial (possibly 0) and $q(x)$ would have to have degree 1, i.e. $q(x) = ax + b$. But then, by comparing coefficients of x^2, $2a = 1$ and this cannot happen modulo 4.

17.2.4 Exercise

1. Let $f(x) = x^2 + 2x + 1$ and let $g(x) = x - 1$ both in $\mathbb{Q}[x]$. Find polynomials $q(x), r(x) \in \mathbb{Q}[x]$ with $\deg(r(x)) < \deg(g(x)) = 1$ such that $f(x) = g(x)q(x) + r(x)$.

2. Let $f(x) = 2x^2 + x + 1$ and let $g(x) = x + 1$ both in $R[x]$, where $R = \mathbb{Z}/3\mathbb{Z}$ (i.e. we are interpreting the coefficients modulo 3. Find polynomials $q(x), r(x) \in R[x]$ with $\deg(r(x)) < \deg(g(x)) = 1$ such that $f(x) = g(x)q(x) + r(x)$.

17.2.5 Corollary Let F be a field. Let $0 \neq f(x) \in F[x]$ and $\alpha \in F$. Then the evaluation of $f(x)$ at $x = \alpha$ is zero, i.e. $f(\alpha) = 0$, if and only if there exists a polynomial $q(x) \in F[x]$ such that $f(x) = (x - \alpha)q(x)$.

Proof. Suppose $f(\alpha) = 0$. By Propostion 17.2.2 there exists $q(x), r(x) \in F[x]$ with $\deg(r(x)) < \deg(x - \alpha) = 1$ such that $f(x) = (x - \alpha)q(x) + r(x)$. Notice that this means that $r(x) = \beta$ for some $\beta \in F$. But then $0 = f(\alpha) = (\alpha - \alpha)q(\alpha) + \beta$ giving $\beta = 0$.

Conversely if there exists $q(x) \in F[x]$ such that $f(x) = (x - \alpha)q(x)$ then $f(\alpha) = (\alpha - \alpha)g(\alpha) = 0$. $\diamondsuit$

17.2.6 Exercise The factor theorem tells us that if $f(x) \in \mathbb{C}[x]$ and $f(i) = 0$, where i is the usual square root of -1, then $f(x) = (x - i)g(x)$ for some $g(x) \in \mathbb{C}[x]$. Use the factor theorem to prove that if $f(x) \in \mathbb{R}[x]$ then $f(i) = 0$ if and only if $f(x) = (x^2 - 1)h(x)$ for some $h(x) \in \mathbb{R}[x]$.

17.2.7 Corollary Let F be a field. Let I be an ideal in $F[x]$. Then there exists $f(x) \in F[x]$ such that $I = \{f(x)g(x) \mid g(x) \in F[x]\}$.

Proof. If $I = 0$ then the result is clearly true with $f(x) = 0$.

If $I \neq 0$ let $0 \neq f(x) \in I$ have smallest possible degree among non-zero polynomials in I. Since I is an ideal we have $\{f(x)g(x) \mid g(x) \in F[x]\} \subseteq I$.

Now suppose $h(x) \in I$. By part 4 of Proposition 17.2.2 there exist $q(x), r(x) \in F[x]$ with $\deg(r(x)) < \deg(f(x))$ and $h(x) = f(x)q(x) + r(x)$. But then $r(x) = h(x) - f(x)q(x) \in I$ (since both $h(x) \in I$ and $-f(x)q(x) \in I$). Since $r(x)$ has degree less than $f(x)$ it must be the zero polynomial. Therefore $h(x) = f(x)q(x) \in \{f(x)g(x) \mid g \in F[x]\}$. This means that $I \subseteq \{f(x)g(x) \mid g(x) \in F[x]\}$ and since we have the relevant inclusions both ways round we can conclude that $\{f(x)g(x) \mid g(x) \in F[x]\} = I$. $\diamondsuit$

From now on if $f(x) \in F(x)$ we'll denote the subset of $F[x]$ given by

$$\{f(x)g(x) \mid g(x) \in F[x]\}$$

by $f(x)F[x]$. Note that such a subset is always an ideal of $F[x]$.

17.2.8 Examples

1. $I = \{f(x) \mid f(0) = 0\}$ is an ideal of $\mathbb{R}[x]$. It's actually those polynomials with constant term of 0, those which you can factor an x out of. This means that $I = \{xg(x) \mid g(x) \in \mathbb{R}[x]\} = x\mathbb{R}[x]$.

2. $I = \{f(x) \mid f(1) = 0\}$ is an ideal of $\mathbb{R}[x]$. By Corollary 17.2.5, it's actually those polynomials which have $x - 1$ as a factor. This means that

$$I = \{(x-1)g(x) \mid g(x) \in \mathbb{R}[x]\} = (x-1)\mathbb{R}[x].$$

17.2.9 Definition Let $0 \neq f(x) \in F[x]$. Then $f(x)$ is said to be irreducible over F (or, equivalently, irreducible in $F[x]$) if whenever$f(x) = g(x)h(x)$ with $g(x), h(x) \in F[x]$ then either $g(x)$ or $h(x)$ is a constant polynomial (but not both). $\diamondsuit$

Note that the above definition says that we can't have both $g(x)$ and $h(x)$ being constant polynomials which means that an irreducible polynomial $f(x)$ cannot be a constant polynomial and, since $f(x) \neq 0$ either, it has to have degree greater than or equal to 1.

Since the non-zero constant polynomials in $F[x]$ are precisely the units, being irreducible is the same as saying that if $0 \neq f(x) = g(x)h(x)$ with $g(x), h(x) \in F[x]$ then either $g(x)$ is a unit or $h(x)$ is a unit, but not both.

Probably the easiest way of all to think about this is that $0 \neq f(x) \in F[x]$ is irreducible if it has degree at least one and you cannot write it as the product of two polynomials in $F[x]$ both with smaller degree. Convince yourself that the definition above is equivalent to this.

Think of being an irreducible polynomial in $F[x]$ as being analagous to being a prime integer. An irreducible polynomial has degree at least 1 and cannot be 'broken down' into a product of 'smaller' polynomials (where we think of the degree as giving the size), just as a prime integer cannot be broken down into a product of smaller integers.

Irreducibility depends on the field, hence we have 'irreducible over F' rather than just 'irreducible'. Some of the examples below illustrate this.

17.2.10 Examples

1. For any field F, any polynomial with degree 1 is irreducible over F. To see this, suppose $a, b \in F$ with $a \neq 0$ and $ax + b = f(x)g(x)$. Then by

Proposition 17.2.2 either deg$f(x) = 0$ and $\deg(g(x)) = 1$ or $\deg(f(x)) = 1$ and $\deg(g(x)) = 0$. In the first case $f(x)$ is a unit and $g(x)$ is not a unit and in the second $g(x)$ is a unit and $f(x)$ is not a unit.

2. $f(x) = x^2 + 1$ is irreducible over $\mathbb{R}$. If not we would have
$$x^2 + 1 = (ax + b)(cx + d)$$
for some real numbers a, b, c, d with a and c both non-zero. But then $-\frac{b}{a}$ would be a real number satisfying $x^2+1 = 0$. However $f(x)$ is not irreducible over $\mathbb{C}$ because $f(x) = (x + i)(x - i)$.

3. $f(x) = x^2 - 2$ is irreducible over $\mathbb{Q}$. If not we would have
$$x^2 - 2 = (ax + b)(cx + d)$$
for some rational numbers a, b, c, d with a and c both non-zero. But then $-\frac{b}{a}$ would be a rational number satisfying $x^2-2 = 0$, i.e. a rational square root of 2. However $f(x)$ is not irreducible over $\mathbb{R}$ because $f(x) = (x+\sqrt{2})(x-\sqrt{2})$.

4. Let $a, b, c \in \mathbb{R}$. Then $f(x) = ax^2 + bx + c$ is irreducible over $\mathbb{R}$ if $b^2 < 4ac$. If not we would have $ax^2 + bx + c = (rx + s)(ux + v)$ where r, s, u, v are real numbers with r, u both non-zero. But then $-\frac{s}{r}$ would be a real number satisfying $ax^2 + bx + c = 0$.

We'll now cover some definitions and results that will set us up to prove the existence and uniqueness of factorisation of polynomials in $F[x]$ where F is a field.

17.2.11 Definition Let F be a field. Let $f(x), g(x) \in F[x]$. We say that $f(x)$ divides $g(x)$ and write $f(x)|g(x)$ if there exists $h(x) \in F[x]$ such that $f(x)h(x) = g(x)$. ♢

17.2.12 Definition Let F be a field. Let $f(x), g(x) \in F[x]$ be non-zero polynomials. We say that $f(x)$ and $g(x)$ are relatively prime if whenever $0 \neq h(x) \in F[x]$ with $h(x)|f(x)$ and $h(x)|g(x)$ then $\deg(h(x)) = 0$, i.e. $h(x)$ is a unit. ♢

Any unit divides any polynomial and so the above is essentially requiring that $f(x)$ and $g(x)$ have no divisors in common other then the ones that will always exist, in particular they have no common divisors which are polynomials with degree greater than 0.

The next theorem is the equivalent of Bezout's lemma for integers but for polynomials in $F[x]$ where F is a field.

17.2.13 Theorem Let F be a field. Let $f(x), g(x) \in F[x]$ be non-zero polynomails with $f(x)$ and $g(x)$ relatively prime. Then there exist $h^*(x), k^*(x) \in F[x]$ such that

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

Proof. Let $I = \{f(x)h(x) + g(x)k(x) \mid h(x), k(x) \in F[x]\}$. Then I is an ideal of $F[x]$ as follows. First we'll show that I is a subgroup of $(F[x], +)$. We'll use the usual (a),(b),(c) subgroup test below.

(a) The zero polynomial is in I because $0 = f(x) \times 0 + g(x) \times 0$.

(b) If polynomials $j_1(x), j_2(x) \in I$ then $j_1(x) = f(x)h_1(x) + g(x)k_1(x)$ and $j_2(x) = f(x)h_2(x) + g(x)k_2(x)$. Then

$$j_1(x) + j_2(x) = f(x)[h_1(x) + h_2(x)] + g(x)[k_1(x) + k_2(x)]$$

and since $h_1(x) + h_2(x) \in F[x]$ and $k_1(x) + k_2(x) \in F[x]$ then means that $j_1(x) + j_2(x) \in I$.

(c) If the polynomial $j(x) \in F[x]$ then $j(x) = f(x)h(x) + g(x)k(x)$ for some polynomials $h(x), k(x) \in F[x]$. Then $-j(x) = f(x)(-h(x)) + g(x)(-k(x))$ and, because $-h(x), -k(x) \in F[x]$ this means that $-j(x) \in I$.

Now we need to show that if $j(x) \in I$ and $r(x) \in F[x]$ then $j(x)r(x) \in I$ (there's no need to show that $r(x)j(x) \in I$ separately because $F[x]$ is a commutative ring).

Since $j(x) \in I$, $j(x) = f(x)h(x) + g(x)k(x)$ for some polynomials $h(x), k(x) \in F[x]$. But then

$$j(x)r(x) = f(x)h(x)r(x) + g(x)k(x)r(x)$$

and since both $h(x)r(x) \in F[x]$ and $k(x)r(x) \in F[x]$ this means the $j(x)r(x) \in I$.

This completes the proof that I is an ideal of $F[x]$ and so by Corallary 17.2.7 there exists $j(x) \in F[x]$ such that $I = j(x)F[x]$.

Since $f(x), g(x) \in I = j(x)F[x]$ there are polynomials $k_1(x), h_1(x) \in F[x]$ such that $f(x) = j(x)k_1(x)$ and $g(x) = j(x)h_1(x)$. This means that $j(x)|f(x)$ and $j(x)|g(x)$ and, since $f(x)$ and $g(x)$ are relatively prime, it follows that $j(x)$ is a unit and that $I = F[x]$ by Corollary 15.2.9.

Since $1 \in F[x] = I$, this means that there exist $h(x), k(x) \in F[x]$ such that $f(x)h(x) + g(x)k(x) = 1$. $\Diamond$

If p is a prime integer which divides the product ab of integers a and b then either p divides a or p divides b. Next we'll prove the equivalent statement for irreducible polynomials.

17.2.14 Theorem Let F be a field. Let $f(x) \in F[x]$ be an irreducible over F. Suppose that $g(x), h(x) \in F[x]$ and $f(x)|g(x)h(x)$. Then $f(x)|g(x)$ or $f(x)|h(x)$.

Proof. We will show that if $f(x)$ does not divide $g(x)$ then it must divide $h(x)$ which will give the conclusion we need.

Suppose $f(x)$ does not divide $g(x)$. Then $f(x)$ and $g(x)$ are relatively prime as follows.

Suppose $j(x)|f(x)$ and $j(x)|g(x)$. Then $f(x) = j(x)k_1(x)$ and $g(x) = j(x)l_1(x)$ for some $k_1(x), l_1(x) \in F[x]$. Since $f(x)$ is irreducible one of $j(x)$ or $k_1(x)$ is a non-zero constant polynomial. If the latter, $k_1(x) = \alpha \neq 0$ where $\alpha \in F$ then $j(x) = \alpha^{-1}f(x)$. But then $g(x) = j(x)l_1(x) = \alpha^{-1}f(x)l_1(x)$ and $f(x)$ divides $g(x)$, a contradiction. So $j(x)$ is a non-zero constant polynomial, a unit. This means that $f(x)$ and $g(x)$ are relatively prime.

By Theorem 17.2.13 there exist polynomials $k(x), m(x) \in F[x]$ such that $f(x)k(x) + g(x)m(x) = 1$. Then

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

Since $f(x)$ divides $g(x)h(x)$, $f(x)$ divides the right hand side of the above, so $f(x)$ divides $h(x)$. $\Diamond$

17.2.15 Corollary Let F be a field. Let $f(x) \in F[x]$ be an irreducible polynomial. Suppose $g_1(x), g_2(x), \ldots, g_n(x) \in F[x]$ and $f(x)|g_1(x)g_2(x)\ldots g_n(x)$. Then there exists i with $1 \leq i \leq n$ such that $f(x)|g_i(x)$.

Proof. By induction on n using Theorem 17.2.14. $\Diamond$

17.2.16 Theorem Let F be a field. Let $f(x) \in F[x]$ with $\deg(f(x)) \geq 1$. Then $f(x)$ is expressible as a product of polynomials which are irreducible over F. This expression is unique in the following sense. Suppose

$$f(x) = g_1(x)g_2(x)\ldots g_m(x) = h_1(x)h_2(x)\ldots h_n(x)$$

where $g_i(x), h_j(x)$ are all irreducible over F then $m = n$ and, after possibly renumbering the $h_j(x)$ polynomials, $g_1(x) = a_1 h_1(x)$, $g_2(x) = a_2 h_2(x)$,... $g_n(x) = a_n h_n(x)$, where $0 \neq a_i \in F$, i.e. each a_i is a unit in $F[x]$.

We say the factorisation is 'unique up to multiplication by units'.

Proof.

1. Existence of factorisation. We proceed by induction on $\deg(f(x))$.

   If $\deg(f(x)) = 1$ then $f(x)$ itself is irreducible and so it is its own factorisation! So the result is true if $\deg(f(x)) = 1$.

   Now assume that we have polynomial $f(x) \in F[x]$ with $\deg(f(x)) > 1$ and, for an induction argument, that any polynomial with degree greater than or equal to 1 and with smaller degree than that of $f(x)$ has a factorisation as a product of polynomials which are irreducible in $F[x]$.

   If $f(x)$ is irreducible we have our factorisation and can stop. If $f(x)$ is not irreducible then there exist $g(x), h(x) \in F[x]$ with $\deg(g(x)) < \deg(f(x))$ and $\deg(h(x)) < \deg(f(x))$. By induction we can express both $g(x)$ and $h(x)$ as products of irreducible polynomials in $F[x]$ and therefore the same is true for $f(x) = g(x)h(x)$, by putting those two products together.

2. Uniqueness of factorisation. We now proceed by induction on the number of irreducible factors, m, on the left hand side in the statement of the theorem (i.e. the number of polynomials $g_i(x)$).

   Suppose there is only one polynomial on the left hand side, i.e. $f(x) = g_1(x) = h_1(x)h_2(x)\dots h_n(x)$ where $g_1(x), h_i(x)$ are all irreducible over F. But then, since $g_1(x)$ is irreducible, there can only be one polynomial on the right hand side, i.e. $n = 1$ and $f(x) = g_1(x) = h_1(x)$ and we are done in the case $m = 1$.

   Suppose $f(x) = g_1(x)g_2(x)\dots g_m(x) = h_1(x)h_2(x)\dots h_n(x)$ where $g_i(x), h_j(x)$ are all irreducible and, again for an induction argument, assume the result is true for any situation where there are fewer than m polynomials on the left hand side.

We have $g_1(x)|h_1(x)h_2(x)\dots h_n(x)$ so by Corallary 17.2.15 $g_1(x)$ divides $h_i(x)$ for some i. Renumbering if necessary, we may assume without loss of generality, that $i = 1$, in other words that $g_1(x)|h_1(x)$. So $h_1(x) = k(x)g_1(x)$ for some $k(x) \in F[x]$. Since $h_1(x)$ is irreducible either $\deg(k(x)) = 0$ or $\deg(g_1(x)) = 0$ but not both. But $g_1(x)$ is irreducible and so $\deg(g(x)) > 0$ which means that $\deg(k(x)) = 0$, i.e. $k(x)$ is a non-zero constant polynomial, a unit in $F[x]$. Let $k(x) = a \in F$.

We then have $f(x) = g_1(x)g_2(x)\dots g_m(x) = ag_1(x)h_2(x)\dots h_n(x)$. This gives

$$g_1(x)[g_2(x)\dots g_m(x) - ah_2(x)\dots h_n(x)] = 0.$$

Since $g_1(x) \neq 0$, by Proposition 17.2.2, $g_2(x)\dots g_m(x) - ah_2(x)\dots h_n(x) = 0$ and $g_2(x)\dots g_m(x) = ah_2(x)\dots h_n(x)$. By induction we must have that $n-1 = m-1$ (which implies than $n = m$ and that, after possibly renumbering the $h_j(x)$ polynomials, $g_2(x) = a_2h_2(x)$, $g_3(x) = a_3h_3(x)$,... $g_n(x) = a_nh_n(x)$, where $0 \neq a_i \in F$. This, together with the already established $g_1(x) = a^{-1}h_1(x)$, gives the result. ◊

17.3 Finding the irreducible polynomials in $F[x]$

We have seen that whether a polynomial is irreducible certainly depends on the field F. For example $x^2 + 1$ is irreducible over $\mathbb{R}$ but not irreducible over $\mathbb{C}$. Look back at Example 17.2.10 for the details.

Here we'll briefly discuss which polynomials are irreducible over the fields $\mathbb{C}$, $\mathbb{R}$ and $\mathbb{Q}$.

The fundamental theorem of algebra, stated without proof below, will help us with this.

17.3.1 Theorem (Fundamental Theorem of Algebra) Let $f(x) \in \mathbb{C}[x]$ be a polynomial with $\deg(f(x)) \geq 1$. Then there exists $\alpha \in \mathbb{C}$ such that $f(\alpha) = 0$. In other words, the equation $f(x) = 0$ has a root in the complex numbers. ◊

Any proof of this will involve concepts from analysis. This is because the real numbers, from which the complex numbers are constructed, are an object defined in terms of concepts from analysis (the completeness axiom which is about the existence of supremums or about certain sequences converging).

17.3.2 Examples We can use the Fundamental Theorem of Algebra to work out precisely what the irreducible polynomials over $\mathbb{C}$ and $\mathbb{R}$ are.

1. In $\mathbb{C}[x]$ the irreducible polynomials are precisely those with degree 1. You can see this as follows.

   We know that degree 1 polynomials in $F[x]$ are irreducible over F for any field, so certainly degree 1 polynomials in $\mathbb{C}[x]$ are irreducible over $\mathbb{C}$. Suppose $f(x) \in \mathbb{C}[x]$ is irreducible and $\deg(f(x)) > 1$. By the Fundamental Theorem of Algebra there exists $\alpha \in C$ such that $f(\alpha) = 0$. Then, by Corollary 17.2.5, $f(x)$ factorises as $(x - \alpha)g(x)$ for some $g(x) \in \mathbb{C}[x]$. But this can't happen because $\deg(g(x))$ would have to then be greater than 0 and $f(x)$ is irreducible over $\mathbb{C}$.

2. In $\mathbb{R}[x]$ the irreducible polynomials are precisely those which either have degree 1 or those with degree 2 of the form $ax^2 + bx + c$ where $a, b, c \in \mathbb{R}$ with $b^2 - 4ac < 0$. Here is a justification of this.

   It's clear that the polynomials in $\mathbb{R}[x]$ which either have degree 1 or those with degree 2 of the form $ax^2 + bx + c$ where $a, b, c \in R$ with $b^2 - 4ac < 0$ are irreducible over $\mathbb{R}$. It's also clear that polynomials $ax^2 + bx + c$ where $a, b, c \in \mathbb{R}$ with $b^2 - 4ac \geq 0$ are not irreducible over $\mathbb{R}$ because they have real roots and will factorise into a product of degree 1 polynomials.

   Suppose that $f(x) \in \mathbb{R}[x]$ has degree greater than 2. Then there exists $\alpha \in \mathbb{C}$ such that $f(\alpha) = 0$. But then we also have $f(\overline{\alpha}) = 0$ where $\overline{\alpha}$ is the complex conjugate of α. You may remember this result from school/college maths, it's dependent on the coefficients of $f(x)$ being real numbers. By repeated application of the factor theorem this means that $f(x) = (x - \alpha)(x - \overline{\alpha})g(x)$ for some $g(x) \in \mathbb{C}[x]$ with degree greater than 0.

   But $(x-\alpha)(x-\overline{\alpha}) = x^2-(\alpha+\overline{\alpha})x+\alpha\overline{\alpha} \in \mathbb{R}[x]$ because $\alpha+\overline{\alpha} \in \mathbb{R}$ and $\alpha\overline{\alpha} \in \mathbb{R}$.

   By 'division with remainder' applied in $\mathbb{R}[x]$ there exists $q(x), r(x) \in \mathbb{R}[x]$ with $\deg(r(x)) < 2$ such that

   $$f(x) = (x - \alpha)(x - \overline{\alpha})q(x) + r(x).$$

   This means that

   $$(x - \alpha)(x - \overline{\alpha})(g(x) - q(x)) = r(x).$$

If $g(x) - q(x) \neq 0$ then the degree of left hand side of the above is at least 2 whereas the degree of the right hand side $r(x)$ is less than 2 which would be a contradiction. It follows that $g(x) - q(x) = 0$ and $g(x) = q(x) \in \mathbb{R}[x]$ and so $f(x)$ is not irreducible over $\mathbb{R}$. $\Diamond$

For polynomials in $\mathbb{Q}[x]$ the question of deciding whether a given polynomial is irreducible can be very difficult. Few criteria exist for this but here is one of them, which we state without proof.

17.3.3 Theorem (Eisenstein's criterion for irreducibility) Let

$$f(x) = a_n x^n + a_{n-1}x^{n-1} + \cdots + a_1 x + a_0 \in \mathbb{Z}[x].$$

Suppose there exists a prime number p such that

- $p|a_i$ for each i with $0 \leq i < n$
- p does not divide a_n
- p^2 does not divide a_0.

Then $f(x)$ is irreducible over $\mathbb{Q}$. $\Diamond$

We use this in the following exercise, whicha also shows that there are irreducible polynomials over $\mathbb{Q}$ with arbitrarily large degree (although clearly this fact could be established in a more direct way by construction using the theorem above, e.g. $x^n + 5x^{n-1} + 5x^{n-2} + ... + 5x + 5$ is irreducible over $\mathbb{Q}$ for any positive integer n by the above with $p = 5$).

17.3.4 Exercise Prove that $f(x) = x^{p-1} + x^{p-2} + \cdots + x + 1$ is irreducible over $\mathbb{Q}$ for any prime p by the following steps.

1. Show that if p is prime and $0 < k < p$ then $\binom{n}{k} = \dfrac{p!}{k!(p-k)!}$ is a multiple of p.
2. Prove that $f(x) \in \mathbb{Q}[x]$ is irreducible over $\mathbb{Q}$ if and only if $f(x+1) \in \mathbb{Q}[x]$ (i.e. the polynomial obtained from $f(x)$ by replacing x by $x+1$ and then multiplying out) is irreducible over $\mathbb{Q}$.
3. Show that, since, $f(x) = \dfrac{x^p - 1}{x - 1}$,

$$f(x+1) = \frac{(x+1)^p - 1}{x} = x^{p-1} + \binom{p}{1}x^{p-2} + \binom{p}{2}x^{p-3} + \cdots + \binom{p}{p-1}.$$

4. Use Eisenstein's critierion to deduce that $f(x+1)$ is irreducible over $\mathbb{Q}$ and so $f(x)$ is irreducible over $\mathbb{Q}$.

Chapter 18 - Cosets and quotient rings

There are many exciting things about cosets! You can get cosets in groups and you can also get cosets in rings. You use cosets in groups to prove Lagrange's theorem. However we will only look at cosets in rings. We'll see that we can make a new ring from an existing ring where the elements are cosets. Understanding this (potentially simpler) new ring can then help us to understand the original ring. But let's not get ahead of ourselves, here's the definition of a coset with respect to an ideal in a ring.

18.1 What is a coset of an ideal in a ring?

18.1.1 Definition Let R be a ring and I an ideal of R. Let r be an element of R. We call the set

$$r + I = \{r + x \mid x \in I\}$$

a *coset of* I in R.

18.1.2 Example The set of even integers $2\mathbb{Z}$ is an ideal of $\mathbb{Z}$. What are the cosets of $2\mathbb{Z}$ in $\mathbb{Z}$? Let's compute a few:

$$\begin{aligned}
0 + 2\mathbb{Z} &= \{\dots, 0 + (-4), 0 + (-2), 0 + 0, 0 + 2, 0 + 4, \dots\} = \{\dots, -4, -2, 0, 2, 4, \dots\};\\
1 + 2\mathbb{Z} &= \{\dots, 1 + (-4), 1 + (-2), 1 + 0, 1 + 2, 1 + 4, \dots\} = \{\dots, -3, -1, 1, 3, 5, \dots\};\\
2 + 2\mathbb{Z} &= \{\dots, 2 + (-4), 2 + (-2), 2 + 0, 2 + 2, 2 + 4, \dots\} = \{\dots, -4, -2, 0, 2, 4, \dots\};\\
3 + 2\mathbb{Z} &= \{\dots, 3 + (-4), 3 + (-2), 3 + 0, 3 + 2, 3 + 4, \dots\} = \{\dots, -3, -1, 1, 3, 5, \dots\}.
\end{aligned}$$

You'll quickly discover that

$$\cdots = -4 + 2\mathbb{Z} = -2 + 2\mathbb{Z} = 2\mathbb{Z} = 2 + 2\mathbb{Z} = 4 + 2\mathbb{Z} = \dots$$

and

$$\cdots = -3 + 2\mathbb{Z} = -1 + 2\mathbb{Z} = 1 + 2\mathbb{Z} = 3 + 2\mathbb{Z} = \dots.$$

So the ideal $2\mathbb{Z}$ has two cosets in $\mathbb{Z}$, which happen to be $2\mathbb{Z}$ itself, and $1 + 2\mathbb{Z}$ which is the set of odd integers. $\Diamond$

18.1.3 Exercise You know that $\mathbb{Z}^2$ is a ring. Let

$$2\mathbb{Z}^2 = \{(2a, 2b) \mid a, b \in \mathbb{Z}\}.$$

In otherwords, $2\mathbb{Z}^2$ is the set of vectors in $\mathbb{Z}^2$ with both coordinates even. Check that $2\mathbb{Z}^2$ is an ideal of $\mathbb{Z}^2$, having four cosets. What are they?

18.1.4 Example Now let's look at the index of the trivial ideal $\{0\}$ of the ring $\mathbb{Z}$. Note that

$$a + \{0\} = \{a\}.$$

So the cosets of $\{0\}$ in $\mathbb{Z}$ are

$$\ldots, \{-2\},\ \{-1\},\ \{0\},\ \{1\},\ \{2\}, \ldots$$

$\diamondsuit$

Here comes the key result in this section.

18.1.5 Lemma Let R be a ring and let I be an ideal. Let x, $y \in R$ so that $x + I$ and $y + I$ are cosets. Then

$$x + I = y + I \text{ if and only if } x - y \in I.$$

Proof. Note that this is and if and only if statement and therefore requires proof in both directions.

Starting with 'left implies right', suppose $x + I = y + I$. Then, since $x = x + 0$ and $0 \in I$, $x \in x + I = y + I = \{y + h \mid h \in I\}$. Therefore there exist $h \in I$ such that $x = y + h$. But then $x - y = h$. This means that $x - y \in I$.

Now for 'right implies left'. Now suppose that $x - y \in I$. Let $h^* = x - y$. Let $y + h$ where $h \in I$ be an arbitrary element of $y + I$. Then $y + h = x - h^* + h = x + (h - h^*)$ and so $y + h \in x + I$. This gives that $y + I \subseteq x + I$.

Continuing, let $x + h$ where $h \in I$ be an arbtrary element of $x + I$. Then $x + h = (y + h^*) + h = y + (h^* + h)$. Since $h^* + h \in I$ this means that $x + h \in y + I$ and that $x + I \subseteq y + I$.

The two inclusions $y + I \subseteq x + I$ and $x + I \subseteq y + I$ imply that $x + I = y + I$.$\diamondsuit$

One way to think about the cosets in R with respect to I is that they take an element of a ring, x, and they bring into one coset, $x + I$, all elements that differ from that element by an element of I. The next two examples illustrate this.

18.1.6 Example Let $X = \{(a, 0) \mid a \in \mathbb{R}\}$. Then X is an ideal of the ring $\mathbb{R}^2$, it is the x-axis.

Given $(a_1, b_1) \in \mathbb{R}^2$, the coset $(a_1, b_1) + X = \{(a_1, b_1) + (a, 0) \mid a \in \mathbb{R}\} = \{(a_1 + a, b_1) \mid a \in \mathbb{R}\}$. Because, by choosing a appropriately, $a_1 + a$ can be made to equal any real number, this is just all the points with y-coordinate b_1.

Furthermore, given (a_1, b_1) and (a_2, b_2) in $\mathbb{R}^2$, what does it mean for the coset $(a_1, b_1) + X$ to be the same as the coset $(a_2, b_2) + X$?

By Lemma , $(a_1, b_1) + X = (a_2, b_2) + X$ if and only if $(a_1 - a_2, b_1 - b_2)$ belongs to X. This happens if and only if $b_1 - b_2 = 0$. So the two cosets $(a_1, b_1) + X$ and $(a_2, b_2) + X$ are equal if and only (a_1, b_1) and (a_2, b_2) have have the same y-coordinate.

$\Diamond$

18.1.7 Example Let $R = \mathbb{R}[x]$. Let $I = \{f(x) \in \mathbb{R}[x] \mid f(0) = 0\}$. This is the set of all polynomials with real coefficients and with no constant term (or whose constant term is 0). I is an ideal of $\mathbb{R}[x]$.

Given $g(x) \in R[x]$ the coset $g(x) + I = \{g(x) + f(x) \mid f(0) = 0\}$. Because the constant term of $f(x)$ is zero this will be all the polynomials which have the same constant term as $g(x)$.

Now suppose $g(x)$, $h(x) \in \mathbb{R}[x]$. What does it mean for the cosets $g(x) + I$ and $h(x) + I$ to be equal?

By Lemma 18.1.6, $g(x) + I = h(x) + I$ if and only if $g(x) - h(x) \in I$. Suppose[11]

$$g = a_0 + a_1 x + \cdots + a_n x^n, \qquad h = b_0 + b_1 x + \cdots + b_n x^n,$$

where $a_0, \ldots, a_n$ and $b_0, \ldots, b_n$ are real numbers. Then $g(x) - h(x) \in I$ if and only if $a_0 - b_0 = 0$ if and only if $a_0 = b_0$ (i.e. g and h have the same constant term). Therefore $g(x) + I$ and $h(x) + I$ are the same cosets if and only the constant term of $g(x)$ equals the constant term of $h(x)$. $\Diamond$

18.2 Quotient rings

18.2.1 Definition Let R be a ring and I an ideal of R. We define the quotient ring R/I to be the set of cosets

$$R/I = \{x + I \mid x \in R\}$$

[11] It seems that we're writing $f(x)$ and $g(x)$ both as polynomials of the same degree n; this looks wrong as there is no reason to suppose that g and h have the same degree. But looks can be misleading. Here we're in fact writing $g(x)$ and $h(x)$ as polynomials of degree *at most* n. For example, if $g = 2 + 7x$ and $h = 4 - 3x + 2x^3$ then we can take $n = 3$ and let $a_0 = 2$, $a_1 = 7$, $a_2 = a_3 = 0$, and $b_0 = 4$, $b_1 = -3$, $b_2 = 0$, $b_3 = 2$

with addition and multiplication defined by

$$(x+I)+(y+I)=(x+y)+I. \tag{12}$$

$$(x+I)(y+I)=(xy)+I. \tag{13}$$

◊

Note carefully that the elements in R/I are themselves cosets and so the addition and multiplication defined tell us how to add and multiply two cosets to get another coset! This is why quotient rings can take a bit of getting used to.

We will now go on to prove that R/I is a ring. There is a more serious point which is that we need to show that the operations (12) and (13) are *well-defined.*

What does this mean? We know that cosets can have more than one 'name'. For example in $\mathbb{Z}/5\mathbb{Z}$ we have $1+5\mathbb{Z}=6+5\mathbb{Z}$ so this coset goes both by the name '$1+5\mathbb{Z}$' and by the name '$6+5\mathbb{Z}$'. The definition above uses this name. So we'd better make sure that the definition is independent of this choice of name. Precisely, we need to check that this is true:

If

$$a+I=a'+I \text{ and } b+I=b'+I,$$

then

$$(a+b)+I=(a'+b')+I \text{ and } (ab)+I=(a'b')+I.$$

The following theorem checks this.

18.2.2 Theorem - the addition and multiplication in $(R/I,+,.)$ are well defined. Let $(R,+,.)$ be an ring and I an ideal of R. Let a, a', b, b' be elements of R such that in R/I we have

$$a+I=a'+I \text{ and } b+I=b'+I,$$

then

$$(a+b)+I=(a'+b')+I \text{ and } (ab)+I=(a'b')+I$$

Proof. Since $a+I=a'+I$ and $b+I=b'+I$ in R/I by Lemma 18.1.6 $a-a'=h_1$ and $b-b'=h_2$ where h_1, $h_2 \in I$. Thus (and note that the commutativity of the addtion is used in the step below):

$$(a+b)-(a'+b')=(a-a')+(b-b')=h_1+h_2.$$

As $(I,+)$ is a subgroup of $(R,+)$ containing h_1 and h_2, we know that the sum h_1+h_2 belongs to H. Thus, by Lemma 18.1.6 again, the cosets $(a+b)+I$ and $(a'+b')+I$ are equal.

Now for the multiplication. We have

$$ab-a'b'=ab+ab'-ab'-a'b'=a(b-b')-b'(a'-b').$$

With $a-a'=h_1 \in I$ and $b-b'=h_2 \in I$ as before this can be written as

$$ab-a'b'=ah_2+b'h_1.$$

But, since I is an ideal and $h_1, h_2, \in I$ we have $ah_2+b'h_1 \in I$ or $ab-a'b' \in I$. Thus, by Lemma 18.1.6 once again, the cosets $(ab)+I$ and $(a'b')+I$ are equal. $\Diamond$

We need to check one more thing: that R/I is indeed a ring!

18.2.3 Theorem Let $(R,+,.)$ be ring and I an ideal of R. Then $(R/I,+,.)$ is a ring.

Proof. We have to check the defining properties for a ring.

The addition in $(R/I,+,.)$ is closed because if $a+I$ and $b+I$ are two cosets then their sum $(a+I)+(b+I)=(a+b)+I$ is another coset.

The additive identity element in $(R/I,+)$ is the coset $0+I$ (note that as a set this coset is equal to I). This is because, if $a \in R$,

$$(a+I)+(0+I)=(a+0)+I=a+I=(0+a)+I=(0+I)+(a+I).$$

Given $a \in R$, the additive inverse of the coset $a+I$ is the coset $(-a)+I$. This is because

$$(a+I)+((-a)+I)=(a+(-a))+I=0+I=((-a)+a)+I=((-a)+I)+(a+I)$$

and $0+I$ is the identity element in $(R/I,+)$.

The addition is associative because if $a, b, c \in R$ then

$$[(a+I)+(b+I)]+(c+I) = ((a+b)+I)+(c+I) = ((a+b)+c)+I$$

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

The addition is abelian because if a, $b \in R$. Then

$$\begin{aligned}(b+I)+(a+I) &= (b+a)+I && \text{from the definition of addition in } R/I \\ &= (a+b)+I && b+a = a+b \text{ as } (R,+) \text{ is abelian} \\ &= (a+I)+(b+I) && \text{from the definition of addition in } R/I.\end{aligned} \tag{14}$$

Now for the multiplication related properties.

The multiplication in R/I is closed because if $a+I$ and $b+I$ are two cosets then their product $(a+I)(b+I) = (ab)+I$ is another coset.

The multiplicative identity element in R/I is the coset $1+I$. This is because, if $a \in R$,

$$(a+I)(1+I) = (a.1)+I = a+I = (1.a)+I = (1+I)(a+I).$$

The addition is associative because if $a, b, c \in R$ then

$$[(a+I)(b+I)](c+I) = ((ab)+I)(c+I) = ((ab)c)+I$$

$$= (a(bc))+I = (a+I)((bc)+I) = (a+I)[(b+I)(c+I)].$$

The distributive properties hold because if if $a, b, c \in R$ then

$$[(a+I)+(b+I)](c+I) = ((a+b)+I)(c+I) = ((a+b)c)+I$$

$$= (ac+bc)+I = ((ac)+I)+((bc)+I) = (a+I)(c+I)+(b+I)(c+I).$$

Similarly $(a+I)[(b+I)+(c+I)] = (a+I)(b+I)+(a+I)(c+I)$. ◊

18.2.4 Example (which turns out to be very familiar!) Let $m \geq 2$ be an integer. We know that $m\mathbb{Z}$ is an idea of $\mathbb{Z}$ (consisting of the multiples of m). Lets think about the quotient ring $(\mathbb{Z}/m\mathbb{Z}, +, x)$. We're about to find out that we've seen this before.

Recalling *Foundations* again, you defined the congruence class modulo m of an integer a. This was $[a]_m = \{b \in \mathbb{Z} \mid b \equiv a \pmod{m}\}$. It turns out that the congruence class $[a]_m$ is equal to the coset $a + m\mathbb{Z}$ as follows:

Let $c \in [a]_m$. Then $c \equiv a \pmod{m}$ which means that $c - a$ is a multiple of m. So $c - a = mk$ for some integer k. Then $c = a + mk$ and $c \in (a + m\mathbb{Z})$. Therefore $[a]_m \subseteq a + m\mathbb{Z}$.

Conversely, if $d \in (a+m\mathbb{Z})$ then $d = a+ml$ for some $l \in \mathbb{Z}$ and $d-a$ is a multiple of m. This means that $d \equiv a \pmod{m}$ and $d \in [a]_m$. Therefore $a+m\mathbb{Z} \subseteq [a]_m$.

Since both $[a]_m \subseteq a + m\mathbb{Z}$ and $a + m\mathbb{Z} \subseteq [a]_m$ we have that $a + m\mathbb{Z} = [a]_m$.

In *Foundations* you we saw an addition and multiplation on the congruence classes as

$$[a]_m +_m [b]_m = [a + b]_m \text{ and } [a]_m \times_m [b]_m = [ab]_m$$

Notice that we could now write these in coset notation as

$$(a + m\mathbb{Z}) + (b + m\mathbb{Z}) = (a + b) + m\mathbb{Z} \text{ and } (a + m\mathbb{Z})(b + m\mathbb{Z}) = (ab) + m\mathbb{Z}$$

and that this exactly how addition and multiplication in the quotient ring $(\mathbb{Z}/m\mathbb{Z}, +, .)$ was defined in Definition 18.2.1.

So the congruence classes modulo n under addition and mutiplication is *exactly the same ring* as the quotient ring $(\mathbb{Z}/m\mathbb{Z}, +)$! $\Diamond$

Importantly, this means that taking the quotient ring $(R/I, +, .)$ in the case when $R = \mathbb{Z}$ and $I = n\mathbb{Z}$ corresponds exactly to ring of conjugacy classes modulo n under addition. Thus, taking a quotient ring can be seen as a generalisation of defining an addition on conjugacy classes modulo n in the sense that they coincide in this particular instance but that taking quotient rings can be applied more widely, as we'll see in the next example.

18.2.5 Example Let $f(x)$ be a polynomial in $R = F[x]$ where F is a field. Let $I = f(x)R$. Let's think about the quotient ring R/I.

We have $I = f(x)R = \{f(x)h(x) \mid h(x) \in F[x]\}$ which is precisely those polynomials in $F[x]$ which are multiples of $f(x)$.

Cosets take the form $g(x) + I$ where $g(x) \in F[x]$. If $g_1(x), g_2(x) \in F[x]$ then the cosets $g_1(x)+I = g_2(x)+I$ if and only if $g_1(x) - g_2(x) \in I$, i.e. if $g_1(x) - g_2(x)$ is a multiple of $f(x)$.

Now suppose $g(x) + I$ is a coset where $\deg(g(x)) \geq \deg(f(x))$. By division with remainder there exists $q(x), r(x) \in F[x]$ with $\deg(r(x)) < \deg(f(x))$ such that $g(x) = f(x)q(x) + r(x)$. But then $g(x) - r(x) \in I$ and $g(x) + I = r(x) + I$. In fact $r(x)$ is unique (but we didn't prove this). This means there is a unique 'canonical' way of writing $g(x) + I$ as $r(x) + I$ where $\deg(r(x)) < \deg(f(x))$.

This is entirely analogous to there being a canonical way to write a coset in $\mathbb{Z}/n\mathbb{Z}$. We would tend to write $1 + 3\mathbb{Z}$ rather then $4 + 3\mathbb{Z}$ or $-5 + 3\mathbb{Z}$ for example.

Let's take a specific example. Let's take $f(x) = x^2 + 3x + 2$ and $R = \mathbb{R}[x]$. So $I = f(x)\mathbb{R}[x]$. The quotient ring in question is then R/I and $(2x^4 + x^3 - 3x + 1) + I$ is an element of the quotient ring.

Applying 'division with remainder'using $2x^4 + x^3 - 3x + 1 + I$ and $x^2 + 3x + 2$ we get

$$2x^4 + x^3 - 3x + 1 = (x^2 + 3x + 2)(2x^2 - 5x + 11) + (-26x - 21)$$

so that the polynomial $2x^2 - 5x + 1$ is the quotient and the polynomial $-26x - 21$ is the remainder. Notice that $(x^2 + 3x + 2)(2x^2 - 5x + 11) = f(x)(2x^2 - 5x + 11) \in I$, being a multiple of $f(x)$, and so

$$2x^4 + x^3 - 3x + 1 - (-26x - 21) \in I.$$

This means that we have equality of cosets:

$$(2x^4 + x^3 - 3x + 1) + I = (-26x - 21) + I$$

and $(-26x - 21) + I$ is the canonical form of the coset $(2x^4 + x^3 - 3x + 1) + I$.

This idea is particularly useful when multiplying two cosets. When perfoming multiplication in this quotient ring i.e.,

$$(g(x) + I)(h(x) + I) = g(x)h(x) + I,$$

it's often the case the $g(x)h(x)$ has degree higher than that of $f(x)$ and so taking the remainder upon dividing it by $f(x)$ will lead to use being able to write $g(x)h(x) + I$ back in canonical form. This is entirely analogous to something like

$$(3 + 5\mathbb{Z})(4 + 5\mathbb{Z}) = 12 + 5\mathbb{Z} = 2 + 5\mathbb{Z}$$

where the coset $12 + 5\mathbb{Z}$ has been reduced back to canonical form by taking the remainder when 12 is divided by 5. $\Diamond$

18.2.6 Example Something interesting happens when we take $f(x) = x^2 + 1$ and $R = \mathbb{R}[x]$ in the previous example.

As we've seen before in the ring R/I where $I = f(x)R$ the additive identity is $0 + I$ and the multiplicative identiy is $1 + I$. The multiplicative identity has additive inverse of $-1 + I$.

But note that since $x^2 + 1 = x^2 - (-1) \in I$ we have the following equality of cosets

$$x^2 + I = -1 + I.$$

This means that

$$(x + I)(x + I) = -1 + I.$$

We have a square root for the negative of the additive identity, a square root of -1, in this ring! $\diamondsuit$

It turns out that, in the previous example, R/I is 'the same' as the complex numbers. We'll formalise this in the next (and final) chapter.

Chapter 19 - Ring Homomorphisms

19.1 What is a ring homomorphism?

In a very similar way to a group homomorphism, a ring homomorphism is a function from a ring to another ring which 'preserves' both the addition and the multiplication. Here is the definition.

19.1.1 Definition Let $(R_1, +_{R_1}, \times_{R_1})$ and $(R_2, +_{R_2}, \times_{R_2})$ be rings. We say that the function $\phi : R_1 \to R_2$ is a *(ring) homomorphism* if $\phi(1_{R_1}) = 1_{R_2}$ and it satisfies

$$\phi(r +_{R_1} s) = \phi(r) +_{R_2} \phi(s) \text{ and } \phi(r \times_{R_1} s) = \phi(r) \times_{R_2} \phi(s)$$

for all r, s in R_1.

If, in addition, ϕ is a bijection. Then it is said to be an *(ring) isomorphism.* ♢

Let's look at some examples.

19.1.2 Example Let $R_1 = \mathbb{Z}$ and let $R_2 = \mathbb{Z}/n\mathbb{Z}$. Define a function $\phi : R_1 \to R_2$ as follows:

$$\phi(m) = m + n\mathbb{Z} \text{ (formerly known as } [m]_n).$$

To check that this is a homomorphism we need to show that $\phi(1_{R_1}) = 1_{R_2}$ then that $\phi(k) + \phi(l) = \phi(k + l)$ and that $\phi(k)\phi(l) = \phi(kl)$ for all integers k, l.

It's clear that $1_{R_1} = 1$ and $1_{R_2} = 1 + n\mathbb{Z}$ and $\phi(1) = 1 + n\mathbb{Z}$, i.e. $\phi(1_{R_1}) = 1_{R_2}$.

We have

$$\phi(k) + \phi(l) = (k + n\mathbb{Z}) + (l + n\mathbb{Z}) = (k + l) + n\mathbb{Z} = \phi(k + l)$$

and

$$\phi(k)\phi(l) = (k + n\mathbb{Z})(l + n\mathbb{Z}) = (kl) + n\mathbb{Z} = \phi(k + l)$$

So ϕ is a homomorphism. ♢

19.1.3 Example Let $R_1 = \mathbb{C}[x]$ and let $R_2 = \mathbb{C}$. Define a function $\phi : R_1 \to R_2$ as follows:

$$\phi(f(x)) = f(i).$$

In other words, ϕ evaluates $f(x) \in \mathbb{C}[x]$ at $x = i$, where i is the usual square root of -1.

To check that this is a homomorphism we need to show that, $\phi(1) = 1$, $\phi(f(x)+g(x)) = \phi(f(x)) + \phi(g(x))$ and that $\phi(f(x)g(x)) = \phi(f(x))\phi(g(x))$ for all polynomials $f(x), g(x) \in \mathbb{C}[x]$.

It's clear that if you evaluate the polynomial 1 at any complex number you just get the number 1.

We also have

$$\phi(f(x) + g(x)) = f(i) + g(i) = \phi(f(x)) + \phi(g(x))$$

and

$$\phi(f(x)g(x)) = f(i)g(i) = \phi(f(x))\phi(g(x))$$

So ϕ is a homomorphism. ♢

19.1.4 Example This one is very similar to the previous one. Let $R_1 = \mathbb{R}[x]$ and let $R_2 = \mathbb{C}$. Define a function $\phi : R_1 \to R_2$ as follows:

$$\phi(f(x)) = f(i).$$

Again, ϕ evaluates $f(x) \in \mathbb{R}[x]$ at $x = i$.

ϕ is a homomorphism exactly as in the previous example. ♢

You may be wondering what the real difference is between the homomorphims in the previous two examples. This will be answered soon.

19.1.5 Exercse Let

$$F = \left\{ \begin{pmatrix} a & b \\ -b & a \end{pmatrix} \mid a, b \in \mathbb{R} \right\}.$$

1. Show that F is a field (under the usual addition and multiplication of matrices). (**Hint:** Begin by showing that F is a subring of $M_{2\times 2}(\mathbb{R})$. You need to also show that F is commutative and that every non-zero element has a multiplicative inverse in F.)

2. Let $\phi : F \to \mathbb{C}$ be given by $\phi\begin{pmatrix} a & b \\ -b & a \end{pmatrix} = a + bi$. Show that ϕ is a ring isomorphism

3. Show that
$$F' = \left\{ \begin{pmatrix} a & b \\ -b & a \end{pmatrix} \mid a, b \in \mathbb{C} \right\}$$
is not a field.

19.2 The kernel and the image of a homomorphism

As for group homomorphism we have a kernel and an image. You will learn much more about the kernel and the image of ring homomorphisms in future modules covering rings. Just as for group homomorphisms and groups. They are fundamental in understanding exactly how rings are related.

19.2.1 Definitions Let R_1 and R_1 be rings and let $\phi : R_1 \to R_2$ be a ring homomorphism. Then

1. Ker $\phi = \{r \in R_1 \mid \phi(r) = 0_{R_2}\}$ is called the *kernel* of ϕ.
2. Im $\phi = \{s \in R_2 \mid \text{there exists } r \in R_1 \text{ with } \phi(r) = s\} = \{\phi(r) \mid r \in R_1\}$ is called the *image* of ϕ

You may be expecting the kernel and the image of a ring homomophims to be subrings of the rings they live inside (by analogy with the corresponding result for groups, Theorem 13.2.2). This is not the case. Although the image is a subring (of R_2 in the above definition) the kernel is in fact an ideal (of R_1).

19.2.2 Theorem Let R_1 and R_2 be rings and let $\phi : R_1 \to R_2$ be a homomorphism. Then

1. Ker $\phi = \{g \in G \mid \phi(g) = 1\}$ is an ideal of R_1
2. Im $\phi = \{\theta(g) \mid g \in G\}$ is a subring of R_2.

Proof.

1. Noting that ϕ is a homomorphism between the groups $(\mathbb{R}_1, +)$ and $(R_2, +)$ we see that Ker ϕ is a subgroup of $(\mathbb{R}_1, +)$ by the corresponding theorem about group homomorphims, Theorem 13.2.2.

   To show that Ker ϕ is an ideal take $x \in$ Ker ϕ and $r \in R$. Then $\phi(xr) = \phi(x)\phi(r) = 0.r = 0$ so $xr \in$ Ker ϕ. Similarly $rx \in$ Ker ϕ .

2. Im $\phi = \{\theta(r) \mid r \in R_1$ is an additive subgroup of R_2, again by considering ϕ as a group homomorphm the groups $(\mathbb{R}_1, +)$ and $(R_2, +)$. All that is left to check is that $1_{R_2} \in$ Im ϕ and that if $s_1, s_2 \in$ Im ϕ then $s_1s_2 \in$ Im ϕ.

   $1_{R_2} \in$ Im ϕ because $\phi(1_{R_1}) = 1_{R_2}$, as we see in the definition of a ring homomorphism.

   If $s_1, s_2 \in$ Im ϕ then $s_1 = \phi(r_1), s_2 = \phi(r_2$ for some $r_1, r_2 \in R_1$. Then

   $$\phi(r_1r_2) = \phi(r_1)\phi(r_2) = s_1s_2$$

   and so $s_1s_2 \in$ Im ϕ. ♢

Let's finish by looking at some examples.

19.2.3 Example For the ring homomorphism in Example 19.1.2,

$$\text{Ker } \phi = \{k \in \ Z \mid k + n\mathbb{Z} = 0 + n\mathbb{Z}\} = n\mathbb{Z},$$

the multiples of n. Convince yourself that Im $\phi = \mathbb{Z}/n\mathbb{Z}$.

19.2.4 Example For the ring homomorphism from $\mathbb{C}[x]$ to $\mathbb{C}$ in Example 19.1.3 suppose $f(x) \in$ Ker ϕ.

Then $f(i) = 0$. This is true if any only $f(x) = (x-i)g(x)$ for some $g(x) \in \mathbb{C}[x]$. Therefore Ker $\phi = (x - i)\mathbb{C}[x]$.

Given $a + bi \in \mathbb{C}$, where $a, b \in \mathbb{R}$, its clear that $a + bi = f(i)$ where $f(x)$ is the constant polynomial $a + bi \in \mathbb{C}[x]$. This means that ϕ is surjective and Im $\phi = \mathbb{C}$.

19.2.5 Example For the ring homomorphism from $\mathbb{R}[x]$ to $\mathbb{C}$ in Example 19.1.4 suppose $f(x) \in$ Ker ϕ.

Then $f(i) = 0$. This is true if any only $f(x) = (x^2-1)g(x)$ for some $g(x) \in \mathbb{R}[x]$. Therefore Ker $\phi = (x^2 - 1)\mathbb{C}[x]$ by Exercise 17.2.6.

Given $a+bi \in \mathbb{C}$, where $a,b \in \mathbb{R}$, its clear that $a+bi = f(i)$ where $f(x)$ is the polynomial $bx+a \in \mathbb{C}[x]$. This means that ϕ is surjective and Im $\phi = \mathbb{C}$.

19.2.6 Exercise Let $G = \mathbb{R}^2$ and let $H = \mathbb{R}^2$. Define a function $\phi : G \to H$ as follows:

$$\phi(x,y) = (x,x).$$

Show that ϕ is a ring homomorphism and calculate Ker ϕ and Im ϕ.

19.2.7 Exercise Let $G = \mathbb{R}^2$ and let $H = \mathbb{R}^2$. Define a function $\phi : G \to H$ as follows:

$$\phi(x,y) = (x,0).$$

Give a reason that ϕ is not a ring homomorphism.

19.2.8 Exercise Let R_1 and R_2 be groups and let $\phi : R_1 \to R_2$ be a ring homomorphism. Show that ϕ is injective if and only if Ker $\phi = \{0_{R_1}\}$.

19.2.9 Exercise Let R_1 and R_2 be groups and let $\phi : R_1 \to R_2$ be an injective ring homomorphism. Show that R_2 contains a subring which is isomorphic to R_1.

19.2.10 Example $\mathbb{R}[x]/(x^2+1)\mathbb{R}[x]$ is isomorphic to $\mathbb{C}$ as follows.

The isomorphism, ϕ, needs to map and element of $\mathbb{R}[x]/(x^2+1)\mathbb{R}[x]$, i.e. a coset of the form $f(x) + (x^2+1)\mathbb{R}[x]$ (where $f(x) \in \mathbb{R}[x]$) to a complex number. Letting $I = (x^2+1)\mathbb{R}[x]$, we define it as follows

$$\phi(f(x)+I) = f(i)$$

.

Consider the following argument, for $f(x), g(x) \in \mathbb{R}[x]$,

$$\phi(f(x)+I) = \phi(g(x)+I) \iff f(i) = g(i) \iff f(i) - g(i) = 0$$

$$\iff \text{both } f(i) - g(i) = 0 \text{ and } f(-i) - g(-i) = 0$$
$$\iff f(x) - g(x) = (x^2+1)h(x) \text{ for some } h(x) \in \mathbb{R}[x]$$
$$\iff f(x) - g(x) \in I \iff f(x) + I = g(x) + I.$$

Following the implications from right to left in this tells us that ϕ is well-defined. Following the implicatons from left to right then tells us that ϕ is injective.

You can check that this is a ring homomorphism and also surjective (and hence an isomorphism). $\Diamond$

For all the examples of ring homomorphism in this chapter it's possible to show that the quotient ring $R/\text{Ker } \phi$ is isomorphic (as a ring) to $\text{Im } \phi$. That's because this is true for any ring homomorphism. However that story will be left for a future algebra module.

Many thanks and good luck, Richard.