MT1003 Pure and Applied Mathematics
Pure Lecture Notes[1]

Tom Coleman[2]

July 23, 2018

[2]Adapted from previous years lecture notes by Dr Martyn Quick.

Contents

Introduction to pure mathematics **3**

What does that symbol mean? A brief introduction to set theory 4

What does that word mean? A glossary of pure mathematics 7

How can I do pure mathematics? 10

1 Number theory **12**

1.1 Divisibility of integers 12

1.1.1 Divisors and remainders 12

1.1.2 Positional notation 15

1.2 Greatest common divisors and the Euclidean algorithm 18

1.3 Primes 24

1.3.1 The Fundamental Theorem of Arithmetic 24

1.3.2 Properties of primes 27

1.4 Congruences and modular arithmetic 31

1.5 Linear Diophantine equations 37

1.6 Higher order Diophantine equations 46

2 Functions and relations **52**

2.1 Functions 52

2.2 Relations 58

3 Graph theory **68**

3.1 Graphs and digraphs . . . 69
3.1.1 Directed graphs . . . 69
3.1.2 Graphs . . . 74
3.2 Eulerian and Hamiltonian graphs . . . 82
3.2.1 Eulerian walks and circuits . . . 82
3.2.2 Hamiltonian paths and cycles . . . 85
3.3 Planar graphs . . . 88

4 Group theory **99**

4.1 Permutations . . . 100
4.1.1 Starting out with permutations . . . 100
4.1.2 Cycle notation . . . 103
4.1.3 Properties of permutations . . . 107
4.2 Groups . . . 111
4.2.1 Defining a group . . . 112
4.2.2 Examples of groups . . . 114
4.2.3 Properties of groups . . . 122

Introduction to pure mathematics

What is pure mathematics?

In previous studies, you may have come across pure (or 'core') mathematics as the development of mathematical techniques for application in other problems. However, this sort of mathematics is taught in MT1002; so what is 'pure' mathematics?

Roughly, pure mathematics is the study of 'abstract concepts' of number, space, structure and change. Pure mathematics acts as a foundation; from which more specialised areas of mathematics can be developed and used as tools to understand the universe around us.

The study of pure mathematics typically follows the following format:

- An abstract definition is made concerning some mathematical object;
- An investigation is made into these objects that only depend on the details of the definition;
- The results made in this investigation are used to describe any example that satisfies the definition, either applied to solving problems or used to differentiate that object from others that may share the same definition.

This is roughly how the course will go. Initially, the idea of abstracting concepts may take some getting used to; but this idea of abstracting concepts is both the power and the beauty of pure mathematics.

In this short introductory chapter, there is a brief introduction to the language of set theory and a glossary of words often used in pure mathematics; both of which will be used liberally throughout these notes. In addition, there is a list of top tips in order to help you embrace the study of pure mathematics. I would strongly recommend reading this before continuing through the lecture notes.

What does that symbol mean? A brief introduction to set theory

In this course, a **set** is a collection of objects. A set can be anything; from a set of numbers to a set of saxophonists. A object a in a set A is called an **element** of that set, and is written $a \in A$. Two sets are equal exactly when they have the same elements. It is important to see that elements are **not** the same as sets; so the set containing the element a is **not** the same as a. There is a special set that contains **no** elements; this is called the **empty set** and is written using the symbol $\varnothing$.

Mathematicians use braces $\{,\}$ to write sets. So, for instance, the set A containing the elements a, b, c, d is written as

$$A = \{a, b, c, d\}$$

When you write a set, the order in which you write the elements doesn't matter. This means that

$$A = \{a, b, c, d\} = \{a, c, d, b\}$$

are the same sets. Also, repeating the same element when writing a set doesn't matter either. This means that

$$A = \{a, b, c, d\} = \{a, a, a, a, b, b, b, c, c, d\}$$

are the same sets. As mentioned above, sets are equal exactly when they have the same elements; so the two sets

$$A = \{a, b, c, d\} \quad \text{and} \quad B = \{b, c, d\}$$

are not the same.

Other special sets are sets of numbers. Table 1 contains a list of the five most common of these, each with an individual symbol. You should learn all of these and be comfortable with their usage.

Sometimes, you may want to write a specific set that has too many elements to write comfortably (like the sets of numbers in Table 1). To do this, you can use something called **set-builder notation**. Set-builder notation consists of writing a set in two parts; mathematical objects, and properties those objects satisfy. This is usually written with a

set	name	explanation
$\mathbb{N} = \{1, 2, 3, \ldots\}$	**natural numbers**	These are non-negative whole numbers. Some sources may say that 0 is a natural number; in this course, 0 is not a natural number.
$\mathbb{Z} = \{\ldots, -2, -1, 0, 1, \ldots\}$	**integers**	This is the set of all whole numbers.
$\mathbb{Q} = \left\{ \frac{a}{b} \ : \ a, b \in \mathbb{Z},\ b \neq 0 \right\}$	**rational numbers**	This is the set of all fractions, where the numerator and denominator are both whole numbers.
$\mathbb{R}$	**real numbers**	Informally, this is the set of all possible decimal numbers.
$\mathbb{C} = \{a + bi \ : \ a, b \in \mathbb{R}\}$	**complex numbers**	Here, $i^2 = -1$. Complex numbers will not be considered in this course.

Table 1: Sets of numbers

colon : or a bar | separating these two components; so you could write

$$\text{set} = \{\text{objects} \ : \ \text{properties}\}$$

You can think of the : or | as replacing the words 'such that'. For instance, 'all integers such that their squares are greater than fifteen' is a set, and can be written as

$$S = \{x \in \mathbb{Z} \ : \ x^2 > 15\}$$

(where $\mathbb{Z}$ is the set of all integers; see Table 1). You can notice here that this set has infinitely many elements, and so would take a while to write out without this concept! The description of $\mathbb{Q}$ in Table 1 is written using set-builder notation. If you do not fully understand how a set is built, then writing the set out in words may help you to understand the problem.

Sets can be contained in other sets. If A and B are two sets then you can say that A is a **subset** of B if every element of A is also an element of B. If this happens, you can write that $A \subseteq B$. For example,

$$\{b, c, d\} \subseteq \{a, b, c, d\}$$

You can see that as every element of A is an element of A, then $A \subseteq A$. Similarly, as every element of $\varnothing$ (there are none) is an element of A, then $\varnothing \subseteq A$. If $A \subseteq B$ and $B \subseteq A$ then they have the same elements and so $A = B$.

The **union** of A and B is the set that contains all elements of A and all elements of B. The union of A and B is written as $A \cup B$. In set-builder notation, you can write this as

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

The **intersection** of A and B is the set of all elements that are in *both* A and B. The intersection of A and B is written as $A \cap B$. In set-builder notation, you can write this as

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

For instance,

$$\{a, b, c\} \cup \{c, d, e\} = \{a, b, c, d, e\} \quad \text{and} \quad \{a, b, c\} \cap \{c, d, e\} = \{c\}$$

The **Cartesian product** $A \times B$ is defined to be the set of all ordered pairs (a, b) with the first element in A and the second element in B; in set-builder notation, this is:

$$A \times B = \{(a, b) \ : \ a \in A, \, b \in B\}$$

Finally, there are two symbols that are used to 'quantify' elements of sets. The first of these is $\exists$ which means 'there exists'. This symbol is used to say that there is an object in some set that satisfies some mathematical property. For instance, $\exists a \in A$ such that $a^2 = 1$ means there exists an element a of the set A where a^2 is equal to 1. The second of these is $\forall$ which means 'for all'. This symbol is used to say that all mathematical objects in a set satisfies some property. For instance, $n \geq 0$ $(\forall n \in \mathbb{N})$ means that n is greater than 0 for all natural numbers n.

What does that word mean? A glossary of pure mathematics

definition These are precise statements of the meaning of a concept. Definitions provide a starting point to mathematics; by writing down these formal statements, you can then go on to prove results about those definitions in question.

result Definitions are all well and good, but they are boring on their own! Mathematics relies on showing **results**; statements of fact that are logically shown (**proved**; see below) from the definitions. Results can be expressed in the form

> If H is true then C is true

where H is known as the **hypothesis** (or **assumptions**) of the result and C is known as the **conclusion** of the result. There are **five** types of result in this course, and they are:

theorem A major mathematical result.

lemma Lemmas are (usually) minor mathematical results shown in most cases to be used in proving theorems.

proposition A mathematical result that is too important to be a lemma but not important enough to be a theorem.

corollary A corollary is a direct consequence of a theorem or proposition.

claim A claim is a result that needs a proof, but is too minor to be a lemma. You will usually find claims inside the proof of a theorem.

converse If a result looks like

> If H is true then C is true

then the **converse** of a result is the statement

> If C is true then H is true

The converse of a statement may or may not be true! For instance, all natural numbers are positive, but not all positive numbers are natural numbers (for instance, 1.5 is positive but not a whole number).

proof A proof of a result is a series of logical steps that start with the hypothesis H of the result and end with the conclusion C. These logical steps can include (but are certainly not limited to) the hypotheses of a theorem, definitions, use of previous steps in the proof, use of a previously proved result, or standard mathematical manipulation. Every mathematical result (theorem, lemma, proposition, corollary) should have a proof.

Proofs of results come in many different shapes and sizes; from one line consequences of theorems (some included in this course) to tens of thousands of pages (not included in this course). Here are five common types of proof.

direct proof In a **direct proof**, you can start with the hypothesis H and aim straight for the conclusion C. Usually, you will use definitions and previously proved results in a direct proof. Theorem 1.1.4 contains examples of direct proofs.

proof by induction A proof by induction is usually used in results that look like

For all $n \geq x$, then C holds

They come in two distinct parts. The first is a **base case**; showing that the statement is true for the lowest possible value of n (usually $n = 1$, but not always). The second is an **inductive case**: here, you would assume that the statement holds for $n = k$ (an assumption known as the **inductive hypothesis**), and then prove that the statement holds for the case where $n = k + 1$ using the inductive hypothesis. If both the base case and the inductive case are true, then the result is also true. See Theorem 1.1.6 for an example of a proof by induction.

proof by contradiction In a proof by contradiction, you assume that the hypothesis H is true but that the conclusion C is **false**. Then you can use an argument to get to a mathematical statement that is definitely wrong (like $0 = 1$, or $5 < 2$). What this does is then show that

If H is true then C cannot be false

which is exactly the same as if H is true then C is true.

contrapositive proof If a result looks like

If H is true then C is true

then the **contrapositive** of this result is

If C is false then H is false

They are exactly the same result. Sometimes it is easier to prove the contrapositive of a result (by any of the means above) than it is to prove the result in its original form.

counterexample You see a mathematical statement that you have decided is wrong. You can prove that it *is* false by using a **counterexample**. Here, a counterexample involves a single mathematical object that satisfies the hypothesis H of a result but **not** the conclusion C. You only need to find **one** counterexample to show that the result is not true; don't be tempted to work generally!

Here are some other common mathematical words and phrases, along with what they mean.

if and only if If and only if results are actually **two** results in one. Saying "A if and only if B" is the same as saying

"if A ... then B" **and** "if B ... then A".

These two statements are sometimes called **directions** of the theorem; the "if" direction (if A ... then B) and the "only if" direction (if B ... then A). To prove an if and only if statement, you must prove **both** directions of the theorem. The order in which you prove the directions does not matter; but it is important to make sure you prove both.

the following are equivalent This is exactly the same as saying if and only if. 'The following are equivalent' may be used if there is more than one statement in the result. For instance the statement

The following are equivalent: A, B, C, D

means

A if and only if B if and only if C if and only if D

To prove this statement, you only need to prove the statements in a 'cycle'; so to prove the above 'the following are equivalent' statement, it is enough to show

If A then B, if B then C, if C then D, if D then A.

in full generality This means that you need to use general terms rather than specific examples in your working. For instance, if you are proving a statement about all integers, then you would need to use something like $a \in \mathbb{Z}$ rather than a specific number like 7 or -16.

without loss of generality This means you can make a further assumption about the working in question without taking away from the generality of the result you are trying to show. You should be very careful when using this phrase; sometimes, students say this while unwittingly adding an extra assumption to the hypothesis! This then only proves half of the theorem.

remark This is a short statement following a proof of a result about something that is of interest, but not enough to be a result.

□, ■, QED These symbols/words signify the end of a proof. This course uses □, but also feel free to use ■; think of it as a pat on the back! However, try to avoid using QED; while iconic, it is rarely used and hardly any contemporary mathematicians use it to end a proof.

How can I do pure mathematics?

- **Learn all of the definitions**. This is the most important point; to be able to understand what you are asked to show, you need to have a firm grasp of the definitions involved in the statement. The same applies to the writing of proofs. For instance, if you are asked to prove that some integer is congruent to 0 modulo 5; there would be no chance of doing so unless you understand what congruent modulo 5 means.

- **Read through questions carefully, highlighting definitions, jargon, assumptions and the conclusion**. This is important. For instance, if the theorem is an 'if and only if' statement, then you would need to prove both directions (if you don't, the work is only half done). By highlighting the assumptions in the question, you know what you are allowed to work with. Writing down the precise meanings of the hypothesis and conclusion helps to reinforce the definitions you have learned and gives you guidance on how to prove the statement.

- **Don't assume the conclusion is true and work backwards**. Doing this does **not** prove anything. For instance, if you were to assume that the conclusion was true, and then show that $1 = 1$ or something similar, then the working is backwards. The conclusion should only be written in your working at the **very end** of your proof.

- **Keep it simple**. It may be tempting to use a proof by contradiction, or a contrapositive proof; these are shiny new tools for an undergraduate mathematician. However, it may be that these ways of proving things are too elaborate for the statement in

question. A common mistake is to assume that the conclusion C is false, and then start from the hypothesis H and then prove that C is in fact true. This is actually a direct proof, and the assumption for a contradiction was not needed.

- **A good proof is a mixture of words and symbols**. This is one of the hardest skills to learn. The inclination may be to prove something using mathematical symbols entirely; while concise, it does not offer the reasoning behind your steps like words do. On the other hand, a proof consisting entirely of words may communicate your reasoning clearly, but lack the mathematical clarity to really mean what you are trying to say. A good proof therefore uses both symbols (for clarity and mathematical precision) and words (for demonstrating your understanding and explaining the method of your proof).

- **Always check your work**. This is especially important in pure mathematics. Read through your proof carefully, asking 'why?' at every step. Make sure you only assume definitions, previously proved work and the hypothesis; never assume the conclusion is true. Finally, you should make sure that the end of the proof is really the same as the conclusion of the statement. Don't leave a proof half-done!

- **And finally: practice, practice, practice**.

Chapter 1

Number theory

1.1 Divisibility of integers

1.1.1 Divisors and remainders

You can do addition, subtraction and multiplication of integers, and end up with another integer. However, division is **not** always defined for a pair of integers; dividing two integers does not necessarily give an integer. For instance, $18/198$ is not an integer. The course begins with a well-known property of the integers. It says you can always divide two integers using a remainder.

Fact 1.1.1. *For every two integers a and b with $b > 0$ there exist unique integers q and r such that*

$$a = bq + r \text{ and } 0 \leq r < b.$$

Here, q is called the **quotient** *and r is called the* **remainder**. □

Remark. This fact will not be proved here. However, it can be proved starting from the standard Peano axioms for number theory (this is beyond the scope of the present course), or via a proof by induction.

The remarkable thing about Fact 1.1.1 is that many important properties of integers follow from this result.

Example 1.1.2. Dividing 42 by 8 with remainder gives $42 = 8 \cdot 5 + 2$. Here, 5 is the quotient and 2 is the remainder.

Dividing 55 by 5 with remainder gives $55 = 5 \cdot 11 + 0$. Here, 11 is the quotient and 0 is the remainder.

Definition 1.1.3. For two integers a and b with $b \neq 0$, say that b **divides** a or that a **is divisible by** b if there exists an integer q such that $a = bq$. This is written as $b \mid a$. In this case, you can say that b is a **divisor** of a.

If a is *not* divisible by b, then this is sometimes written as $b \nmid a$. You can also say that b is **not** a divisor of a.

You can see from Example 1.1.2 that $5 \mid 55$, but $8 \nmid 42$. In fact, if you divide a by b with remainder, then you can say that b divides a if the remainder is 0.

Once you have a definition, you can move on to finding properties of the mathematical objects concerned with the definition. Here, you have been introduced to the idea of a divisor of an integer; now you can use this idea to prove several useful consequences of this idea. This leads on to the first result of the course.

Theorem 1.1.4 (Basic properties)**.** *Let a, b, c, d, x, y be integers. Then the following statements hold:*

(i) $a \mid 0$, $1 \mid a$, $a \mid a$.

(ii) If $a \mid 1$, *then* $a = \pm 1$.

(iii) If $a \mid b$ *and* $c \mid d$, *then* $ac \mid bd$.

(iv) If $a \mid b$ *and* $b \mid c$ *then* $a \mid c$.

(v) If $a \mid b$ *and* $b \mid a$, *then* $a = \pm b$.

(vi) If $a \mid b$ *and* $a \mid c$, *then* $a \mid (bx + cy)$.

Proof. (i) Given the integer a, you can write that

$$0 \cdot a = 0$$

and as 0 is an integer, $a \mid 0$ by Definition 1.1.3. To show that $1 \mid a$ and $a \mid a$, you can say that

$$1 \cdot a = a$$

As a is an integer, it follows from Definition 1.1.3 that $1 \mid a$. Similarly, as 1 is an integer, you can say that $a \mid a$.

(v) Suppose that $a \mid b$ and $b \mid a$. By Definition 1.1.3, this means that there exist integers q and r such that $b = aq$ and $a = br$. Using the fact that $a = br$, you can write that

$$a = aqr$$

Rearranging this equation gives that

$$a - aqr = a(1 - qr) = 0$$

As this happens, you know that either $a = 0$ or $(1 - qr) = 0$; so now you can check both cases. If $a = 0$, then

$$b = aq = 0 \cdot q = 0$$

and so $a = b$. If $(1 - qr) = 0$, then $qr = 1$. Here, $q \mid 1$ by Definition 1.1.3 and so $q = \pm 1$ by part (ii) of this theorem. This means that $b = aq = \pm a$. So in either case, $b = \pm a$, which completes the proof. □

Using Fact 1.1.1 and Definition 1.1.3, you can prove a nice theorem that will be useful later on when considering sums of squares (see Section 1.4).

Theorem 1.1.5. *Let a be an integer. Then a^2 is either divisible by 4 or it has remainder 1 when divided by 8.*

Proof. The proof involves checking a series of cases. By Fact 1.1.1, the possible remainders of a when divided by 4 are 0, 1, 2 and 3. So this means that a can be written as either $4k, 4k + 1, 4k + 2$ or $4k + 3$ for some integer k.

You can then check each of these cases in turn.

- For $a = 4k$, it follows that

$$a^2 = (4k)^2 = 16k^2 = 4 \cdot 4k^2$$

  As $4k^2$ is an integer, 4 divides a^2 by Definition 1.1.3.

- For $a = 4k + 1$, you can see that

$$a^2 = (4k + 1)^2 = 16k^2 + 8k + 1 = 8 \cdot (2k^2 + k) + 1$$

so in this case, a^2 is a multiple of 8 plus 1; this is exactly the same as having remainder 1 when divided by 8.

- For $a = 4k + 2$, you can write that

$$a^2 = (4k + 2)^2 = 16k^2 + 16k + 4 = 4 \cdot (4k^2 + 4k + 1)$$

As $4k^2 + 4k + 1$ is an integer, it follows that 4 divides a^2 by Definition 1.1.3.

- Finally, if $a = 4k + 3$ then

$$a^2 = (4k + 3)^2 = 16k^2 + 24k + 9 = 8 \cdot (2k^2 + 3k + 1) + 1$$

and this is similar to the case where $a = 4k + 1$; it has remainder 1 when divided by 8.

□

1.1.2 Positional notation

Fact 1.1.1 is one of the most important statements in mathematics. Not only does it govern the division of integers, but it also has an effect on the way that humanity expresses numbers. For instance, you can express this year (2018) by writing

$$2018 = 2 \cdot 1000 + 0 \cdot 100 + 1 \cdot 10 + 8 \cdot 1 = 2 \cdot 10^3 + 0 \cdot 10^2 + 1 \cdot 10^1 + 8 \cdot 10^0$$

This idea of writing integers in this way follows from Fact 1.1.1, and the mathematical name for it is **positional notation**. As in all of pure mathematics however, you need to prove that you can really write integers in this way, and not just for the numbers 2018 and 10. This is the subject of the next theorem.

Theorem 1.1.6 (Positional notation)**.** *Let $b > 1$ be a fixed integer. Then every positive integer a can be written as*

$$a = d_{n-1}b^{n-1} + d_{n-2}b^{n-2} + \ldots + d_1 b + d_0,$$

where $n \geq 1$ and $0 \leq d_i < b$ for all $i = 0, \ldots, n - 1$.

Moreover, n and all d_i are uniquely determined.

Proof. The proof of the theorem is by induction on a. Here, you will need to show that there **exists** a representation of a as stated in the theorem, before going on to show that this representation is **unique**.

Base case: For $a = 1$ (and, more generally, when $a < b$) there is nothing to prove; you can take $d_0 = a$. This representation is unique in the integers, so the base case holds.

Inductive case: For $a > 1$, assume that the statement of the theorem holds for every positive integer less than a; this is the **inductive hypothesis**. To begin, divide a by b with remainder (this is Fact 1.1.1) to get:

$$a = qb + d_0$$

where $0 \leq d_0 < b$. Since $b > 1$, it must be true that $q < a$. By the inductive hypothesis, q can be written as:

$$q = d_{n-1}b^{n-2} + d_{n-2}b^{n-3} + \ldots + d_1,$$

with $0 \leq d_i < b$ for all i between 1 and $n-1$. This gives

$$a = d_{n-1}b^{n-1} + d_{n-2}b^{n-2} + \ldots + d_1 b + d_0$$

as required. This takes care of the existence portion of the proof.

It remains to show that this representation of a is unique. To do this, assume that a can also be written as

$$a = f_{m-1}b^{m-1} + f_{m-2}b^{m-2} + \ldots + f_1 b + f_0$$

for some integer m and $0 \leq f_i < b$ for $i = 0, ..., m-1$. (Note that you should not use the same letters as above here!) Then you can write that

$$\begin{aligned} a &= f_{m-1}b^{m-1} + f_{m-2}b^{m-2} + \ldots + f_1 b + f_0 \\ &= b(f_{m-1}b^{m-2} + f_{m-2}b^{m-3} + \ldots + f_1) + f_0 \\ &= bq_1 + f_0. \end{aligned}$$

Now, Fact 1.1.1 implies that $f_0 = d_0$; it then follows that $q_1 = q$. Using the inductive hypothesis gives $m = n$ and $d_i = f_i$ for all i. □

Notation. In the above theorem, b is called the **base**, and d_i are called the **digits**. In this case, you can write

$$a = \overline{(d_{n-1}d_{n-2}\ldots d_1 d_0)}_b$$

This expression is called the **representation of a in base b**. When there is no danger of confusion (usually when using base 10), the representation can just be written as $a = d_{n-1}d_{n-2}\dots d_1d_0$.

Remark. Different bases have different names; base 10 is *decimal*, base 2 is *binary*, base 3 is *ternary*, base 8 is *octal* and base 16 is *hexadecimal*. Base 60 (used by the Babylonians) is known as *sexagesimal*.

The proof of Theorem 1.1.6 is useful; not only does it prove the theorem, but it also tells you you how to find the representation of an integer a in base b. You can then write down an algorithm for this process.

Algorithm 1.1.7 (Writing numbers in different bases)**.** Let a and $b > 1$ be integers. To find the representation of a in base b:

Step 1: Divide a by b: $a = qb + r$;

Step 2: Set r to be the last digit of the expansion;

Step 3: Rename: $a := q$;

Step 4: Repeat the above steps until $a = 0$.

Essentially, what this says is to take a number a, and divide by b many times, keeping track of the remainders, until you get to 0. Here are two examples that demonstrate Algorithm 1.1.7.

Example 1.1.8. Suppose you are asked to write 53 in base 7; to do this, you can use Algorithm 1.1.7. So here,

$$53 = 7 \cdot 7 + 4$$

and so 4 is the last digit of the expansion. Setting $a = 7$ and repeating the process gives

$$7 = 7 \cdot 1 + 0$$

and so 0 is the second to last digit of the expansion. Finally, setting $a = 1$ gives

$$1 = 7 \cdot 0 + 1$$

and so 1 is the next digit of the expansion. As q in this case is zero, you can stop and write

$$53 = \overline{104}_7$$

You can check your answer by writing out the expansion in the form of Theorem 1.1.6; so here

$$\overline{104}_7 = 1 \cdot 7^2 + 0 \cdot 7^1 + 4 \cdot 7^0 = 49 + 0 + 4 = 53$$

Now suppose you are asked to write 53 in base 3. Using Algorithm 1.1.7 gives:

$$\begin{aligned} 53 &= 3 \cdot 17 + 2 \\ 17 &= 3 \cdot 5 + 2 \\ 5 &= 3 \cdot 1 + 2 \\ 1 &= 3 \cdot 0 + 1 \end{aligned}$$

So $53 = \overline{1222}_3$; you can check this answer using Theorem 1.1.6:

$$\overline{1222}_3 = 1 \cdot 3^3 + 2 \cdot 3^2 + 2 \cdot 3^1 + 2 \cdot 3^0 = 27 + 18 + 6 + 2 = 53$$

1.2 Greatest common divisors and the Euclidean algorithm

Definition 1.2.1. For two integers a and b (at least one of which is not 0), their **greatest common divisor** $\gcd(a, b)$ is the largest *positive* integer d which divides a and b. In other words, it is the unique positive number with the following properties:

- $d \mid a$ and $d \mid b$;
- If $c \mid a$ and $c \mid b$, then $c \mid d$.

If $\gcd(a, b) = 1$ then a and b are said to be **coprime**.

Remark. Some definitions of $\gcd$ may state that $c \leq d$. This is also perfectly valid as a definition of the greatest common divisor of two integers. However, the definition given in Definition 1.2.1 can be generalised to other mathematical structures known as *commutative rings* (see MT3505).

For relatively small integers a and b, you can calculate the greatest common divisor of a and b by just writing it down.

Example 1.2.2. Here, $\gcd(12, 42) = 6$, $\gcd(8, 20) = 4$, and $\gcd(13, 14) = 1$. As the greatest common divisor of 13 and 14 is 1, you can say that 13 and 14 are coprime.

However, for larger integers, finding their greatest common divisor becomes much harder. Where would you begin if you were asked to find the greatest common divisor of 444 and 903? How about 1107 and 496? Even writing down the divisors of these would take a while. Therefore, it would be useful if there was an advanced tool that allows you to find the greatest common divisor of two numbers without reference to their divisors. Thankfully, such a tool exists and is called the **Euclidean algorithm**.

Since $\gcd(a,b) = \gcd(b,a)$, you can define the algorithm for $a \geq b$ without losing any information. (In pure mathematics, this is sometimes called **without loss of generality**.)

Algorithm 1.2.3 (Euclidean Algorithm)**.** Let $a \geq b$ be integers. To find the greatest common divisor $\gcd(a,b)$ of a and b, you can do the following:

- **Step 1:** Define $a_1 = a,\ \ b_1 = b$.
  Divide a_1 by b_1 with remainder to get $a_1 = b_1 q_1 + r_1$.
- **Step** n**:** Define $a_n = b_{n-1},\ \ b_n = r_{n-1}$.
  Divide a_n by b_n to get $a_n = b_n q_n + r_n$.
- Repeat until $r_k = 0$.
- The last non-zero remainder r_{k-1} is $\gcd(a,b)$.

Remark. It is important to see the difference between the Euclidean algorithm (Algorithm 1.2.3) and the algorithm for representing a number a in base b (Algorithm 1.1.7). In each step of the Euclidean algorithm you are changing the number you divide by, whereas in the algorithm for writing a number in base b you are dividing by b at every step.

Example 1.2.4. Suppose you are asked to find the $\gcd$ of 444 and 903, as stated above. You can use the Euclidean algorithm (Algorithm 1.2.3) to do this, by starting with $a_1 = 903$ and $b_1 = 444$. Dividing 903 by 444 with remainder gives

$$\underbrace{903}_{a_1} = \underbrace{444}_{b_1} \cdot 2 + \underbrace{15}_{r_1}$$

At the next step of the algorithm, you set $a_2 = b_1 = 444$ and $b_2 = r_1 = 15$. Dividing 444 by 15 with remainder gives

$$\underbrace{444}_{a_2} = \underbrace{15}_{b_2} \cdot 29 + \underbrace{9}_{r_2}$$

As the remainder r_2 is still non-zero, you must keep going. Now, setting $a_3 = b_2 = 15$ and

$b_3 = r_2 = 9$, and dividing 15 by 9 with remainder gives

$$\underbrace{15}_{a_3} = \underbrace{9}_{b_3} \cdot 1 + \underbrace{6}_{r_3}$$

You can set $a_4 = b_3 = 9$ and $b_4 = r_3 = 6$ and divide by remainder again to get

$$\underbrace{9}_{a_4} = \underbrace{6}_{b_4} \cdot 1 + \underbrace{3}_{r_4}$$

Finally, setting $a_5 = b_4 = 6$ and $b_5 = r_4 = 3$ and dividing gives

$$\underbrace{6}_{a_5} = \underbrace{3}_{b_5} \cdot 2 + \underbrace{0}_{r_5}$$

As $r_5 = 0$, you can stop here. Algorithm 1.2.3 then states that the greatest common divisor of 444 and 903 is the last non-zero remainder; in this case, it is $r_4 = 3$. So you can write $\gcd(903, 444) = 3$.

You can write this working out without a lot of explanation; provided you set out your calculations clearly enough. For instance, using the Euclidean algorithm to find $\gcd(1107, 496)$:

$$\begin{aligned} 1107 &= 496 \cdot 2 + 115 \\ 496 &= 115 \cdot 4 + 36 \\ 115 &= 36 \cdot 3 + 7 \\ 36 &= 7 \cdot 5 + 1 \\ 7 &= 1 \cdot 7 + 0 \end{aligned}$$

and so $\gcd(1107, 496) = 1$ and therefore the numbers are coprime (see Definition 1.2.1).

Finding the greatest common divisor of two integers a and b using the Euclidean algorithm is appealing. However, how do you know that the Euclidean algorithm actually outputs $\gcd(a, b)$ as Algorithm 1.2.3 claims? This is something that you need to prove. The first step in doing this is the following result about greatest common divisors.

Lemma 1.2.5. *If a, b, q, r are integers satisfying $a = qb + r$ then $\gcd(a, b) = \gcd(b, r)$.*

Proof. Let $d_1 = \gcd(a, b)$ and $d_2 = \gcd(b, r)$. As $d_2 \mid b$ and $d_2 \mid r$, it follows that $d_2 \mid a$ by Theorem 1.1.4 (vi); so then $d_2 \mid d_1$ by the definition of $\gcd(a, b)$. By writing $r = a - qb$,

and as $d_1 \mid a$ and $d_1 \mid b$, then $d_1 \mid r$ by Theorem 1.1.4 (vi); so $d_1 \mid d_2$ by definition of $\gcd(b, r)$.

As $d_1 \mid d_2$ and $d_2 \mid d_1$, together with the fact that $d_1, d_2 > 0$, you can say that $d_1 = d_2$ by Theorem 1.1.4 (v). □

Theorem 1.2.6. *The Euclidean algorithm works; given positive integers $a \geq b$, you can apply the Euclidean algorithm (Algorithm 1.2.3) to find $\gcd(a, b)$.*

Proof. Applying the Euclidean algorithm to $a \geq b$ gives a sequence $(a_1, b_1), \ldots, (a_k, b_k)$ of pairs of non-negative integers, defined by $a_1 = a$ and $b_1 = b$, and for $n \geq 2$ by

$$a_n = b_{n-1} \quad \text{and} \quad b_n = r_{n-1}$$

where $a_{n-1} = b_{n-1}q_{n-1} + r_{n-1}$ and $0 \leq r_{n-1} < b_{n-1}$ by Fact 1.1.1. First of all, you can notice that as $r_{n-1} < b_{n-1}$, then $b_1 > b_2 > \ldots > b_k$. As each b_i is non-negative, this means that the process stops at some point. You can use Lemma 1.2.5 and the definition of a_n and b_n to say that

$$\gcd(a_{n-1}, b_{n-1}) = \gcd(b_{n-1}, r_{n-1}) = \gcd(a_n, b_n)$$

This means that

$$\gcd(a_1, b_1) = \gcd(a_2, b_2) = \ldots = \gcd(a_k, b_k)$$

At the last stage of the Euclidean algorithm, $r_k = 0$; which means that $a_k = b_k q_k$ and therefore $b_k \mid a_k$. Now this implies that $\gcd(a_k, b_k) = b_k = r_{k-1}$ and so, as $\gcd(a, b) = \gcd(a_1, b_1) = \gcd(a_k, b_k)$, you can conclude that $\gcd(a, b) = r_{k-1}$. □

An important corollary of the Euclidean algorithm follows as a result of Bézout.

Corollary 1.2.7 (Bézout's Lemma)**.** *Suppose that $a \geq b$ are positive integers. Then there exist integers x and y such that*

$$\gcd(a, b) = xa + yb$$

Proof. Using the Euclidean algorithm (Algorithm 1.2.3), you can generate a sequence of pairs of integers

$$(a, b) = (a_1, b_1),\ (a_2, b_2),\ \ldots\ , (a_k, b_k)$$

The idea here is to show that there exist integers x_n, y_n, z_n, t_n such that

$$a_n = x_n a + y_n b \quad \text{and} \quad b_n = z_n a + t_n b$$

for all $n \in \mathbb{N}$; and then note that $b_k = r_{k-1} = \gcd(a, b)$ is the last non-zero remainder, meaning that $b_k = \gcd(a, b)$ can be expressed in this form. You can do this via induction on the length k of the sequence of pairs of integers as generated above.

Base case $(k = 1)$: Here, you can write that $a_1 = a = a \cdot 1 + b \cdot 0$ and $b_1 = b = a \cdot 0 + b \cdot 1$. So the base case holds.

Inductive case $(k = n)$: Suppose that $n = 1$. You can assume that

$$a_{n-1} = x_{n-1} a + y_{n-1} b \quad \text{and} \quad b_{n-1} = z_{n-1} a + t_{n-1} b$$

for some integers $x_{n-1}, y_{n-1}, z_{n-1}, t_{n-1}$; this is the inductive hypothesis. The steps for the Euclidean algorithm say that $a_n = b_{n-1} = z_{n-1} a + t_{n-1} b$; and so you can set $x_n = z_{n-1}$ and $y_n = t_{n-1}$ and use the inductive hypothesis to show that the statement is true for a_n. It remains to show the result for b_n.

Rearranging the $(n-1)$th step of the Euclidean algorithm tells you that $r_{n-1} = a_{n-1} - b_{n-1} q_{n-1}$. As $b_n = r_{n-1}$ by Algorithm 1.2.3, you can say that $b_n = a_{n-1} - b_{n-1} q_{n-1}$. By the inductive hypothesis $a_{n-1} = x_{n-1} a + y_{n-1} b$ and $b_{n-1} = z_{n-1} a + t_{n-1} b$; so you can substitute these in to the expression for b_n to get:

$$b_n = x_{n-1} a + y_{n-1} b - q_{n-1}(z_{n-1} a + t_{n-1} b)$$

You can then rearrange this to get:

$$b_n = (x_{n-1} - q_{n-1} z_{n-1}) a + (y_{n-1} - q_{n-1} z_{n-1}) b$$

Setting $(x_{n-1} - q_{n-1} z_{n-1}) = z_n$ and $(y_{n-1} - q_{n-1} z_{n-1}) = t_n$ shows that the result holds. □

There are other and arguably nicer-looking proofs of Bézout's lemma (Corollary 1.2.7). However, the reason this one has been included is that it demonstrates a method to find integers x, y such that $ax + by = \gcd(a, b)$. The next example shows this method in action.

Example 1.2.8. It was shown in Example 1.2.4 that $\gcd(1107, 496) = 1$. In fact, you can reverse the Euclidean algorithm given in Example 1.2.4 to find integers x, y such that

$1 = 1107x + 496y$. Here is a picture of the end of this process:

$$\begin{aligned} 1 &= 36 - 7 \cdot 5 \\ &= 36 - (115 - 36 \cdot 3) \cdot 5 && = 36 \cdot 16 - 115 \cdot 5 \\ &= (496 - 115 \cdot 4) \cdot 16 - 115 \cdot 5 && = 496 \cdot 16 - 115 \cdot 69 \\ &= 496 \cdot 16 - (1107 - 496 \cdot 2) \cdot 69 && = 496 \cdot 154 - 1107 \cdot 69 \end{aligned}$$

and so $x = -69$ and $y = 154$.

In fact, for positive integers a, b, there is a way to compute x, y such that $\gcd(a, b) = xa + by$ **at the same time** as doing the Euclidean algorithm to find $\gcd(a, b)$. This method is known as the **extended Euclidean algorithm**; a picture of the end of this process is given below for $a = 903$ and $b = 444$ (as in Example 1.2.4).

$$\begin{array}{rcr|rcr} a & = & 903 & & & \\ 2b & = & 888 & 444 & = & b \\ \hline -2b + a & = & 15 & 435 & = & -58b + 29a \\ \hline 59b - 29a & = & 9 & 9 & = & 59b - 29a \\ \hline -61b + 30a & = & 6 & 6 & = & -61b + 30a \\ \hline & & 6 & 3 & = & 120b - 59a \\ \hline & & 0 & & & \end{array}$$

and so $120 \cdot 444 - 59 \cdot 903 = 3 = \gcd(903, 444)$.

The final result of the section is a nice consequence of Bézout's lemma (Corollary 1.2.7)

Corollary 1.2.9. *Let r, s, t be integers. If $\gcd(r, s) = 1$ and $r \mid st$ then $r \mid t$.*

Proof. By Corollary 1.2.7, there exists integers x and y such that $xr + ys = 1$. By Definition 1.1.3, $r \mid st$ means that there exists an integer u such that $ru = st$. Putting these together gives:

$$t = t \cdot 1 = t(xr + ys) = xrt + yst$$

As $ru = st$, you can write

$$t = xrt + yst = xrt + yru = r(xt + yu)$$

and so there exists an integer $xt + yu$ such that $r(xt + yu) = t$; therefore $r \mid t$ by

Definition 1.1.3. □

1.3 Primes

1.3.1 The Fundamental Theorem of Arithmetic

Definition 1.3.1. An integer $p > 1$ is called a **prime number** if its only positive divisors are 1 and p.

Example 1.3.2. The first few primes are: $2, 3, 5, 7, 11, 13, 17, 19, 23, 29, 31, 37, 41, \ldots$

The largest currently known prime is $2^{77232917} - 1$ (Jonathan Pace of GIMPS, December 2017). It has over 23 million decimal digits and ends in a 1.

Primes play an important role in dividing numbers in the sense that every number can be written as a unique product of primes. The next result outlines some basic properties of primes dividing products of integers, in anticipation of proving the **Fundamental Theorem of Arithmetic** (Theorem 1.3.4).

Lemma 1.3.3. *Let p be a prime number. Then the following statements are true:*

(i) Let a, b be integers. If $p \mid ab$ then $p \mid a$ or $p \mid b$.

(ii) Let $a_1, a_2, ..., a_s$ be integers. If $p \mid a_1 a_2 \ldots a_s$ then $p \mid a_i$ for some i.

(iii) If $p \mid q_1 q_2 \ldots q_t$ and each q_i is a prime, then $p = q_j$ for some j.

Proof. (i) If $p \mid a$ then there is nothing to prove. If $p \nmid a$ then as the only divisors of p are 1 and p, it follows that $\gcd(p, a) = 1$. You can then say that $p \mid b$ by Corollary 1.2.9.

(ii) You can prove this by induction on s, the length of the product.

Base case: If $s = 1$ there is nothing to prove. If $s = 2$, the statement is exactly the same as part (i).

Inductive case: For $s > 2$, assume that the statement holds for products of length $r < s$; this is the inductive hypothesis. By (i), it follows that $p \mid a_1 a_2 \ldots a_{s-1}$ or $p \mid a_s$; you need to consider both cases here. In the first case, $p \mid a_i$ for $i = 1, ..., s-1$ by the inductive hypothesis. In the second case, $p \mid a_s$ and so you are done.

(iii) As each q_i is prime it has no divisors other than 1 and q_i. The result then follows from part (ii). □

The result that every positive integer can be written as a unique product of primes is one of the most important results in mathematics.

Theorem 1.3.4 (Fundamental Theorem of Arithmetic)**.** *Every integer $n > 1$ can be written in the form*

$$n = p_1^{k_1} p_2^{k_2} \dots p_r^{k_r},$$

where $p_1 < p_2 < \dots < p_r$ are primes and all k_j are positive integers. This product of primes is unique.

Proof. Here, you will need to show that there **exists** a product of primes $p_1 < p_2 < \dots < p_r$ such that $n = p_1^{k_1} p_2^{k_2} \dots p_r^{k_r}$, and that this stated product of primes is **unique**.

You can use induction on $n > 1$ to show that there exists such a product of primes.

Base case: $(n = 2)$ If $n = 2$ then $n = 2^1$; as 2 is prime, the base case is proved.

Inductive case: $(n = k)$ Assume that the statement of the theorem holds for all $1 < r \leq k$; the idea is to prove the statement for the integer $k + 1$. There are two cases to consider.

- If $k + 1$ is a prime then the statement holds.
- If $k+1$ is not prime then it can be written as $k+1 = st$, for two integers $1 < s, t < k$. Using the inductive hypothesis, both s and t can be expressed as a product of primes; therefore, so can $k + 1$.

It remains to show uniqueness of this product of primes. To do this, assume that n can be expressed as a product of prime powers in two different ways:

$$n = p_1^{k_1} p_2^{k_2} \dots p_r^{k_r} = q_1^{l_1} q_2^{l_2} \dots q_t^{l_t}. \tag{1.1}$$

It is important to notice that each of the primes could be different, the powers could be different, and the length of the product could be different. Here,

$$p_i \mid n = q_1^{l_1} q_2^{l_2} \dots q_t^{l_t},$$

so you can say that $p_i = q_j$ for some j by Lemma 1.3.3 (iii). Similarly, each q_l is equal to some p_m, and so

$$r = t,\ p_1 = q_1, \dots,\ p_r = q_r.$$

So the length of products are the same and the primes are the same; the only thing you

now need to check are the values of the powers. Here, Equation 1.1 becomes

$$p_1^{k_1} p_2^{k_2} \dots p_r^{k_r} = p_1^{l_1} p_2^{l_2} \dots p_r^{l_r}. \tag{1.2}$$

Assume for a contradiction that $k_i \neq l_i$ for some $i = 1, ..., r$. One of these is therefore larger than the other; so you can take $k_i > l_i$ without loss of generality. Dividing both sides of Equation 1.2 by $p_i^{l_i}$ gives:

$$p_1^{k_1} p_2^{k_2} \dots p_{i-1}^{k_{i-1}} p_i^{k_i - l_i} p_{i+1}^{k_{i+1}} \dots p_r^{k_r} = p_1^{l_1} p_2^{l_2} \dots p_{i-1}^{l_{i-1}} p_{i+1}^{l_{i+1}} \dots p_r^{l_r}.$$

By using Lemma 1.3.3 (iii), you can say that p_i is equal to one of $p_1, \dots, p_{i-1}, p_{i+1}, \dots, p_r$; which is a contradiction as all these primes are meant to be different. So the original assumption was wrong; therefore $k_i = l_i$ for all $i = 1, ..., r$. This means that the expressions for n given in Equation 1.1 are the same and the proof is complete. □

Remark. Following this theorem, the expression

$$n = p_1^{k_1} p_2^{k_2} \dots p_r^{k_r},$$

where $p_1 < p_2 < \dots < p_r$ are primes and all k_j are positive integers is known as the **unique prime decomposition of** n.

Example 1.3.5. For an example of a 'unique' prime decomposition, $180 = 2^2 \cdot 3^2 \cdot 5$. You could also write $180 = 5 \cdot 3^2 \cdot 2^2$ as well, changing the order of the product, but here the primes are not written in ascending order as Theorem 1.3.4 requires.

It is possible to use the unique prime decomposition of an integer n to find all the divisors of n. Here, you can take integers m, n such that $m \mid n$ and write

$$n = p_1^{k_1} p_2^{k_2} \dots p_r^{k_r}$$

and

$$m = q_1^{l_1} q_2^{l_2} \dots q_t^{l_t}.$$

By Lemma 1.3.3 (iii) each q_i is equal to some p_j, and that $l_i \leq k_j$ in this case. So the divisors of n are precisely the numbers of the form

$$p_1^{s_1} p_2^{s_2} \dots p_r^{s_r},$$

where $0 \leq s_i \leq k_i$, $i = 1, \dots, r$.

Example 1.3.6. It was shown in Example 1.3.5 that the prime factorisation of 180 is $2^2 \cdot 3^2 \cdot 5$. You can use this to write down all the divisors of 180, which are:

$$\begin{array}{llllll} 1 = 2^0 \cdot 3^0 \cdot 5^0 & 9 = 2^0 \cdot 3^2 \cdot 5^0 & 15 = 2^0 \cdot 3^1 \cdot 5^1 \\ 2 = 2^1 \cdot 3^0 \cdot 5^0 & 18 = 2^1 \cdot 3^2 \cdot 5^0 & 30 = 2^1 \cdot 3^1 \cdot 5^1 \\ 4 = 2^2 \cdot 3^0 \cdot 5^0 & 36 = 2^2 \cdot 3^2 \cdot 5^0 & 60 = 2^2 \cdot 3^1 \cdot 5^1 \\ 3 = 2^0 \cdot 3^1 \cdot 5^0 & 5 = 2^0 \cdot 3^0 \cdot 5^1 & 45 = 2^0 \cdot 3^2 \cdot 5^1 \\ 6 = 2^1 \cdot 3^1 \cdot 5^0 & 10 = 2^1 \cdot 3^0 \cdot 5^1 & 90 = 2^1 \cdot 3^2 \cdot 5^1 \\ 12 = 2^2 \cdot 3^2 \cdot 5^0 & 20 = 2^2 \cdot 3^0 \cdot 5^1 & 180 = 2^2 \cdot 3^2 \cdot 5^1. \end{array}$$

1.3.2 Properties of primes

The Fundamental Theorem of Arithmetic (Theorem 1.3.4) describes the importance of prime numbers in number theory. Consequently, prime numbers are an area of significant interest for mathematicians both throughout history and today. Many deep results have been proved about prime numbers, often involving surprising and eye-catching methods. On the other hand, there are many questions that are easy to state but have not yet been answered.

The first result of this chapter was known by the ancient world, and appears in Euclid's *Elements*.

Theorem 1.3.7 (Euclid)**.** *There are infinitely many primes.*

Proof. Euclid's proof of this statement is by contradiction. Assume for a contradiction that $p_1, p_2, \ldots, p_n$ are **all** the prime numbers that exist. Now, consider the number

$$P = p_1 p_2 \ldots p_n + 1.$$

By the Fundamental Theorem of Arithmetic (Theorem 1.3.4), P is divisible by some prime p_i in the list of all primes. But then

$$p_i \mid P - p_1 p_2 \ldots p_n = 1$$

by Theorem 1.1.4 (vi). This is a contradiction as this implies $p_i = \pm 1$ by Theorem 1.1.4 (ii). So the original assumption that there are finitely many primes is wrong; meaning that there are infinitely many primes. □

You can adapt this argument to show that there are infinitely many primes of a certain form.

For instance, every prime number greater than 2 must have one of the forms $4k+1$ or $4k+3$ (because the numbers of the form $4k$ and $4k+2$ are even), but it is not immediately clear if there are infinitely many primes of **each** of these forms. For instance, there could be finitely many primes of the form $4k+3$ and infinitely many of the form $4k+1$. The next result takes care of one of these cases.

Theorem 1.3.8. *There are infinitely many primes of the form* $4k+3$.

Proof. The proof is similar to that of Theorem 1.3.7. Assume for a contradiction that $p_1, p_2, \ldots, p_n$ are all the primes of the form $4k+3$. First, write

$$N = 4p_1p_2 \ldots p_n - 1$$

As $-1 = -4+3$, you can substitute this in and factorise to get

$$N = 4p_1p_2 \ldots p_n - 1 = 4p_1p_2 \ldots p_n - 4 + 3 = 4(p_1p_2 \ldots p_n - 1) + 3.$$

So N can be written in the form $4n+3$ for some integer n. At this stage, it is important to note that any product of numbers of the form $4k+1$ again has that form; for instance

$$(4k+1)(4l+1) = 4(4kl + k + l) + 1.$$

This fact means that any product M of primes of the form $4k+1$ must have the form $M = 4m+1$. As $N = 4n+3$ for some integer n, this means that N must have at least one prime divisor of the form $4k+3$; say p_i. But then as p_i divides both N and $4p_1p_2 \ldots p_n$, it follows from Theorem 1.1.4 (vi) that

$$p_i \mid 4p_1p_2 \ldots p_n - N = 4p_1p_2 \ldots p_n - (4(p_1p_2 \ldots p_n - 1) + 3) = 1$$

which is a contradiction as $p_i \neq \pm 1$. □

Remark. You can modify the above argument to show that there are infinitely many primes of the form $6k+5$ (see tutorial sheet).

In fact, there are also infinitely many primes of the forms $4k+1$ and $6k+1$, but this is harder to prove. This is because the product of primes of the form $4k+3$ (for instance) may have the form $4k+1$. For example, $3 = 4 \cdot 0 + 3$ and $7 = 4 \cdot 1 + 3$, but $3 \cdot 7 = 21 = 4 \cdot 5 + 1$. So you cannot adapt the proof of Theorem 1.3.8 directly to prove there are infinitely many primes of this form.

However, all of these results are special cases of the following theorem proved by Dirichlet:

Theorem 1.3.9 (Dirichlet 1837)**.** *If a and b are coprime (see Definition 1.2.1) positive integers then there are infinitely many primes of the form $ak + b$ ($k = 0, 1, 2, \ldots$).*

Proof. Sadly not covered by the ideas in this course, requiring calculus(!) and analytic number theory. □

Remark. Note that this result does not claim that *every* number of the form $6k + 1$ (for instance) is prime; a counterexample is $6 \cdot 4 + 1 = 25 = 5 \cdot 5$.

There is no known simple formula which would give only prime numbers. For some time mathematicians believed that the polynomial

$$f(n) = n^2 + n + 41$$

is such a formula, having been checked for $n = 1, 2, \ldots 39$; all of which turned out to be prime. But

$$f(40) = 40^2 + 40 + 41 = 40 \cdot 41 + 41 = 41^2$$

is not prime.

This extends to the following theorem.

Theorem 1.3.10. *There is no non-constant polynomial $f(n)$ with integer coefficients which takes on only prime values for all non-negative integers n.*

Proof. As with every other theorem in this section that has been proved so far, the proof is by contradiction. Assume for a contradiction that

$$f(n) = a_k n^k + \ldots + a_1 n + a_0$$

is a polynomial such that $f(n)$ is prime for all $n \geq 0$.. Then $f(0) = a_0$ is a prime, and so is

$$f(ta_0) = a_k a_0^k t^k + \ldots + a_1 a_0 t + a_0$$

for all $t = 1, 2, \ldots$ by assumption. As a_0 divides every term of the above expression, $a_0 \mid f(ta_0)$ by Theorem 1.1.4 (vi). Since $f(ta_0)$ is assumed to be prime, it follows that $f(ta_0) = a_0$ for all $t = 1, 2, \ldots$. Therefore the polynomial $f(n)$ takes the value a_0 infinitely many times, and so it must be the constant polynomial with that value. This is a contradiction as $f(n)$ was assumed to be non-constant; so there is no such polynomial f. □

It is therefore not very easy to generate prime numbers. You can ask a different question then; what about the **distribution** of prime numbers? There are many more results about this question, one of which we give as the next theorem. You will need the result that for all natural numbers n:

$$1 + 2 + \ldots + 2^{n-1} = 2^n - 1$$

in the proof; this can be proved by induction. (Try it!)

Theorem 1.3.11. *If p_n is the nth prime then*

$$p_n \leq 2^{2^{n-1}}.$$

Proof. The proof is by induction on the number of primes n. So here, $p_1 = 2$, $p_2 = 3$ and so on.

Base case: If $n = 1$, then

$$p_1 = 2 \leq 2^{2^{1-1}} = 2$$

so the statement is true.

Inductive case: Assume that the statement holds for p_n; this is the inductive hypothesis. The aim is to prove that the statement holds for p_{n+1}. The inductive case relies on the proof of Theorem 1.3.7; as $p_1 p_2 \ldots p_n + 1$ is not divisible by any of $p_1, p_2, \cdots p_n$, it follows that $p_{n+1} \leq p_1 p_2 \ldots p_n + 1$ as it must have a prime divisor by the Fundamental Theorem of Arithmetic Theorem 1.3.4. By this and the inductive hypothesis, you can write that

$$\begin{aligned} p_{n+1} &\leq p_1 p_2 \ldots p_n + 1 \\ &\leq 2 \cdot 2^2 \cdot \ldots \cdot 2^{2^{n-1}} + 1 \end{aligned}$$

Grouping the powers together and using the result that $1 + 2 + \ldots + 2^{n-1} = 2^n - 1$ gives:

$$p_{n+1} \leq 2 \cdot 2^2 \cdot \ldots \cdot 2^{2^{n-1}} + 1 = 2^{1+2+\ldots+2^{n-1}} + 1 = 2^{2^n - 1} + 1$$

Finally, as $1 \leq 2^{2^{n-1}}$ for all natural numbers n, you can write that

$$p_{n+1} \leq 2^{2^n - 1} + 1 \leq 2^{2^n - 1} + 2^{2^n - 1} = 2^{2^n}$$

completing the proof. □

Whilst this result is true, it's not a very good bound for the nth prime number, and so not a good estimation of distribution. For instance, it says that the 5th prime number (which

is 11) is less than $2^{2^{5-1}} = 2^{16} = 65536$.

A much stronger (and harder to prove) result is that

$$\lim_{n\to\infty} \frac{n \log n}{p_n} = 1.$$

which says (very roughly) that the occurrence of the nth prime is roughly nearby $n \log n$. This is one of the equivalent formulations of the famous **Prime Number Theorem**; proved in 1896 by Hadamard and de la Vallée Poussin.

Primes are fundamental to number theory, yet not every question about them has been answered. Some famous open problems regarding the prime numbers are:

Goldbach conjecture : Is it true that every even number greater than two can be written as the sum of two prime numbers?

Twin prime conjecture : Is it true that there are infinitely many primes p such that $p+2$ is a prime as well?

Mersenne prime conjecture : Are there infinitely many primes of the form $2^n - 1$?

In some sense the Prime Number Theorem and the above questions (in case of positive answers) say that there are many prime numbers and that they occur regularly. Here is a result which seems to say the opposite.

Theorem 1.3.12. *For every $n > 0$ there is a sequence of n consecutive* composite *numbers.*

Proof. Take the sequence $(n+1)! + 2, (n+1)! + 3, \ldots, (n+1)! + (n+1)$; then for $i = 2, ..., (n+1)$, it follows that $i \mid (n+1)! + i$ by Theorem 1.1.4 (vi). □

1.4 Congruences and modular arithmetic

Definition 1.4.1. Let n be an integer with $n > 1$. Say that two integers a and b are **congruent modulo** n if $a - b$ is divisible by n. This is written as $a \equiv b \pmod{n}$.

This means that $a \equiv b \pmod{n}$ if and only if $n \mid (a - b)$. This definition is quite difficult to handle; so the following equivalent definition is also often used.

Theorem 1.4.2. *Let a, b be integers and take another integer $n > 1$. Then $a \equiv b \pmod{n}$ if and only if a and b have the same remainder r on dividing by n.*

Proof. Suppose that a and b have the same remainder r on division by n. By Fact 1.1.1, this means that there exist q, q' such that $a = qn + r$ and $b = q'n + r$. Then you can write that

$$a - b = qn + r - (q'n + r) = qn - q'n = (q - q')n$$

and so $n \mid (a - b)$ by Definition 1.1.3.

Now, assume that $n \mid (a - b)$. By Definition 1.1.3, there exists an integer k such that $nk = (a - b)$; you can rearrange this to get $a = b + kn$. You can use Fact 1.1.1 to say that $b = qn + r$ for some q and r. Substituting this into $a = b + kn$ gives

$$a = b + kn = qn + r + kn = (q + k)n + r$$

and so a has the same remainder as b on division by n. □

Theorem 1.4.2 means that two integers a, b are congruent modulo n if they have the same remainder upon divison by n. Importantly, if $a = qn + r$ where $0 \leq r < n$, then $a \equiv r \pmod{n}$. Here are a few examples that demonstrate this fact.

Example 1.4.3. As $14 = 2 \cdot 5 + 4$, then $14 \equiv 4 \pmod{5}$. As $-21 = -3 \cdot 9 + 6$ and $60 = 6 \cdot 9 + 6$, it follows that $-21 \equiv 60 \pmod{9} \equiv 6 \pmod{9}$. It is very important to notice that $21 \equiv 3 \pmod{9}$. As $7 = 1 \cdot 4 + 3$, it follows that $7 \not\equiv 1 \pmod{4}$.

As with the definition of divisor in Section 1.1, the idea now is to use Definition 1.4.1 (and Theorem 1.4.2) to obtain results about congruence modulo n. The first results follow from Definition 1.4.1 and Theorem 1.4.2.

Corollary 1.4.4. *(1) Let a, b be integers and take another integer $n > 1$. Then $a \equiv 0 \pmod{n}$ if and only if $n \mid a$.*
(2) Let a, b be integers. Then $a \equiv b \pmod{10}$ if and only if the decimal expansions of a and b end in the same digit r.

Proof. The proof of this is in the tutorial sheet. These are both if and only if statements, so you need to prove both the forward and the converse directions of each statement.

□

The second property of congruence modulo n says that if $a \equiv b \pmod{n}$, then a and b are 'somehow the same'. This shows that congruence modulo n has some similarity to equality.

Theorem 1.4.5. *Suppose that a, b, c are integers and $n > 1$ is a positive integer. Then the following hold:*

(1) $a \equiv a \ (mod\ n)$;

(2) if $a \equiv b \ (mod\ n)$, *then* $b \equiv a \ (mod\ n)$, *and;*

(3) if $a \equiv b \ (mod\ n)$ *and* $b \equiv c \ (mod\ n)$, *then* $a \equiv c \ (mod\ n)$.

Remark. These properties have names: property (1) is known as **reflexivity**, property (2) is known as **symmetry**, and property (3) is known as **transitivity**. These properties are the defining feature of an **equivalence relation**, and you can find out more about these in Chapter 2.

Proof. (1) As $a - a = 0$, and $n \mid 0$ for any $n > 1$, it follows that $n \mid a - a$ and so $a \equiv a \pmod{n}$ by Definition 1.4.1.

(2) Suppose that $a \equiv b \pmod{n}$; by Definition 1.4.1, this means that $n \mid (a - b)$. By Definition 1.1.3, there exists an integer k such that $a - b = kn$. As $b - a = -(a - b)$, you can write that

$$b - a = -(a - b) = -(kn) = -kn$$

and so $n \mid b - a$. This means that $b \equiv a \pmod{n}$ by definition.

(3) Assume that $a \equiv b \pmod{n}$ and $b \equiv c \pmod{n}$; by Definition 1.4.1, this means that $n \mid (a - b)$ and $n \mid (b - c)$. By Theorem 1.1.4 (vi), it follows that

$$n \mid (a - b) + (b - c) = (a - c)$$

and so $a \equiv c \pmod{n}$. □

The third property of congruence modulo n is perhaps the most important of all; you can perform addition and multiplication modulo n. This is known as **modular arithmetic**, and is an important tool in many areas of mathematics.

Theorem 1.4.6 (Modular arithmetic)**.** *Suppose that $n > 1$ is an integer, and let a, b, c, d, k be integers with $k \geq 0$.*

(i) If $a \equiv b \ (mod\ n)$ *and* $c \equiv d \ (mod\ n)$ *then*

$$a + c \equiv b + d \ (mod\ n) \quad \textit{and} \quad ac \equiv bd \ (mod\ n)$$

(ii) If $a \equiv b \ (mod\ n)$ *then* $a^k \equiv b^k \ (mod\ n)$.

Proof. (i) By the assumptions, it follows that $n \mid (a-b)$ and $n \mid (c-d)$. By Theorem 1.1.4 (vi), it follows that

$$n \mid (a-b)+(c-d) = (a+c)-(b+d)$$

and so $n \mid (a+c)-(b+d)$. Therefore $(a+c) \equiv (b+d) \pmod{n}$ by Definition 1.4.1. Furthermore, as b, c are integers, it follows from Definition 1.1.3 that $n \mid (a-b)c$ and $n \mid (c-d)b$. By Theorem 1.1.4 (vi), it follows that

$$n \mid (a-b)c + (c-d)b$$

You can expand the brackets and cancel remaining terms to get

$$\begin{aligned} n \mid (a-b)c + (c-d)b &= ac - bc + bc - bd \\ &= ac - bd \end{aligned}$$

and so $ac \equiv bd \pmod{n}$ by Definition 1.4.1.

(ii) The proof is by induction on k. Assume that $a \equiv b \pmod{n}$.

Base case: If $k = 1$ then $a^1 \equiv b^1 \pmod{n}$ which is the same as saying $a \equiv b \pmod{n}$.

Inductive case: Assume that the statement holds for $a^k \equiv b^k \pmod{n}$. You can use part (i) of this theorem with $a = c$ and $b = d$ to demonstrate that

$$a^k \cdot a \equiv b^k \cdot b \pmod{n}$$

and so $a^{k+1} \equiv b^{k+1} \pmod{n}$. □

Modular arithmetic is incredibly useful in a number of ways, as will be demonstrated following this corollary. This important result generalises an observation made in Example 1.4.3.

Corollary 1.4.7. *Suppose that* $a \equiv r \ (mod\ n)$, *where* $0 \leq r < n$. *Then* $-a \equiv -r \ (mod\ n)$ *and so* $-a$ *is congruent to* $n - r$ *modulo* n.

Proof. You can see that for any $n > 1$ that $-1 = 0 \cdot n - 1 \equiv -1 \pmod{n}$. Setting $c = d = -1$ and using Theorem 1.4.6 (i) gives

$$a \cdot (-1) \equiv r \cdot (-1) \pmod{n}$$

and so $-a \equiv -r \pmod{n}$.

As $n \mid n$, it follows that $n \equiv 0 \pmod{n}$ by Corollary 1.4.4 (1). This means that $0 \equiv n \pmod{n}$ by Theorem 1.4.5 (2). Therefore, using Theorem 1.4.6 (i) gives that

$$-a = 0 - a \equiv n - r \pmod{n}$$

as required. □

Remark. By multiplying through the second statement of Corollary 1.4.7 by -1, you can say that $a \equiv -n + r \pmod{n}$, where $0 \leq r < n$. This can be very useful as demonstrated in the following examples.

Example 1.4.8. You can use the techniques of modular arithmetic to work out 517 modulo 7. You know that $517 = 51 \cdot 10 + 7$, and that $51 \equiv 2 \pmod{7}$. You can then say that

$$517 = 51 \cdot 10 + 7 \equiv (2)(10) + 7 \equiv 27 \equiv 6 \pmod{7}$$

You can use modular arithmetic to show that an unreasonably large number is divisible by a smaller number by using Corollary 1.4.4 (i). For instance, let's show that $2^{70} + 3^{70}$ is divisible by 13. The way to do this is to break down the powers of the two terms 2^{70} and 3^{70}. By looking for a power of 2 or 3 that is congruent to 1 or -1 modulo 13, you can use Theorem 1.4.6 (ii) to raise 1 or -1 to a huge power modulo 13; which will either be 1 or -1. So here,

$$2^6 = 64 \equiv -1 \pmod{13}$$

and so you can use Theorem 1.4.6 (ii) to say that

$$2^{66} = (2^6)^{11} \equiv (-1)^{11} \equiv -1 \pmod{13}$$

As $2^{70} = 2^{66} \cdot 2^4$, and $2^4 = 16 \equiv 3 \pmod{13}$, you can use Theorem 1.4.6 (i) to write that

$$2^{70} = 2^{66} \cdot 2^4 \equiv (-1) \cdot 3 \equiv -3 \pmod{13}$$

Now you can move onto 3^{70}. Here, $3^3 = 27 \equiv 1 \pmod{13}$ and using Theorem 1.4.6 (ii) gives

$$3^{69} = (3^3)^{23} \equiv (1)^{23} \equiv 1 \pmod{13}$$

and so

$$3^{70} = 3 \cdot 3^{69} \equiv 3 \cdot 1 \equiv 3 \pmod{13}$$

Finally, you can use Theorem 1.4.6 (i) to write that

$$2^{70} + 3^{70} \equiv -3 + 3 \equiv 0 \pmod{13}$$

and so $2^{70} + 3^{70}$ is divisible by 13 by Corollary 1.4.4 (i).

Finally, you can use modular arithmetic to work out the last digit of a absurdly big number. This is done by finding the value modulo 10 and using Corollary 1.4.4 (ii). For instance, let's find the last digit of 3^{1729}. You can use the same technique as above; find a power of 3 that is congruent to 1 or -1 modulo 10, and then raise this again to get near 3^{1729}. So here, $3^2 = 9 \equiv -1 \pmod{10}$ and so you can write

$$3^{1728} = (3^2)^{864} \equiv (-1)^{864} \equiv 1 \pmod{10}$$

This means that

$$3^{1729} = 3 \cdot 3^{1728} \equiv 3 \cdot 1 \equiv 3 \pmod{3}$$

and so the last digit of 3^{1729} is 3.

There is a nice way to tell whether or not a number a is divisible by 3 or 9. You can add together the digits of a to get the **digital root** of a. Then if the digital root of a is divisible by 3 (or 9), then a is divisible by 3 (or 9). This fact follows from the following result:

Theorem 1.4.9. *Every positive integer a is congruent to the digital root of a modulo 9.*

Proof. First, you can see that $10 \equiv 1 \pmod{9}$. So Theorem 1.4.6 (ii) means that

$$10^k \equiv 1^k \equiv 1 \pmod{9}$$

Using positional notation (Theorem 1.1.6), you can write $a = \overline{(d_n d_{n-1} \dots d_1 d_0)}_{10}$ and so

$$a = d_n \cdot 10^n + d_{n-1} \cdot 10^{n-1} + \ldots + d_1 \cdot 10^1 + d_0 \cdot 10^0$$

Reducing modulo 9 gives

$$\begin{aligned} a &= d_n \cdot 10^n + d_{n-1} \cdot 10^{n-1} + \ldots + d_1 \cdot 10^1 + d_0 \cdot 10^0 \\ &\equiv d_n \cdot 1 + d_{n-1} \cdot 1 + \ldots + d_1 \cdot 1 + d_0 \cdot 1 \pmod{9} \\ &\equiv d_n + d_{n-1} + \ldots + d_1 + d_0 \pmod{9} \end{aligned}$$

and this is the digital root of a. □

Remark. You can use a similar method to prove the same theorem for modulo 3.

Using Theorem 1.4.9 together with Corollary 1.4.4 (i) gives the last result of the section.

Corollary 1.4.10. *A positive integer a is divisible by 9 if and only if the digital root of a is divisible by 9.* □

Remark. The same result holds for 3 in place of 9.

1.5 Linear Diophantine equations

A **Diophantine equation** is an equation involving integer coefficients. If an equation is called a Diophantine equation, then you should always be searching for **integer** solutions to the equation. Here is a Diophantine equation with one variable x:

$$3x + 5 = 14$$

This is quite boring, with solution $x = 3$. Diophantine equations get much more interesting when you consider two variables.

Example 1.5.1. You are given the equation

$$x + y = 3$$

Now this is more interesting! For every choice of integer x, there is a unique solution for y; this is $y = 3 - x$. This means that the equation has infinitely many solutions. All of the solutions can be expressed as a set of pairs of integers (x, y). So you can write

$$S_1 = \{(x, 3 - x) \ : \ x \in \mathbb{Z}\}$$

as the set of solutions to the Diophantine equation $x + y = 3$.

So a Diophantine equation in two variables can have infinitely many solutions. Here is another Diophantine equation, but this time there is more to be thought about here.

Example 1.5.2. Now, say you are given the equation

$$x + 2y = 5$$

Here, a solution to x must be an odd number, as $2y$ is an even number. So if x is an even number, then this equation has no solutions This means that you can't write a general solution in terms of any integer x. However, what you can do is say that for any choice of y there is always a solution for x; this is $x = 5 - 2y$. So this equation does have infinitely many solutions, with the solution set

$$S_2 = \{(5 - 2y, y) \ : \ y \in \mathbb{Z}\}$$

Is it always true that a Diophantine equation in two variables has infinitely many solutions? The next example says that it is not true.

Example 1.5.3. Now, consider the equation

$$5x + 10y = 3$$

This equation has no solutions. This is because the left-hand side is always a multiple of 5, and the right-hand side is **not** a multiple of 5.

These examples seem to imply that the existence of solutions for Diophantine equations in two variables rest on the properties of the coefficients. To examine this in more detail, it would be useful to have a formal definition of a Diophantine equation in two variables.

Definition 1.5.4. A **linear Diophantine equation in two variables** is an equation of the form

$$ax + by = c \tag{1.3}$$

where a, b, c are integers with $a, b \neq 0$.

You can ask the following questions of a linear Diophantine equation in two variables:

- What values of a, b, c mean that Equation 1.3 has integer solutions?
- If Equation 1.3 has integer solutions, how many pairs of integer solutions does it have?
- Is there an expression for all of the solutions to Equation 1.3?

You can guess that Example 1.5.3 means that the common divisors of a and b have some relevance.

Is there a solution?

For ease of use, here is Equation 1.3 again:

$$ax + by = c \tag{1.3}$$

where $a, b, c \in \mathbb{Z}$ and $a, b \neq 0$. The claim is the following.

Claim. Equation 1.3 has a solution if and only if $\gcd(a, b)$ divides c.

You can see here that this is an if and only if statement; so you will need to prove both directions of the result.

Proof. Suppose that Equation 1.3 has a solution; call this (x_0, y_0). This means that $ax_0 + by_0 = c$. Let $d = \gcd(a, b)$. As d divides both a and b by Definition 1.2.1, it divides $ax_0 + by_0 = c$ by Theorem 1.1.4 (vi). So $d \mid c$ and this direction of the proof holds.

Now assume that $d = \gcd(a, b)$ divides c; as this happens, there exists an integer c_1 such that $c = dc_1$. Now, Bézout's Lemma (Corollary 1.2.7) says that there exist integers u, v such that

$$d = au + bv$$

You can multiply this equation through by c_1 to get

$$dc_1 = auc_1 + bvc_1$$

As $dc_1 = c$, you can write that

$$c = a(uc_1) + b(vc_1)$$

and so (uc_1, vc_1) is an integer solution to Equation 1.3. □

How many solutions are there?

Suppose that Equation 1.3 does have an integer solution (x_0, y_0); so $ax_0 + by_0 = c$. As $d = \gcd(a, b)$ divides both a and b, this means that you can write

$$a = da_1 \quad \text{and} \quad b = db_1$$

for integers a_1, b_1. Now, for any integer t, write

$$x = x_0 + b_1 t \quad \text{and} \quad y = y_0 - a_1 t$$

You can substitute these into Equation 1.3 and simplify to get that

$$\begin{aligned} ax + by &= a(x_0 + b_1 t) + b(y_0 - a_1 t) \\ &= ax_0 + ab_1 t + by_0 - ba_1 t \\ &= (ax_0 + by_0) + (ab_1 t - ba_1 t) \end{aligned}$$

As $ax_0 + by_0 = c$, and since $a = da_1$ and $b = db_1$, you can write

$$\begin{aligned} ax + by &= (ax_0 + by_0) + (ab_1 t - ba_1 t) \\ &= (c) + (da_1 b_1 t - db_1 a_1 t) = c + 0 = c \end{aligned}$$

and so $(x_0 + b_1 t, y_0 - a_1 t)$ is a solution to Equation 1.3 **for any** $t \in \mathbb{Z}$. This means that if Equation 1.3 has at least one solution, then it has infinitely many solutions.

Have all the solutions been found?

If Equation 1.3 has a solution (x_0, y_0), then it follows from Section 1.5 that $d = \gcd(a, b)$ divides c. Furthermore, it has infinitely many solutions of the form $(x_0 + b_1 t, y_0 - a_1 t)$, where t is an integer and where $da_1 = a$ and $db_1 = b$. The question is; are these all of the solutions?

Suppose that (x_0, y_0) is a solution to Equation 1.3, and take (x, y) to be any other solution. This means that

$$ax_0 + by_0 = c = ax + by$$

Rearranging the equation and substituting $a = da_1$ and $b = db_1$ gives

$$da_1(x - x_0) = db_1(y - y_0)$$

You can cancel d from both sides to get that

$$a_1(x - x_0) = b_1(y - y_0) \tag{1.4}$$

As $d = \gcd(a, b)$, and $a_1 = a/d$ and $b_1 = b/d$, then $\gcd(a_1, b_1) = 1$ (see Tutorial Sheet 2). The above equation suggests that $a_1 \mid b_1(y - y_0)$; so as $\gcd(a_1, b_1)$, then Corollary 1.2.9

suggests that $a_1 \mid (y - y_0)$. This means that $a_1 t = (y - y_0)$ for some integer t. You can substitute this into Equation 1.4 to get

$$a_1(x - x_0) = b_1 a_1 t$$

Cancelling a_1 from both sides gives that $b_1 t = (x - x_0)$. Therefore

$$x = x_0 + b_1 t \quad \text{and} \quad y = y_0 - a_1 t$$

so all solutions to Equation 1.3 are of this form.

Collecting these results together give:

Theorem 1.5.5 (Solutions to linear Diophantine equations)**.** *Let $a, b, c \in \mathbb{Z}$, with $a, b \neq 0$.*

(i) The linear Diophantine equation with two variables

$$ax + by = c$$

has a solution if and only if $d = \gcd(a, b)$ divides c.

(ii) If d divides c, then a solution can be found by determining u, v such that $d = ua + vb$ and then setting

$$x_0 = uc/d \quad \textit{and} \quad y_0 = vc/d$$

(iii) All other solutions are given by

$$x = x_0 + (b/d)t \quad \textit{and} \quad y = y_0 - (a/d)t$$

for $t \in \mathbb{Z}$. □

Using Theorem 1.5.5 to get solutions to linear Diophantine equations depend on finding $d = \gcd(a, b)$ **and** if d does divide c, determining integers u, v such that $au + bv = c$. Therefore, this is a perfect opportunity to use the extended Euclidean algorithm as seen in Example 1.2.8.

Example 1.5.6. Suppose you are asked to find all solutions to the linear Diophantine equation

$$144x + 84y = 60 \tag{1.5}$$

Here. $a = 144, b = 84$ and $c = 60$. The idea is to find $\gcd(144, 84) = d$ and if d divides 60, find numbers u, v such that $144u + 84v = 60$. As mentioned above, you can use the extended Euclidean algorithm with $a = 144$ and $b = 84$ to get:

$$\begin{array}{rcl|rcl}
a & = & 144 & & & \\
b & = & 84 & 84 & = & b \\
\hline
-b + a & = & 60 & 60 & = & -b + a \\
4b - 2a & = & 48 & 24 & = & 2b - a \\
\hline
-5b + 3a & = & 12 & 24 & = & -10b + 6a \\
 & & & \hline
 & & & 0 & &
\end{array}$$

So here, $\gcd(144, 84) = 12$. As 12 does divide 60, there exist solutions to Equation 1.5 by Theorem 1.5.5 (i). By the working of the extended Euclidean algorithm above, you know that $3 \cdot 144 - 5 \cdot 84 = 12$; this means you can take $u = 3$ and $v = 5$. From Theorem 1.5.5 (ii), you can write

$$x_0 = uc/d = 3 \cdot \frac{60}{12} = 15 \quad \text{and} \quad y_0 = vc/d = -5 \cdot \frac{60}{12} = -25$$

You can use your values for x_0 and y_0, together with Theorem 1.5.5 (iii), to write expressions for all solutions to Equation 1.5:

$$\begin{aligned}
x &= x_0 + (b/d)t = 15 + (84/12)t = 15 + 7t \\
y &= y_0 - (a/d)t = -25 - (144/12)t = -25 - 12t
\end{aligned}$$

and these are for $t \in \mathbb{Z}$.

Example 1.5.7. Little Timmy Cauchy has been saving his pocket change for a blow-out trip to the world famous Cantor Confectionery sweet shop; the only place with an uncountable range of sweet tasting goodies! Anyway, Timmy spent all of his £1.32 on 12 items; a mixture of Fig Newton and Choco Leibniz biscuits. Each Fig Newton costs 3p more than a Choco Leibniz. As to his taste, Timmy buys more Fig Newtons than Choco Leibniz. How many of each item did Timmy buy?

This is a Diophantine equation in disguise. Let x be the number of Fig Newtons that Timmy bought; therefore, the number of Choco Leibniz bought is $12 - x$. You can use y to write the cost of a Fig Newton; so $y - 3$ is the cost of a Choco Leibniz. From the information

above, you can write that

$$xy + (12 - x)(y - 3) = 132$$

You can then expand out the brackets and rearrange this equation to get that

$$\begin{aligned} 132 &= xy + 12y - 36 - xy + 3x \\ 168 &= 3x + 12y \\ 56 &= x + 4y \end{aligned}$$

The solutions to this equation are therefore $x = 56 - 4t$ for $t \in \mathbb{Z}$. However, Timmy bought more Fig Newtons than Choco Leibniz; so here, $x > 6$ and $x < 12$. You can summarise this in an inequality, using the fact that $x = 56 - 4t$, to get

$$6 < 56 - 4t < 12$$

You can rearrange this, taking care with the minus signs, to get

$$44 < 4t < 50$$

and so

$$11 < t < \frac{50}{4}$$

The only integer that t can be here is 12. Therefore, you can write that $y = 12$ and $x = 56 - 4 \cdot 12 = 8$. This means that Timmy bought 8 Fig Newtons at 12p each and 4 Choco Leibniz at 9p each. Whether or not this reflects little Timmy's opinion on the origins of calculus is unknown.

Example 1.5.8. Cantor Confectionery is looking to export its tasty baked goods to supermarkets via their Continuum Courier service. It's looking to ship boxes of Euler Eclairs (costing £5 each) and Dedekind Donuts (costing £7 each). Boxes cannot be split (not even the Dedekind Donuts). You are running the export centre at Cantor Confectionery, and you would like to write a foolproof guide for supermarkets to follow. What values can supermarkets put aside to order these highly desirable goods?

Again, this is a Diophantine equation in disguise. Let e be the number of boxes of Euler Eclairs and d be the number of boxes of Dedekind Donuts. The question is; for what values of c does

$$5e + 7d = c$$

have **non-negative** solutions d, e? (Despite your encouragement, supermarkets cannot

order negative amounts of perfectly baked goods.) As $\gcd(5,7)=1$, you can say that the equation $5e+7d$ has solutions by Theorem 1.5.5; but some of these may be negative. You can follow the standard procedure (as in Example 1.5.6) to find solutions for Diophantine equations; by using the extended Euclidean algorithm on 5 and 7. You can then adjust the method to ensure that you only get non-negative solutions.

So here

$$\begin{array}{rcl|lcl} d & = & 7 & & & \\ e & = & 5 & 5 & = & e \\ \hline -e+d & = & 2 & 4 & = & -2e+2d \\ & = & 2 & 1 & = & 3e-2d \\ \hline & & 0 & & & \end{array}$$

So you can take $u=3$ and $v=-2$, as in Theorem 1.5.5. A solution to $5e+7d=c$ is therefore given by

$$x_0 = 3c \quad \text{and} \quad y_0 = -2c$$

You can then use Theorem 1.5.5 to write the general solution as

$$x = 3c-7t \quad \text{and} \quad y = -2c+5t$$

At this point, you can look for non-negative solutions. To do this, you need

$$3c-7t \geq 0 \quad \text{and} \quad -2c+5t \geq 0$$

You can make t the subject in both inequalities to get $t \leq 3c/7$ and $t \geq 2c/5$. So t must be between $2c/5$ and $3c/7$; this means that there exists at least one integer between $2c/5$ and $3c/7$. How big does c need to be to make this happen? You can find this out by finding the difference between the two bounds $2c/5$ and $3c/7$. So

$$\frac{3c}{7} - \frac{2c}{5} = \frac{15c-14c}{35} = \frac{c}{35}$$

This means that if c is greater than 35, the gap between $2c/5$ and $3c/7$ is greater than 1, and so there is guaranteed to be an integer between them. So non-negative solutions exist for $c \geq 35$.

You can conclude that any value of £35 or greater is enough to buy a decent mixture of Euler Eclairs and Dedekind Donuts. (Values smaller than £35 will have to be checked by hand... and Mr Cantor doesn't particularly fancy that job.)

The main point here is that d, e are coprime. If this happens, then there is always some point where all integers can be written as a non-negative sum of d, e.

Theorem 1.5.9. *Let a, b be coprime positive integers. Then every number $c \geq ab$ can be expressed as $xa + yb$ where x, y are integers greater than 0.*

Proof. The following proof has been kindly contributed by Aimee Bebbington (with minor editorial corrections).

> Here, $\gcd(a, b) = 1$. Solutions to the Diophantine equation $ax + by = c$ exist if $\gcd(a, b)$ divides c (by Theorem 1.5.5). 1 divides c for all integers c, therefore solutions always exist.
>
> We will set $\gcd(a, b) = d$.
>
> There are solutions of the form
>
> $$x_0 = x + (uc)/d = x + uc$$
>
> and
>
> $$y_0 = y - (vc)/d = y - vc$$
>
> as $d = 1$.
>
> All other solutions are given by
>
> $$x = uc + (b/d)t = uc + bt$$
>
> and
>
> $$y = vc - (a/d)t = vc - at$$
>
> For non-negative integer solutions,
>
> $$uc + bt \geq 0 \quad \text{and} \quad vc - at \geq 0$$
>
> We also know that $ua + bv = 1$.
>
> Rearranging gives
>
> $$(-uc)/b \leq t \leq (vc)/a$$
>
> To find integer solutions in this range, $[(vc)/a] + [(uc)/b] \geq 1$. Adding these gives

$$[c(bv + au)]/ab \geq 1$$

Then, since $bv + au = 1$, it follows that $c \geq ab$.

□

1.6 Higher order Diophantine equations

Theorem 1.5.5 details how to find all solutions for a given linear Diophantine equation in two variables $ax + by = c$, given that a solution exists. However, not all Diophantine equations behave this nicely. This final section on number theory considers **higher-order** Diophantine equations. These are equations where at least one variable is raised to a power of 2 or greater.

Pythagorean triples

Here is an example of a higher-order Diophantine equation:

$$x^2 + y^2 = z^2 \tag{1.6}$$

This is the equation underpinning Pythagoras' theorem. A solution (x, y, z) to this equation, where x, y, z are integers, corresponds to a right-angled triangle with integer side lengths. Since right-angled triangles are important, it would be nice to be able to find all integer solutions for this equation. These are known as **Pythagorean triples**.

Definition 1.6.1. Let x, y, z be integers. A solution (x, y, z) of Equation 1.6 is known as a **Pythagorean triple**. If $x, y, z > 0$ and $\gcd(x, y, z) = 1$, then a solution (x, y, z) of Equation 1.6 is known as a **primitive Pythagorean triple**.

Primitive Pythagorean triples play an important role in finding all Pythagorean triples. Suppose that (x, y, z) is a Pythagorean triple with $\gcd(x, y, z) = d > 1$. Then you can write that $x = dx_1, y = dy_1$ and $z = dz_1$. You can then substitute this into Equation 1.6 and get

$$(dx_1)^2 + (dy_1)^2 = (dz_1)^2$$
$$d^2(x_1^2 + y_1^2) = d^2 z_1^2$$

Dividing by d^2 gives $x_1^2+y_1^2=z_1^2$, where (from a result in a tutorial sheet) $\gcd(x_1,y_1,z_1)=1$. By Definition 1.6.1, (x_1,y_1,z_1) is a primitive Pythagorean triple, and $(x,y,z)=(dx_1,dy_1,dz_1)$ is an integer multiple of (x_1,y_1,z_1). What this means is that if you want find all Pythagorean triples, you only have to find all **primitive** Pythagorean triples, and then multiply through by some integer. Therefore, the main goal of the section is to find out how to construct a primitive Pythagorean triple, and show that. Before this though, here are three useful results.

Lemma 1.6.2. *Let $a,b,c,n\in\mathbb{N}$. If $ab=c^n$ and $\gcd(a,b)=1$, then both a and b are nth powers.*

Proof. Using the Fundamental Theorem of Arithmetic Theorem 1.3.4, you can write the prime decomposition of the natural number c:

$$c=p_1^{k_1}p_2^{k_2}\dots p_r^{k_r}$$

Raising this to the power of n gives:

$$ab=c^n=p_1^{nk_1}p_2^{nk_2}\dots p_r^{nk_r}$$

As $\gcd(a,b)$, each prime p_i in the prime decomposition of c divides exactly one of a or b. It follows then that both a and b are both nth powers of a product of prime powers. □

The next result uses Theorem 1.1.5; which states that any square number is either divisible by 4 or has remainder 1 when divided by 8. As you have studied congruences in Section 1.4,you can rephrase Theorem 1.1.5 to say that any square number is congruent to either $0\pmod 4$ or $1\pmod 8$.

Lemma 1.6.3. *If a and b are odd integers, then a^2+b^2 is not a square number.*

Proof. If a and b are both odd, then a^2 and b^2 are both odd; so $a^2\equiv 1\pmod 8$ and $b^2\equiv 1\pmod 8$ by Theorem 1.1.5. This means that $a^2+b^2\equiv 2\pmod 8$; as this happens, it must be that $a^2+b^2\equiv 2\pmod 4$. So a^2+b^2 is not a square number by Theorem 1.1.5. □

Corollary 1.6.4. *If (x,y,z) is a* primitive *Pythagorean triple, then exactly one of x and y is even and the other is odd. Therefore, z is also odd.*

Proof. As (x,y,z) is a primitive Pythagorean triple, then $\gcd(x,y,z)=1$. If both x and y were even, then z would also be even; so $\gcd(x,y,z)\geq 2$, which is a contradiction.

Lemma 1.6.3 states that both of x and y can't be odd. Therefore, either x or y is even and the other is odd. □

These three results provide a starting point for the aim of this section, which is to determine all primitive Pythagorean triples. To do this, you can first prove that every primitive Pythagorean triple can be expressed as some integer equations, and then showing that these integer equations always give a primitive Pythagorean triple as an answer. Doing this ensures that you have found all primitive Pythagorean triples.

Is every primitive Pythagorean triple the result of some equations?

To begin with, assume that (x, y, z) is a primitive Pythagorean triple. Corollary 1.6.4 says that only one of the three integers in a primitive Pythagorean triple is even. Since this can be either x or y, you can state without loss of generality that the even number is x. This means that y, z are both odd. From this, you can say that both $z + y$ and $z - y$ are even, and so there exist integers r, u such that

$$z + y = 2r \quad \text{and} \quad z - y = 2u \tag{1.7}$$

Adding these two expressions together gives $2z = 2r + 2u$, and so dividing through by 2 gives $z = r + u$. Subtracting $z + y = 2r$ from $z - y = 2u$ gives that $-2y = 2u - 2r$, and so dividing through by -2 gives $y = r - u$. This leads into a claim.

Claim. $\gcd(r, u) = 1$.

Proof. Let $d = \gcd(r, u)$. Assume for a contradiction that $d > 1$. By the Fundamental Theorem of Arithmetic (Theorem 1.3.4) there exists a prime p dividing d. As p divides d, p also divides r and u; so you can use Theorem 1.1.4 (vi) to say that

$$p \mid r + u = z \quad \text{and} \quad p \mid r - u = y$$

Now, Theorem 1.1.4 (iii) states that as $p \mid z$ then $p^2 \mid z^2$, and as $p \mid y$ then $p \mid y^2$. By Theorem 1.1.4 (vi), it follows that $p^2 \mid z^2 - y^2 = x^2$. By Lemma 1.3.3 (i), you can see that $p \mid x$ and so $\gcd(x, y, z) \geq p$. This is a contradiction since $\gcd(x, y, z) = 1$, as (x, y, z) is a primitive Pythagorean triple. □

You have expressions for y, z in terms of r and u; the idea now is to express x in terms of r and u. Equation 1.7 says that $z + y = 2r$ and $z - y = 2u$; dividing both of these

expressions by 2 gives

$$r = \frac{z+y}{2} \quad \text{and} \quad u = \frac{z-y}{2}$$

As $(z+y)(z-y) = z^2 - y^2 = x^2$, multiplying r and u together seems like a logical thing to do. Doing this gives

$$ru = \left(\frac{z+y}{2}\right)\left(\frac{z-y}{2}\right) = \frac{z^2-y^2}{4} = \frac{x^2}{4}$$

As x is even, $c = x/2$ is an integer and so $ru = c^2$. As $\gcd(r,u) = 1$, Lemma 1.6.2 says that both r and u are squares; therefore you can write

$$r = s^2 \qquad \text{and} \qquad u = t^2$$

Furthermore, as $\gcd(r,u) = \gcd(s^2,t^2) = 1$, it follows that $\gcd(s,t) = 1$. Collecting together all of these give the three integer equations;

$$x = \sqrt{4ru} = 2st \tag{1.8}$$

$$y = r - u = s^2 - t^2 \tag{1.9}$$

$$z = r + u = s^2 + t^2 \tag{1.10}$$

where $\gcd(s,t) = 1$ and s and t are not both odd (as this would mean that x, y, z are all even numbers).

Do these equations always give primitive Pythagorean triples?

Suppose that (x,y,z) are given by the above equations 1.8, 1.9, and 1.10. The idea now is to show that (x,y,z) is a primitive Pythagorean triple. First, you can put equations 1.8 and 1.9 into the initial equation Equation 1.6 to get

$$\begin{aligned} x^2 + y^2 &= 4s^2t^2 + (s^2 - t^2)^2 \\ &= 4s^2t^2 + s^4 - 2s^2t^2 + t^4 \\ &= s^4 + 2s^2t^2 + t^4 \\ &= (s^2 + t^2)^2 = z^2 \end{aligned}$$

Therefore (x,y,z) is a Pythagorean triple. To show that it is a primitive Pythagorean triple, you need to prove that $\gcd(x,y,z) = 1$. As $\gcd(s,t) = 1$ and s and t are not both even, it follows that $\gcd(x,y) = \gcd(x,z) = 1$; as in this case no number that divides $2st$ also

divides $s^2 \pm t^2$ (apart from 1). So all that needs to be done is to show that $\gcd(y, z) = 1$. As $\gcd(s, t) = 1$ and s and t are not both odd, this means that both y and z are odd. Now, assume for a contradiction that $\gcd(y, z) > 1$. Let p be a prime dividing both y and z; since both y and z are odd, this means that p must be odd. By Theorem 1.1.4 (vi) and Equation 1.7 it follows that

$$p \mid (z + y) = 2s^2 \quad \text{and} \quad p \mid (z - y) = 2t^2$$

Now, you can use Lemma 1.3.3 (i) to see that $p \mid s$ and $p \mid t$ and so $\gcd(s, t) \geq p$; this is a contradiction as $\gcd(s, t) = 1$. So the original assumption that $\gcd(y, z) > 1$, and so y and z are coprime. Following the discussion above, this means that $\gcd(x, y, z) = 1$ and so (x, y, z) is a primitive Pythagorean triple.

All solutions to Equation 1.6

This means that you have proved the following theorem concerning primitive Pythagorean triples.

Theorem 1.6.5. *All the solutions of the equation*

$$x^2 + y^2 = z^2$$

where

$$x, y, z > 0, \quad \gcd(x, y, z) = 1 \quad \textit{and} \quad 2 \mid x$$

are given by

$$x = 2st \tag{1.11}$$
$$y = s^2 - t^2 \tag{1.12}$$
$$z = s^2 + t^2 \tag{1.13}$$

where $s > t > 0$ are integers such that $\gcd(s, t) = 1$ and s and t are not both odd.

All other solutions of the equation can be obtained from this by multiplying through by an integer d, interchanging x and y, and changing the sign of any x, y, z. □

Example 1.6.6. Let (x, y, z) be a primitive Pythagorean triple. Then (x, y, z) are determined by Theorem 1.6.5. Now suppose that neither s nor t are divisible by 3; if this happens,

then $x = 2st$ is not divisible by 3 either. This assumption means that $s \equiv \pm 1 \pmod 3$ and $t \equiv \pm 1 \pmod 3$, so $s^2 \equiv t^2 \equiv 1 \pmod 3$.

This means that

$$y = s^2 - t^2 \equiv 1 - 1 \equiv 0 \pmod 3$$

and so if $3 \nmid s$ and $3 \nmid t$, then $3 \mid y$.

You can then say that if (x, y, z) is a primitive Pythagorean triple, then exactly one of x and y is divisible by 3.

As it turns out, higher-order Diophantine equations tend to be a lot harder than Equation 1.6. Here is a result known to Pierre de Fermat:

Theorem 1.6.7. *The equation $x^4 + y^4 = z^2$ has no positive integer solutions.*

Proof sketch. The idea is to assume for a contradiction that (x_0, y_0, z_0) is a solution with z_0 as small as possible. If this happens, then (x_0^2, y_0^2, z_0) is a primitive Pythagorean triple. Using Theorem 1.6.5 will eventually reduce to a solution to the original equation with $z < z_0$, which is a contradiction. □

Remark. For more details, please click on this link to the Maths StackExchange.

Corollary 1.6.8. *The equation $x^4 + y^4 = z^4$ has no positive integer solutions.*

Proof. If (x_0, y_0, z_0) were a solution to $x^4 + y^4 = z^4$, then (x_0, y_0, z_0^2) would be a solution to $x^4 + y^4 = z^2$; which Theorem 1.6.7 says is impossible. □

Finally, Corollary 1.6.8 has been generalised; but it took a while!

Theorem 1.6.9 (Fermat's Last Theorem)**.** *Let $n > 2$ be an integer. Then $x^n + y^n = z^n$ has no integer solutions.*

Proof. A truly marvellous proof, but this space is too small to contain it. This problem resisted all attempts of a proof for over three hundred years until Andrew Wiles presented a solution using the theory of elliptic curves over the rational numbers. For more information on Fermat's Last Theorem, please click on this link to the Wikipedia page. □

Chapter 2

Functions and relations

Often in pure mathematics, you will be forced to deal with sets. You know that sets have elements, and a set is determined by those elements: see the introductory chapter for a brief introduction to set theory.

Comparing two different sets is a fundamental tool in pure mathematics; for instance, finding out if two sets are not the same size can determine whether the two sets do not share a common property. The mathematical tool that allows us to compare sets is the idea of a **function**.

Comparing elements of sets is also a fundamental tool. For example, how can you tell whether an integer a is larger in value than another integer b? Or whether or not two elements of a set are equal? Defining a way to compare two elements of a set (or even an element of one set to an element of another set) is the idea behind a **relation**.

This chapter of the notes is an introduction to the theory of functions and relations. These concepts can be applied to the material you have seen in the course so far (such as modular arithmetic) and also to further material in the course (such as graphs in Chapter 3 and groups in Chapter 4).

2.1 Functions

You will have met functions in other courses. In particular, you were maybe interested in performing calculus on these functions; particularly if these functions sent real numbers to other real numbers. This part of the course is interested in more abstract properties of functions.

Definition 2.1.1. Let X, Y be two sets. A **function** $f : X \to Y$ is a 'rule' that associates every element $x \in X$ to some $y \in Y$. This 'rule' is written as

$$f : x \mapsto f(x)$$

for $x \in X$. In this case, say that f **maps** the element $x \in X$ to $f(x) \in Y$.

For a function $f : X \to Y$, the set X is called the **domain** of f and Y is called the **codomain** of f.

Finally, the set

$$f(X) = \{f(x) \ : \ x \in X\}$$

is called the **image** of f.

Remark. Notice that $f(x)$ is **determined** by x; so for $y, z \in Y$, then $f(x) = y$ and $f(x) = z$ means that $y = z$. This means that a 'rule' defining a function must map each input to **exactly one** output.

For a function $f : X \to Y$, it follows from the definition that the image $f(X)$ of f is a subset of the codomain Y of f. Notice that these may **not** be the same sets!

Other names for a function include 'map' and 'transformation'.

The crucial thing to notice about functions is that they are defined by three things: the 'rule', the domain, and the codomain. If you change just one of these things, the function (and therefore the properties of the function) changes!

Example 2.1.2. Here are some examples of functions.

(a) Here, you can define the function

$$\begin{aligned} f &: \mathbb{R} \to \mathbb{R} \\ &x \mapsto 3x^2 \end{aligned}$$

This function maps a real number x to three times the square of x. The domain and codomain of f is $\mathbb{R}$; the image of f is the set $\{3x^2 \ : \ x \in \mathbb{R}\}$. Notice that the image of f is **not** equal to the codomain $\mathbb{R}$ of f; as (for instance) there are no negative numbers in $f(X)$.

(b) Now, consider the following function $g : \mathbb{Z} \to \mathbb{N}$ given by

$$g(n) = \begin{cases} n \text{ if } n \geq 0 \\ 13 \text{ if } n < 0 \end{cases}$$

This is a perfectly valid function, as the 'rule' assigns exactly one output $g(n)$ to every input n.

(c) (Important: square roots) Take $h_0 : \mathbb{R} \to \mathbb{R}$ with rule defined by $h_0(x) = \pm\sqrt{x}$, the square roots of x. Here, h_0 is not a function as there is not a unique output $h_0(x)$ for every input $x \in \mathbb{R}$; for instance, do you take $h_0(4)$ to be 2 or -2?

To try and fix this issue, you could define $h_1 : \mathbb{R} \to \mathbb{R}$ with rule defined by $h_1(x) = \sqrt{x}$, the positive square root of x. Here, h_1 is not a function as the 'rule' is not defined for every element of the domain; this is because $h(-1)$ is not a real number.

What could you do here to get a function that gives you the positive square root of a positive real number? You could redefine the domain. Define the set

$$\mathbb{R}^{\geq 0} = \{x \in \mathbb{R} \ : \ x \geq 0\}$$

Now, take $h_2 : \mathbb{R}^{\geq 0} \to \mathbb{R}$ to be defined by the rule $h_2(x) = \sqrt{x}$. This is a function which maps every positive number x to its positive square root $\sqrt{x}$. However, the image of this function is **not** equal to the codomain $\mathbb{R}$; for instance, there is no input x such that $h_2(x) = -1$.

To fix *this* situation, you could define $h_3 : \mathbb{R}^{\geq 0} \to \mathbb{R}^{\geq 0}$ with the rule $h_3(x) = \sqrt{x}$. This is a function where the image of h_3 is equal to the codomain $\mathbb{R}^{\geq 0}$ of h_3; so for every $y \in \mathbb{R}^{\geq 0}$ there exists an $x \in \mathbb{R}^{\geq 0}$ such that $f_3(x) = y$.

So the domain and codomain of a function really matter! Here, h_2 and h_3 are **not** the same function. In fact, they have have entirely different properties: see Example 2.1.6 for more details.

The fact that for every $y \in \mathbb{R}^{\geq 0}$ there exists $x \in \mathbb{R}^{\geq 0}$ such that $h_3(x) = y$ is one of the key properties a function can have. This is generalised in the next definition.

Definition 2.1.3. Let X, Y be sets and take $f : X \to Y$ to be a function from X to Y.

(i) Say that f is **injective** (or **one-to-one**, or even **one-one**) if $x \neq y$ means that $f(x) \neq f(y)$ for all $x, y \in X$. Equivalently, f is injective if $f(x) = f(y)$ implies that $x = y$ for all $x, y \in X$. (See Figure 2.1 for a picture of an example.)

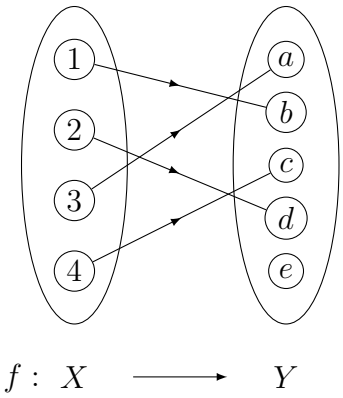


Figure 2.1: A injective function $f : X \to Y$

(ii) Say that f is **surjective** (or **onto**) if for all $y \in Y$ there exists $x \in X$ such that $f(x) = y$. Equivalently, f is surjective if the image $f(X)$ of f is equal to the codomain Y of f. (See Figure 2.2 for a picture of an example.)

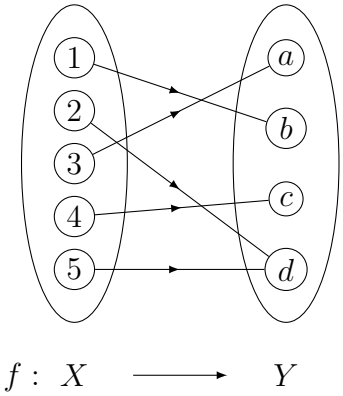


Figure 2.2: A surjective function $f : X \to Y$

(iii) Say that f is **bijective** if f is both injective and surjective.

More informally, you could say that

- f is injective if different elements in X always map to different elements in Y.
- f is surjective if every element of Y has an element of X that maps to it.

- f is bijective if every element of Y has a **unique** element of X that maps to it.

Example 2.1.4. You are given the function $c : \mathbb{R} \to \mathbb{R}$ defined by the rule $c(x) = x^3$. This function c is both injective and surjective.

c **injective** Let $x, y \in \mathbb{R}$ such that $c(x) = c(y)$. Then $x^3 = y^3$; so you can take cube roots of both sides to say that $x = y$. This means that $c(x) = c(y)$ implies that $x = y$; so c is injective by Definition 2.1.3 (i).

c **surjective** Now suppose that $y \in \mathbb{R}$. The idea is to find some x in $\mathbb{R}$ such that $c(x) = y$. What you can do here is to set $x = \sqrt[3]{y}$; so $c(x) = (\sqrt[3]{y})^3 = y$. This means that for all y in the codomain $\mathbb{R}$ of c there exists x in the domain $\mathbb{R}$ of c such that $c(x) = y$; so c is surjective by Definition 2.1.3 (ii).

As c is both injective and surjective, it follows that c is bijective by Definition 2.1.3 (iii).

Example 2.1.5. In Example 2.1.2 (i), you were given the function $f : \mathbb{R} \to \mathbb{R}$ such that $f(x) = x^2$. Let's see whether f is injective or surjective.

Here, $f(2) = 2^2 = 4$; but also, $f(-2) = (-2)^2 = 4$. This means that $f(2) = f(-2)$ but $2 \neq -2$. Therefore, f is **not** injective by Definition 2.1.3 (i). Additionally, there is no real number x such that x^2 is -5; so not every element y of the codomain $\mathbb{R}$ of f is mapped to by an element x of the domain $\mathbb{R}$ of f. This means that f is **not** surjective by Definition 2.1.3 (ii).

Example 2.1.6. This example is designed to see that by changing the codomain of the function, you can change the properties of the function.

In Example 2.1.2, you defined the set $\mathbb{R}^{\geq 0}$ to be the set of all real numbers greater than 0. You can take the function $h_2 : \mathbb{R}^{\geq 0} \to \mathbb{R}$ to be $h_2(x) = \sqrt{x}$, as defined in Example 2.1.2. You can show that this function is injective, but not surjective:

h_2 **injective** Let $x, y \in \mathbb{R}^{\geq 0}$ such that $h_2(x) = h_2(y)$. Then $\sqrt{x} = \sqrt{y}$; so you can square both sides to say that $x = y$. This means if $h_2(x) = h_2(y)$ then $x = y$; so h_2 is injective by Definition 2.1.3 (i).

h_2 **not surjective** Take -1 to be an element in the codomain $\mathbb{R}$ of h_2. There is no real number x such that $h_2(x) = \sqrt{x} = -1$. So not every element y of the codomain $\mathbb{R}$ of h_2 is mapped to by an element x of the domain $\mathbb{R}^{\geq 0}$ of h_2. This means that h_2 is **not** surjective by Definition 2.1.3 (ii).

Now, suppose you are given $h_3 : \mathbb{R}^{\geq 0} \to \mathbb{R}^{\geq 0}$ with the rule $h_3(x) = \sqrt{x}$. This has exactly the same rule as h_2; but the codomain is different. You can actually see here that h_3 is both injective **and** surjective:

h_3 **injective** The proof of this is exactly the same as the proof that h_2 was injective.

h_3 **surjective** Suppose that y is an element of the codomain $\mathbb{R}^{\geq 0}$. You can take $x = y^2$ in the domain $\mathbb{R}^{\geq 0}$; this means that $h_3(x) = \sqrt{y^2} = y$. Therefore, every element y of the codomain $\mathbb{R}^{\geq 0}$ of h_3 is mapped to by an element x of the domain $\mathbb{R}^{\geq 0}$ of h_3. So h_2 is surjective by Definition 2.1.3 (ii).

This means that h_3 is a bijection as it is both injective and surjective.

So by changing the codomain of a function, you can alter its properties: here h_2 is not a bijection but h_3 is!

The idea of a bijective function is one of the most important in mathematics. Suppose that $f : X \to Y$ is a bijective function. This means that for every element y of Y, there exists a **unique** element x of X such that $f(x) = y$. In other words, you can pair two elements of X and Y; this is known as a **one-to-one correspondence**.

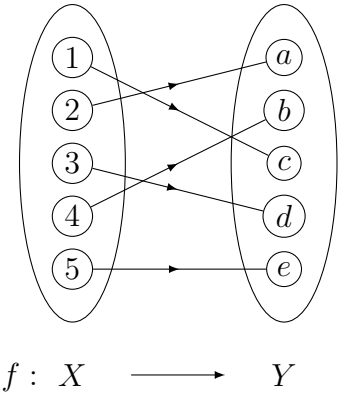


Figure 2.3: A bijective function $f : X \to Y$: this is a one-to-one correspondence. Note how X has the same number of elements as Y.

A consequence of this is that X and Y have the **same number of elements**; this number of elements is known as the **cardinality** or **size** of the set. This idea is one of the most surprising in mathematics. You can prove that there is a bijective function f from the natural numbers $\mathbb{N}$ to the **rational numbers** $\mathbb{Q}$. This means that there are as many

natural numbers as there are fractions! However, there is **no** bijection between the natural numbers $\mathbb{N}$ and the real numbers $\mathbb{R}$; so the size of the set $\mathbb{R}$ is definitely larger than the size of the set $\mathbb{N}$. This implies that there are different sizes of infinity.

2.2 Relations

In the first chapter, you saw that the Cartesian product of two sets A and B is given by the set

$$A \times B = \{(a, b) \ : \ a \in A, \, b \in B\}$$

You can also take the Cartesian product of a set A with itself: these are all the pairs (a_1, a_2) where both a_1, a_2 are elements of A. You can imagine 'relating' a_1 to a_2 by some rule on A. In fact, the idea of a relation is much more general than this.

Definition 2.2.1. Let X be a set. A **relation** R on a set X is any subset of $X \times X$. If the pair $(x, y) \in R$, then you can write that xRy.

Remark. More formally, this is the definition of a **binary** relation because it relates just two elements of the set X. Other types of relation exist; for instance, a **ternary relation** T is a subset of $X^3 = X \times X \times X$. Also, you can have a relation S between two sets X and Y; here, S is any subset of $X \times Y$. However, this brief introduction to relations only looks at binary relations R on a set X.

Example 2.2.2. Relations are everywhere in mathematics: here are a few examples.

(1) $\leq$ is a relation on $\mathbb{Q}$. As a set of ordered pairs, this relation is

$$\leq = \{(x, y) \ : \ x \leq y\} \subseteq \mathbb{Q} \times \mathbb{Q}$$

Similarly, $<$ is a relation on $\mathbb{Q}$.

(2) $=$ is a relation on $\mathbb{C}$. In fact, $=$ is a relation on **any** set X, given by the set of ordered pairs $\{(x, x) \ : \ x = x\} \subseteq X \times X$.

(3) The division sign $|$ is a relation on $\mathbb{Z}$.

(4) For any $n > 1$, congruence modulo n is a relation on $\mathbb{Z}$.

(5) If $\mathcal{P}(X)$ is the set of all subsets of X, then $\subseteq$ is a relation on $\mathcal{P}(X)$.

The idea of a relation generalises all these well-known ideas from mathematics. You can see that this motivates the notation xRy if x and y are related by R.

You can draw diagrams to represent any relation R on any set X. You could do this by drawing a 'node' (a dot) for every element x of X; then if xRy, draw an arrow from the node representing x to the node representing y. (This diagram is known as a **directed graph**: see the next chapter for more on directed graphs.)

Example 2.2.3. You are given the set $X = \{1, 2, 3, 4, 5, 6\}$ together with the relation $|$ on X; so $a \mid b$ if and only if a divides b.

You can use the properties of divisibility to write this relation as a set of ordered pairs. As 1 divides every element of X (Theorem 1.1.4 (i)), it follows that $(1, x) \in |$ for all $x \in X$. Similarly, as $x \mid x$ for all $x \in X$, then $(x, x) \in |$. Finally, $(2, 4)$, $(2, 6)$ and $(3, 6)$ are all also in $|$. The diagram representing this relation $|$ is given in Figure 2.4:

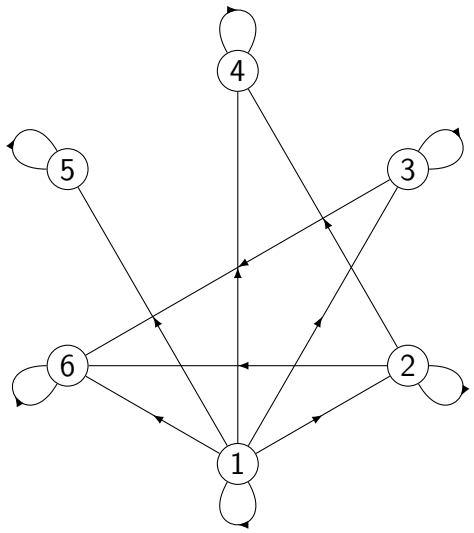


Figure 2.4: The relation $|$ on the set $X = \{1, 2, 3, 4, 5, 6\}$

Example 2.2.4. Now, suppose you are given the set $X = \{1, 2, \ldots, 12\}$. Define the relation R on X by xRy if and only if 2 appears the same number of times in the factorisation of x and y. So

- for any odd number, 2 appears in the prime factorisation 0 times;
- for $2, 6, 10$, 2 appears in the prime factorisation once;

- for 4 and 12, 2 appears in the prime factorisation twice, and;
- 2 appears three times in the prime factorisation of 8.

Furthermore xRx for all x in X; so every node has a loop. Also, for any numbers $x, y \in X$, it follows that if xRy then yRx; so there needs to be two arrows on each edge in the diagram. The diagram representing this relation R is given in Figure 2.5:

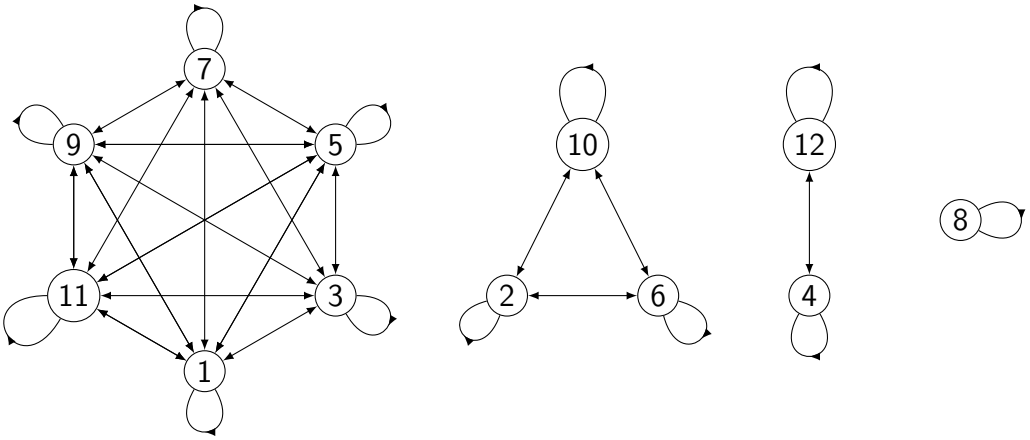


Figure 2.5: The relation R on the set $X = \{1, 2, \ldots, 12\}$

It was stated in Example 2.2.2 (5) that for $n > 1$, then congruence modulo n is a relation on $\mathbb{Z}$. In Theorem 1.4.5, you proved that the relation of congruence modulo n satisfied three useful properties. You can generalise these properties for any relation R on a set X.

Definition 2.2.5. Let R be a relation on the set X.

(i) R is **reflexive** if for all $x \in X$ it follows that xRx;

(ii) R is **symmetric** if for all $x, y \in X$, whenever xRy then yRx;

(iii) R is **antisymmetric** if for all $x, y \in X$, whenever xRy and yRx then $x = y$, and;

(iv) R is **transitive** if for all $x, y, z \in X$, whenever xRy and yRz then xRz.

Example 2.2.6. You can look at the examples of relations in Example 2.2.2 and decide which of these properties they satisfy.

(1) The relation $\leq$ on $\mathbb{Q}$ is reflexive, antisymmetric and transitive; it is not symmetric.

Reflexive: For all $a \in \mathbb{Q}$, it follows that $a \leq a$.

Antisymmetric: Let $a, b \in \mathbb{Q}$ and suppose that $a \leq b$ and $b \leq a$. Then if this happens, $a = b$.

Not symmetric: A counterexample will do. Here, $5 \leq 7$ in $\mathbb{Q}$, but $7 \not\leq 5$ in $\mathbb{Q}$.

Transitive: Let $a, b, c \in \mathbb{Q}$, and suppose that $a \leq b$ and $b \leq c$. Then $a \leq c$.

However, the relation $<$ is **not** reflexive; as $0 \not< 0$.

(2) $=$ is reflexive, symmetric, transitive and antisymmetric on any set X.

(3) Let's consider the division relation on $\mathbb{Z}$. By various properties and consequences of Theorem 1.1.4, it is reflexive and transitive, but is neither symmetric (since $2 \mid 4$ but $4 \nmid 2$) nor antisymmetric (since $-2 \mid 2$ and $2 \mid -2$, but $2 \neq -2$). However, the relation $\mid$ on the set $\mathbb{N}$ of natural numbers is antisymmetric, since if $a \mid b$ and $b \mid a$ in $\mathbb{N}$, then $a = b$ (since $-b \notin \mathbb{N}$) by Theorem 1.1.4. So changing the base set of the relation can change the properties of the relation; much like changing the codomain of a function.

(4) For any $n > 1$, the relation of congruence modulo n on $\mathbb{Z}$ is reflexive, symmetric and transitive by Theorem 1.4.5. It is not antisymmetric for any n, since $0 \equiv n \pmod{n}$ and $n \equiv 0 \pmod{n}$ but $0 \neq n$.

(5) The relation $\subseteq$ on the set $\mathcal{P}(X)$ of all subsets of a set X is reflexive, antisymmetric and transitive.

(6) Example 2.2.4 described the relation R on the set $X = \{1, 2, \ldots, 12\}$ to be defined by xRy if and only if 2 appears the same number of times in the factorisation of x and y. You can show that this relation is reflexive, symmetric and transitive, but not antisymmetric.

Reflexive: Let $x \in X$. As 2 appears the same number of times in prime factorisation of x and x it follows that xRx.

Not antisymmetric: You can use a counterexample. Here $2R10$ and $10R2$, but $2 \neq 10$.

Symmetric: Let $x, y \in X$, and assume that xRy. As 2 appears the same number of times in prime factorisation of x and y it follows that 2 appears the same number of times in prime factorisation of y and x. Therefore yRx and R is symmetric.

Transitive: Let $x, y, z \in X$, and suppose that xRy and yRz. As 2 appears the same number of times in prime factorisation of x and y, and the same holds for y and

z, it follows that 2 appears the same number of times in prime factorisation of x and z. Therefore xRz and the relation R is transitive.

The properties of reflexivity, symmetry and transitivity are **independent** properties; you can write down examples of relations that satisfies any collection of these but not the others. However, you can see that if a relation (that is not equality) is symmetric, then it is not antisymmetric (and vice versa).

In fact, there is a special name for relations that are reflexive, transitive, and either symmetric or antisymmetric.

Definition 2.2.7. Let R be a relation on a set X.

- If R is reflexive, symmetric and transitive, then you can say that R is an **equivalence relation**.
- If R is reflexive, antisymmetric and transitive, then you can say that R is a **partial order**.

Remark. Equivalence relations are a generalisation of the equality relation on $\mathbb{Z}$. Similarly, partial orders are a natural generalisation of the $\leq$ relation on $\mathbb{Z}$.

Example 2.2.8. You can look at some of the relations covered in Example 2.2.6 for examples of equivalence relations and partial orders. Examples of equivalence relations include:

- congruence modulo n on $\mathbb{Z}$;
- equality on any set X, and;
- the relation R on X examined in Example 2.2.3.

Examples of partial orders include:

- $\leq$ on $\mathbb{Q}$ (but *not* $<$ on $\mathbb{Q}$);
- equality on any set X
- the division relation $|$ on $\mathbb{N}$ (but *not* on $\mathbb{Z}$), and;
- the containment relation $\subseteq$ on the set $\mathcal{P}(X)$ of all subsets of a set X.

Equivalence relations play an important role in mathematics because they collect together things that are somehow the 'same'. You can then group together things that are the 'same' into collections of elements of the set the equivalence relation is defined on. This concept is shown in Example 2.2.4. You can notice how there are four different portions to the diagram in Figure 2.5; these correspond to four different collections of elements.

However, as with everything in this course, this idea needs to be formally defined.

Definition 2.2.9. Let R be an equivalence relation on a set X, and take $x \in X$. You can define the **equivalence class** of x to be:

$$[x] = \{y \in X \ : \ xRy\}$$

This is the set of all elements y of X that are related to x by R. You can notice that $[x] \subseteq X$ for all $x \in X$.

Example 2.2.10. You can write down the equivalence classes for the equivalence relation R on $X = \{1, 2, \ldots, 12\}$ as defined in Example 2.2.4. Here,

$$\begin{aligned}
[1] &= \{x \in X \ : \ xR1\} = \{1, 3, 5, 7, 9, 11\} \\
[2] &= \{x \in X \ : \ xR2\} = \{2, 6, 10\} \\
[4] &= \{x \in X \ : \ xR4\} = \{4, 12\} \\
[8] &= \{x \in X \ : \ xR8\} = \{8\}
\end{aligned}$$

You can write down some basic properties of equivalence classes, based on the definition of an equivalence relation (Definition 2.2.7).

Theorem 2.2.11. *Let R be an equivalence relation on a set X. Then the following are true:*

(a) $x \in [x]$ for all $x \in X$;

(b) $\bigcup_{x \in X} [x] = X$. That is, the union of all the equivalence classes $[x]$ is X, and;

(c) If $x, y \in X$, then either $[x] = [y]$ or $[x] \cap [y] = \varnothing$.

Proof. (a) By Definition 2.2.7, the relation R is reflexive and so xRx for all $x \in X$. So you can say that $x \in [x]$ by Definition 2.2.9.

(b) Since $x \in [x]$ for all $x \in X$ by part (a), then the union of all equivalence classes $[x]$ contains all elements of X; so $X \subseteq \bigcup_{x \in X} [x]$. As every equivalence class $[x]$ is a subset

of X (by Definition 2.2.9), then the union of all equivalence classes is a subset of X; so $\bigcup_{x\in X}[x] \subseteq X$. Since these sets contain each other as subsets, they have the same elements and so they are equal.

(c) Let $x, y \in X$, and suppose that $[x] \cap [y] \neq \varnothing$. It is enough to show that in this case, then $[x] = [y]$. Since the intersection of $[x]$ and $[y]$ is non-empty, then there exists some element $z \in X$ such that $x \in [x] \cap [y]$. By Definition 2.2.9, this means that

$$xRz \quad \text{and} \quad yRz$$

As R is symmetric by Definition 2.2.7, then zRy. Since R is transitive by the same definition, it follows that xRy. You can use this to prove that $[x]$ and $[y]$ have the same elements. You could do this by showing that every element $u \in [x]$ is also in $[y]$, and by showing that every element $v \in [y]$ is also in $[x]$.

- Suppose that $u \in [x]$. By Definition 2.2.9, this means that xRu. Since R is symmetric, it follows that as xRy, then yRx. So by transitivity of R, then yRu and therefore $u \in [y]$ by Definition 2.2.9.
- Now assume that $v \in [y]$; therefore, vRy. By transitivity of R, it follows that vRx. Since R is symmetric, then xRv and so $v \in [x]$ by Definition 2.2.9.

So as $[x]$ and $[y]$ have the same elements, they are equal as sets.

□

You can notice that by part (b), X is the union of equivalence classes of some equivalence relation R on X. You also know that any two equivalence classes $[x]$ and $[y]$ are either equal or their intersection is empty (in this case, you can say that the sets are **disjoint**) by part (c). This means that you can assert the following result:

Corollary 2.2.12. *If R is an equivalence relation on a set X, then X is the disjoint union of the equivalence classes of R on X.*

Remark. If this happens, you can say that this is a **partition** of X. So any equivalence relation R on a set X partitions that set.

Example 2.2.13. You can define a relation R on $\mathbb{Q}$ by the following:

$$xRy \quad \text{if and only if} \quad x - y \in \mathbb{Z}$$

You can see that it is an equivalence relation by checking that it is reflexive, symmetric and transitive (Definition 2.2.7):

Reflexive: Let $x \in \mathbb{Q}$. Then as $x - x = 0$, and since $0 \in \mathbb{Z}$, you can say that xRx for all $x \in \mathbb{Q}$.

Symmetric: Let $x, y \in \mathbb{Q}$, and assume that xRy. By definition of R, it follows that $x - y = z \in \mathbb{Z}$. Then you could say that

$$y - x = -(x - y) = -z \in \mathbb{Z}$$

Since this happens, you can say that yRx and so R is symmetric.

Transitive: Let $x, y, z \in \mathbb{Q}$, and suppose that xRy and yRz. This means that $x - y = a \in \mathbb{Z}$ and $y - z = b \in \mathbb{Z}$. So then $x - z = (x - y) + (y - z) = a + b \in \mathbb{Z}$. Therefore, you can say that xRz and so R is transitive.

So in this case, R is an equivalence relation. You can now try and write down (or *describe*) the equivalence classes. First of all, you can see that $x - 0 \in \mathbb{Z}$ if and only if x is an integer to begin with. This means that $xR0$ if and only if $x \in \mathbb{Z}$. You can use Definition 2.2.9 to write that

$$[0] = \mathbb{Z}$$

You can then generalise this; for instance:

$$\left[\frac{1}{2}\right] = \left\{\ldots, -\frac{3}{2}, -\frac{1}{2}, \frac{1}{2}, \frac{3}{2}, \ldots\right\}$$

Finally, you can say that for any $q \in \mathbb{Q}$:

$$[q] = \{q + a \ : \ a \in \mathbb{Z}\}$$

and these are all the equivalence classes.

It was shown in Theorem 1.4.5 that for any $n > 1$, the relation of congruence modulo n is reflexive, symmetric and transitive; and so it is an equivalence relation by Definition 2.2.7. Armed with your new knowledge of equivalence classes, you can ask yourself about the equivalence classes of congruence modulo n. Here, reminding yourself of Theorem 1.4.2 might be useful; it says that $a \equiv b \pmod{n}$ if and only if a and b have the same remainder r upon division by n.

Theorem 2.2.14. *Let $n > 1$. The equivalence relation of congruence modulo n has precisely n equivalence classes; these are*

$$[r] = \{kn + r \ : \ k \in \mathbb{Z}\}$$

for $r = 0, 1, \ldots n - 1$.

Remark. These equivalence classes are known as **congruence classes modulo** n.

Proof. Let $0 \leq r < n$. By Definition 2.2.9, it follows that $a \in [r]$ if and only if $r \equiv a \pmod{n}$. You can use Theorem 1.4.2 to say that $r \equiv a \pmod{n}$ if and only if r and a have the same remainder upon dividing by n. By Fact 1.1.1, you can say that this happens if and only if $a = kn + r$ for some $k \in \mathbb{Z}$. This means you can write

$$[r] = \{kn + r \ : \ k \in \mathbb{Z}\}$$

and so the equivalence classes are of this form.

You now need to check that every element of $\mathbb{Z}$ is in a congruence class modulo n. If $a \in \mathbb{Z}$ then a has remainder r upon division by n, where $0 \leq r < n$; this is by Fact 1.1.1. By Theorem 1.4.2, this means that $a \equiv r \pmod{n}$. This proves that $a \in [r]$ by Definition 2.2.9. Therefore, every element a in $\mathbb{Z}$ is in one of the equivalence classes $[0], [1], \ldots, [n-1]$.

Finally, you can check these are **all** the equivalence classes. So if $[r] = [s]$ where $0 \leq r, s < n$, then you can say that $r \equiv s \pmod{n}$. By Theorem 1.4.2, you can say that r and s have the same remainder upon division by n. Since they are both between 0 and n, they must be the same; so $r = s$. This means that there are exactly n equivalence classes $[0], [1], \ldots, [n-1]$. □

You can then restate Theorem 1.4.6 (modular arithmetic) in terms of congruence classes modulo n. Part (i) of this theorem stated that:

> If $a \equiv b \pmod{n}$ and $c \equiv d \pmod{n}$ then
>
> $$a + c \equiv b + d \pmod{n} \quad \text{and} \quad ac \equiv bd \pmod{n}$$

You can think of addition and multiplication as acting on the **congruence classes** modulo n as well as the integers. You can interpret addition on congruence classes to be

$$[a] = [b] \text{ and } [c] = [d] \quad \text{implies that} \quad [a + c] \text{ and } [b + d]$$

and multiplication as

$$[a] = [b] \text{ and } [c] = [d] \quad \text{implies that} \quad [ac] \text{ and } [bd]$$

You can use this to define operations on the congruence classes by

$$[a] + [b] = [a + b] \quad \text{and } [a] \cdot [b] = [ab]$$

Here, you should replace $a + b$ and ab by the integers in the range $0 \leq r < n$, where these integers are congruent to $a + b$ and ab modulo n.

Example 2.2.15. Here, let $n = 6$. You can write $0, 1, 2, 3, 4, 5$ to be the congruence classes modulo 6, instead of $[0], [1], \ldots$. You can then compute the following tables:

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

$\cdot$	0	1	2	3	4	5
0	0	0	0	0	0	0
1	0	1	2	3	4	5
2	0	2	4	0	2	4
3	0	3	0	3	0	3
4	0	4	2	0	4	2
5	0	5	4	3	2	1

This is modular arithmetic. This plays an important role in abstract algebra, and could be studied further in later courses.

Chapter 3

Graph theory

In Chapter 2, you learned about **relations**, and how to represent them as diagrams (see Example 2.2.4 for an example).

Relations are not the only thing that can be represented by such a diagram. Take a look at Figure 3.1 below: what do you think it could mean?

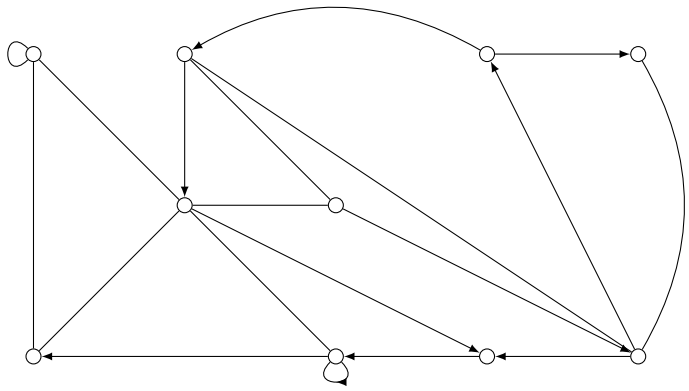


Figure 3.1: What could this be?

It's possible that this diagram could be viewed as a road map, with circles representing towns and the lines between them representing roads between them (with arrows signifying one way roads). It's equally possible that this is the diagram of a social network; with circles representing users, arrows between users representing follows and lines between users representing friendship. It could be that this diagram is of a power grid, or a local area network. From this, you can say that studying the *diagram* could help in understanding what they could represent and how to fix issues. The study of these diagrams is known as **graph theory**, and is the subject of this chapter; focusing on definitions and properties of graphs and digraphs.

3.1 Graphs and digraphs

3.1.1 Directed graphs

Definition 3.1.1. A **directed graph** (or **digraph**) $\Gamma = (V, E)$ consists of a set V of points and a set $E \subseteq V \times V$ of ordered pairs with elements from V. Elements of V are called **vertices** and elements of E are called **edges**.

Example 3.1.2. Let $V = \{1, 2, 3, 4, 5\}$ and let E be the set

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

You can draw $\Gamma = (V, E)$ by drawing 'nodes' to represent the vertices and 'arrows' to represent the edges. You can notice that this is exactly the same process as drawing the relation diagram in Example 2.2.4 (please see Definition 3.1.12 for more details).

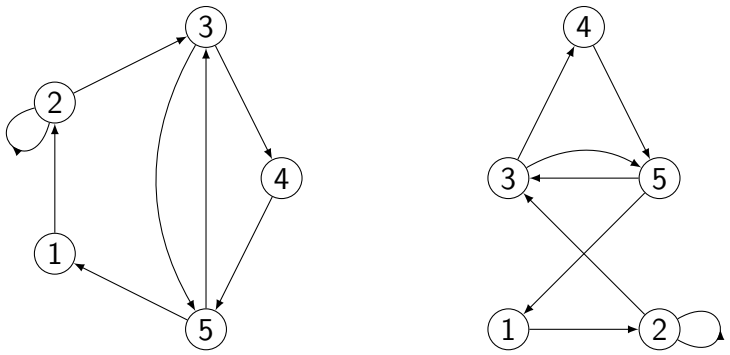


Figure 3.2: Γ in Example 3.1.2 drawn in two different ways

You can think of this digraph Γ as being one of the pictures, rather than the sets given in the definition. This is why elements of V are called 'vertices' and elements of E are 'edges'.

Both of these pictures in Figure 3.2 represent the same digraph $\Gamma = (V, E)$; they have the same vertices (just moved around a little), and the connections between those vertices are the same. As you can see however, they are not **exactly** the same; because they look different. So this leads to an important question– when are two digraphs the 'same'? The next definitions will make this idea more formal.

Definition 3.1.3. Let $\Gamma = (V, E)$ be a directed graph. Say that a vertex $u \in V$ is **adjacent** to $v \in V$ if and only if $(u, v) \in E$ is an edge of Γ.

You can use this to describe when two graphs are the 'same'.

Definition 3.1.4. Let $\Gamma = (V, E)$ and $\Gamma' = (V', E')$ be two directed graphs. Say that Γ and Γ' are **isomorphic** if there is a bijection $f : V \to V'$ such that for all pairs of vertices $u, v \in V$:

$$(u, v) \in E \text{ if and only if } (f(u), f(v)) \in E'$$

that is, u, v are adjacent in Γ if and only if $f(u), f(v)$ are adjacent in Γ'.

If this happens, you can write that $\Gamma \cong \Gamma'$.

So two digraphs are isomorphic if they have the same vertices (up to the names of the vertices) and the edges between those vertices are the same; this is exactly the situation in Figure 3.2. In graph theory, you do not need to distinguish between isomorphic graphs, as any two graphs that are isomorphic share exactly the same properties. (In fact, isomorphism of graphs is an equivalence relation.)

However, deciding when two digraphs are isomorphic is quite a difficult problem in general (even for computers). What you can do to make this problem easier is to look at mathematical objects that are related to digraphs Γ and Γ', and then see if the mathematical objects are the same to help decide if Γ is isomorphic to Γ'. This leads on to the next definition, where the mathematical object is a *matrix*.

Definition 3.1.5. Let $\Gamma = (V, E)$ be a directed graph with finite vertex set $V = \{v_1, \dots v_n\}$. The **adjacency matrix** of Γ is the $n \times n$ matrix $A(\Gamma)$ with

$$(A(\Gamma))_{ij} = \begin{cases} 1 \text{ if } (v_i, v_j) \in E \\ 0 \text{ if } (v_i, v_j) \notin E \end{cases}$$

that is, the (i, j)th entry of $A(\Gamma)$ is 1 if v_i is adjacent to v_j in Γ, and 0 if v_i is not adjacent to v_j in Γ.

Example 3.1.6. Here, you are given the following directed graph $\Gamma = (V, E)$:

You can notice here that the vertices of Γ are labelled with letters and not with numbers. To get around this, you could set a to be the first column and row, b to be the second

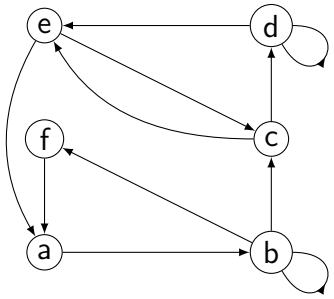


Figure 3.3: Γ in Example 3.1.6

column and row, and so on. Rows come first in the matrix; so if there is an edge from f to a, this corresponds to a 1 in the **sixth** row and **first** column (based on the rules you have just defined). This means that the adjacency matrix of Γ is:

$$A(\Gamma) = \begin{pmatrix} 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 1 & 1 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

You can notice that the adjacency matrix of a digraph Γ contains all of the important information about Γ; how many vertices there are (number of rows/columns) and the adjacencies between those vertices (entries of matrix). So the following result is true:

Lemma 3.1.7. *Two directed graphs with finitely many vertices are isomorphic if and only if they have identical adjacency matrices following a relabelling of the vertices.* □

Walks

Remember the diagram from Figure 3.1? If you viewed it as a road map, then you'd be perfectly comfortable in travelling between the destinations (vertices) along the given routes (edges). In graph theory, travelling between vertices along edges is known as a **walk**.

Definition 3.1.8. Let $\Gamma = (V, E)$ be a directed graph. A **walk** in Γ is a sequence of vertices and edges where each edge is directed from the vertex before it to the vertex following it.

A walk has the form

$$v_0 e_1 v_1 e_2 v_2 \dots v_{n-1} e_n v_n$$

where each edge $e_i = (v_{i-1}, v_i)$. The **length** of the walk is the number of edges in the walk.

Remark. You don't need to write the vertices in the sequence. This is because the edges of a walk tell you which vertices you have visited along the way.

Now, if you went back to viewing Figure 3.1 as a road map, you might wish to plan ahead for your trip. For instance, you may not want to go along the same route twice for the sake of scenery. You may also want to end back up at your starting destination, instead of ending somewhere else. These are special types of walk, and they are discussed in the next definition.

Definition 3.1.9. Once again, let $\Gamma = (V, E)$ be a directed graph.

(1) A **path** in Γ is a walk in which no *vertex* appears more than once.

(2) A **circuit** in Γ is a walk in which the first vertex and the last vertex are the same. This is also called a **closed** walk.

Example 3.1.10. You are given the digraph Γ from Example 3.1.6: see Figure 3.3 for a picture of Γ.

The following sequence of edges is a walk of length 6 in Γ. The edges of this walk are represented in red in Figure 3.4, with the start and end vertices also highlighted.

$$(a, b),\ (b, b),\ (b, c),\ (c, e),\ (e, c),\ (c, d)$$

The following sequence of edges is a path of length 5 in Γ; the edges of this path are represented in green in Figure 3.4.

$$(f, a),\ (a, b),\ (b, c),\ (c, d),\ (d, e)$$

The following sequence of edges is a circuit of length 4 in Γ; this is represented in blue in Figure 3.4.

$$(c, d),\ (d, d),\ (d, e),\ (e, c)$$

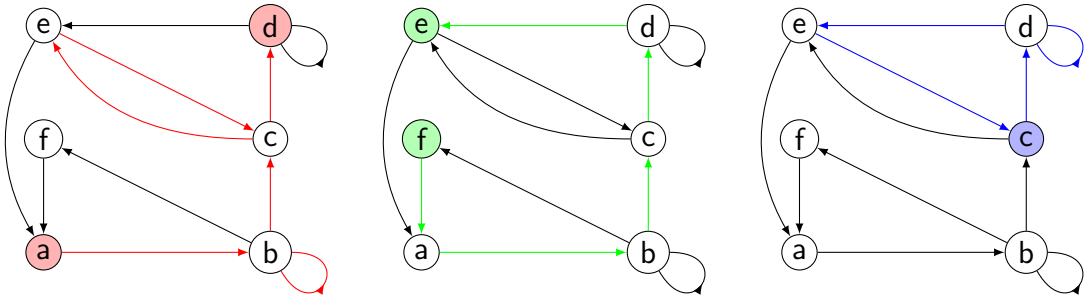


Figure 3.4: The walk (red), path (green) and circuit (blue) in Γ from Example 3.1.10

As it turns out, the number of walks from one vertex to another in a digraph Γ is closely related to powers of the adjacency matrix $A(\Gamma)$ (see Definition 3.1.5).

Theorem 3.1.11. *Let $\Gamma = (V, E)$ be a directed graph with finite vertex set $\{v_1, \dots v_n\}$ and adjacency matrix $A = A(\Gamma)$. Then the (i, j)th entry of A^m is the number of walks of length m from v_i to v_j in Γ.*

Proof. The proof is by induction on the natural number m. (This is a common proof technique in graph theory.) Throughout, you can write $A = (a_{ij})$, where a_{ij} represents the (i, j)th entry of A.

Base case: Suppose that $m = 1$. You can say that the walks of length 1 in Γ are exactly the edges $(v_i, v_j) \in E$ of Γ; so there is a walk of length 1 from v_i to v_j if and only if $a_{ij} = 1$. Since there is at most one edge between any two vertices of Γ, the base case is done.

Inductive case: Assume for the inductive hypothesis that the statement holds for $A^m = B = (b_{ij})$; so b_{ij} represents the number of walks of length m from v_i to v_j.

Consider the matrix $BA = C = (c_{ij})$. By the rules of matrix multiplication, it follows that

$$c_{ij} = \sum_{k=1}^{n} b_{ik} a_{kj}$$

By the inductive hypothesis, b_{ik} is the number of walks of length m from v_i to v_k. You can extend any of these to a walk of length $m + 1$ from v_i to v_j if and only if there is an edge from v_k to v_j in Γ, which is true if and only if $a_{kj} = 1$ and not zero. Therefore, the term $b_{ik} a_{kj}$ is equal to the number of walks of length $m + 1$ from v_i to v_j that stops at v_k at the mth step. You can then say that the sum of all these terms

$$c_{ij} = \sum_{k=1}^{n} b_{ik} a_{kj}$$

is the **total** number of walks from v_i to v_j of length $m+1$. □

3.1.2 Graphs

You can see from Example 2.2.4 that every relation R on a set X can be represented as a digraph $\Gamma(R) = (X, R)$: where X is the set of vertices and the relation R is the set of edges. Conversely, every digraph $\Gamma = (V, E)$ is a relation E on a set V. Because this is a one-to-one correspondence (see the end of Section 2.1), properties of relations R correspond to properties of digraph Γ. Here is a definition that will help to show this connection.

Definition 3.1.12. If $\Gamma = (V, E)$ is a directed graph, then an edge of the form (v, v) is called a **loop**. A digraph Γ is said to be **loop-free** if it has no loops.

This concept relates back to the definition of a **reflexive relation** (see Definition 2.2.7). It follows that if R is reflexive then **every** vertex of the corresponding digraph $\Gamma(R)$ has a loop. Conversely, if a digraph $\Gamma(R)$ is loop-free, then $(v, v) \notin R$ for all $v \in X$; this means that the relation R represented by $\Gamma(R)$ is **irreflexive**.

You can then ask yourself about other properties given in Definition 2.2.7. What about symmetry, for instance? This is an important property, and the type of digraph $\Gamma(R)$ that represents a symmetric relation R is given in the next definition.

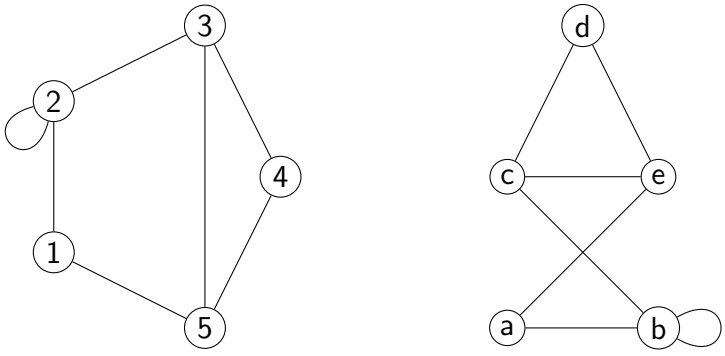
Definition 3.1.13. Let $\Gamma = (V, E)$ be a directed graph with the property that whenever (v_i, v_j) is an edge then (v_j, v_i) is an edge. Then $\Gamma = (V, E)$ is known as an **undirected graph**; or just a **graph**.

In this case, you can write $\{v_i, v_j\}$ to be the edge between v_i and v_j.

If a relation R is symmetric, then $\Gamma(R)$ is a graph. If you have a graph Γ, you can save time and energy by drawing the two directed edges between v_i and v_j as a single undirected edge (with no arrow).

Because every graph is also a directed graph, all the definitions from the previous section (isomorphism, adjacency matrix...) are also valid for graphs. Here's an example.

Example 3.1.14. You are given the following graphs Γ and Γ':



Γ Γ'

Figure 3.5: Γ and Γ' in Example 3.1.14

You can define a function $f : \Gamma \to \Gamma'$ by sending 1 to a, 2 to b and so on. This bijection proves that Γ and Γ' are isomorphic. To write the adjacency matrix of Γ, you should note that if $\{v_i, v_j\} \in E$, then $a_{ij} = 1$ **and** $a_{ji} = 1$. So the adjacency matrix of Γ is:

$$A(\Gamma) = \begin{pmatrix} 0 & 1 & 0 & 0 & 1 \\ 1 & 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 & 1 \\ 0 & 0 & 1 & 0 & 1 \\ 1 & 0 & 1 & 1 & 0 \end{pmatrix}$$

You can notice that this matrix $A(\Gamma)$ is **symmetric**; with $a_{ij} = a_{ji}$ for all i, j.

A path of length 3 in Γ' could be the sequence

$$\{c, d\},\ \{d, e\},\ \{e, a\}$$

Sometimes, graphs will be allowed to have more than one edge between some pair of vertices. If this happens, then you can say that the graph has **multiple edges** and call it a **multigraph** (see Figure 3.6 for an example).

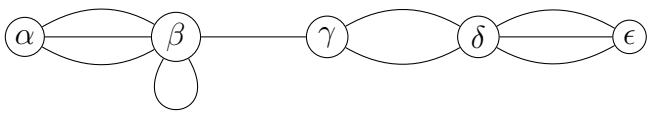


Figure 3.6: A multigraph

The following definition restricts the attention of this section to a more standard notion of graph.

Definition 3.1.15. A graph Γ is **simple** if Γ has no multiple edges or loops.

Remark. If a relation R is irreflexive and symmetric, then $\Gamma(R)$ is a simple graph.

Here are some important families of simple graphs.

Example 3.1.16. Let n be a natural number.

(1) The **complete graph** K_n is the simple graph with n vertices in which every pair of distinct vertices is adjacent. See Figure 3.7 for pictures of K_n for small values of n, and Figure 3.8 for a drawing of K_{12}.

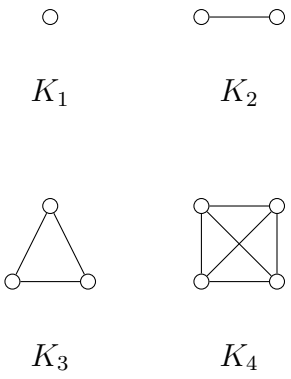


Figure 3.7: Complete graphs for small n

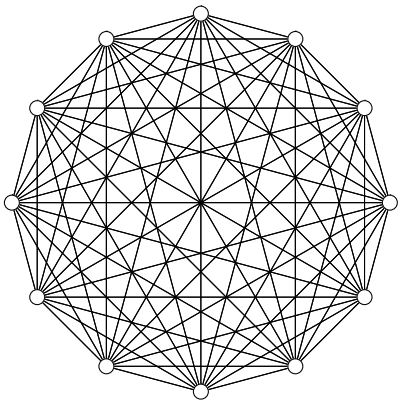


Figure 3.8: The complete graph K_{12}

Here, K_n has $n(n-1)/2$ edges for all natural numbers n. You should see Lemma 3.1.19 for confirmation of this result.

(2) The **empty graph** N_n on n vertices is the graph with n vertices and **no edges**. See Figure 3.9 for an example where $n = 12$.

(3) The **cycle graph** C_n of length n is the graph with vertices $V = \{v_1, \ldots, v_n\}$ and edges

$$E = \{(v_1, v_2),\ (v_2, v_3),\ \ldots,\ (v_{n-1}, v_n), (v_n, v_1)\}$$

See Figure 3.10 for an example where $n = 12$.

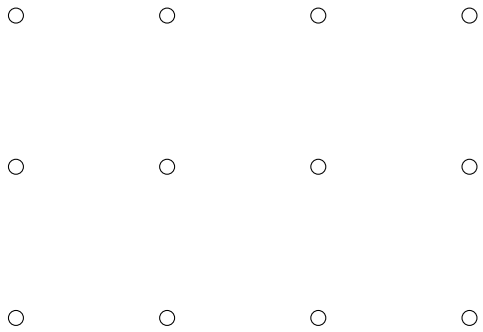


Figure 3.9: The empty graph N_{12}

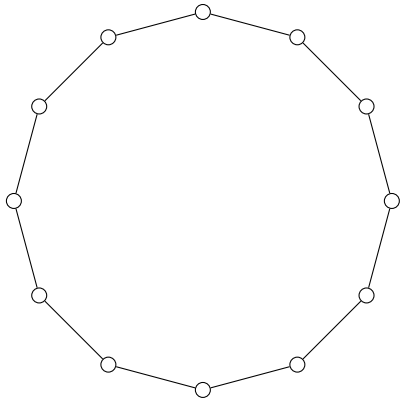


Figure 3.10: The cycle graph C_{12}

Degrees

Definition 3.1.17. Let $\Gamma = (V, E)$ be a graph. The **degree** of a vertex v is the number of edges that are 'incident' to v. You can write $d_\Gamma(v)$ to mean the degree of the vertex v in the graph Γ. If $d_\Gamma(v) = n$ for all $v \in V$ and some $n \in \mathbb{N} \cup \{0\}$, then the graph Γ is called **regular** or n-**regular**.

Remark (Non-examinable remark). There is such a thing as the 'degree' of a vertex in a *digraph* $\Gamma = (V, E)$. The **indegree** of a vertex v in a digraph Γ is the number of edges (u, v) for $u \in V$, and the **outdegree** of v in Γ is the number of edges (v, u) for $u \in V$. If Γ is a graph, then these numbers are the same.

Example 3.1.18. Here are some examples of degrees in graphs, and some regular graphs.

(1) Figure 3.11 reproduces Γ from Example 3.1.14:

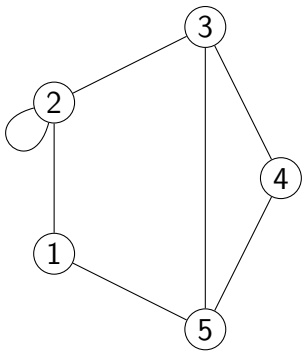


Figure 3.11: Γ from Example 3.1.14

You could count the edges going into each vertex to find out the degrees. So as there are two edges going into vertex 1, it follows that $d_\Gamma(1) = 2$. There are two edges and a loop going into vertex 2 of Γ; here, the loop is counted **twice**. (Have a think about why this is the case.) Therefore, you can say that $d_\Gamma(2) = 4$. You can do this for the other vertices of Γ to find that:

$$d_\Gamma(3) = 3 \qquad d_\Gamma(4) = 2 \qquad d_\Gamma(5) = 3$$

You can also say that Γ is **not** regular.

(2) From Example 3.1.16, the complete graph K_n is n-regular for all n. Every cycle graph C_n is 2-regular, and a graph is 0-regular if and only if it is isomorphic to an empty graph N_n.

Here is a nice result that enabled the statement of the amount of edges of K_n in Example 3.1.16.

Lemma 3.1.19 (Handshake lemma (Euler))**.** *Let $\Gamma = (V, E)$ be a finite graph, with $|E|$ denoting the number of edges in Γ. Then*

$$\sum_{v \in V} d_\Gamma(v) = 2|E|$$

That is, the sum of degrees in a graph Γ is equal to twice the number of edges of Γ.

Proof (non-examinable). You can count the number of 'incident pairs' (v, e) (with $v \in V$ and $e \in E$), where e contains v as an endpoint, in two different ways. Here, the vertex v belongs to $d_\Gamma(v)$ many incident pairs (v, e), and so the number of incident pairs is the sum of all degrees. On the other hand, every edge belongs to exactly two incident pairs (v, e) (as edges have two ends); so the number of incident pairs is also twice the total number of edges $|E|$ of Γ. □

So in the case of the complete graph K_n, there are n vertices, each with degree $(n-1)$. This means that the sum of degrees is $n(n-1)$; by the handshake lemma Lemma 3.1.19, it follows that $n(n-1) = 2|E|$. You can divide both sides of this equation by 2 to get that the number of edges in K_n is $n(n-1)/2$.

Trees

You have seen that some graphs have edges connecting vertices (like Γ in Example 3.1.14), and some graphs do not have edges connecting vertices (like the empty graphs N_n from Example 3.1.16. Some graphs have circuits (like the cycle graphs C_n from Example 3.1.16), and some graphs do not (the empty graphs again). Let's look at these properties in more detail; using the idea of **paths** and **circuits** from Definition 3.1.9.

Definition 3.1.20. A graph Γ is **connected** if there is a path between any two distinct vertices of Γ, and **disconnected** otherwise.

For example, the complete graphs K_n and the cycle graphs C_n are connected for all $n \geq 2$. The empty graph N_n is disconnected for every natural number n. Here is a special kind of connected graph, known as a **tree**.

Definition 3.1.21. A simple graph Γ is called a **tree** if it is connected and contains no circuits.

Example 3.1.22. Here are some examples of trees.

(1) The following graphs are all non-isomorphic trees on five vertices; these are given in Figure 3.12.

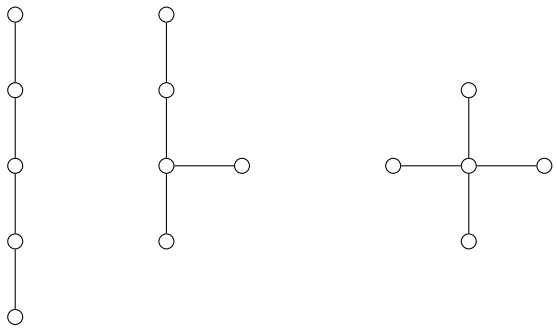


Figure 3.12: The three non-isomorphic trees on five vertices

A disjoint union of trees is called a **forest**. A vertex of degree 1 is called a **leaf**.

(2) A fun example of a tree is the **star graph** $K_{1,n}$ on $n+1$ vertices. (The choice of notation will be explained later.) This is the graph on vertex set $V = \{v_1, v_2, v_3, \ldots, v_{n+1}\}$ with edge set

$$E = \{(v_1, v_j) \ : \ 2 \leq j \leq n+1\}$$

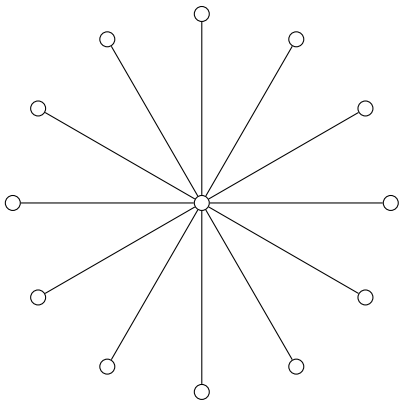


Figure 3.13: The star graph $K_{1,12}$

. See Figure 3.13 for the star graph $K_{1,12}$ on $12+1$ vertices.

(3) Γ from Example 3.1.14 is not a tree; it contains a loop, and so is not a simple graph.

There are two main properties of trees that finish the section.

Theorem 3.1.23. *Let T be a tree and suppose that u, v are distinct vertices in T. Then there is a unique path starting at u and finishing at v.*

Proof. Since T is connected by Definition 3.1.21, there is at least one path from u to v by definition of connected (Definition 3.1.20). Now suppose for a contradiction that there is more than one path connecting u and v; then from these paths, you can go from u to v using one path and from v to u using another path. At some point in this walk from u to u, there is a circuit; which contradicts the fact that T is a tree. □

Theorem 3.1.24. *Let $T = (V, E)$ be a tree. Then the number of edges $|E|$ is equal to the number of vertices $|V|$* **minus** *one; in symbols, $|E| = |V| - 1$.*

Proof. Let $|V| = n$. The proof is by induction on n.

Base case: Suppose that $n = 1$. In this case, the graph T is a single vertex. Since it is a tree, it is simple by Definition 3.1.21 and so T has no edges. Therefore, $|E| = 0 = |V| - 1$ and so the base case is true.

Inductive case: Assume for the inductive hypothesis that the statement holds for all trees T' with $r < n$ vertices. Pick any edge $\{u, v\}$ in T. You can delete this edge to get a new graph $\Gamma = (V, E \setminus \{u, v\})$. As there is a unique path from u to v in T by Theorem 3.1.23, it follows that by deleting the edge $\{u, v\}$ there is no path from u to v any more. This means that Γ is disconnected by Definition 3.1.20. Since T is connected (Definition 3.1.21),

then deleting $\{u, v\}$ means that Γ consists of two 'connected components' T_1 and T_2 (see Figure 3.14 of an example of how this works).

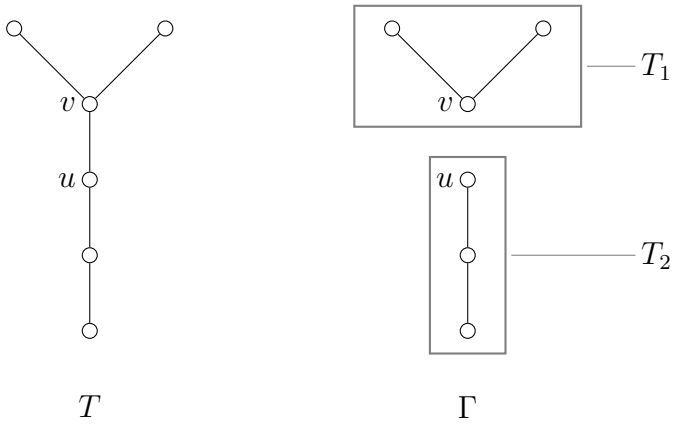


Figure 3.14: Example of edge deletion in a tree T

Both T_1 and T_2 must be trees as deleting an edge from the tree T does not introduce a loop or a circuit. Let n_1 be the number of vertices in T_1 and n_2 be the number of vertice in T_2; as you have not deleted any vertices, it follows that $n_1 + n_2 = |V| = n$. Importantly, this means that $n_1, n_2 < n$ and so the inductive hypothesis says that T_1 has $n_1 - 1$ edges and T_2 has $n_2 - 1$ edges. Since you deleted a single edge $\{u, v\}$, it follows that the number of edges in T is given by:

$$|E| = \underbrace{(n_1 - 1)}_{\text{edges in } T_1} + \underbrace{(n_2 - 1)}_{\text{edges in } T_2} + \underbrace{1}_{\{u,v\}} = n_1 + n_2 - 1 = n - 1$$

and this completes the proof.

□

3.2 Eulerian and Hamiltonian graphs

3.2.1 Eulerian walks and circuits

This section begins with a trip back in time to the 18th century and the city of Königsberg (now Kaliningrad) on the coast of the Baltic Sea. You've been given a map of the city at the time in Figure 3.15.

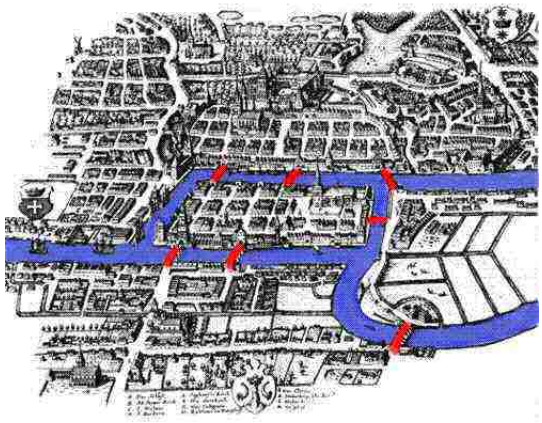


Figure 3.15: The city of Königsberg

The locals of Königsberg have been wondering for years whether or not it is possible to walk around the city and cross each of its seven bridges (highlighted in red) exactly once, and return to the position in which they started (although they're not too fussy about it).

This is a problem that can be solved using graph theory. You can view a road map as a graph, with vertices corresponding to places on the map and edges as roads between those places. Here, you can view the city of Königsberg as a graph; with regions of the city as the vertices $\{a, b, c, d\}$ and the bridges between them as the edges. A multigraph $\mathcal{K}$ corresponding to the map of Königsberg is given in Figure 3.16.

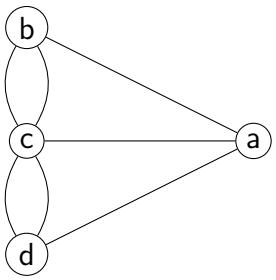


Figure 3.16: The city of Königsberg; a multigraph $\mathcal{K}$

Using this, you can translate the local problem into the language of graph theory. The

problem of finding a route crossing each bridge in Figure 3.15 exactly once corresponds to finding a walk (in the sense of Definition 3.1.8) in the associated multigraph $\mathcal{K}$ in Figure 3.16 visiting each edge exactly once. An extra problem on top of this would be to end up in the section of the city in which you started. Solving the problem with this additional condition corresponds to finding a *circuit* (as in Definition 3.1.9) in $\mathcal{K}$ that visits each edge exactly once.

You can formalise these as properties of graphs with the following definitions:

Definition 3.2.1. A graph $\Gamma = (V, E)$ is called **semi-Eulerian** if there is a walk in Γ that passes through every vertex $v \in V$ and includes every edge $e \in E$ *exactly* once. Such a walk is known as an **Eulerian walk**.

Informally, a graph Γ has an Eulerian walk if there is a way to draw the edges of Γ without taking your pen off the paper (or chalk on the board, or stick in the sand...) and without going over the same edge twice.

Definition 3.2.2. A graph $\Gamma = (V, E)$ is called **Eulerian** if there is a *circuit* in Γ that passes through every vertex $v \in V$ and includes every edge $e \in E$ exactly once. Such a circuit is known as an Eulerian circuit or **Euler tour**.

It took one of the greatest mathematicians in history to solve the plight of the people in Königsberg; Leonhard Euler. The following theorem of his is widely recognised today to be the first ever mathematical result about graph theory, and relates Eulerian and semi-Eulerian graphs to the idea of the degree of a vertex in Definition 3.1.15.

Theorem 3.2.3 (Euler)**.** *Let $\Gamma = (V, E)$ be a connected graph. Then:*

(1) Γ is semi-Eulerian if and only if Γ has at most two vertices with odd degree.

(2) Γ is Eulerian if and only if every vertex in Γ has even degree. □

The (multi)graph of the city of Königsberg in Figure 3.16 has four vertices of odd degree; therefore, it is neither Eulerian nor semi-Eulerian by Theorem 3.2.3. The denizens of Königsberg are destined to be disappointed until the end of time! (or until somebody builds another bridge...)

Example 3.2.4. (1) Every cycle graph C_n with $n \geq 2$ is Eulerian. This is because every vertex in a cycle graph has degree 2, and so satisfies Theorem 3.2.3 (2).

(2) Every complete graph K_n with $n \geq 2$ is Eulerian if and only if n is **odd**. This is because every vertex in K_n has degree $n-1$; and $n-1$ is even if and only if n is odd. This means that K_{12} from Figure 3.8 is not Eulerian; it would be painful to check without Theorem 3.2.3!

Here's another famous problem involving the theory of Eulerian and semi-Eulerian graphs.

Example 3.2.5 (The envelope problem)**.** In Figure 3.17, you are given three graphs A, B, C that look like envelopes. The first graph A looks like a 'double open' envelope, the second graph B an 'open' envelope and the third graph C a 'closed' envelope (you can recognise C as K_4, the complete graph on 4 vertices Example 3.1.16). The idea of the problem is to try and draw these graphs **without taking your pen off the paper**.

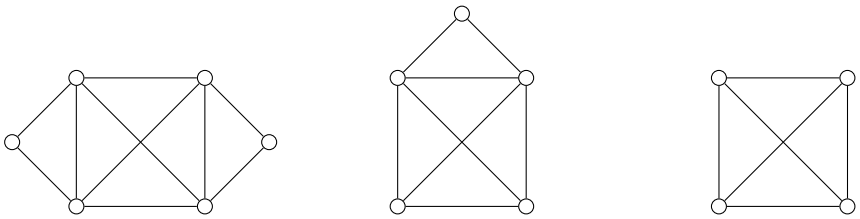


Figure 3.17: The envelope graphs A, B, C

Similar to the bridges, this problem is exactly the same as deciding whether or not the graphs are semi-Eulerian. So you can use Theorem 3.2.3 to decide whether or not the problem has solutions or not. This involves finding the degrees of each vertex in A, B, C in turn.

Graph A: In this case, there are four vertices with degree 4 and two vertices with degree 2. Since all the vertices in A have even degree, you can say that A is Eulerian by Theorem 3.2.3.

Graph B: In this case, there are two vertices with degree 4, two vertices with degree 3 and one vertex with degree 2. Since there are exactly two vertices in B with odd degree, you can say that B is semi-Eulerian by Theorem 3.2.3.

Graph C: In this case, all four vertices have degree 3. So C is neither Eulerian nor semi-Eulerian by Theorem 3.2.3.

Therefore, you can draw A and B without taking your pen off the page; but you can't draw C like this.

3.2.2 Hamiltonian paths and cycles

Away from the hustle and bustle of 18th century Königsberg, a 19th century mathematician called William Hamilton developed a game that he later sold to a toy maker in Dublin. The game consisted of a wooden regular dodecahedron with the twenty corners labelled with names of famous cities. The object of the game was to find a path along the edges of the solid so that you would visit each city exactly once.

This is a similar problem to the bridges in Königsberg; except you are asked to visit every **vertex** and not every edge exactly once. You can use the techniques of graph theory to help; see Figure 3.18 for a graph D of a dodecahedron that has been 'flattened' into the plane.

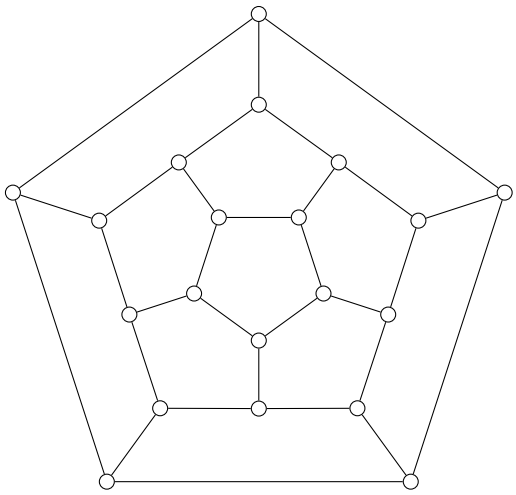


Figure 3.18: The flattened dodecahedron D

Once again, you can define a type of circuit that visits each vertex exactly once:

Definition 3.2.6. Let $\Gamma = (V, E)$ be a graph. A **Hamiltonian circuit** is a circuit that passes through every *vertex* exactly once (with only the start/end point repeated). A graph Γ is called **Hamiltonian** if Γ has a Hamiltonian circuit.

Example 3.2.7. (1) Every cycle graph C_n with $n \geq 2$ is Hamiltonian; the graph itself is a Hamiltonian circuit. You can also say that every complete graph K_n with $n \geq 2$ is Hamiltonian (you can find a circuit!). This means that K_4 is an example of a graph that is Hamiltonian but **not** Eulerian.

(2) The flattened dodecahedron D in Figure 3.18 is a Hamiltonian graph. An example of a Hamiltonian circuit for D is given in Figure 3.19.

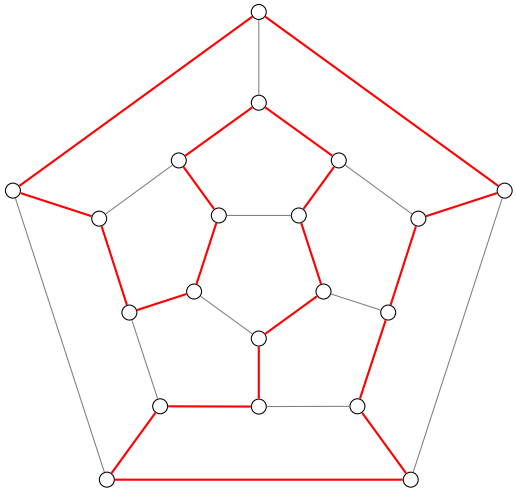


Figure 3.19: The flattened dodecahedron D with a Hamiltonian cycle in red

(3) In Figure 3.20 you are given the 'bow-tie' graph $\mathcal{B}$ (no prizes for guessing why it is so called):

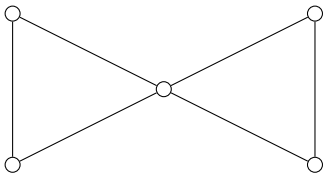


Figure 3.20: The bow-tie graph $\mathcal{B}$

As each vertex of $\mathcal{B}$ has even degree, you can use Theorem 3.2.3 to say that $\mathcal{B}$ is a Eulerian graph. However, $\mathcal{B}$ is **not** a Hamiltonian graph. If you start a walk in one of the four corners, you must travel through the centre vertex twice in order to complete a circuit. If you start a walk in the centre vertex towards one side, then you need to travel through it again to reach the corners on the other side. So $\mathcal{B}$ is an example of a graph that is Eulerian but not Hamiltonian.

Example 3.2.7 gives some examples of Hamiltonian graphs. However, the way to *prove* that a graph is Hamiltonian appears to be to find a cycle. Conversely, how do you prove that a graph isn't Hamiltonian? The argument for the bow-tie graph $\mathcal{B}$ in Example 3.2.7 is a bit flimsy. In the same way that a graph Γ is Eulerian if and only if every vertex of Γ has even degree, it would be useful to have an if and only if statement saying that a graph Γ to be Hamiltonian. However, this is an unsolved problem in mathematics; no such statement has been found!

Here however, is a theorem that says if a graph Γ has some condition, then it is Hamiltonian.

Theorem 3.2.8 (Dirac 1952)**.** *Let $\Gamma = (V, E)$ be a simple graph with n vertices and suppose that $d_\Gamma(v) \geq n/2$ for every vertex $v \in V$. Then Γ is a Hamiltonian graph.*

Proof. Suppose for a contradiction that Γ is not a Hamiltonian graph.

The first step of the proof is to notice that if you add more edges to Γ, then you will eventually get to a Hamiltonian graph. This is because the complete graph K_n is Hamiltonian Example 3.2.7, and if you add in all possible edges to Γ you will get to K_n. The crucial point is that at some stage during this process, you will reach a graph Γ' that is *not* Hamiltonian, but when you add in one more edge to Γ', you get a Hamiltonian graph.

The idea is to work on Γ', which is not Hamiltonian. If you can show there is a contradiction concerning Γ', then this will ensure a contradiction involving Γ; this is because Γ' was created under the assumption that Γ is not Hamiltonian.

First off, you can notice $d_{\Gamma'}(v) \geq d_\Gamma(v) \geq n/2$. Let $\{v_1, v_2\}$ be the edge that, when added to Γ', makes the graph Hamiltonian. This means that $\{v_1, v_2\}$ is part of a Hamiltonian circuit; which could look like:

$$v_1 \to v_2 \to v_3 \to \cdots \to v_{i-1} \to v_i \to \cdots \to v_n \to v_1$$

Now, if for some $3 \leq i \leq n$ there are edges $\{v_2, v_i\}$ and $\{v_1, v_{i-1}\}$ in Γ', then you can write a new Hamiltonian circuit:

$$v_2 \to v_i \to v_{i+1} \to \cdots \to v_n \to v_1 \to v_{i-1} \to v_{i-2} \to \cdots \to v_3 \to v_2$$

Since this circuit does not involve the edge $\{v_1, v_2\}$, it must have been in Γ'; this contradicts the assumption that Γ' is not Hamiltonian.

What does this mean? This means that it's not true that $\{v_1, v_{i-1}\}$ and $\{v_2, v_i\}$ are both edges in Γ' for $i = 3, \ldots, n$. You can translate this information into a statement about the adjacency matrix $A(\Gamma') = (a_{kl})$ of Γ'. So as $\{v_1, v_{i-1}\}$ and $\{v_2, v_i\}$ are not both edges in Γ', you can say that:

$$a_{1,i-1} + a_{2,i} \leq 1$$

for $3 \leq i \leq n$. (Commas have been added to separate entries.) You can then sum over all the i to get that

$$\sum_{i=3}^{n} a_{1,i-1} + \sum_{i=3}^{n} a_{2,i} \leq n - 2$$

Next, you can notice that as Γ is simple it has no loops (Definition 3.1.15); so $a_{11} = a_{22} = 0$. As Γ' does not contain the edge $\{v_1, v_2\}$, this means that $a_{12} = a_{21} = 0$. So you can add these terms to the sums to get:

$$\sum_{i=1}^{n} a_{1,i-1} + \sum_{i=1}^{n} a_{2,i} \leq n - 2$$

You know that $\{v_1, v_n\}$ is an edge in Γ' from the Hamiltonian circuit, and so $a_{1n} = 1$. Incorporating this in to the first term of the sum and treating both sides of the inequality the same gives:

$$\sum_{i=1}^{n} a_{1,i} + \sum_{i=1}^{n} a_{2,i} \leq n - 1$$

You can ask yourself; what do these sums represent? They count the number of non-zero terms $a_{1,k}$ and $a_{2,l}$ in the adjacency matrix; so they count the number of edges involving v_1 and v_2. This means that the sums are the degrees of v_1 and v_2 in Γ', and so you can write:

$$d_{\Gamma'}(v_1) + d_{\Gamma'}(v_2) \leq n - 1$$

But by the assumption on Γ':

$$d_{\Gamma'}(v_1) + d_{\Gamma'}(v_2) \geq n/2 + n/2 = n$$

This is a contradiction, and so Γ must be Hamiltonian. □

Remark. You can see that the converse of this theorem is **not** true. The flattened dodecahedron D is Hamiltonian, but it has 20 vertices and the degree of any vertex v in D is 3, which is definitely less than $20/2 = 10$. So this is not the if and only if statement that is coveted by graph theorists everywhere.

3.3 Planar graphs

Definition 3.3.1. A graph Γ is a **plane graph** if it is drawn in the plane with its edges intersecting only at vertices of Γ'.

A graph Γ is **planar** if it is *isomorphic* to a plane graph Γ'. It follows that every plane graph is a planar graph.

As with most definitions in graph theory, examples really help in understanding them!

Example 3.3.2. (1) You are given some graphs in Figure 3.21; these are some examples of plane (and therefore planar) graphs.

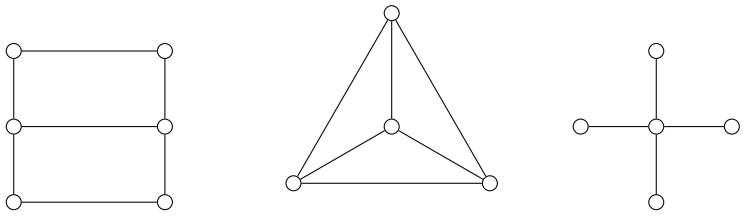


Figure 3.21: Examples of plane graphs

The middle graph is a tetrahedron flattened into the plane; similar to the flattened dodecahedron in Figure 3.18. You can notice that the graph on the right is a tree; all trees are planar graphs.

(2) K_4 is a planar graph. You can see from Figure 3.22 that K_4 is isomorphic to the plane graph Γ, and so K_4 is a planar graph by Definition 3.3.1.

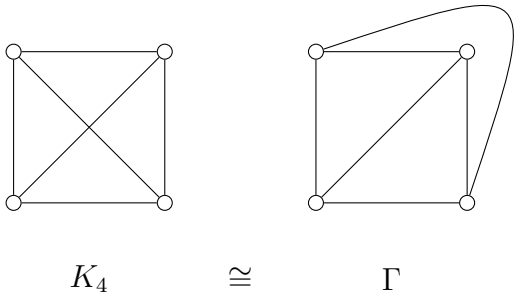


Figure 3.22: K_4 is planar

In fact, the complete graph is planar for $n = 1, 2, 3$ as well; look at the graphs drawn in Example 3.1.16! You can then ask yourself– is K_n planar for all natural numbers n? The next result answers this question.

Theorem 3.3.3. K_5 *is* **not** *planar.*

Proof. Let's try and draw K_5. First of all, there are five vertices $\{1, 2, 3, 4, 5\}$ and so you can start with the pentagon drawn in Figure 3.23:

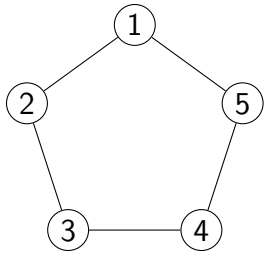


Figure 3.23: Drawing K_5 in the plane: stage one

Since a complete graph contains an edge between every pair of vertices, there must be an edge between vertices 1 and 3. There is a choice here; the edge can be inside the pentagon, or it can be outside, around the vertex 2. Let's put it inside the pentagon (see Figure 3.24).

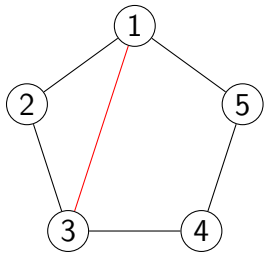


Figure 3.24: Drawing K_5 in the plane, stage two: new edge in red

Now, let's add in a few more edges. The edge between 2 and 5 must be outside the pentagon, as if it was inside it would cross the edge between 1 and 3. The edge between 1 and 4 must be inside the pentagon, as then it would cross the edge between 2 and 5. Finally, the edge between 2 and 4 must be outside the pentagon, as otherwise it would cross the edge between 1 and 4. This process is illustrated in Figure 3.25.

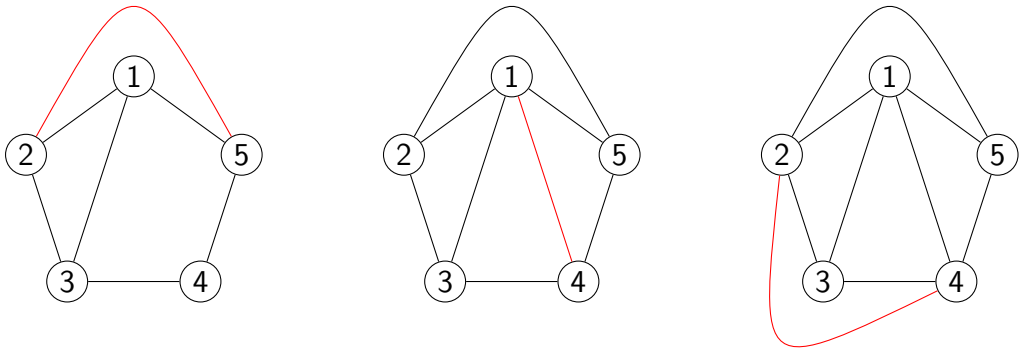


Figure 3.25: Drawing K_5 in the plane, stages three, four, five: new edges in red

You still need to draw the edge between 3 and 5. This cannot be outside the pentagon as

it would cross the edge between 2 and 4. It cannot be inside the pentagon as it would cross the edge between 1 and 4. Therefore, K_5 is not isomorphic to a plane graph; so K_5 is not planar. □

This is one famous example of a non-planar graph. The other famous example is found by looking at another type of graph.

Definition 3.3.4. A graph $\Gamma = (V, E)$ is called **bipartite** if $V = V_1 \cup V_2$ with $V_1 \cap V_2 = \varnothing$ and every edge of Γ has one vertex in V_1 and the other vertex in V_2.

Example 3.3.5. (1) The graph in Figure 3.26 is an example of a bipartite graph, with vertices in V_1 coloured in red and vertices in V_2 coloured in blue.

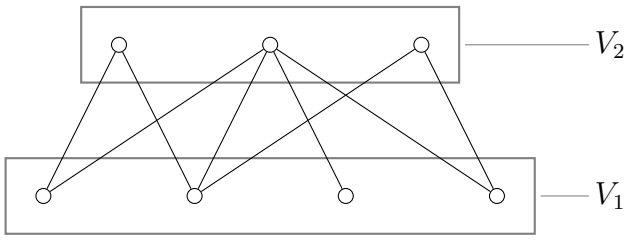


Figure 3.26: A bipartite graph, with V_1 and V_2 highlighted

(2) If every vertex in V_1 is joined to every vertex in V_2, you have what is called a **complete bipartite graph**. You can write $K_{m,n}$ for the complete bipartite graph with $|V_1| = m$ and $|V_2| = n$. The complete bipartite graph $K_{3,3}$ is drawn in Figure 3.27.

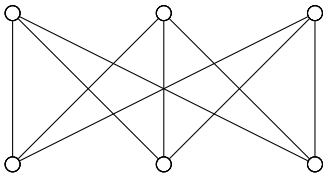


Figure 3.27: The complete bipartite graph $K_{3,3}$

You can notice that this explains the notation $K_{1,n}$ for the star graph with $n+1$ vertices in Example 3.1.22; it is a special example of a complete bipartite graph.

The reason for picking $K_{3,3}$ as the drawn example in Figure 3.27 is justified in this next result.

Theorem 3.3.6. *The complete bipartite graph $K_{3,3}$ is not planar.*

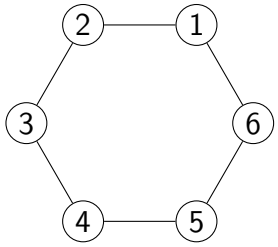


Figure 3.28: Drawing $K_{3,3}$ in the plane; the hexagonal circuit

Proof. Let $V_1 = \{1, 3, 5\}$ and let $V_2 = \{2, 4, 6\}$. Since $K_{3,3}$ contains a hexagonal circuit, you can draw one of these first and try and add in the remaining edges; this tactic is similar to the proof of Theorem 3.3.3. Drawing such a circuit in Figure 3.28 gives:

In order to draw a graph isomorphic to $K_{3,3}$, there are three edges left to add: $\{1, 4\}$, $\{3, 6\}$ and $\{5, 2\}$. Any two of these edges are either both inside the hexagon (in which case they cross) or both are outside the hexagon (in which case they cross). So it is impossible to draw $K_{3,3}$ in the plane. □

Remark. This is why you cannot solve the famous 'three utilities' problem (click link for details).

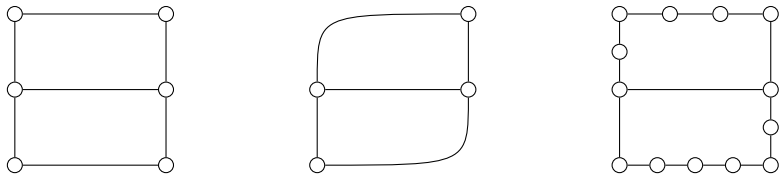
The significance of Theorems 3.3.3 and 3.3.6 is going to be important in a resulting classification of non-planar graphs. But first, you need a definition to express this if and only if statement.

Definition 3.3.7. Two graphs $\Gamma = (V, E)$ and $\Gamma' = (V', E')$ are called **homeomorphic** if Γ' can be obtained from Γ by inserting or deleting a number of vertices of degree 2.

You can think of homeomorphic graphs as being the same 'shape'. Adding or deleting vertices of degree 2 somehow doesn't change the overall 'shape' of the graph; as evidenced in th next example.

Example 3.3.8. The following three graphs $\Gamma_1, \Gamma_2, \Gamma_3$ given in Figure 3.29 are homeomorphic to each other. Γ_2 is obtained from Γ_1 by deleting the top left and bottom right vertices. Γ_3 is obtained from Γ_1 by the addition of a number of vertices.

Similar to Theorem 3.2.3, you can now give an if and only if statement that classifies non-planar (and therefore planar) graphs.



Γ_1 Γ_2 Γ_3

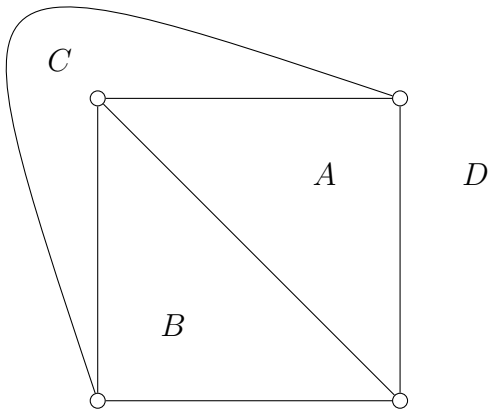
Figure 3.29: Γ_1, Γ_2 and Γ_3 are homeomorphic to each other

Theorem 3.3.9 (Kuratowski 1930)**.** *Let $\Gamma = (V, E)$ be a graph. Then Γ is non-planar if and only if it there is a graph $\Gamma' = (V', E')$ with $V' \subseteq V$ and $E' \subseteq E$ where Γ' is homeomorphic to either K_5 or $K_{3,3}$.* □

Proof. As homeomorphism does not affect whether or not a graph is planar or not. It follows that if there exists such a Γ' contained in Γ that is homeomorphic to K_5 or $K_{3,3}$, then Γ is not planar as K_5 is not planar (Theorem 3.3.3) and as $K_{3,3}$ is not planar (Theorem 3.3.6).

The other direction is much harder, and is not included here. □

You can notice that planar graphs divide the plane into a number of regions. For example, Figure 3.30 contains a plane graph Γ' isomorphic to K_4. You can notice that the spaces that are contained by edges have been given names A, B, C and that there is a D on the outside.



Γ'

Figure 3.30: Regions of Γ'

The regions obtained around a planar graph Γ are known as **faces**. The region outside the graph Γ is called the **infinite face** of Γ. The reason why these regions are called faces is due to the application of this next important theorem to polyhedra such as the dodecahedron.

Theorem 3.3.10 (Euler)**.** *Let $\Gamma = (V, E)$ be a connected planar graph which may not be simple. Let v, e and f be the number of vertices, edges and faces of Γ respectively. Then*

$$v - e + f = 2$$

Proof. This proof is by induction on the number of edges e.

Base case: Suppose that $e = 0$. In this case, the graph Γ is a single vertex. This means that $v = 1$, and the number of faces f is also 1; this is the infinite face. As $1 - 0 + 1 = 2$, the statement holds when $e = 0$.

Inductive case: Assume for the inductive hypothesis that the statement holds for all connected planar graphs with fewer than e edges.

Select some edge m in Γ and delete that edge. This then gives a graph $\Gamma = (V', E')$ with v' vertices, e' edges and f' faces. Since $e' = e - 1 < e$, it follows that $v' - e' + f' = 2$ by the inductive hypothesis.

There are three possibilities for the edge m; either m is a loop, m joins two distinct vertices of Γ', or m is incident to only one vertex in Γ' (so that when m is deleted from Γ, it leaves a vertex of degree 0, and you have to remove this to satisfy the hypothesis of the theorem.). You can look at each of these possibilities in turn.

Case 1 (m is a loop): Adding this loop back to recreate Γ introduces another face, so $f' + 1 = f$. This does not change the number of vertices, so $v' = v$. As $e = e' + 1$, you can now write

$$v - e + f = v' - (e' + 1) + (f' + 1) = v' - e' + f' = 2$$

and so the statement is true in this case. (See Figure 3.31 for an example of this process.)

Case 2 (m joins two distinct vertices of Γ): Adding m back splits one of the faces of Γ' into two faces; this means that $f' + 1 = f$. Since you do not add in any more vertices,

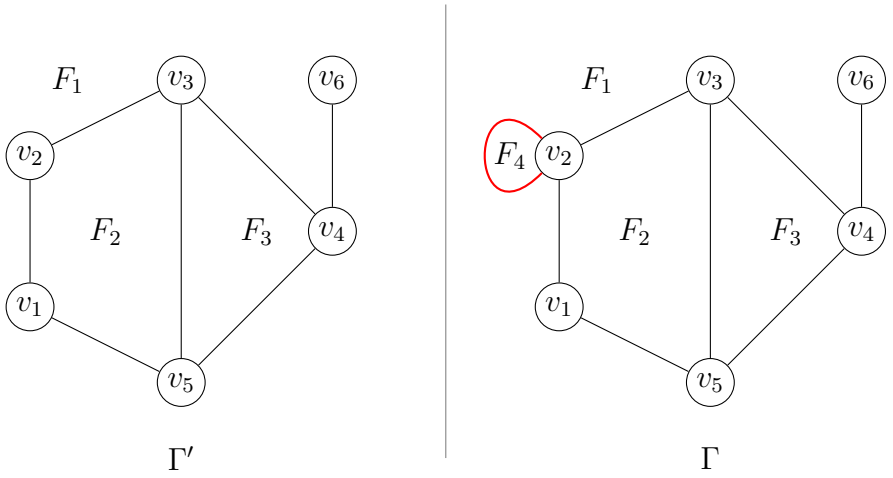


Figure 3.31: Example of process in Case 1, added edge in red

it follows that $v = v'$. As $e = e' + 1$, you can write that

$$v - e + f = v' - (e' + 1) + (f' + 1) = v' - e' + f' = 2$$

and so the statement is true in this case. (See Figure 3.32 for an example of this process.)

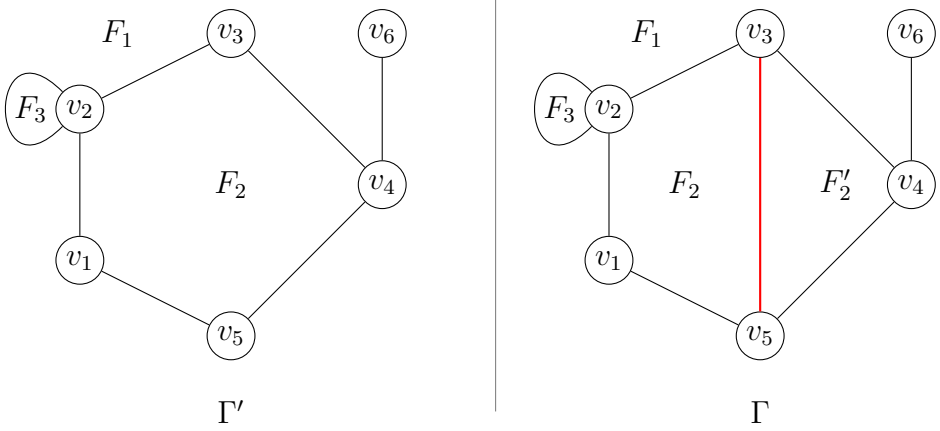


Figure 3.32: Example of process in Case 2, added edge in red

Case 3 (m incident to one vertex in Γ'): Adding this edge back into the graph also adds in an extra vertex; because this vertex was deleted along with the edge (a reason for this is given before the case analysis). So in this case $v = v' + 1$. You can see that

as this new edge does not bound a region, then $f = f'$. By the fact that $e = e' + 1$ you can write

$$v - e + f = (v' + 1) - (e' + 1) + f' = v' - e' + f' = 2$$

and so the statement is true in this case. (See Figure 3.33 for an example of this process.)

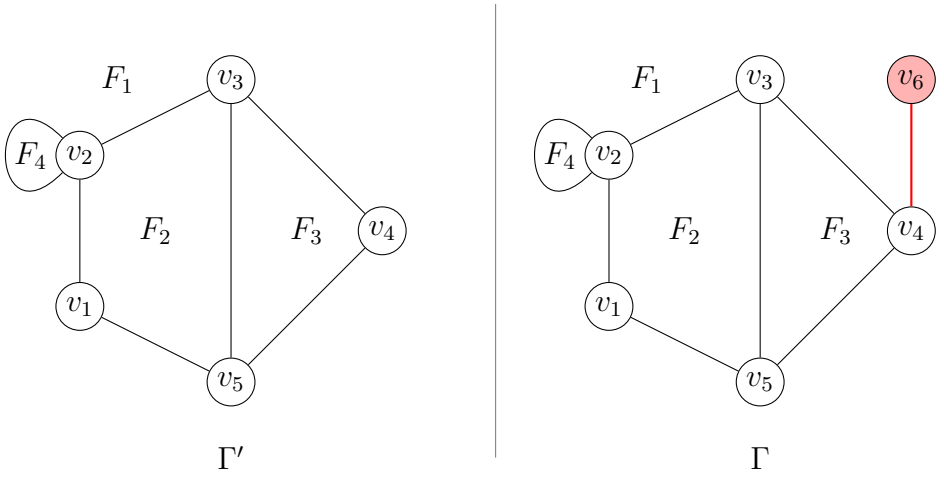


Figure 3.33: Example of process in Case 2, added edge and vertex in red

Since these are the only possible cases, the proof is complete. □

Remark (Non-examinable remark). Why 2? This is due to the fact that the surface on which you draw planar graphs has no holes that you can draw edges through. (You are drawing them on the *plane*; not a torus, for instance.) This is an important concept in the mathematical study of topology known as the Euler characteristic (clickable link).

Furthermore, this is one of at least **twenty** different proofs of this formula; you can follow this link to find some more. Some are harder than others.

Example 3.3.11. Consider the flattened dodecahedron as in Figure 3.18. A dodecahedron is so named as it has twelve faces. You can count the number of regions in Figure 3.18 to see that there are twelve. Don't forget about the infinite face!

In fact, **any** polyhedron can be 'flattened' in the plane like the dodecahedron and the tetrahedron (seen in Figure 3.21). So Euler's formula for the vertices, edges and faces of polyhedrons can be derived from Theorem 3.3.10.

You can use Euler's formula in Theorem 3.3.10 to prove a useful corollary.

Corollary 3.3.12. *Let Γ be a connected, simple, planar graph with v vertices, $e \geq 2$ edges and f faces. Then $3f \leq 2e$ and $e \leq 3v - 6$.*

Proof. You can assume that Γ is a plane graph in this case. By the assumption that Γ is simple (see Definition 3.1.15), it follows that Γ has neither loops nor multiple edges. This means that the only way a face is bounded by fewer than three edges is if that face is the infinite face and Γ is the graph ○—○—○. In this case, $3f \leq 2e$ (as $3 \leq 4$) and $e \leq 3v - 6$ (as $2 \leq 3$).

So you can assume that each face is bounded by at least three edges. Now, add up the number of edges around each face; this gives

$$\sum_{\text{all faces } F} (\text{edges bounding } F) \geq \sum_{\text{all faces } F} 3 = 3f$$

Here, each edge lies on at most two faces, and so the sum over all faces F of edges bounding F is twice the number of edges. Therefore, $2e \geq 3f$.

By Euler's formula Theorem 3.3.10, it follows that $v - e + f = 2$. Multiplying this through by 3 and rearranging gives:

$$3e = 3v + 3f - 6$$

You have just proved that $3f \leq 2e$; you can then say that

$$3e \leq 3v + 2e - 6$$

and subtracting $2e$ from both sides of this inequality gives $e \leq 3v - 6$. □

Finally in this chapter, you can use Corollary 3.3.12 together with Euler's formula Theorem 3.3.10 to re-prove some results that otherwise took some lengthy arguments to do.

Example 3.3.13. You are given the complete graph K_5 on five vertices and the complete bipartite graph $K_{3,3}$ in Figure 3.34.

You showed in Theorem 3.3.3 and Theorem 3.3.6 that these graphs are non-planar; you can also prove these results by using Corollary 3.3.12 and Euler's theorem.

Here, K_5 has five vertices and ten edges. But $3v - 6 = 9 \leq 10$; this is impossible for a planar graph due to Corollary 3.3.12. Counting the vertices and edges of $K_{3,3}$ gives six vertices and nine edges. If this graph were planar, then each face would have to be bound

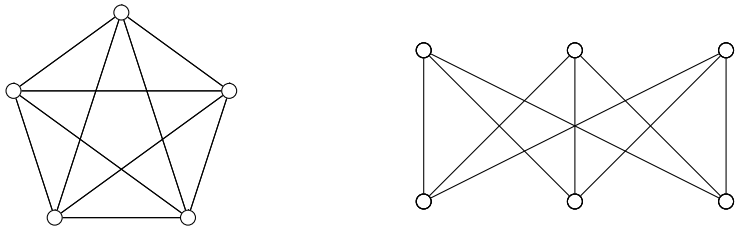


Figure 3.34: The complete graph K_5 (on left) and the complete bipartite graph $K_{3,3}$ (on right)

by at least **four** edges; this is because bipartite graphs can't have triangles in them. You can adapt the argument in the proof of Corollary 3.3.12 to say that

$$4f \leq 2e = 18$$

But Euler's formula gives $f = e - v + 2 = 5$; contradicting the above statement.

Chapter 4

Group theory

Let X, Y be sets. You have seen in Chapter 2 that a **bijective function** $f : X \to Y$ is defined to be a function that is injective and surjective. You may also recall that such a function is called a **one-to-one correspondence**.

Bijective functions $f : X \to X$ from a set to itself play an important role in the study of **symmetry**. The first concepts of symmetry you may have come across are the reflection and rotation of shapes. You can view these as bijective functions from the set of corners of a shape to itself that preserve the structure of the shape.

Crucially, you can collect together bijective functions from a set to itself. If you compose a bijective function $f : X \to X$ with a bijective function $g : X \to X$, you get another bijective function $h = g \circ f : X \to X$. This means that the set $\mathrm{Sym}(X)$ of **all** bijective functions is what is called **closed** under multiplication. In fact, this set $\mathrm{Sym}(X)$ is a mathematical structure called a **group**. Groups are important ideas in mathematics because they are used to study the symmetry of mathematical objects such as a graph.

This final chapter of the course looks at the theory of groups, by first looking at **permutations** and **permutation groups** before moving towards the more abstract idea of a group.

4.1 Permutations

4.1.1 Starting out with permutations

First of all, you should look at the section title and ask: what's a permutation? Here is the definition:

Definition 4.1.1. Let X be a non-empty set. A bijective function $f : X \to X$ is known as a **permutation** of X.

So whenever you see the phrase 'permutation of X', you should think 'bijective function from X to itself'. In this course, X will be a **finite** set of n elements, usually written $X = \{1, 2, \ldots, n\}$. Here, whenever the word 'element' is referred to in these notes **without** a set next to it, you should read 'element of the set $\{1, 2, \ldots, n\}$'.

Definition 4.1.2. Let $X = \{1, 2, \ldots, n\}$ be a finite set. The set of all permutations of X is called the **symmetric group** $\mathrm{Sym}(n)$ on n symbols.

Remark. It is common for the symmetric group $\mathrm{Sym}(n)$ to be called different things, such as S_n or even $\mathfrak{S}_n$.

(Non-examinable) You can define the symmetric group on an infinite set, such as the natural numbers. This then represents the collection of all bijective functions from the natural numbers to the natural numbers.

Notation (Important: writing permutations). If $f \in \mathrm{Sym}(n)$ is a permutation of the set $X = \{1, 2, \ldots n\}$, then you can write xf for the image of the element $x \in X$ after application of f. This is different from your usual function notation!

The reason this is useful is that it makes the composition of permutations much easier, and eliminates any confusion. For instance, the permutation fg means 'apply f first and then apply g'. This is a more natural way of reading composition of functions.

If $f \in \mathrm{Sym}(n)$, then you can write f in **two row notation**. To do this, first write the elements of $X = \{1, 2, \ldots n\}$ in a line. Then for every $k \in \{1, 2, \ldots n\}$, write kf (the image of k under f) is directly below k. This means that you can write the permutation $f \in \mathrm{Sym}(n)$ like this:

$$\begin{pmatrix} 1 & 2 & 3 & \ldots & n \\ 1f & 2f & 3f & \ldots & nf \end{pmatrix}$$

Finally, if $kf = k$ for some $k \in \{1, 2, \ldots, n\}$, then you can say that the permutation f **fixes** the point k.

Example 4.1.3. You can write the permutation $f \in \mathrm{Sym}(4)$ sending 1 to 2, 2 to 4, 3 to 1 and 4 to 3 as:

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

You can notice here that as f is a bijective function, all of the elements of $X = \{1, 2, \ldots, n\}$ must occur **exactly once** in the second row of the permutation. This is why they are called permutations; elements of $\mathrm{Sym}(n)$ reorder the elements of $X = \{1, 2, \ldots n\}$.

Example 4.1.4. [**Important: composing permutations**] Let $f, g \in \mathrm{Sym}(4)$ be given by the permutations

$$f = \begin{pmatrix} 1 & 2 & 3 & 4 \\ 2 & 4 & 1 & 3 \end{pmatrix} \qquad g = \begin{pmatrix} 1 & 2 & 3 & 4 \\ 3 & 2 & 4 & 1 \end{pmatrix}$$

Suppose you were asked to find the composition of the permutation f with the permutation g; in the notation above, this is fg. You can use two row notation to help you calculate this composition.

You can start by writing the composition you are trying to find:

$$fg = \begin{pmatrix} 1 & 2 & 3 & 4 \\ 2 & 4 & 1 & 3 \end{pmatrix} \circ \begin{pmatrix} 1 & 2 & 3 & 4 \\ 3 & 2 & 4 & 1 \end{pmatrix} =$$

Now, f sends 1 to 2 and g sends 2 to 2. So the composite permutation fg sends 1 to 2 to 2; therefore, fg maps 1 to 2. (Notice that the first 2 is not needed; only the start and end points are.) You can write this in two row notation as follows:

$$fg = \begin{pmatrix} 1 & 2 & 3 & 4 \\ 2 & 4 & 1 & 3 \end{pmatrix} \circ \begin{pmatrix} 1 & 2 & 3 & 4 \\ 3 & 2 & 4 & 1 \end{pmatrix} = \begin{pmatrix} 1 & 2 & 3 & 4 \\ 2 & & & \end{pmatrix}$$

As f sends 2 to 4 and g sends 4 to 1, it follows that fg sends 2 to 4 to 1; so it maps 2 to 1. (Notice that the 4 is not needed; only the start and end points are.) You can write this in to get

$$fg = \begin{pmatrix} 1 & 2 & 3 & 4 \\ 2 & 4 & 1 & 3 \end{pmatrix} \circ \begin{pmatrix} 1 & 2 & 3 & 4 \\ 3 & 2 & 4 & 1 \end{pmatrix} = \begin{pmatrix} 1 & 2 & 3 & 4 \\ 2 & 1 & & as \end{pmatrix}$$

Now, you can see that as f send 3 to 1 and g sends 1 to 3, it follows that fg sends 3 to 3. You can write this in:

$$fg = \begin{pmatrix} 1 & 2 & 3 & 4 \\ 2 & 4 & 1 & 3 \end{pmatrix} \circ \begin{pmatrix} 1 & 2 & 3 & 4 \\ 3 & 2 & 4 & 1 \end{pmatrix} = \begin{pmatrix} 1 & 2 & 3 & 4 \\ 2 & 1 & 3 & \end{pmatrix}$$

Finally, as f sends 4 to 3 and g sends 3 to 4, then fg maps 4 to 4 and so you can write:

$$fg = \begin{pmatrix} 1 & 2 & 3 & 4 \\ 2 & 4 & 1 & 3 \end{pmatrix} \circ \begin{pmatrix} 1 & 2 & 3 & 4 \\ 3 & 2 & 4 & 1 \end{pmatrix} = \begin{pmatrix} 1 & 2 & 3 & 4 \\ 2 & 1 & 3 & 4 \end{pmatrix}$$

So you've found fg. Notice you could find this another way: all elements of $X = \{1, 2, 3, 4\}$ appear exactly once in the bottom row, you can fill in the 4 as it is the remaining element of $\{1, 2, 3, 4\}$.

You may also be asked to find gf; the composition of the permutation g with the permutation f. You can use this process to find gf, which is:

$$gf = \begin{pmatrix} 1 & 2 & 3 & 4 \\ 3 & 2 & 4 & 1 \end{pmatrix} \circ \begin{pmatrix} 1 & 2 & 3 & 4 \\ 2 & 4 & 1 & 3 \end{pmatrix} = \begin{pmatrix} 1 & 2 & 3 & 4 \\ 1 & 4 & 3 & 2 \end{pmatrix}$$

It is really important to notice that:

$$fg = \begin{pmatrix} 1 & 2 & 3 & 4 \\ 2 & 1 & 3 & 4 \end{pmatrix} \neq \begin{pmatrix} 1 & 2 & 3 & 4 \\ 1 & 4 & 3 & 2 \end{pmatrix} = gf$$

Like matrix multiplication, the composition of two permutations is not necessarily commutative.

Sometimes, this will be referred to as finding the **product** of two permutations f and g.

There is a special permutation that does nothing to the elements of $\{1, 2, \ldots, n\}$; this is the subject of the next definition.

Definition 4.1.5. The **identity permutation** $e \in \mathrm{Sym}(n)$ is defined to be the permutation

such that $ke = k$ for all $k \in \{1, 2, \dots n\}$. It is written in two row notation as

$$\begin{pmatrix} 1 & 2 & 3 & \dots & n \\ 1 & 2 & 3 & \dots & n \end{pmatrix}$$

Informally, the identity permutation e is one that doesn't move any elements of $X = \{1, 2, \dots, n\}$.

Remark. You can show that for any $f \in \mathrm{Sym}(n)$, then $f \circ e = f = e \circ f$, where e is the identity permutation in $\mathrm{Sym}(n)$.

The final note in this section deals with the amount of possible permutations of a set $X = \{1, 2, \dots, n\}$.

Lemma 4.1.6. *Let $n \in \mathbb{N}$. Then there are $n!$ many permutations in $\mathrm{Sym}(n)$.*

Proof. The proof relies on the fact that this is the collection of bijective functions from $X = \{1, 2, \dots, n\}$ to itself. Suppose that you are choosing a permutation f; the idea of the proof is to count how many ways you can do this.

Here, there are n many points of X that you could send 1 to using f. How many points are there to send 2 to? As the function you are constructing is bijective, you cannot send 2 to the same point that you send 1 to. Therefore, there are $n - 1$ choices for the image of 2 under f. Similarly, there are $n - 2$ choices for the image of 3, and so on until there is just 1 choice for the image of n under f.

The total number of choices for f is given by multiplying together the expressions for all the choices; this is

$$n \cdot (n-1) \cdot (n-2) \cdot \ldots \cdot 2 \cdot 1 = n!$$

So there are $n!$ choices for f and therefore $n!$ many permutations in $\mathrm{Sym}(n)$. □

4.1.2 Cycle notation

Two row notation is useful for composing permutations, but it is quite inefficient; particularly if you are working with larger sets $X = \{1, 2, \dots n\}$. You could represent permutations in a different, shorter way. For instance, the permutation $f \in \mathrm{Sym}(8)$ given by:

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

is long to write, and also sends the elements 1, 3 and 5 of $\{1, 2, \ldots, 8\}$ to themselves (they are **fixed**; see important notation on page 100 in Subsection 4.1.1). It would be a good thing to have a shorter way to write f without writing $1, 3, 5$ out twice. This motivates the definition of **cycle notation**.

Definition 4.1.7. Let $x_1, x_2, \ldots, x_r$ be r distinct elements of $\{1, 2, \ldots, n\}$ (so here, $1 \leq r \leq n$). Then the r-**cycle** $(x_1\, x_2\, \ldots\, x_r)$ is the permutation in $\mathrm{Sym}(n)$ which maps

$$x_1 \mapsto x_2, \quad x_2 \mapsto x_3, \quad \ldots x_{r-1} \mapsto x_r, \quad x_r \mapsto x_1$$

and fixes all other points in $\{1, 2, \ldots, n\}$.

Remark. You could represent the cycle in a circular picture (similar to a **cycle graph** in Example 3.1.16); a picture of this is given in Figure 4.1.

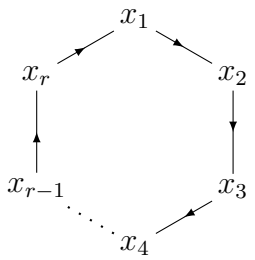


Figure 4.1: Drawing $(x_1\, x_2\, \ldots\, x_r)$

You can say that as long as these are the elements contained in the cycle, it doesn't matter where you start. So it follows that $(x_2\, x_3\, \ldots\, x_r\, x_1)$ and $(x_{r-1}\, x_r\, x_1\, \ldots\, x_{r-2})$ (for instance) are the same cycle as $(x_1\, x_2\, \ldots\, x_r)$.

Example 4.1.8. (1) The permutation $f \in \mathrm{Sym}(5)$ given by

$$f = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 \\ 1 & 3 & 5 & 4 & 2 \end{pmatrix}$$

fixes 1 and 4 and sends 2 to 3, 3 to 5 and 5 to 2. This is a 3-cycle on the elements 2, 3 and 5; so you can say that $f = (2\,3\,5)$.

(2) What about the identity permutation $e \in \mathrm{Sym}(n)$? There are no cycles here as every element $k \in \{1, 2, \ldots, n\}$ is fixed, so how could you write this in cycle notation? The answer is to notice that e sends 1 to 1 and fixes all the other elements; so you can write $e = (1)$. (Notice that you could write $e = (k)$ for any element $k \in \{1, 2, \ldots, n\}$; but to avoid confusion, probably best to stick to $e = (1)$.)

However, this hasn't quite solved our problem yet. You could notice that the permutation $f \in \mathrm{Sym}(8)$ given at the start of the section by:

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

has two different cycles inside; $(2\,4)$ and $(6\,7\,8)$. So how could you write f using cycle notation? The next two statements deal with this.

Definition 4.1.9. Two cycles $(x_1\,x_2\,\ldots\,x_r)$ and $(y_1\,y_2\,\ldots\,y_s)$ in $\mathrm{Sym}(n)$ are **disjoint** if no element in $\{1, 2, \ldots, n\}$ is moved by both cycles.

If these cycles do not represent the identity permutation, then they are not of length 1 (see Example 4.1.8 (2)) and so $2 \leq r, s \leq n$. If this happens, then the condition can be expressed as

$$\{x_1, x_2, \ldots, x_n\} \cap \{y_1, y_2, \ldots, y_s\} = \varnothing$$

Notice that these are disjoint sets (see remark before Corollary 2.2.12), which explains the name of the definition.

This leads into the following theorem. In the theory of permutations, the word 'composition' and 'product' can be used interchangeably. Sometimes, it is nicer to say product than composition and vice versa.

Theorem 4.1.10. *Let $n \in \mathbb{N}$. Then every permutation in $\mathrm{Sym}(n)$ can be written as a* **product of disjoint cycles**.

Proof. Omitted: it's not too hard, but Theorem 4.1.10 is perhaps best explained using examples. □

Example 4.1.11. (1) Let $f \in \mathrm{Sym}(8)$ be given by:

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

It was noted above that f has two cycles $(2\,4)$ and $(6\,7\,8)$, and fixes $1, 3$ and 5. You can write f as the **product** of these two cycles by

$$f = (2\,4)(6\,7\,8)$$

(2) Now, let $g \in \mathrm{Sym}(9)$ be given by:

$$g = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 \\ 1 & 4 & 3 & 2 & 5 & 7 & 8 & 6 & 9 \end{pmatrix}$$

You can notice that g has exactly the same cycles as f above, and fixes the elements $1, 3, 5$ and 9 of $\{1, 2, \ldots, 9\}$. So you can write $g = (2\,4)(6\,7\,8)$, but it is important to notice that $g \neq f$; as they are different functions (the domain of f and g is not the same). You should always state which symmetric group a particular permutation is in if you write it using cycle notation.

(3) Finally, let $h \in \mathrm{Sym}(10)$ be given by:

$$h = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 \\ 9 & 6 & 2 & 5 & 7 & 3 & 10 & 4 & 1 & 8 \end{pmatrix}$$

Here, h does not fix any points of the set $\{1, 2, \ldots, 10\}$. There are three cycles of h: 1 to 9 and back to 1; 2 to 6 to 3 and back to 2; and finally, 4 to 5 to 7 to 10 to 8 and then back to 4. This means you can write h in cycle notation as

$$h = (1\,9)(2\,6\,3)(4\,5\,7\,10\,8)$$

Notice that the order in which you write elements of $\{1, 2, \ldots, n\}$ inside a cycle really matters; here, $(2\,3\,6) \neq (2\,6\,3)$. However, you can start the cycle with any of these, as long as you preserve the order; so here, $(2\,3\,6) = (6\,2\,3) = (3\,6\,2)$.

You can compose two permutations in written in cycle notation as well.

Example 4.1.12. Let $f, g \in \mathrm{Sym}(5)$ be given by $f = (4\,5\,3)$ and $g = (1\,2\,3\,4\,5)$. By Definition 4.1.7, this means that f fixes the elements 1 and 2. You can work out fg in cycle notation by following images. Here, f sends 1 to 1 and g sends 1 to 2. This means that fg sends 1 to 2 and you can write

$$fg = (1\,2$$

Now, you can look at the next element in the cycle, which is 2. Here f sends 2 to 2 and g

sends 2 to 3; so you can say that g sends 2 to 3. Therefore:

$$fg = (1\,2\,3$$

The next element in the cycle is 3; since f sends 3 to 4 and g sends 4 to 5, this means that fg sends 3 to 5 and you can write

$$fg = (1\,2\,3\,5$$

Now look at the next element in the cycle, which is 5. Here f sends 5 to 3 and g sends 3 to 4; so fg maps 5 to 4, giving

$$fg = (1\,2\,3\,5\,4$$

Finally, f maps 4 to 5 and g maps 5 to 1; so fg maps 4 to 1. Since 1 is at the start of the cycle, you can close the brackets, giving

$$fg = (1\,2\,3\,5\,4)$$

as your final answer.

Before looking at properties of permutations, you can say something nice about 'generating' permutations. You know that every permutation of a set $\{1, 2, \ldots, n\}$ can be written as a product of disjoint cycles; this is Theorem 4.1.10. In fact, you can say something a little more general using cycles of length two.

Definition 4.1.13. A cycle of length two is called a **transposition**.

Corollary 4.1.14. *Every permutation can be expressed as a product of transpositions.*

Sketch of proof. Let $(x_1\,x_2\,\ldots\,x_r)$ be an r-cycle. Then you can say that:

$$(x_1\,x_2\,\ldots\,x_r) = (x_1\,x_2)(x_1\,x_3)\ldots(x_1\,x_r)$$

and the result follows from Theorem 4.1.10. □

4.1.3 Properties of permutations

Writing permutations as a product of disjoint cycles using Theorem 4.1.10 is a useful tool. Not only does it save time, energy and ink, but it can also tell you important information

about the permutation you are working on. Before this however, you need the following statements:

Definition 4.1.15. Let $n \in \mathbb{N}$. Two permutations $f, g \in \mathrm{Sym}(n)$ **commute** if $fg = gf$.

Remark. Notice that this property is **not** true in general; see Example 4.1.4 for a counterexample.

You can notice in Example 4.1.11 (3) that the order matters in writing elements of $\{1, 2, \ldots, n\}$ in a cycle; since (for instance) $(2\,3\,6) \neq (2\,6\,3)$. However, the order in which you write the *cycles* doesn't matter so much. This is demonstrated in the following lemma, which will not be proved here.

Lemma 4.1.16. *Disjoint cycles commute.* □

Example 4.1.17. Like in Example 4.1.11 (3), let $h \in \mathrm{Sym}(10)$ be given by:

$$h = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 \\ 9 & 6 & 2 & 5 & 7 & 3 & 10 & 4 & 1 & 8 \end{pmatrix}$$

You worked out that $h = (1\,9)(2\,6\,3)(4\,5\,7\,10\,8)$; by Lemma 4.1.16, you can also say that $h = (2\,6\,3)(1\,9)(4\,5\,7\,10\,8)$ or $h = (4\,5\,7\,10\,8)(2\,6\,3)(1\,9)$.

You can notice that the cycle $(2\,6\,3)$ commutes with h: you can say this as

$$\begin{aligned}(2\,6\,3)h &= (2\,6\,3)(1\,9)(2\,6\,3)(4\,5\,7\,10\,8) \\ &= (1\,9)(2\,6\,3)(2\,6\,3)(4\,5\,7\,10\,8) \\ &= (1\,9)(2\,6\,3)(4\,5\,7\,10\,8)(2\,6\,3) = h(2\,6\,3)\end{aligned}$$

Similarly to how you would multiply an integer a by itself k times to get a^k, you can compose a permutation f with itself k times to get another permutation. So here

$$f^k = \underbrace{f \circ f \circ f \circ \ldots \circ f}_{k \text{ times}}$$

It is a property of permutations that for any natural number n and for any permutation $f \in \mathrm{Sym}(n)$ then there exists a $k \in \mathbb{N}$ such that $f^k = e$, where e is the identity permutation. The smallest such k has a special name:

Definition 4.1.18. Let $n \in \mathbb{N}$ and suppose that $f \in \mathrm{Sym}(n)$. The **order** of the permutation f is the smallest natural number k such that $f^k = e$, where e is the identity permutation.

This is an important definition in the theory of permutations, so you can ask: how can I calculate this? Well, what you could do is calculate f, f^2, f^3, and so on, until you find a k such that $f^k = e$. The problem with this is that it can be very long, particularly if the permutation is defined on a set with lots of elements. For instance, would you want to calculate powers of h in Example 4.1.17? Here's a result that will help, using the fact that you can write a permutation as a product of disjoint cycles.

Theorem 4.1.19. *Let $n \in \mathbb{N}$ and suppose that $f \in \mathrm{Sym}(n)$ is written as a product of disjoint cycles*

$$f = f_1 f_2 \dots f_m$$

Then the order of f is the lowest common multiple of the lengths of the disjoint cycles.

Proof. Write f in disjoint cycle notation:

$$f = f_1 f_2 \dots f_m$$

Suppose that a cycle $f_i = (x_1 x_2 \dots x_r)$ (with $1 \leq i \leq m$) is of length r. What is the order of f_i on the set $\{x_1, \dots x_r\}$? Here, you could see that f_i^2 sends x_1 to x_3, and f_i^3 sends x_1 to x_4. It follows that f_i^{r-1} sends x_1 to x_r, and so

$$x_1 f_i^r = x_1 f_i^{r-1} \circ f_i = x_r f_i = x_1$$

So $x_1 f_i^r = x_1$. You can use a similar argument to show that $x_j f_i^r = x_j$ for all $1 \leq j \leq r$. Therefore, f_i^r is the identity permutation on the set $\{x_1, \dots x_r\}$; so the order of f_i is r.

Now, consider $f = f_1 f_2 \dots f_m$. You can raise f to the power of k; since disjoint cycles commute by Lemma 4.1.16, you can write that:

$$f^k = f_1^k f_2^k \dots f_m^k$$

You can say that f^k is the identity permutation e if and only if f_i^k is the identity permutation for all $1 \leq i \leq m$. This means that the smallest such k is the lowest common multiple of the orders of the disjoint cycles f_i. You showed that the order of f_i is the same as its length: so the result is proved. □

The final thing you can do with permutations is to 'undo' them. In fact, this is one of the defininings properties of permutations:

Definition 4.1.20. Let $f \in \mathrm{Sym}(n)$ be a permutation. Then the **inverse** of f is the *unique* permutation f^{-1} such that $f \circ f^{-1} = e = f^{-1} \circ f$, where e is the identity permutation.

Informally, you can think of the inverse f^{-1} of the permutation f as 'reversing' the effects of f. This can be seen in the first of two methods to find the inverse of a permutation.

Example 4.1.21 (Method: Two row notation). Let $h \in \mathrm{Sym}(10)$ be as in Example 4.1.11 (3) and Example 4.1.17:

$$h = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 \\ 9 & 6 & 2 & 5 & 7 & 3 & 10 & 4 & 1 & 8 \end{pmatrix}$$

You can swap over the rows of h to get:

$$h^{-1} = \begin{pmatrix} 9 & 6 & 2 & 5 & 7 & 3 & 10 & 4 & 1 & 8 \\ 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 \end{pmatrix}$$

While this is technically the inverse permutation of h, it looks quite messy. You can now rearrange the columns of this to get

$$h^{-1} = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 \\ 9 & 3 & 6 & 8 & 4 & 2 & 5 & 10 & 1 & 7 \end{pmatrix}$$

which looks much nicer. You could (and should) check your answer by composing the two permutations h and h^{-1} to see if you get the identity permutation.

$$h \circ h^{-1} = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 \\ 9 & 6 & 2 & 5 & 7 & 3 & 10 & 4 & 1 & 8 \end{pmatrix} \circ \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 \\ 9 & 3 & 6 & 8 & 4 & 2 & 5 & 10 & 1 & 7 \end{pmatrix} = e$$

If your permutation is already written in disjoint cycle notation, then you don't need to worry; if you can invert one cycle, you can invert them all.

So if $g = (x_1\, x_2\, \ldots\, x_r)$, then the inverse permutation g^{-1} reverses this cycle: g^{-1} sends x_r to x_{r-1}, x_{r_1} to x_{r-2}, and so on, all the way up to sending x_2 to x_1 and x_1 to x_r. You can

write this using cycle notation as

$$g^{-1} = (x_r\, x_{r-1}\, \ldots\, x_2\, x_1)$$

Since you can start the cycle at any point in this sequence, you can write

$$g^{-1} = (x_1\, x_r\, x_{r-1}\, \ldots\, x_2)$$

So to invert a cycle g, you could reverse the entire cycle and move the first element of $\{1, 2, \ldots, n\}$ in the cycle g (or the final element of $\{1, 2, \ldots, n\}$ in the reversed cycle) to the start. Let's see an example of this process in action.

Example 4.1.22 (Method: Disjoint cycle notation). Once again, let $h \in \mathrm{Sym}(10)$ be the permutation

$$h = (1\,9)(2\,6\,3)(4\,5\,7\,10\,8)$$

Reversing these cycles gives:

$$h^{-1} = (9\,1)(3\,6\,2)(8\,10\,7\,5\,4)$$

Moving the last elements of $\{1, 2, \ldots, n\}$ in each of these cycles to the start of their respective cycles gives:

$$h^{-1} = (9\,1)(2\,3\,6)(4\,8\,10\,7\,5)$$

You can notice that this is the same answer you gave for h^{-1} in Example 4.1.21.

4.2 Groups

In the previous section, you have seen many properties of elements of the collection $\mathrm{Sym}(n)$ of all permutations on a set $\{1, 2, \ldots, n\}$. For instance, you know from Example 4.1.4 that the composition fg of two permutations $f, g \in \mathrm{Sym}(n)$ is again a permutation; in addition, the permutation fg may not be the same as the permutation gf. You also know from Definition 4.1.5 that there is an identity permutation $e \in \mathrm{Sym}(n)$ such that $f \circ e = f = e \circ f$ for any $f \in \mathrm{Sym}(n)$. Furthermore, you know from Definition 4.1.20 that for every $f \in \mathrm{Sym}(n)$, there exists an inverse permutation f^{-1} such that $f \circ f^{-1} = e = f^{-1}$. Finally, if you try and calculate the composition of more than two permutations it doesn't matter how you do it; in symbols, you can say that $(f \circ g) \circ h = f \circ (g \circ h)$ for all $f, g, h \in \mathrm{Sym}(n)$.

As with other mathematical concepts you have seen, the idea here is to try and study different structures with the same properties; this is another example of generalisation. This is done by writing down strict mathematical definition and trying to prove some results. A set with these properties is called a **group**, and the study of groups is called **group theory**. You can show (see Example 4.2.8) that $\mathrm{Sym}(n)$ is a group.

The final section of the course introduces the idea of a group, and then moves on to some examples of groups in mathematics, as well as some initial theorems.

4.2.1 Defining a group

First of all, you can notice that the composition of permutations is common to all of the properties of $\mathrm{Sym}(n)$ listed above. So in order to generalise these properties, you need a way to combine two elements of a set to create a new one.

Definition 4.2.1. Let X be a set. A **binary operation** $\star$ is a function $\star : X \times X \to X$ defined by

$$\star : (x, y) \mapsto x \star y$$

You can think of a binary operation as a method for combining two elements of a set. Examples of binary operations on the set $\mathbb{Z}$ include addition $+$ and multiplication $\cdot$.

This definition is needed to express the following:

Definition 4.2.2. Let G be a set and let $\star$ be a binary operation on G. Say that the pair $(G, \star)$ is a **group** if $(G, \star)$ satisfies the following three axioms:

associativity: For all $a, b, c \in G$, then:

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

identity: There exists $e \in G$, called the **identity element**, such that for all $a \in G$:

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

inverse: For all $a \in G$ there exists an element $a^{-1} \in G$ such that

$$a \star a^{-1} = e = a^{-1} \star a$$

You can say that a^{-1} is the **inverse** of a.

Remarks. Sometimes, the identity element of a group is called $1 \in G$; be careful about this!

Some mathematicians list 'closure' as one of the axioms of group theory. This says that if $a, b \in G$, then $a \star b \in G$; which you've already built in to the definition of a binary operation in Definition 4.2.1.

You can notice that I've used the word 'the' to describe the identity element of a group and the inverse element of some $a \in G$; implying that these are unique in a group. This is a fact that will need to be proved:

Theorem 4.2.3. *Let $(G, \star)$ be a group. Then:*

(1) The identity element e of G is unique.

(2) The inverse a^{-1} of an element a in G is uniquely determined by a.

Proof. In order to show that something is unique in mathematics, you should take two different items with the same properties and show that they are the same.

(1) Suppose that e and f are two identity elements for the group $(G, \star)$. Then as e is an identity, $e \star f = f$ by Definition 4.2.2. Similarly, as f is an identity, then $e \star f = e$ by Definition 4.2.2. So $e = e \star f = f$, and therefore $e = f$.

(2) Now assume that x, y are both inverses for the element a of G. By Definition 4.2.2, this means that:

$$a \star x = e = x \star a \qquad \text{and} \qquad a \star y = e = y \star a$$

where e is the identity element of G. As this happens, you can say that

$$y = e \star y = (x \star a) \star y$$

Since $(G, \star)$ is a group, it is associative and so:

$$y = e \star y = (x \star a) \star y = x \star (a \star y)$$

As x is an inverse of a, you can write:

$$y = e \star y = (x \star a) \star y = x \star (a \star y) = x \star e = x$$

Therefore, $y = x$ and so they are the same element.

□

Sometimes, you can add some extra properties on top of a group. Here's one of them:

Definition 4.2.4. Let $(G, \star)$ be a group. If $a \star b = b \star a$ for all elements $a, b \in G$, then you can say that G is **abelian** (or sometimes **commutative**).

Remark (Important remark). Most groups $(G, \star)$ are **not** abelian. For instance, the set $\mathrm{Sym}(4)$ together with composition is a group. Here, Example 4.1.4 proves that this group is not abelian.

Finally, the number of elements in a group $(G, \star)$ has a special name.

Definition 4.2.5. The **order** of a group $(G, \star)$ is the number of elements in the set G. You can write the order of G as $|G|$.

Remark (Another important remark). It's important not to confuse the order $|G|$ of a group G for the definition of order of a permutation in Definition 4.1.18.

4.2.2 Examples of groups

Much like equivalence relations in Definition 2.2.9, in order to show that a pair $(G, \star)$ is a group, you need to prove that the three axioms of Definition 4.2.5 (associativity, identity and inverse) hold for the pair $(G, \star)$. In order to prove that the pair $(G, \star)$ is **not** a group, you could find a counterexample to one of the three axioms. Alternatively, you can show that the given operation is **not** a binary operation on the set G; usually by finding a pair $a, b \in G$ where $a \star b \notin G$.

Example 4.2.6. The pair $(\mathbb{Z}, +)$ is a group. You can show this by checking all the axioms in Definition 4.2.2.

associativity: You can notice that for all $a, b, c \in \mathbb{Z}$ that

$$a + (b + c) = (a + b) + c$$

which means that $(\mathbb{Z}, +)$ satisfies the associativity axiom.

identity: Here, $0 \in \mathbb{Z}$ is an element such that for all $a \in \mathbb{Z}$:

$$0 + a = a = a + 0$$

and so 0 is the identity element for $(\mathbb{Z}, +)$.

inverse: Finally, you can notice that for every $a \in \mathbb{Z}$, there exists $(-a)$ in $\mathbb{Z}$ such that

$$a + (-a) = 0 = (-a) + a$$

and here, $-a$ is the inverse element of $a \in \mathbb{Z}$.

Furthermore, as $a + b = b + a$ for all $a, b \in \mathbb{Z}$, you can say that $(\mathbb{Z}, +)$ is an abelian group by Definition 4.2.4.

You can adapt this example to show that the pairs $(\mathbb{Q}, +)$, $(\mathbb{R}, +)$ and $(\mathbb{C}, +)$ are all groups.

Remark. It's really important to remember that when you are working with the group $(\mathbb{Z}, +)$, you **can't** refer to multiplication. The triple $(\mathbb{Z}, +, \cdot)$ is an entirely different structure called a **ring** (see MT3505).

Example 4.2.7. (1) You can the set of all **positive** rational numbers by the set

$$\mathbb{Q}^+ = \{q \in \mathbb{Q} \ : \ q > 0\}$$

The pair $(\mathbb{Q}^+, \cdot)$ is a group.

associativity: For all $a, b, c \in \mathbb{Q}^+$ that

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

which means that $(\mathbb{Q}^+, \cdot)$ satisfies the associativity axiom.

identity: Here, $1 \in \mathbb{Q}^+$ is an element such that for all $a \in \mathbb{Q}^+$:

$$1 \cdot a = a = a \cdot 1$$

and so 1 is the identity element for $(\mathbb{Q}^+, \cdot)$.

inverse: Finally, you can notice that for every $x = p/q \in \mathbb{Q}^+$, then you can define $1/x = q/p \in \mathbb{Q}^+$. This element $1/x$ is such that

$$x \cdot (1/x) = 1 = (1/x) \cdot x$$

As $1 \in \mathbb{Q}^+$ is the identity element of $(\mathbb{Q}^+, \cdot)$, you can say that $1/x$ is the inverse element of $a \in \mathbb{Q}^+$.

You can adapt this example to prove that $(\mathbb{R}^+, \cdot)$ is also a group. Also, you can show that the pairs $(\mathbb{Q} \setminus \{0\}, \cdot)$, $(\mathbb{R} \setminus \{0\}, \cdot)$ and $(\mathbb{C} \setminus \{0\}, \cdot)$ are all groups.

(2) The pair $(\mathbb{Z}^+, \cdot)$ is not a group. This pair satisfies the associativity and the identity axioms, but does not satisfy the inverse axiom. A counterexample to this could be that there is no integer $a \in \mathbb{Z}^+$ such that $2 \cdot a = 1$.

(3) You can define the set $\mathbb{R}^-$ to be the set of all **negative** real numbers. Here, the pair $(\mathbb{R}, \cdot)$ is not a group as multiplication is not a binary operation on the set $\mathbb{R}^-$. You can see this by noticing that $-2, -3 \in \mathbb{R}^-$ but that $(-2) \cdot (-3) = 6 \notin \mathbb{R}^-$. This contradicts Definition 4.2.1; multiplication is not a function from $\mathbb{R}^- \times \mathbb{R}^-$ to $\mathbb{R}^-$.

Example 4.2.8. You know from Section 4.1 that $\mathrm{Sym}(n)$ is the set of permutations on the set $X = \{1, 2, \ldots, n\}$. You can use Definition 4.2.2 to show that the pair $(\mathrm{Sym}(n), \circ)$ is a group.

associativity: Let $f, g, h \in \mathrm{Sym}(n)$, and let $x \in X = \{1, 2, \ldots, n\}$. Then $xf(gh)$ means you first apply f to x, then apply gh to xf; this is the same as applying g and then h. So $xf(gh) = ((xf)g)h$. Similarly, $x(fg)h$ means that you first apply f then g to x, then apply h; so $x(fg)h = ((xf)g)h)$. Therefore $xf(gh) = x(fg)h$ for all $x \in X$; so the functions are the same, giving $(fg)h = f(gh)$ for all $f, g, h \in \mathrm{Sym}(n)$.

identity: Here, the identity permutation $e \in \mathrm{Sym}(n)$ is an element such that for all $f \in \mathrm{Sym}(n)$:

$$e \circ f = f = f \circ e$$

(by remark following Definition 4.1.5) and so e is the identity element for $(\mathrm{Sym}(n), \circ)$.

inverse: Finally, you can use Definition 4.1.20 to say that for any $f \in \mathrm{Sym}(n)$, there exists an $f^{-1} \in \mathrm{Sym}(n)$ such that is such that

$$f \circ f^{-1} = e = f^{-1} \circ f$$

As the identity permutation e is the identity element of $(\mathrm{Sym}(n), \circ)$, you can say that f^{-1} is the inverse element of $f \in \mathrm{Sym}(n)$.

Multiplication tables

When you were younger, you may have learned your times tables. As you have seen in Example 4.2.7, the set of all positive real numbers together with multiplication forms a group. Like a times table presents the products of some of the elements in this group in a table form, it follows that you can expand this to the idea of groups.

What you can do then is draw a multiplication table for a group. This is sometimes called a **Cayley table**, after noted group theorist Arthur Cayley. To draw a Cayley table, you can write the elements of a group $(G, \star)$ along the first row and the first column, and fill in the entry of row labelled a and column labelled b with the element $a \star b \in G$.

For example, here is a group of order 2. You can notice that this group has an identity element e, and the element a is its own inverse.

$\star$	e	a
e	e	a
a	a	e

For another example, here's a group $(G, \star)$ of order 3. You can notice here that the elements a, b, e of the group $(G, \star)$ appear exactly once in every row and exactly once in every column of the Cayley table. This is similar to how a Sudoku puzzle is completed.

$\star$	e	a	b
e	e	a	b
a	a	b	e
b	b	e	a

So when you fill in a multiplication table for a group $(G, \star)$, you should remember the following things:

- The identity $e \in G$ should be at the start, and that $e \star a = a = a \star e$ for all $a \in G$:
- Every element of G must appear exactly once in every column and exactly once in every row.

You can use multiplication tables to quickly find inverses of elements in a group. Here's an example of a presentation of a group as a multiplication table.

Example 4.2.9. You are given the set $G = \{1, a, b, c, d, e\}$ with binary operation $\star$ defined in Table 4.1:

Here, you can see that 1 is the identity element of the group G; as $1 \star g = g = g \star 1$ for all $g \in G$. You can also read off the inverse of any particular element easily; here, as $a \star b = 1 = b \star a$, you can say that the inverse of a in G is the element b. Finally, you can see that this group is not abelian, as $c \star d = b \neq a = d \star c$.

$\star$	1	a	b	c	d	e
1	1	a	b	c	d	e
a	a	b	1	d	e	c
b	b	1	a	e	c	d
c	c	e	d	1	b	a
d	d	c	e	a	1	b
e	e	d	c	b	a	1

Table 4.1: $(G, \star)$ in Example 4.2.9

Here are some more examples of where multiplication tables are useful. Remember from Definition 1.4.1 that a is congruent to b modulo n (where $n > 1$) if and only if $n \mid a-b$. You can also remember from Theorem 1.4.6 that you can add and multiply modulo n. You also know from Theorem 1.4.5 and Definition 2.2.9 that congruence modulo n is an equivalence relation, and from Theorem 2.2.14 that the equivalence classes are given by

$$[r] = \{kn + r \ : \ k \in \mathbb{Z}\}$$

for $r = 0, 1, \ldots n-1$.

You can write the set of equivalence classes modulo n by the notation $\mathbb{Z}/n\mathbb{Z}$ (or $\mathbb{Z}_n$). You can also define addition and multiplication on this set of equivalence classes by the rules

$$\underbrace{[a] + [b]}_{\text{adding in } \mathbb{Z}/n\mathbb{Z}} = \underbrace{[a+b]}_{\text{adding mod n}}$$

and

$$\underbrace{[a] \cdot [b]}_{\text{multiplying in } \mathbb{Z}/n\mathbb{Z}} = \underbrace{[a+b]}_{\text{multiplying mod n}}$$

For clarity, you can drop the brackets. In the case of addition, you can express this knowledge in a table. For instance, if $n = 6$, then

$$\mathbb{Z}/n\mathbb{Z} = \{0, 1, 2, 3, 4, 5\}$$

Examples of addition include $4 + 5 = 3$ in $\mathbb{Z}/n\mathbb{Z}$; the full table of addition is given by:

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

In fact, you can say something more:

Example 4.2.10. Here, $(\mathbb{Z}/n\mathbb{Z}, +)$ is a group for all natural numbers $n > 1$. You can check the axioms in turn:

associativity: As addition is associative in $\mathbb{Z}$, this property is inherited by addition in $\mathbb{Z}/n\mathbb{Z}$. So for all $a, b, c \in \mathbb{Z}/n\mathbb{Z}$, then

$$a + (b + c) = (a + b) + c$$

which means that $(\mathbb{Z}/n\mathbb{Z}, +)$ satisfies the associativity axiom.

identity: Here, $0 \in \mathbb{Z}/n\mathbb{Z}$ is an element such that for all $a \in \mathbb{Z}/n\mathbb{Z}$:

$$0 + a = a = a + 0$$

and so 0 is the identity element for $(\mathbb{Z}/n\mathbb{Z}, +)$.

inverse: Finally, you can notice that for every $a \in \mathbb{Z}/n\mathbb{Z}$, there exists $(n - a)$ in $\mathbb{Z}/n\mathbb{Z}$ such that

$$a + (n - a) = 0 = (n - a) + a$$

and here, $n - a$ is the inverse element of $a \in \mathbb{Z}/n\mathbb{Z}$.

However, this set $\mathbb{Z}/n\mathbb{Z}$ (with $n > 1$) is not a group with the operation of modular multiplication. This is because there is no element $a \in \mathbb{Z}/n\mathbb{Z}$ such that $0 \cdot a = 1$ in $\mathbb{Z}/n\mathbb{Z}$. You can try and take away the 0 (like in Example 4.2.8), but even then you are not guaranteed to get a group. For instance, $\mathbb{Z}/6\mathbb{Z} \setminus \{0\}$ is not a group as multiplication is not a binary operation on this set: here, $2 \cdot 3 = 0$. In fact, you can characterise exactly when $\mathbb{Z}/n\mathbb{Z}$ is a group.

Theorem 4.2.11. *$(\mathbb{Z}/n\mathbb{Z} \setminus \{0\}, \cdot)$ is a group if and only if n is a prime.*

Proof. For the forward direction, you can prove the contrapositive statement. Suppose that n is **not** a prime; then there exists $1 < a, b < n$ such that $ab = p$. Then you can say that $ab = 0$; so the operation of multiplication is not a binary operation on $\mathbb{Z}/n\mathbb{Z} \setminus \{0\}$. This means that $(\mathbb{Z}/n\mathbb{Z} \setminus \{0\}, \cdot)$ is not a group.

For the converse direction, suppose that n is a prime. You can use the properties of prime numbers to prove that $(\mathbb{Z}/n\mathbb{Z} \setminus \{0\}, \cdot)$ is a group. First of all, you can notice that if $a, b \in \mathbb{Z}/n\mathbb{Z} \setminus \{0\}$, then $a, b < p$ and $p \nmid a$ and $p \nmid b$. This means that $p \nmid ab$ by the contrapositive statement Lemma 1.3.3 (i) and so $\cdot$ is a binary operation on $\mathbb{Z}/n\mathbb{Z} \setminus \{0\}$. You can now check the axioms for a group as in Definition 4.2.2:

associativity: Once again, as multiplication is associative in $\mathbb{Z}$, this property is inherited by multiplication in $\mathbb{Z}/n\mathbb{Z} \setminus \{0\}$. So for all $a, b, c \in \mathbb{Z}/n\mathbb{Z} \setminus \{0\}$, then

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

which means that $(\mathbb{Z}/n\mathbb{Z} \setminus \{0\}, \cdot)$ satisfies the associativity axiom.

identity: Here, $1 \in \mathbb{Z}/n\mathbb{Z} \setminus \{0\}$ is an element such that for all $a \in \mathbb{Z}/n\mathbb{Z}$:

$$1 \cdot a = a = a \cdot 1$$

and so 1 is the identity element for $(\mathbb{Z}/n\mathbb{Z} \setminus \{0\}, \cdot)$.

inverse: Finally, you can notice that for every $a \in \mathbb{Z}/n\mathbb{Z} \setminus \{0\}$, then $\gcd(a, p) = 1$. So by Bézout's lemma Corollary 1.2.7, you can find integers u and v such that $au + vp = 1$. Therefore, you can say that $vp = 1 - au$ and so by Definition 1.4.1, this means that $au \equiv 1 \pmod{p}$. Therefore, r (where $r \equiv u \pmod{p}$ and $0 < r < p$) is an element in $\mathbb{Z}/n\mathbb{Z} \setminus \{0\}$ such that

$$a \cdot r = 1 = r \cdot a$$

and here, r is the inverse element of $a \in \mathbb{Z}/n\mathbb{Z} \setminus \{0\}$.

□

Finally in this section, here are some more examples of groups that appear in the wilderness of mathematics.

Example 4.2.12. In your previous studies, you may have looked at properties of 2×2 matrices with entries in the real numbers $\mathbb{R}$. You can then wonder; is the set of all 2×2 matrices together with multiplication a group?

The multiplication of two 2×2 matrices is again a 2×2 matrix, so this really is a binary operation. Matrix multiplication is associative, so this axiom is satisfied. You can show that the 2×2 identity matrix

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

is a matrix such that $I_2 A = A = A I_2$ for all 2×2 matrices A; so this is a suitable identity element.

However, you could recognise that not every 2×2 matrix is invertible. For instance, there is no matrix $\left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right)$ such that:

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

A result from the theory of matrices says that a matrix A is invertible if and only if the determinant of A does not equal 0. You can then look at the set:

$$\mathrm{GL}(2, \mathbb{R}) = \left\{ \begin{pmatrix} a & b \\ c & d \end{pmatrix} \ : \ ad - bc \neq 0 \right\}$$

of all matrices with non-zero determinant.

If $A \in \mathrm{GL}(2, \mathbb{R})$ then the inverse of A is given by

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

and $AA^{-1} = I_2 = A^{-1}A$. Therefore, $(\mathrm{GL}(2, \mathbb{R}), \cdot)$ is a group. This is called the **general linear group of** 2×2 **matrices over** $\mathbb{R}$.

Example 4.2.13. Finally, it was mentioned in the introduction to this section that groups help to study the symmetries of shapes. You can look at a square and ask about its symmetres.

You can do nothing to the square; this is a symmetry. You could rotate it by $\pi/2$, π or $3\pi/2$ radians; these are **rotational** symmetries of the square (see Figure 4.3 for an example).

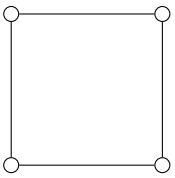


Figure 4.2: A square

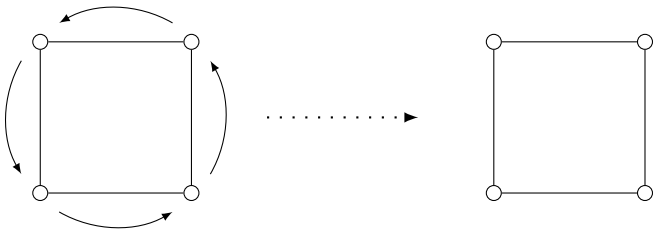


Figure 4.3: Rotating the square

You can also reflect the square along four lines of reflection. These are illustrated in Figure 4.4.

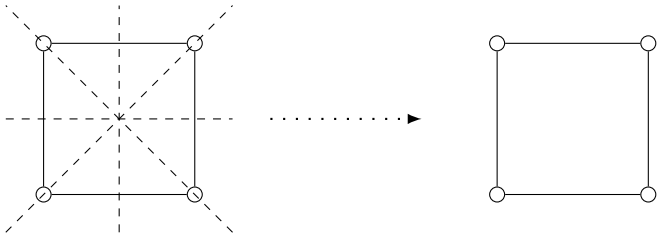


Figure 4.4: Reflecting the square. Dashed lines indicate reflection.

So there are 8 possible symmetries of the square. In general, there are $2n$ symmetries of a regular n-gon. The groups that express the symmetries of 2-dimensional shapes such as these are called **dihedral groups**.

If you label each of the vertices of the square $1, 2, 3, 4$, you can see that every symmetry of the square corresponds to a permutation of the set $\{1, 2, 3, 4\}$. In fact, every element of any group corresponds to a permutation contained in some symmetric group; this is a deep and significant result known as **Cayley's theorem**.

4.2.3 Properties of groups

This final subsection of the course is devoted to proving some further results about group theory.

Definition 4.2.14. Let $(G, \star)$ be a group, and suppose that $a \in G$. For $n \in \mathbb{N}$, you can define the nth power of a in G to be

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

Negative powers can be defined as inverses of nth powers: for instance

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

Finally, define $a^0 = e$, the identity element of $(G, \star)$.

As groups are associative, you can use the standard laws of indices in a group. So here,

$$a^m \star a^n = a^{m+n} \qquad \text{and} \qquad (a^m)^n = a^{mn}$$

Groups exist where every element of the group is a power of some element a. They play an important role in group theory, so it would be appropriate to give them a special name:

Definition 4.2.15. A group $(G, \star)$ is called **cyclic** if there exists an element $a \in G$ such that every element $g \in G$ can be written as $g = a^n$ for some $n \in \mathbb{Z}$. Such an element a is called a **generator** of G.

Cyclic groups are among the most common groups you can find. For instance, for every natural number $n \in \mathbb{N}$, there exists a cyclic group $(G, \star)$ with order n (where the order is defined in Definition 4.2.5). Another property of cyclic groups is the following. You can remember from Definition 4.2.4 that a group $(G, \star)$ is **abelian** if $a \star b = b \star a$ for all $a, b \in G$.

Lemma 4.2.16. *If $(G, \star)$ is a cyclic group, then $(G, \star)$ is abelian.*

Proof. Let x and y be elements of the group $(G, \star)$, and suppose that a is a generator for this group. As this happens, you can write $x = a^m$ and $y = a^n$ for some integers m, n. Since addition in $\mathbb{Z}$ is commutative, and the laws of indices hold in groups, you can write

$$x \star y = a^m \star a^n = a^{m+n} = a^n \star a^m = y \star x$$

and so $(G, \star)$ is abelian. □

Our study of group theory grew out of the generalisation of the properties of permutations investigated in Section 3.1. There is one of these properties which has not yet been gener-

alised to groups; that of the **order** of a permutation, given in Definition 4.1.18. So here it is:

Definition 4.2.17. Let $(G, \star)$ be a group and let $a \in G$. The **order** of a in G is the least positive integer m (if it exists) such that $a^m = e$, where e is the identity element of G.

If there is no such integer m, then you can say that a has **infinite order**.

You can link the definition of order of the element in a cyclic group $(G, \star)$ to the order of the group itself.

Lemma 4.2.18. *If $(G, \star)$ is a cyclic group with generator a, then the order of a in G is equal to $|G|$.* □

The study of structures such as groups does not just revolve around elements of that structure. It can also involve subsets of a group that carries the structure of a group itself. This is the basis for the next definition.

Definition 4.2.19. Let $(G, \star)$ be a group and suppose that H is a non-empty subset of G. Then H is a **subgroup** of G if and only if $(H, \star)$ is a group. If this happens, you can write that $H \leq G$ to mean that H is a subgroup of G.

Proving that a particular subset of a group is a subgroup is an important tool in group theory. Here's a result that makes this easier to do.

Theorem 4.2.20 (Subgroup test)**.** *Let $(G, \star)$ be a group and suppose that H is a subset of G such that for all $x, y \in H$ then $x \star y \in H$ and $x^{-1} \in H$. Then H is a subgroup of G.*

Proof. You can show this by checking that $\star$ is a binary operation on H and that the three axioms of Definition 4.2.2 hold.

As $(G, \star)$ is a group, it follows that $\star$ is a binary operation on G. Since $H \subset G$, you can say that $\star$ is defined on every element of $H \times H$. As $x \star y \in H$ for all $x, y \in H$, it follows that $\star$ is a binary operation on H.

associativity: Let $a, b, c \in H$. Then $a, b, c \in G$ and

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

is true. So $(H, \star)$ is associative.

identity: Let e be the identity element in G. If $x \in H$, then the statement of the theorem says that x^{-1} is also in H. As $x \star y \in H$ for all $x, y \in H$, it follows that $x \star x^{-1} = e \in H$. Therefore, the identity element of G is in H. As $x \star e = x = e \star x$ for all $x \in G$, it follows that $x \star e = x = e \star x$ for all $x \in H$. So H has an identity element.

inverse: It's given in the statement of the theorem that for all $x \in H$ there exists an element $x^{-1} \in H$ such that

$$x \star x^{-1} = e = x^{-1} \star x$$

This means that H satisfies the inverse axiom.

Therefore, $(H, \star)$ is a group and so is a subgroup of G by Definition 4.2.19. □

Example 4.2.21. (1) You can show that $(\mathbb{Z}, +) \leq (\mathbb{Q}, +) \leq (\mathbb{R}, +) \leq (\mathbb{C}, +)$.

(2) Let $(G, \star) = (\mathbb{R} \setminus \{0\}, \cdot)$ and $H = \mathbb{Q} \setminus \{0\}$ be a subset of $\mathbb{R}$. You can show that $(H, \star) = (\mathbb{Q} \setminus \{0\}, \cdot)$ is a subgroup of $(\mathbb{R} \setminus \{0\}, \cdot)$ by using Theorem 4.2.20. Here, you need to show that $x \cdot y \in \mathbb{Q} \setminus \{0\}$ for all $x, y \mathbb{Q} \setminus \{0\}$ and that for all $x \in \mathbb{Q} \setminus \{0\}$ then $x^{-1} \in \mathbb{Q} \setminus \{0\}$.

First of all, you can notice that if x, y are non-zero rational numbers, then their product $x \cdot y$ is a non-zero rational number. This means that for all $x, y \in \mathbb{Q} \setminus \{0\}$, then $x \cdot y \in \mathbb{Q} \setminus \{0\}$.

Now suppose that $x \in \mathbb{Q} \setminus \{0\}$. This means that you can write $x = p/q$ which is a non-zero rational number. Then $x^{-1} = q/p$ is a non-zero rational number. This is the inverse of x in G, and $x^{-1} \in H$. So you can say that H is a subgroup of G by Theorem 4.2.20.

There is one final theorem about groups and subgroups that simply has to be mentioned.

Theorem 4.2.22 (Lagrange's theorem)**.** *Let $(G, \star)$ be a group with $|G| = n$, where $n \in \mathbb{N}$. If H is a subgroup of G, then $|H|$ divides $|G|$.*

In fact, Lagrange's theorem says a lot more than this; but sadly this is too advanced for the course and I think 125 pages is long enough.

∼Fin∼