Isaac Essential Physics

Explaining Physics

Isaac Physics Team

DRAFT

Work in Progress

Summer 2025

PERIPHYSEOS
PRESS

Contents

Patterns in numbers

What, exactly, is physics? Physics is a part of science, and as such is rooted in exploring Creation (or Nature, if you prefer) through experimentation and observation. However there is much more to it than that. But if biologists study life, and if chemists study materials and reactions, what are physicists left with? Motion, electricity, magnetism, gravity, waves and vibrations, heat, light, the atom, the universe, the strength of materials to be sure – but what unites these disparate topics?

One answer is the way in which these topics are studied – using mathematical methods to spot patterns, and observing the way in which the same numerical recipe may well apply to very different situations – this indicating a fundamental similarity between them. On the face of it there seems very little similarity between gravity and static electricity – however we can use virtually the same mathematical methods to comprehend both, and this belies a fundamental unity to the nature of gravitational and electrostatic forces.

Put more prosaically, physicists measure things and then look for patterns in the numbers. And often they find patterns which are extremely reliable – so reliable that the physicist (or more usually, engineer) can bet their life on the result of many physics calculations with virtually no unease.

This chapter is the keystone for everything which follows. Here we will introduce the patterns we shall be finding in the different topics in the later chapters. While they are, without doubt, mathematical; when students are within earshot it may be better to describe our thinking as part of numeracy — namely the application of maths ideas in common sense ways to what we see around us every day.

1 Introducing the patterns

We will introduce the two most important patterns which we frequently find in measurements, and then something a bit more tricky.

In each case, each column of numbers makes a pair. Can you find what all the pairs have in common? Put differently, can you see how to get the bottom number from the top one?

Each table has a couple of numbers missing - see if you can work out what they are once you think you know the pattern.

Pattern 1: Bigger bigger

What do you do to these...	2	5	10	60	120		4.2
...to get these?	6	15	30	180	360	3000	

Here the missing numbers are 1000 and 12.6.

When the top number gets bigger, so does the bottom number, but in a special way: to get the number in the bottom row (let us call this y), you take the number in the top row (x) and multiply it by 3.

Mathematically we write this

$$y = 3x.$$

This is called a proportional relationship. Notice that

- When the top number doubles, so does the bottom number.
- When you divide the bottom number by the top number, you always get the same answer (here 3). Mathematically $y \div x$ is the same for each pair of numbers.

We will find out more about this sort of pattern on page 9

Pattern 2: Bigger smaller

What do you do to these...	3	4	5	6	8	10	
...to get these?	40	30	24	20	15		10

Here the missing numbers are both 12.

When the top number (x) gets bigger, the bottom number (y) gets smaller, but there is a more specific pattern at work: if you look in any column, you get 120 if you multiply the two numbers together.

Mathematically we write this

$$xy = 120 \quad \text{or} \quad y = 120 \div x.$$

This is called an inversely proportional relationship. Notice that

- When the top number doubles, so the bottom number halves.
- When you multiply the bottom number by the top number, you always get the same answer (here 120).

We will find out more about this sort of pattern on page 11

Pattern 3: A challenge

What do you do to these...	1	2	3	4	5	6	
...to get these?	5	20	45	80	125		500

Here the missing numbers are 180 and 10.

This one is more tricky. If you want a clue, look at how the numbers in the bottom row compare with the squares of the numbers in the top row. We will study this pattern a bit more on page 13.

2 ÷ / each and per

In order to gain an instinctive insight into many of the patterns we spot in our experiments, it is essential that we and our students are completely comfortable with the mathematical and practical meaning of division.

If we have three pies and share them fairly between four people, each person will get

- three quarters $\left(\frac{3}{4}\right)$, or
- 'three shared between four', or
- 'three divided by four' $(3 \div 4$ or $3/4)$, or
- 0.75 pies.

So, for any two numbers represented by a and b,

$$\frac{a}{b} = a/b = a \div b.$$

By extension, if we are sharing 6 kg of rice between 30 people, each will get $6 \text{ kg} \div 30 = \frac{6 \text{ kg}}{30} = \frac{6}{30}\text{kg} = 0.2 \text{ kg}$.

When we read this aloud, we would usually say they get '0.2 kg for each person' or '0.2 kg per person'. The words **each** and **per** often occur when there is division or sharing going on.

Let's give another example. Suppose you travel 150 miles ≈ 240 km in a coach at a steady speed over three hours. How far do you go each hour? We work it out by sharing the distance into three parts — one for each hour. The answer is $150 \div 3 = 50$ miles or $240 \div 3 = 80$ km each hour.

With miles we usually call it 50 miles per hour and abbreviate it 50 mph with the 'p' standing for 'per' meaning 'each' or 'every'.

With kilometres, it is more common to write '80 km each hour' as 80 km/h with the / symbol meaning 'each'. This is helpful because it reminds us that **per**, **each** and **divided by** are all the same thing.

The unit even follows the usual algebraic pattern:

$$\frac{240 \text{ km}}{3 \text{ h}} = \frac{240}{3}\,\frac{\text{km}}{\text{h}} = 80 \text{ km/h},$$

with km/h meaning 'kilometres divided by hours'.

When reading a unit with a '/' in it aloud, you may find it is easier for your students to understand if you say the words **for each** or **every** or **per**. Once students get comfortable and confident with what / means, you will find that they are much happier with many of the ideas you will show them in numeric form, and that the units and equations will make more sense.

3 Multiplication, division and formulae

On the day I wrote this, the price of baking potatoes at a local supermarket was 85 p/kg. This means that if I want to buy a kilo of potatoes, I will need to spend 85 p.

Suppose I am making a meal for lots of friends and I need 2 kg of potatoes. How much will this cost? It would cost 85 p/kg $\times$ 2 kg $=$ 170 p $=$ £1.70.

Notice that the units help us see what is going on. Algebraically,

$$\text{p/kg} \times \text{kg} = \frac{\text{p}}{\text{kg}} \times \text{kg} = \frac{\text{p}}{\cancel{\text{kg}}} \times \cancel{\text{kg}} = \text{p},$$

and so it is clear that when you multiply the cost per kilo with the number of kilos you get a cost (in pence).

Laying the calculation out a different way gives this:

cost (p)	=	cost per kilo (p/kg)	×	mass (kg)
	=	85	×	2

where we have used the correct scientific term, mass, for the thing measured in kilograms (we will explain this properly in page reference needed).

This framework enables us to answer other questions clearly. For example, let's try to answer, "How many potatoes can you buy for £4.25 $=$ 425 p?"

cost (p)	=	cost per kilo (p/kg)	×	mass (kg)
425	=	85	×	

We would work the answer out using $425 \div 85 = 5$ kg as this tells us how many 85 ps there are in £4.25. It tells us what you have to multiply 85 p by to get £4.25.

From a common sense understanding of multiplication and division, we know

- total cost = cost per kilo × mass
- mass = total cost ÷ cost per kilo
- cost per kilo = total cost ÷ mass.

While that might look like three equations, it is really just one fact. So we just remember one statement, and then use common sense (or algebraic re-arrangement if you happen to find that easier).

As we explore our topics, we will find rules like this occur frequently in physics. Some students will want to write them out as equations and re-arrange them to change the subject (that is, alter which thing is on the left of the = sign). However, we think it is clearer if you just write it out once, understand what it means, and then use common sense with numbers.

4 Units as clues

Let's try another question. Suppose 0.6 kg of apples cost £1.20 $= 120$ p. How much is each kilo?

cost (p)	=	cost per kilo (p/kg)	×	mass (kg)
120	=		×	0.6

The answer is 200 p/kg as this is what you get when you divide $£1.20 = 120$ p by 0.6 kg. Alternatively it is the amount of money which becomes $£1.20$ when you multiply it by 0.6.

Notice that the units give a very good clue to what is going on:

$$120 \text{ p} \div 0.6 \text{ kg} = \frac{120 \text{ p}}{0.6 \text{ kg}} = 200 \frac{\text{p}}{\text{kg}} = 200 \text{ p/kg}.$$

In other words, if you want to get a cost per kilo, which will be in p/kg, you need to divide something measured in p by something measured in kg.

Here are some other examples which we will meet later on where the units help explain the meaning of a measurement

- To get a speed in kilometres per hour (km/h) you divide a distance in km by a time in h.
- To get an area in square metres ($\text{m}^2 = \text{m} \times \text{m}$), you multiply two distances in metres together.
- Densities are measured in kilograms per cubic metre (kg/m^3). This means that
  - to calculate a density you divide a mass in kg by a volume in m^3
  - densities give a measurement of the mass of a particular volume of material.
- Pressures are measured in newtons per square metre (N/m^2), where the newton N is the unit of force. This means
  - we calculate pressures by dividing a force in N by an area in m^2
  - pressure quantifies how well focused a force is on a surface.
- A more complicated unit is the metre per second squared (m/s^2). What does this measure given that 'a square second' sounds very strange? Notice that mathematically

$$\text{m/s}^2 = \frac{\text{m}}{\text{s}^2} = \frac{\text{m/s}}{\text{s}}$$

and therefore this measurement is probably some kind of speed divided by a time. This is the unit of acceleration telling you how much extra speed something gains (or loses) each second.

The moral of this story is that if you know the equation, you can work out the unit. Conversely, if you know (or are given) the unit, you often can work out the equation and don't need to memorize it.

5 Rates

Many of the most interesting things in physics change, and we want to study how they change. One of the most important things we often want to know is, 'How quickly does it change?'

This question is answered as a rate of change.

If the thing does not change, we say it is constant, and the rate of change is zero.

Given that we measure time in seconds, a rate of change (often called the rate for short) measures how much change you get each second. An example of a rate in the workplace is a pay rate: the amount you are paid for each hour of work.

Let's look at some examples of rates

- Liquid volume is measured in litres (ℓ). If it takes you two minutes (or 120 seconds) to put 30ℓ of fuel in a tank, then the rate of filling is

$$\frac{30\ell}{120\text{ s}} = 0.25\ \ell\text{/s}.$$

- Distance is measured in metres (m). If a cyclist travels 600 m in two minutes, then the rate of moving is

$$\frac{600\text{ m}}{120\text{ s}} = 5\text{ m/s}.$$

The rate of moving is also called speed so we see that the unit of speed is the metre per second (m/s) meaning that it tells you how far you go each second.

You can usually spot a rate because its unit has /s in it, showing that it is telling you how much something changes each second.

Some measurements are labelled specific. This has a different meaning: how much of the quantity each kilogram (kg) of the material carries. A physicist might call the cost per kilo of some bananas their specific cost. You can spot the unit of a specific quantity as it will have /kg in it.

Constant or uniform?

If two things are constant, it means neither of them changes as time passes. However this does not mean that they are the same as each other. Two cyclists could each be travelling at constant speed, but one could be going at a steady 12 mph while the other keeps going at 20 mph.

In physics, the word uniform describes the situation where a measurement is the same at all places, no matter where you look. When you move a paper clip around in a uniform magnetic field, you will feel the same force of attraction on the paper clip no matter where you put it within that field.

Just because something is uniform, however, does not mean that it is constant. Suppose the magnet slowly gets weaker everywhere at the same rate. The magnetic field is still uniform in this case (at any moment in time the field is equally strong in all places), but it is not constant.

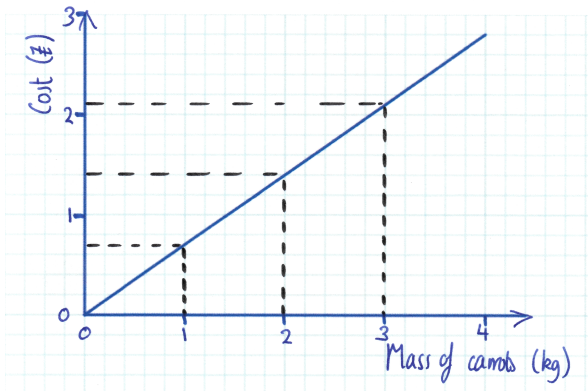
6 Proportionality

Suppose we are only interested in selling carrots at 70 p/kg. The only formula we would need in our trade would be:

cost (p)	=	cost per kilo (p/kg)	×	mass (kg)
	=	70	×	

From this we could work out the cost of a given number of carrots or the number of carrots we could sell for a certain amount of money.

We could plot a graph of the mass of carrots and the total cost. It would look like this:



Notice that

- it doesn't cost anything to buy zero carrots
- each extra kilo of carrots costs an extra 70 p
- if you double the mass of carrots, you double the cost
- if you have only half the money, you can buy half the mass of carrots

In a formula like this, where one of the numbers in the multiplication is fixed, we say that the other two quantities are proportional. This is probably the most powerful and straightforward relationship we can find.

Notice that the graph

- has a straight line,

- the line goes through the (0,0) corner (we call this the origin)
- the line goes up the page by 70 p every time you move to the right by 1 kg

This is the kind of graph you get when two things are proportional. The steepness of the line (called the gradient) always has meaning - here the cost of one kilo of carrots.

Other examples of things which are proportional which we will look at later include

- the distance you travel is proportional to the time of the journey if you go at a fixed speed
- at the same place on Earth, the weight of any object is proportional to the amount of stuff it contains (its mass measured in kilograms).
- the extra length (extension) gained by a spring if you stretch is proportional to the force (providing you don't stretch it too much).

7 Inverse proportionality

Suppose you have 800 people who need to go on a journey. How many buses will you need to book? It depends on the size of the buses. Suppose you can book coaches which will take 50 people each, then you will need $800 \div 50 = 16$ coaches.

However if you could only get hold of 16-seat minibuses, you would need $800 \div 16 = 50$ minibuses.

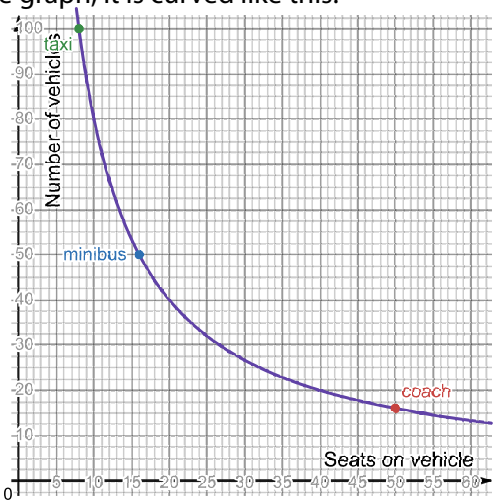
If you had to rely on 8-seat taxis, you would need 100 of them.

Notice that the number of vehicles needed doubled when the size of the vehicle halved. This type of relationship is called inverse proportionality. It is summed up in equations like this one:

people	=	seats on vehicle	×	vehicles
800	=		×	

where the product in the multiplication is fixed, but we can vary the numbers which multiply to get it.

When we plot the graph, it is curved like this:



Lots of graphs are curved, so it is difficult to tell when a curve is of this form. Just because one quantity goes up when another goes down does not mean that the two things are inversely proportional.

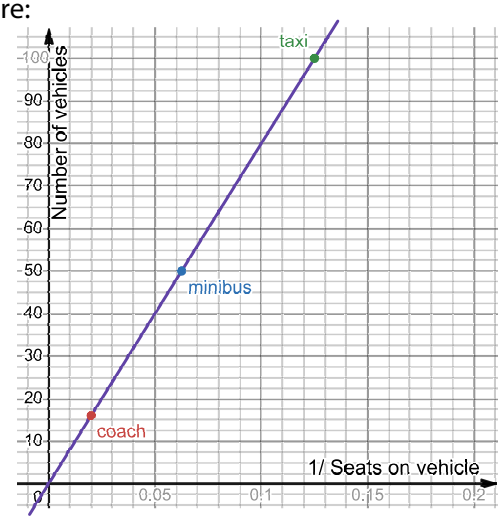
We have two ways of checking if data containing pairs of measurements, for example (x, y), fits an inverse proportional relationship:

- multiply each pair of numbers and compare the products xy for the different pairs. If the products are all the same, then there is inverse proportion.
- plot a graph of y against $1 \div x$. If the line is straight and goes through the origin, then there was an inverse proportional relationship.

To see how these relate, notice that

$$\text{vehicles} = \frac{\text{people}}{\text{seats on vehicle}} = \frac{\text{people} \times 1}{\text{seats on vehicle}} = \text{people} \times \frac{1}{\text{seats on vehicle}}.$$

So if the number of seats on each vehicle halves, $(1 \div \text{seats on vehicle})$ will double, and accordingly the number of vehicles will double too. A graph of this is shown here:



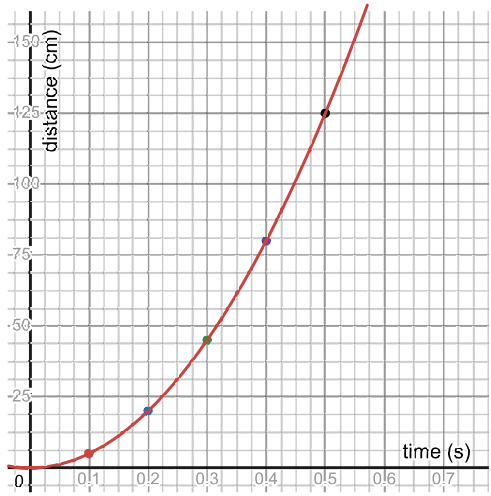
8 More complicated patterns

Once these basic patterns are understood, variations on the theme can be detected without too much extra difficulty. For example, the table below gives the distance fallen by a dropped object in certain amounts of time. This is very similar to the data given in our third pattern-spotting challenge.

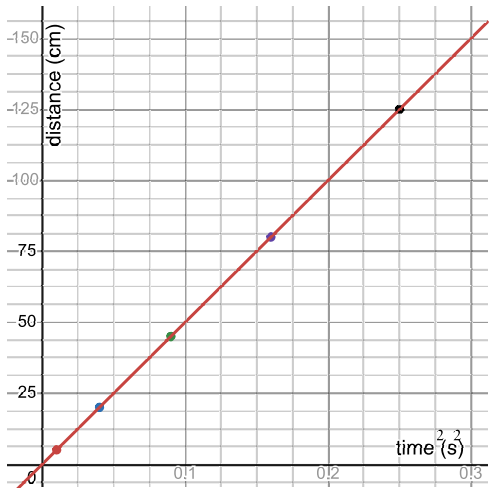
Time $(\mathbf{s})$	0.1	0.2	0.3	0.4	0.5	0.6	
Time^2 $(\mathbf{s}^2)$	0.01	0.04	0.09	0.16	0.25	0.36	
Distance $(\mathbf{cm})$	5	20	45	80	125		500

The answers are 180 and 1.0. The numerical pattern is that if you divide the distance by the square of the time (that is, the time multiplied by itself), you always get 500. Mathematically, this is written $y = 500x^2$. A physicist would say that the distance fallen is proportional to the square of the time. If you double the time, the square of the time quadruples, since $2^2 = 2 \times 2 = 4$, and so the distance also quadruples.

When we plot this on a graph as it is, the graph is curved.



The pattern is clearer if we do not plot the times on the horizontal axis, but plot their squares instead.



This particular pattern was spotted during an ingenious experiment by Galileo, and it led to the understanding of motion that we still use today. Admittedly, this is not all there is to falling – if the object falls for longer, the air begins to impede its passage, and the pattern becomes more complicated.

An identical pattern is found in the strength of wires and cables. The weight which can be safely supported by a cable is proportional to the square of the cable's diameter. This is because if you make a cable twice as wide and

twice as deep, it has four times the cross sectional area, and so can carry four times the load. Engineers call the load carried per square centimetre of cross sectional area the stress of a cable, and use this when deciding which cable to use for a particular job.

As you see in the last paragraph, once you know the pattern, you immediately begin to think 'why should that be the case?' Perhaps you are already beginning to wonder why the distance fallen should quadruple if the time doubles! If you are, you should have no difficulty understanding why physicists get very excited once they spot a pattern. It is as if a curtain hiding some treasure has suddenly slipped, and gives a tantalizing glimpse of what is behind. More will be revealed when we study motion, and you will find the reasoning on page 23.

9 More than one dimension

We live in a three-dimensional world, and frequently it is not enough to measure how long something is, we also wish to specify its orientation. If it falls to you to write the programming for one of the many satellite navigation handsets used in cars, there is no way you can assign a single number to represent the destination of the driver's journey. The driver will not thank you if the device makes them drive exactly the correct distance, but in the wrong direction. Equally, when designing the body of an airliner, it is important not only to know the size of the forces which will be present on the aircraft in flight, but also their direction. Otherwise strengthening beams, wires or braces will be incorrectly fitted.

Legend has it that the solution came to the philosopher Descartes when he observed a fly from his sick bed one day. He realized that he could completely specify the fly's position at a particular instant by saying how far it was from the floor, how far it was from the wall on his left, and how far it was from the wall in front of him. In short by giving three appropriate length measurements, the exact location of the fly could be pinned down. In Descartes' honour this method is called the Cartesian system.

Much digital mapping of the U.K. (such as the maps obtained over the World

Wide Web) uses a similar principle. The Ordnance Survey, which performed the painstaking surveys needed to make the maps, specify a datum point in the English Channel near the Isles of Scilly. Every point in the United Kingdom is then referred to by its distance East from the datum, followed by its distance North. Of course, you would not choose a route to Bristol, say, to start from this datum point. Neither would all your journeys begin with a straight Eastwards leg followed by a sharp turn to the left and a Northwards conclusion. However the two numbers do give the mapping program, and the user, a comprehensive method of specifying the locations of the destination of your journey, and of your current position.

Quantities which incorporate direction are called vector quantities. Position is clearly a vector quantity, and the position of one object relative to another is often called displacement. It is also useful to have a measurement of speed which incorporates directionality. This vector-version of speed is called velocity. Like position, displacement and velocity can not be expressed as single numbers. One number for each dimension in the measurement is needed.

To give an example of a vector quantity in use, think of climbing a '1 in 5' hill. Such an incline is called '1 in 5' because for every 5 metres you move forward horizontally, you climb 1 metre. Road signs sometimes express this gradient as 20% - for every metre you move forwards horizontally, you move 20% of one metre (that is 0.2 m) upwards. Climbing up a '1 in 1' or 100% gradient requires you to ascend a slope inclined at 45° to the horizontal.

Your motion as you move a metre forwards on a 20% gradient can be written

$$\begin{pmatrix} 1.0 \\ 0.2 \end{pmatrix} \text{ m.}$$

The number on the top tells you how far you have moved forwards, horizontally. The number on the bottom tells you how far you have risen vertically. Similarly, if you moved this distance every second, then it would be fair to say that your velocity is

$$\begin{pmatrix} 1.0 \\ 0.2 \end{pmatrix} \text{ m/s.}$$

If you carry on climbing for two minutes (that is, 120 seconds), then your

total displacement will be

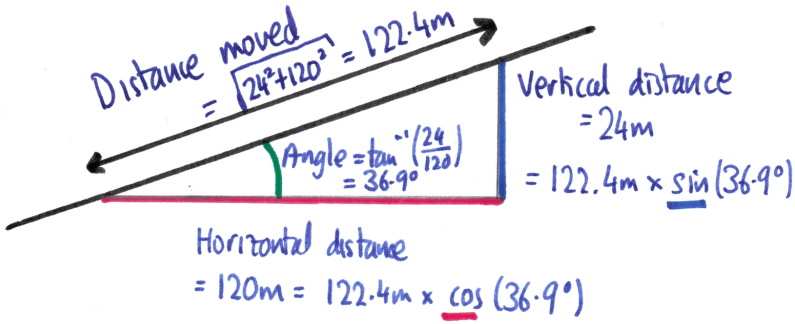
$$\begin{pmatrix} 1.0 \\ 0.2 \end{pmatrix} \text{m/s} \times 120\,\text{s} = \begin{pmatrix} 120 \\ 24 \end{pmatrix} \text{m}.$$

We say that the horizontal component of the journey is 120 m, and the vertical component is 24 m. It may seem more useful to state the information as a total distance of 122.4 m along a slope inclined at 11.3° to the horizontal. Equally, it may seem strange to you that a displacement on the road should be quoted as 3 km East then 4 km North rather than 5 km on a bearing of 037°. Surely the latter is clearer? So why do physicists like the Cartesian system so much? Suppose you climb this hill then climb for 500 m up another one ('3 in 4' this time). What is your total displacement?

Climbing 500 m up a '3 in 4' or 75% gradient is equivalent to moving 400 m forwards and 300 m upwards. Adding this to our first journey gives us an overall total displacement of 120 m $+ 400$ m $= 520$ m forwards and 24 m $+ 300$ m $= 324$ m upwards. However trying to work out the overall displacement as a distance and angle is more tricky. This is why physicists do like Descartes' system of measuring using components. If you want to work out the overall direction and angle, or alternatively you have a direction and angle and wish to work out the components, then you need to use some Trigonometry. If you wish to brush up on this area of mathematics, please look at the end of this chapter.

10 Appendix: A reminder of trigonometry

Let's go back to our journey up the hill. We travelled $\begin{pmatrix} 120 \\ 24 \end{pmatrix}$ m, and this is shown in the diagram below.



How far have you moved in total? This question can be answered using Pythagoras' Theorem, which says that the square of the length of the longest side of a right-angled triangle is equal to the sum of the squares of the lengths of the other two sides. This gives our total distance as $\sqrt{120^2 + 24^2} \approx 122.4$ m.

If you wanted to know the angle of the slope to the horizontal, then Trigonometry comes to your aid. The ratio of the vertical distance to the horizontal distance does not depend on the size of our triangle – only on the angle of the gradient. Mathematicians say that the ratio is a function of the angle, and they call the function the tangent (or $\tan$ for short). To find the angle, we need to find the angle whose tangent is $24/120 = 0.2$, and a calculator gives this as $11.3°$.

But how did we work out the components of the 500 m climb? The answer again uses trigonometry. In this case, the hill makes an angle of $36.9°$ with the horizontal.

The ratio of the vertical distance to the total distance walked (along the diagonal) is also a function of the gradient angle – and this function is called the sine of the angle (or $\sin$ for short). So our vertical distance equals the total distance of 500 m multiplied by the sine of the gradient angle (here $36.9°$), which a calculator gives as 300 m.

The ratio of the horizontal distance to the total distance walked is another function of the angle – called the cosine (or $\cos$ for short). So the horizontal distance equals the total distance of 500 m multiplied by the cosine of the gradient angle (again $36.9°$), which a calculator gives as 400 m.

Describing motion

The largest single part of most physics courses at school is commonly called mechanics. This has nothing to do with fixing cars, but rather groups together all of the physics involved with forces and motion.

As this is rather a large topic, it has historically been broken down into two parts. The first, historically called kinematics, deals with how we describe and measure motion, and is what this chapter is all about.

Merely describing something has little power, however gives us the concepts and framework within which a deeper understanding can be sought. The discipline of dynamics then seeks to explain how motion works, usually in terms of forces.

One of the great challenges which has faced humanity as we have tried to understand mechanics is the realisation that although we almost expect the force on an object to be intrinsically linked to its speed, the force is actually more closely related to acceleration. To appreciate this insight, we first need to make sure that we have a thorough understanding of the difference between the different kinds of measurements we can make on something as it moves.

The approach we take here is first to look at motion along a route where we don't worry about the direction at all. We introduce the ideas of distance, then speed and then acceleration and explore the links between them. Having done this, we then look at how we take into account directionality.

11 Distance, Speed and Acceleration

Fundamental terminology

Let's get this over with as quickly as possible so that we don't lose sight of the wood for the trees.

- Distance measures how far you have moved so far. It is measured in

metres (m).

- Speed measures how quickly you are moving. It tells you how far you go each second, and is calculated as

$$\text{speed} = \frac{\text{distance moved}}{\text{time taken}} \quad \text{in symbols:} \quad v = \frac{s}{t}.$$

  It can be called the rate of change of distance travelled, and is measured in metres per second (m/s). We met this idea on page 8.

- Acceleration measures the rate at which your speed is changing.

$$\text{acceleration} = \frac{\text{change in speed}}{\text{time taken}} \quad \text{in symbols:} \quad a = \frac{v - u}{t}$$

  in the case where your speed has changed from u to v during the time interval t. As acceleration is a rate of change of speed it is measured in (speed units)/s $=$ (m/s) / s $=$ m/s^2. When speaking, we call this unit 'metres per second squared'. If you are moving at a steady speed, then the acceleration is zero.

In everyday speech we usually use the word accelerating to describe something speeding up and use decelerating to describe something slowing down. Be aware that in physics, we often use acceleration as an umbrella term for any kind of change of speed whether up or down. Notice that in the equation for acceleration above, if it is slowing down then v will be less than u, and so $v - u$ and a will both be negative.

When things change

While the ideas shown above cover nearly all of what you need to know in kinematics, there is a wrinkle in the system. Suppose we want to use our first equation $v = s/t$. We might, for example, want to rewrite it as $s = vt$ so we can work out the distance moved by multiplying the speed by the time. For this to work, the speed must be constant. This greatly restricts the situations we can study: we can't look at anything speeding up or slowing down (anything accelerating or decelerating). How boring and useless.

If, on the other hand, the acceleration is constant but not zero, then we can use the second equation $a = (v - u)/t$ well enough, but how do we work out the distance travelled if the speed is changing?

Instantaneous speed

We have two solutions. The first is to measure the distance travelled in a very short time: so short that the speed can't change much in that time. By dividing the little distance by the short period of time, we get the speed at that time. This is called the instantaneous speed. This is the kind of speed registered on the speedometer of a car, train, bus or motorcycle.

In science, small changes get labelled with the Greek letter delta (δ), so a small time interval is δt, and the small distance covered during this time is δs. We can then calculate the instantaneous speed $v = \delta s \div \delta t$ or $v = \delta s / \delta t$.

Average speed

The idea of instantaneous speed allows us to measure the speed of something which is accelerating or decelerating. However, it doesn't help us work out how far something moves while its speed changes.

To do this, we introduce the idea of average speed. Suppose you cycle 30 km in two hours. Your speed almost certainly changed during that time: faster downhill and slower uphill. Nonetheless, you completed the journey in the same time as if you had been going at a steady 15 km/h the whole way. We call this the average speed, where we define

$$\text{average speed} = \frac{\text{total distance}}{\text{total time}}.$$

Suppose you travel at a steady speed of 3 m/s for 5 s and then travel at 6 m/s for the next 5 s. In the first part of the motion you travel 3 m/s $\times$ 5 s $= 15$ m. In the second part you travel 6 m/s $\times$ 5 s $= 30$ m. Overall, you have travelled $15 + 30 = 45$ m in 10 s, so the average speed is 45 m $\div$ 10 s $= 4.5$ m/s. Notice that 4.5 m/s is the average of the two speeds 3 m/s and 6 m/s as $(3 + 6)\ /2 = 4.5$

This situation, where the average speed is the average of the speeds, is only true when the two time periods are the same. So, please do not apply this rule generally.

Suppose we travel 60 m, going at 3 m/s for the first 30 m and at 6 m/s for the remaining half of the distance. The first part of the journey will take 30 m $\div$ 3 m/s $=$ 10 s, while the second will take 30 m $\div$ 6 m/s $=$ 5 s. Overall the journey has taken $10 + 5 = 15$ s, so the average speed is 60 m $\div$ 15 s $=$ 4 m/s, which is not the average of the speeds!

12 Distance and acceleration

Now that we have introduced average speed, it enables us to link the acceleration with the distance travelled.

Distance and acceleration (from rest)

We will start with the situation where something is stationary, and then gets faster at a steady rate until it reaches a top speed v. In other words it gains the same extra speed each second. This means that it has a steady, or constant, acceleration.

During this time, the speed has risen from 0 to v, so the average will be half way between, namely the average of 0 and v, which is $(0 + v)/2 = \frac{1}{2}v$. This means that we can work out the distance travelled:

$$\text{distance travelled} = \text{average speed} \times \text{time} = \frac{v}{2} \times t = \frac{vt}{2}.$$

This is not the only formula we can write for the distance travelled. We can work out the top speed v, by multiplying the speed gained each second by the time itself. In other words,

$$\text{speed gained} = \text{acceleration} \times \text{time} \quad \text{in symbols:} \quad v = at.$$

During this time, the speed has risen from 0 to $v = at$, so the average will be half way between, namely the average of 0 and at, which is $(0 + at)/2 =$

$\frac{1}{2}at$. This means that we can work out the distance travelled:

$$\text{distance travelled} = \text{average speed} \times \text{time} = \frac{at}{2} \times t = \frac{at^2}{2}.$$

Now we can see why the distance travelled by a dropped object depends on the square of the time it has been falling, and fits with the pattern we saw on page 13.

Finally, remembering that $v = at$, it follows that $t = v/a$. Put in words: the time taken is the speed you gain in total divided by how much speed you gain each second. Using this, we may write

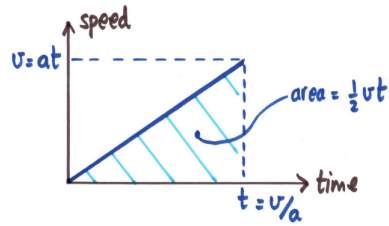
$$\text{distance travelled} = \text{average speed} \times \text{time} = \frac{v}{2} \times \frac{v}{a} = \frac{v^2}{2a}.$$

We therefore have four formulae we can use for describing motion with a steady acceleration from rest:

$$v = at \qquad s = \frac{at^2}{2} \qquad s = \frac{vt}{2} \qquad s = \frac{v^2}{2a}.$$

While you may be happy with the equations, often it is easier to see what is going on by looking at a speed-time graph. On this graph, the speed is plotted on the vertical (y) axis for all relevant times.

The graph below shows constant acceleration from rest (0 m/s) to speed v over a time t.



The steepness of the graph is important: it tells us how quickly the object is speeding up. We measure steepness of a line by dividing the height it rises by the distance it goes horizontally. This is called the gradient of the line. Here the graph has height $v = at$ and width t, so the gradient is given by $v/t = at/t = a$ and tells us the acceleration.

The area of the graph under the line is also important. The region under the line has the shape of a triangle, so we can find its area by multiplying the base width by the height and then dividing by 2. This gives $t \times v/2$ which we already know to be the distance travelled. It turns out that for any motion (even if the acceleration is not steady) we can always work out the distance travelled from the area under the line on its speed-time graph.

Acceleration from one speed to another

Sometimes we want to study motion where something is not stationary to start with. Suppose it starts at speed u, and the speed changes to v during a time interval t.

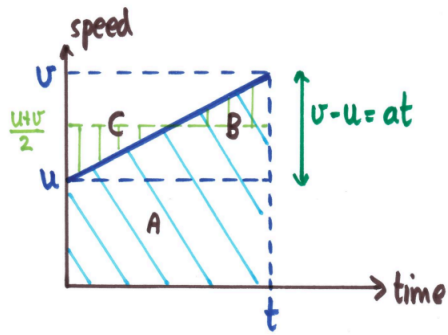
The acceleration formula in this situation is now written

$$\text{speed gained} = \text{acceleration} \times \text{time} \quad \text{in symbols:} \quad v - u = at.$$

It follows that

$$\text{new speed} = \text{old speed} + \text{speed change} \quad \text{in symbols:} \quad v = u + at.$$

The speed-time graph for this motion is shown below.



Notice that, on our graph, the rectangle A has area height $\times$ width $= ut$.

Triangles B and C have the same height and same base width, so they both have the same area. Together, they make a rectangle of area height $\times$ width $= at \times t = at^2$. Therefore B and C each have an area of half this, namely $\frac{1}{2}at^2$.

We have three different ways of working out the distance travelled depending on how we work out the area under the line.

- We could add the area of the rectangle A and the triangle B. This gives the distance travelled as

$$s = ut + \tfrac{1}{2}at^2.$$

- We could subtract the area of the triangle C from the area of the big rectangle containing A, B and C. The big rectangle has height v and width t, so its area is vt. The area under the line is therefore

$$s = vt - \tfrac{1}{2}at^2.$$

- We could work out the area under the line in one go by noticing that its area will be the same as that of a rectangle with width t and a height half way between u and v. This means that the height would be $(u + v)/2$, and so the area would be

$$s = t \times \tfrac{1}{2}(u + v) = \frac{(u + v)\,t}{2}.$$

  This formula makes sense from another perspective as well. The distance travelled will always be equal to the average speed multiplied by the time. As the speed is increasing steadily (at a constant rate) from u to v, the average speed will be the average (mean) of u and v, namely $(u + v)/2$, and so the distance must be this multiplied by t, giving $s = (u + v)t/2$.

We have not finished yet. Sometimes we don't know the time taken, for example when we want to work out the braking distance of a truck and we just know the speeds and the acceleration. We can make an equation which will do this job by remembering that speed change $= v - u = at$. The time will be the speed change $v - u$ divided by the speed change each second. As the speed change each second is the acceleration (a), this means that the time will be $(v - u)/a$.

This gives us $t = (v - u)/a$. We now put this into our most recent equation

for distance

$$s = \frac{(u+v)\,t}{2} = \frac{u+v}{2} \times t = \frac{u+v}{2} \times \frac{v-u}{a},$$

and we then continue algebraically,

$$= \frac{(u+v)\,(v-u)}{2a} = \frac{uv + v^2 - u^2 - uv}{2a} = \frac{v^2 - u^2}{2a}.$$

This equation is often written without a fraction as $2as = v^2 - u^2$, often re-arranged to give $v^2 = u^2 + 2as$.

We have shown these formulae here as students often like solving problems by knowing these formulae, choosing the right one and then substituting the numbers in.

Many problems of this kind involve things falling under gravity. As we will see in our next chapter, near the surface of the Earth, all objects which are reasonably heavy for their size (tennis balls, for example, but not feathers) accelerate downwards due to gravity with an acceleration $a = 9.8$ m/s^2 while they are falling.

Here is an example problem and solution:

Q – *Calculate the distance a tennis ball falls in* 2.5 *s when it is dropped.*

We want s

We know $u = 0$ (ball is dropped, so has no speed at start)

$a = 9.8\,m/s^2$ (acceleration due to Earth's gravity field)

$t = 2.5\,s$ (given in question)

Use $s = ut + \frac{1}{2}at^2 = 0 \times 2.5 + 0.5 \times 9.8 \times 2.5^2$

$= 0 + 30.625 = \underline{31\,m \text{ (2 sig. fig.)}}$

However, it is often more satisfying to solve the problem from first principles. To do this, we only need to remember three ideas (the third is only true if the acceleration is constant):

- speed change = acceleration × time taken,
- distance travelled = average speed × time taken,
- average speed = $\dfrac{\text{final speed} - \text{starting speed}}{2}$.

Let's use these ideas to solve the same problem:

Q – *Calculate the distance a tennis ball falls in* 2.5 *s when it is dropped.*

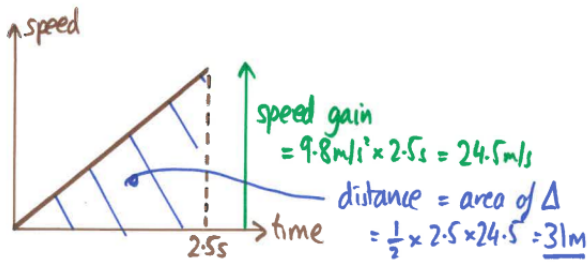
speed change = acceleration × time = 9·8m/s² × 2·5s = 24·5m/s

starting speed = 0m/s final speed = 24·5m/s

so average speed = $\frac{0 + 24.5}{2}$ = 12·25m/s

distance = average speed × time = 12·25m/s × 2·5s = 31m (2sig figs)

This can also be visualized using a speed-time graph.



Here is an opportunity for us to practise our skills with motion calculations. Use the hints below the question to help if you wish, and when you have an answer, compare it with the worked solution on page 39

Q – *Calculate the total distance needed to stop a car from* 12 *m/s (which is slightly less than* 30 *mph) if the driver's reaction time is* $\frac{2}{3}$ *s and the car decelerates at* 6 *m/s*^{2}.

Thoughts:

- We need to answer the question in two parts. First we work out how far the car moves while the driver reacts to the situation. Then we work out how far the car travells while the brakes are applied. The final answer will be the two distances added together.
- While the driver reacts, the car just keeps going at 12 m/s. We need to calculate how far it goes in the reaction time.
- Next, we need to calculate how far a car will go if it starts at 12 m/s and loses 6 m/s of speed each second (that is what a deceleration of 6 m/s^2 means). We might choose to work out the time taken as a first stage.

Now you can look on page 39 to check your answer.

To see why speed on the road is such an important factor in road safety, why not repeat the calculation for a speed of 30 m/s which is about 70 mph? You can check your answer by looking on page 39.

13 When direction matters

When a sports coach analyses a runner's performance, they care about the initial acceleration, and the speeds at the different distances down the track. They are not so worried about the direction of motion (as long as the athlete stayed on the track and was not running backwards). Similarly if you have a journey of 240 km to make, as you judge your progress, you just want to know how much of the journey remains.

However, if we want to know which way to turn at the next junction, or we are controlling a ship and needing to take the effect of wind, tide or current

into account, then directions do matter. If you are on the 32nd floor of a tall building and want to use the lift to get to the ground floor, you are not interested in how far the lift has travelled today, but you do want to know where it is currently so you know how long you are going to have to wait for it.

We are now going to lay a foundation in terms of describing situations where the motion is not always in one direction and where we are more concerned about where something is rather than how far it has travelled.

Let's define our most important terms:

- Displacement measures the current location of something.
- Velocity measures how quickly the displacement is changing.

Displacement

If you need to give the location of your home to a friend who hasn't visited you before, you could get a link off a mapping app or website. Usually, when you look at it really closely, it contains two numbers, such as this location for the Cavendish Laboratory in Cambridge:
`https://www.openstreetmap.org/#map=19/52.209204/0.091225`.

Here, after the label for the magnification (19), the first number 52.209204 is a latitude measurement, telling us how far North of the Equator the laboratory is. The second number 0.091225 is a longitude, telling us how far East of a historically important telescope in London the lab is. Any point North of the equator and East of the special telescope could be located in this way. Nowrich station, for example, has the numbers 52.626758 and 1.307006.

Anfield football stadium in Liverpool and Machu Picchu in Peru both pose a problem. We can't say how far East of the special telescope they are because they are both West of the telescope instead.

Anfield's code comes out `map=19/53.430791/-2.960300`, with Machu Picchu at `map=19/-13.163841/-72.544999`. Did you notice how the application deals with both points being to the West of the special telescope? They use negative numbers. A negative first number (latitude) shows that Machu

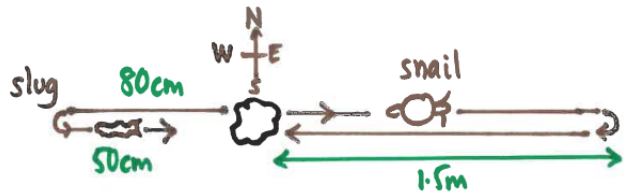
Picchu is South, not North, of the Equator. The negative second numbers (longitudes) show that Anfield and Machu Picchu are both West, not East, of the special telescope.

The two numbers taken together give a location, or a displacement on the surface of the Earth. There are two numbers, because we are looking at the situation in two dimensions (North/South and East/West).

For our studies here, we can illustrate all of the important points using just one dimension, where we consider motion along a line and possibly back again. We can study a lift (elevator) going up and down a building, for example. In this way we just need one number, but it might be negative.

The displacement s measures how far along the line we are from a fixed reference mark. So, for our lift, $s = +6$ m would mean that the lift was 6 m above the ground floor. Negative numbers, like $s = -3$ m would indicate that the lift was 3 m below ground level, in the basement.

Often we are interested in how much the position (or displacement) changes. Suppose a snail starts at a shiny stone, slithers 1.5 m East, and then slithers back to its starting point. The snail has clearly expended effort and moved 3 m in total, however the displacement change of the motion is zero (it it back where it started). So the distance (3 m) and the displacement change (0 m) are different in this case.



A slug which started at the same stone, slithered 80 cm West and then slithered 50 cm East would have moved a total distance of 130 cm. However the total displacement change would be 30 cm West - this is where the slug ends up compared to where it started.

We could use positive numbers for East and negative numbers for West and write the slug's motion as a displacement change of $\delta s = -80$ cm then $\delta s = +50$ cm. These add to give the overall displacement change $\delta s_{\text{total}} = -80 + 50 = -30$ cm, meaning 30 cm West of where it started.

Velocity

Just as displacement is the way of incorporating direction into the idea of distance, velocity does the same for speed. It tells you how much the displacement changes each second:

$$\text{speed} = \frac{\text{displacement change}}{\text{time taken}} \quad \text{in symbols:} \quad v = \frac{\delta s}{\delta t}.$$

It can be called the rate of change of displacement, and is measured in metres per second (m/s).

Unlike speed, which can never be negative, the velocity in a particular direction can be negative if the object is moving the other way (so δs is negative).

When we thought about a lift, we measured our displacements s and our displacement changes δs as positive when the lift went up above the ground. Accordingly, the velocity of the lift will be positive when s is getting bigger and the lift is going upwards. When the lift is going down, s is getting smaller, so the displacement changes δs are negative. This means that the lift will have a negative velocity when it descends.

Suppose our slug always moved with a speed of 2 cm/s. It would take it 80 cm $\div\, 2$ cm/s $= 40$ s to move West from the stone, and 50 cm $\div\, 2$ cm/s $=$ 25 s to move East again. In total it has been moving for $40 + 25 = 65$ s.

We can make the following observations

- For the westward motion, the velocity is 2 cm/s West, so $v = -2$ cm/s.
- For the eastward motion, the velocity is 2 cm/s East, so $v = +2$ cm/s.
- For the journey as a whole, with a total displacement change $\delta s_{\text{total}} = -30$ cm in time $\delta t = 65$ s, we can calculate an average velocity in the same way as we calculated average speeds

$$\frac{\text{total displacement change}}{\text{time taken}} = \frac{-30 \text{ cm}}{65 \text{ s}} \approx -0.46 \text{ cm/s}.$$

Acceleration

On page 20 we defined acceleration as the rate at which the speed changed. So, what do we call the rate at which the velocity changes? It is also called acceleration. There is no special word for the kind of acceleration which takes directionality into account.

Our original (non-directional) acceleration could be positive or negative. Positive accelerations meant that the object was gaining speed, so speeding up. Negative accelerations (or decelerations) described things which were losing speed, so slowing down.

It is more complicated with the new (directional) acceleration. A positive directional acceleration means that the velocity is getting larger (by which we mean it is getting more positive).

Suppose a supermarket trolley is pushed forwards at a speed of 0.5 m/s. Writing 'forwards' as positive, the velocity is $v = +0.5$ m/s. We now push it forwards with a bigger force. This causes both its speed and its velocity to increase (the new velocity might be $+0.9$ m/s for example) so the acceleration is positive.

Now let's imagine what happens if, to start with, we are pulling the trolley backwards at 0.3 m/s. This means the velocity starts at -0.3 m/s. We now push the trolley forwards to stop it. While the trolley has slowed down (it has lost speed), its velocity has changed from -0.3 m/s to 0 m/s, so has **increased**, because 0 is larger than -0.3. So in this case, the acceleration is positive even though the trolley is slowing down.

Similarly, negative accelerations do not always imply that something is slowing down (decelerating). They tell us that the **velocity** is going down. This could be something slowing down while going forward, or it could be something speeding up while going backwards!

The good news

While the incorporation of directionality into our ideas of distance, speed and acceleration to make displacement, velocity and [directional] acceler-

ation has introduced extra complexity, there is good news. The ideas and equations we developed on pages 22 – 26 remain valid and true. The only difference is that we will sometimes be putting negative numbers into the equations, and that the lines on velocity-time graphs will go below the horizontal axis when the velocities are negative.

We can update the principles on page 27 for directional quantities as:

- velocity change = [directional] acceleration × time taken,
- displacement change = average velocity × time taken,
- average velocity = $\dfrac{\text{final velocity} - \text{starting velocity}}{2}$.

We give an example question and solution using those ideas.

Q – *Calculate the height of a tennis ball* 1.5 *s after it was thrown upwards from the ground at* 12 *m/s.*

Use + for upwards, − for downwards. Acceleration $= -9.8\,m/s^2$

velocity change = acceleration × time

$= -9.8\,m/s^2 \times 1.5\,s = -14.7\,m/s$

original velocity $= +12\,m/s$

new velocity $= +12\,m/s - 14.7\,m/s = -2.7\,m/s$

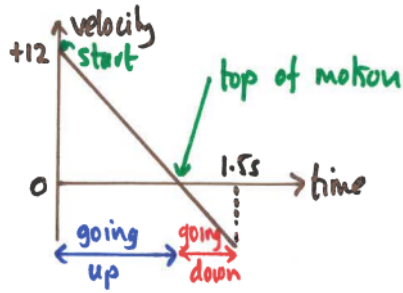
average velocity $= \dfrac{+12 + (-2.7)}{2} = +4.65\,m/s$

displacement change = average velocity × time

$= +4.65\,m/s \times 1.5\,s$

$= +6.975\,m = +7.0\,m$ (2 sig fig)

The velocity-time graph for this motion is here:



More fun with motion graphs

In this section, we are going to work out how to draw the velocity time graphs for two situations. For each question, you have three options. If you think you can give it a go straight away, please do so. Otherwise, read the thoughts and try it then. If you need more help, the answers are given on page 40. Please don't move on to the second question until you have checked your answer to the first one.

Q – *Sketch the velocity-time graph for a ball thrown upwards at* 10 *m/s and caught when it comes back down. Assume that the acceleration of dropped objects* g *is* 10 *m/s*^{2}.

Thoughts:

- Let's take upwards motion to be positive. The initial velocity is then $+10$ m/s. So our graph needs to start at $+10$ m/s on the vertical axis when $t = 0$ s.
- Any falling object accelerates downwards at g. In this question $a = -10$ m/s^2, the negative sign reminding us that the acceleration is downwards.
- If the ball starts with 10 m/s of upwards velocity, and loses 10 m/s of vertical velocity each second (that is what an acceleration of -10 m/s^2 means), it will have lost all of its vertical motion after one second. Therefore $v = 0$ m/s when $t = 1$ s.

- This will be at the top of the motion before it comes down again.
- Given it took one second to rise, it will take an equal time to fall. During this time it will continue to lose velocity at the same rate so the line will continue going down the graph below the horizontal axis.

When you have drawn your graph, do check that

- the gradient (steepness and direction of slope) of your line is the same for both the upward and downward parts of the motion. This needs to be the case as the acceleration is the same in both parts of the motion.
- the area between your line and the horizontal axis (which represents distance travelled) is the same for the upward and downward parts of the motion. This needs to be the case as the ball falls the same distance as it rose.

Now you can look on page 40 to check your answer.

Q – *Sketch the velocity-time graph for a bouncy ball dropped from a height of* 1.25 *m onto a hard floor until it hits the floor for the third time. Assume that at each bounce, its velocity is reversed instantly, so it bounces up with the speed at which it hit the floor. Assume that the acceleration of dropped objects* g *is* 10 *m/s*^{2}.

Thoughts:

- The ball is dropped at the beginning, so $v = 0$ m/s when $t = 0$ s.
- We will take upwards motion to be positive as in the last question. As before, $a = -10$ m/s^2, so the line slopes downwards below the horizontal axis.
- For all parts of the motion where the ball is just responding to gravity (at all times except during the bounces themselves) the line must slope downwards with the same gradient.

- We need to use the methods from earlier in this chapter to work out when and at what speed the ball first hits the floor.
- When the ball hits the floor, it will be going downwards, so the velocity will be negative. At the bounce, the velocity quickly switches over to an equally fast upwards value. So, if the ball hit the floor at 2 m/s, it would bounce up at the same speed, so the velocity would change from -2 m/s to $+2$ m/s.

Now you can look on page 40 to check your answer.

14 Appendix

More dimensions

At GCSE or equivalent courses, a student in the UK will not need to worry about motion in more than one dimension. However motion in more dimensions is important in further study, or in applying these ideas to real situations: I am not aware of any sports, for example, which insist that the ball only goes up and down.

The most important thing when applying our understanding of motion to a situation in more dimensions is only to use the equations or ideas to one dimension at a time.

Suppose we want to use the equation $v = u + at$ or $s = ut + \frac{1}{2}at^2$ to a two dimensional situation where the x axis points horizontally along a field and the y axis points upwards. We write the equations separately for each direction. It often helps to use subscripts to help with the labelling, so u_y means the starting velocity in the upwards direction.

This gives us:

$$v_x = u_x + a_x t \qquad s_x = u_x t + \tfrac{1}{2}a_x t^2$$
$$v_y = u_y + a_y t \qquad s_y = u_y t + \tfrac{1}{2}a_y t^2$$

Let us make some simplifications which are usually acceptable for things like

balls in air (as long as they don't move too fast). The first is that if we ignore any air resistance (the weight is usually much more significant), then there are no horizontal forces, and as we will find in our next chapter, this means that there is no change of horizontal velocity so $a_x = 0$. Vertically, the weight will cause the object to accelerate downwards (in the $-y$ direction) by about 9.81 m/s^2, which we usually call g. Therefore $a_y = -g$. With these changes, we get:

$$v_x = u_x \qquad s_x = u_x t$$
$$v_y = u_y - gt \qquad s_y = u_y t - \tfrac{1}{2} g t^2$$

Here is a worked example using these ideas:

Q – *A basketball is thrown diagonally from the ground at a* 45° *angle with horizontal and vertical velocities of* 5.0 *m/s. How far away will it be when it hits the ground?*

Initial data $u_x = 5$ m/s $u_y = 5$ m/s $a_y = -9.81$ m/s^2 $a_x = 0$

When ball lands $s_y = 0$ $s_y = u_y t + \frac{1}{2} a_y t^2$

$$0 = 5t - \frac{9.81}{2} t^2 = t\left(5 - \frac{9.81}{2} t\right)$$

so either $t = 0$ or $t = \frac{5 \times 2}{9.81} = 1.02$ s

Horizontally $s_x = u_x t = 5 \times 1.02$ s $= 5.10$ m

If you prefer to avoid a quadratic equation and work using the insight that at the top of the motion the ball is neither rising nor falling, you could take a different approach:

Q – *A basketball is thrown diagonally from the ground at a* 45° *angle with horizontal and vertical velocities of* 5.0 *m/s. How far away will it be when it hits the ground?*

When ball reaches top of motion it is neither going up nor down so $v_y = 0$

$v_y = u_y + at$ Here $0 = 5 - 9.81 \times t$

so $5 = 9.81 \times t$ $t = \frac{5}{9.81} = 0.510s$

Time taken for ball to descend will be the same.

Total time = $2 \times 0.510s = 1.02s$

Distance moved horizontally (horizontal displacement)

$s_x = u_x t = 5 \times 1.02s = 5.10m$

We can have more fun with the equation $v^2 = u^2 + 2as$. It works well in any one direction (so the following both work in two dimensions):

$$v_x^2 = u_x^2 + 2a_x s_x \qquad\qquad v_y^2 = u_y^2 + 2a_y s_y.$$

The fun begins when we add them up. This gives us:

$$v_x^2 + v_y^2 = u_x^2 + u_y^2 + 2\left(a_x s_x + a_y s_y\right).$$

By Pythagoras' Theorem, adding $v_x^2 + v_y^2$ gives v^2, the square of the speed (if we use a v without a subscript for the speed). Similarly $u_x^2 + u_y^2 = u^2$. This leads to the equation $v^2 = u^2 + 2\left(a_x s_x + a_y s_y\right)$ which expresses the final speed in terms of the starting speed. This can also be written $v^2 = u^2 + 2\mathbf{a} \cdot \mathbf{s}$ using $\mathbf{a} \cdot \mathbf{s}$ (called a dot or scalar product) as the mathematical shorthand for $a_x s_x + a_y s_y$.

Solutions to questions

Stopping from 12 m/s: question on page 28

1. Work out how far it travels while driver reacts

   distance = speed × time = $12\text{m/s} \times \frac{2}{3}\text{s} = 8\text{m}$

2. Work out how much time is taken for brakes to stop car. Deceleration = $\frac{\text{speed lost}}{\text{time taken}} \Rightarrow 6\text{m/s}^2 = \frac{12\text{m/s}}{t}$

   This means the time taken is $12 \div 6 = 2\text{s}$

3. Work out how far the car goes while braking

   Starting speed = 12m/s Final speed = 0m/s

   Average speed = $\frac{12+0}{2} = 6\text{m/s}$

   Distance = Average speed × time = $6\text{m/s} \times 2\text{s} = 12\text{m}$

Total distance = 8m + 12m = 20m.

Here is the equivalent reasoning for a speed of 30 m/s:

Equivalent calculation for stopping from 30 m/s (about 70mph)

1. Distance during reaction time = $30\text{m/s} \times \frac{2}{3}\text{s} = 20\text{m}$

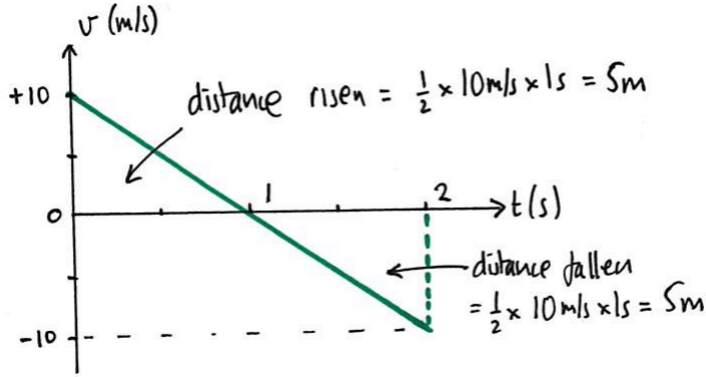
2. Time to stop = $\frac{30\text{m/s}}{6\text{m/s}^2} = 5\text{s}$

3. Distance while braking = $\left(\frac{30+0}{2}\right)\text{m/s} \times 5\text{s} = 15\text{m/s} \times 5\text{s}$
   $= 75\text{m}$

Total distance = 20 + 75 = 95m.

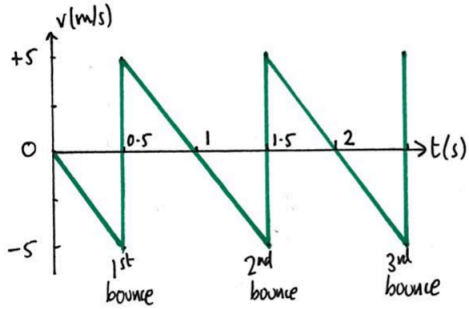
Ball thrown upwards at 10 m/s: question on page 34

The velocity-time graph for this motion is here:



Bouncy ball dropped on floor from height of 1.25 m: question on page 35

The velocity-time graph for this motion is here:



To fall 1·25m with downwards acceleration of $10 m/s^2$
$s = -1·25m$ $a = -10 m/s^2$ Ball dropped so $u = 0 m/s$
$s = ut + \frac{1}{2}at^2$ $-1·25 = 0 - \frac{1}{2} \times 10 \times t^2$ $1·25 = 5t^2$
so $t^2 = \frac{1·25}{5} = 0·25$ so $t = \sqrt{0·25} = 0·5s$ so ball takes 0·5s to fall to ground. In this time it gains $10 m/s^2 \times 0·5s$ = 5m/s of speed.

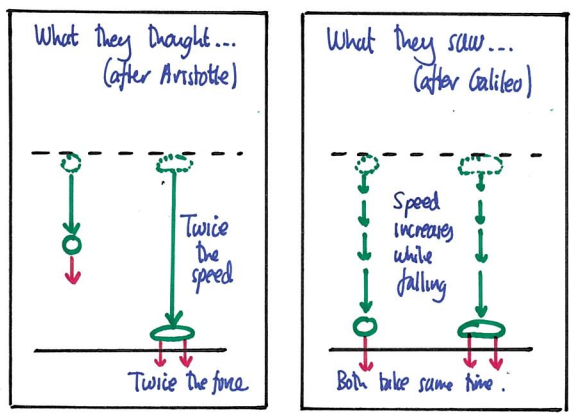
Explaining motion

The previous chapter explained how motion can be described numerically. This framework is a necessary before we can understand motion. Now that the groundwork has been laid, we can begin to explain.

Anything which pushes or pulls something else is said to apply a force to it. Forces are clearly linked to motion: a ball falls because of a gravitational force. A bike won't go unless you push on the pedals, and a lorry can't stop unless the brake pads push on the brake discs. However, the nature of the link is subtle, and is one of the most tricky concepts in all of physics for people to see clearly.

It took centuries of philosophical and scientific thought to come up with this idea: forces **change** motion, but are not needed to **maintain** motion.

Put using the terminology of the last section, people used to think that force, in some way would be related to speed or velocity: a big force would make something go faster than a small force. This is why people thought that if you dropped two weights, where one was twice as heavy as the other, the heavier one would fall at twice the speed.



If this were true, it would mean that if you got two metal lumps, one 100 g and the other 200 g, and dropped them at the same time from the same height above a flat floor, the 100 g lump would take double the time of the 200 g lump to reach the floor as it has half as much weight. A quick experiment shows that this is not right. The two lumps pretty much hit the floor

at the same time. Furthermore, if you film them falling, and play the video back, you notice that the lumps do not fall at a steady speed. They both get faster as they fall further.

The revolution in thinking was to see that forces don't cause speed - forces cause **acceleration**.

Isaac Newton wrote his 'Mathematical Principles of Natural Philosophy,' probably the most important physics book ever written, in 1687. Above all, it describes how motion works. Before beginning his reasoning and explanation, he states his 'axioms, or the laws of motion' - in other words, his guiding principles. We will use these as our starting point too.

15 The first two laws

Newton's first law states:

> **Every body perseveres in its state of being at rest or of moving uniformly straight forward, except insofar as it is compelled to change its state by forces impressed.**
> Projectiles persevere in their motions, except insofar as they are retarded by the resistance of the air and are impelled downwards by the force of gravity.[1]

This was revolutionary. Previously, people tended to think that the natural state of an object is to be stationary. If it is moving, people thought, this is in some way unnatural, and it will slow down and stop all by itself as soon as it can.

This often fits with our experience - if you are cycling and you stop pedalling, the bike slows down - Newton recognized that this is not automatic. The slowing down happens because air (and friction) are actively stopping the bike, not because the bike somehow wants to stop. If you were to remove

[1] This, and all other quotes from Newton's *Mathematical Principles of Natural Philosophy*, is taken from the translation: *Isaac Newton The Principia* by I Bernard Cohen and Anne Whitman, University of California Press, 1999.

the air resistance and friction, then your bike would keep moving all of the way to the destination without further pedalling, as long as it did not need to go up a hill.

The fact that you don't need a continual force to preserve motion is illustrated very well by balls in mid flight - they are no longer in contact with the foot or hand which gave them motion, and yet they carry on moving until something (or someone) else stops them.

This law describes what happens when there is no force acting on an object. This is a very rare situation. It is much more common for the forces on an object to cancel out, for example, if there are equal forward and backwards forces acting at the same time on it. In this case, the object behaves as if there were no force acting on it, and Newton's first law applies very well.

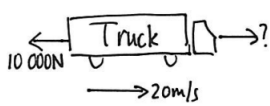
We can use this logic backwards: if something is moving at a constant velocity (its speed and direction are not changing), then we can infer that either there are no forces acting on this object, or the forces are cancelling each other out. Our cyclist, pedalling away against a firm wind to travel at a steady speed of 4 m/s South has constant motion, so the forwards force from the pedalling must be exactly equal to the force from the wind trying to slow them down.

Let's apply this to a couple of examples.

Q – *A phone sits still on a desk. If the Earth's gravity pulls it down with a* 2 *N force, what is the support force from the desk?*

Motion is not changing.
Forces must cancel out.
Support force = 2N

Q – *A truck is driving South at a steady speed of* 20 *m/s. The forces of air resistance and friction add to* $10\,000$ *N. How much force does the engine have to provide?*



Speed is constant.
Motion is not changing.
Forces must cancel out.
Engine force = 10 000N

Let us now consider the situation where the forces on an object do not cancel out. This causes the motion to change. Newton stated, in his second law:

> **A change in motion is proportional to the motive force impressed and takes place along the straight line in which that force is impressed.**
> If some force generates any motion, twice the force will generate twice the motion, and three times the force will generate three times the motion, whether the force is impressed all at once or successively by degrees.

This can make sense - twice the force generates twice the motion. But what does 'twice the motion' look like? In order to answer this, we need to take a step back and introduce two of the most fundamental quantities we need in physics: mass and momentum. We need to go back before even the 'axioms or laws of motion' in Newton's book, to the very beginning, where we find 'definitions.'

16 Mass and Momentum

> **Quantity of motion is a measure of motion that arises from the velocity and the quantity of matter jointly.**
> The motion of a whole is the sum of the motions of all the individual parts, and thus if a body is twice as large as another and has equal velocity there is twice as much motion, and if it has twice the velocity there is four times as much motion.

Newton is pointing out here that if you have a ten kilogram trolley, it is made of ten 1 kg pieces. The motion of the whole trolley is equal to the total motion of the parts, and therefore will be ten times larger than the motion of the individual 1 kg pieces. A different trolley might be larger, say 20 kg, and it would make sense to say that it has twice as much matter.

In modern science, we term 'quantity of matter' mass and we measure it in kilograms (this originally being the mass of a 10 cm $\times$ 10 cm $\times$ 10 cm cube of pure water).

Newton's definition of quantity of motion then becomes mass $\times$ velocity. Like 'quantity of matter,' its name has changed over time and it is now called momentum.

So Newton's second law is telling us that if you apply a force to an object, the object's 'quantity of motion' (ie. momentum) begins to change. As soon as you stop applying the force, the momentum stops changing, and it goes back to its 'natural state' of moving in a straight line at a steady speed.

If the force is in the same direction as the original motion, the motion increases. If the force is in the opposite direction, the motion decreases.

Before we can start calculating things, we need to do one further act of translation. Newton's phrase 'impressed force' is not what we call force, but rather the effect of the force, which is calculated by multiplying the force acting by the time for which it acts. These days this is called impulse.

So, to summarize these results, we have

- Mass (in kg) is the quantity of matter.
- The quantity of motion, momentum = mass $\times$ velocity
- Impressed force, impulse = force $\times$ time
- If a force is applied, the momentum changes.
- The change of momentum $=$ the impulse.

Let us write this mathematically using m for mass, v for velocity, $p = mv$ for momentum, F for force, t for time and the Greek 'delta' Δ to mean 'change

in':

$$Ft = \Delta p = \Delta(mv). \tag{3.1}$$

Now let us imagine that the velocity changes from u to v. This means that the momentum changes from mu to mv. So the change in momentum $\Delta p = mv - mu$. It follows that

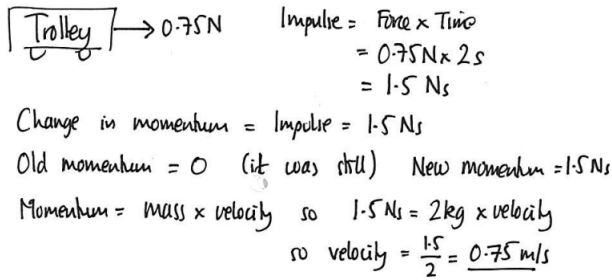
$$\begin{aligned} \text{impulse} &= \text{change of momentum} && (3.2)\\ Ft &= mv - mu = m\,(v-u) && (3.3)\\ F &= \frac{m\,(v-u)}{t} && (3.4)\\ F &= m \times \left(\frac{v-u}{t}\right) = m \times a, && (3.5) \end{aligned}$$

and we have shown that the force is indeed directly related to the acceleration, as $a = (v-u)\,/\,t$ as we wrote on page 20.

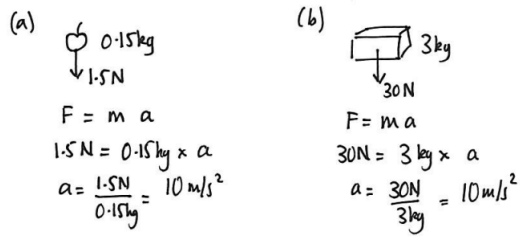
In order for this equation to work well, we need define the force needed to give a unit of mass (one kilogram) one unit of acceleration (1 m/s^2) as 'one unit of force'. Unsurprisingly, this gets called one newton of force, or 1 N for short.

Before moving further, let us use these ideas to solve some problems.

Q – *A* 2 *kg motion trolley with perfect, frictionless bearings sits still on a horizontal school table. A student pulls it horizontally with a* 0.75 *N force for* 2 *s. How fast is it now moving?*



Q – *Calculate the accelerations of (a) a* 0.15 *kg apple being pulled by a* 1.5 *N force; and (b) a* 3 *kg brick being pulled by a* 30 *N force.*



The final example takes us back to the beginning of the chapter where we thought about falling objects. The brick may weigh twenty times as much as the apple, but it has the same acceleration. The idea of mass helps us to explain why. You could even think of the brick as being made of 20 apples-worth of matter, each being pulled downwards by a 1.5 N force. Each part will fall just like the apple, and so the brick falls in the same way.

17 Weight

Everything that has mass is reluctant to change its state of motion. This is sometimes called inertia. It doesn't just mean that a 25 kg sack of builders' sand in a trolley requires a force to start it moving, it also means that a force will be needed to stop it once it is moving.

Inertia is not the only property of mass, though. It turns out that every mass is attracted to every other mass, and we call this phenomenon gravitation or gravity for short. The attraction between the pens on my desk is so small that I don't notice it. This is because the pens don't have much mass. However, the attractive force between the planet Earth and my pen is much larger, because the Earth has a lot of mass. This pulls the pen down towards the ground, and the force called its weight.

Weight is therefore a force, and is measured in newtons, just like friction.

As far as we have measured, mass is proportional to weight. In other words, if you double something's mass, you also double the gravitational force it

experiences. Mass is so closely linked to weight that usually if we want to know how many kilograms of flour we have in a cooking bowl, we put it on scales which actually measure the downwards force of gravity on it (its weight), and the people who make the scales have constructed the readout so that it tells us the mass in kilograms rather than the weight in newtons.

Sometimes people have even tried to use the same unit for mass and for weight to simplify things. However, while this is ok for most of our everyday business, it doesn't work when accurate measurements are needed. The Earth's gravity field is not equally strong at all point. In London, a 1 kg lump has a weight of 9.81 N, in Rio de Janeiro the same lump would have a weight of 9.79 N. This means that although it would take exactly the same force to accelerate it at 1 m/s^2 in both places and although it contains the same amount of stuff in both places, the lump it would be slightly easier to lift in Rio compared to London.

We refer to the weight of an object for each kilogram of its mass as the gravitational field strength (g) at that place. The gravitational field strength in Rio de Janeiro is 9.79 N/kg (9.79 newtons of weight for each kilogram of mass). Using this idea, we write:

$$\text{weight} = \text{mass} \times \text{gravitational field strength} \quad \text{in symbols:} \quad W = mg.$$

If weight is the only force acting on an object, then its acceleration will be equal to

$$a = \frac{F}{m} = \frac{W}{m} = \frac{mg}{m} = g.$$

So the acceleration of a dropped object in m/s^2 is numerically equal to the gravitational field strength in N/kg. This is why in London, where $g = 9.81$ N/kg, dropped objects fall with an acceleration of 9.81 m/s^2 if there are no other significant forces acting.

18 More than one force

If there is more than one force, then each force will contribute to the change in motion of the object. We work out the total change in motion by adding

up the forces. To do this, we need to take direction into account. Suppose an apple has a weight of 1 N and you push up on it with a 2 N force from your hand. The upwards force from your hand is larger, so we expect the apple to accelerate upwards. Given that the upwards force is 1 N larger than the downward force, we expect the apple to move as if there were just one 1 N upwards force. We say that 1 N upwards is the resultant force.

Using the same ideas of directionality as on page 30, we could call upwards forces positive and downwards forces negative. The weight is -1 N and the force from the hand $+2$ N. If we add these, we get the resultant force of $+1$ N, namely one newton upwards.

If the forces balance, then they add to zero, there is zero resultant force, and therefore no change to the motion.

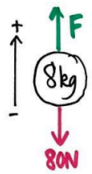
Q – *Calculate the acceleration of a* 0.5 *kg block pushed with a* 7 *N force if there is also* 3 *N of friction.*

Resultant force = 7N − 3N
= 4N

Resultant force = mass × acceleration 4N = 0·5kg × a

$a = \frac{4N}{0.5kg} = 8m/s^2$

Q – *A* 8 *kg bag of shopping has a weight of* 80 *N. How hard must you pull it to accelerate it upwards at* 5 *m/s*^{2}*?*



Resultant force = F − 80N

Resultant force = mass × acceleration = 8kg × 5m/s²
= 40N

so 40N = F − 80N

F = 40N + 80N = 120N

For situations where more than one dimension is involved, we use the methods introduced on page 15. We split each force into components: for example a horizontal component f_x and a vertical component f_y. We add the

horizontal components of all the forces to get the horizontal part of the resultant force F_x. We separately add the vertical components of all the forces to get the vertical part of the resultant force F_y. Using equation 3.5 for each part, and using the notation of page 36, we then work out

$$a_x = \frac{F_x}{m} \qquad a_y = \frac{F_y}{m}. \tag{3.6}$$

For many situations of objects thrown or dropped in air, we can neglect air resistance, so $F_x = 0$ and $F_y = -mg$ is the downwards force of weight. This means that $a_x = 0$ and $a_y = -g$, and this justifies the way we worked on page 37.

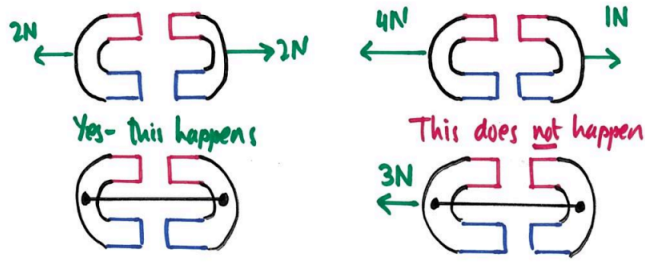
19 Newton's Third Law of Motion

A force always involves two objects - one from which the force pushes on the other. If a loaded trolley hits a wall at speed, the trolley gets stopped by a force from the wall. At the same time there is a force from the trolley pushing on (and probably damaging) the wall.

It turns out that forces always come in pairs like this. When two objects push (or pull) each other, they both feel the force. They both feel the same type of force, they both feel the same strength of force, but they feel forces in opposite directions (if one is pulled up, the other is pulled down).

When a magnet attracts a paperclip, both the paperclip and the magnet have equal forces pulling them together. The paperclip will usually move quicker because it has less mass, so the force of attraction gives it a bigger acceleration.

Why must forces come in pairs? Why must the two forces within the pair be equal in strength but opposite in direction? Look at the diagram below which shows two magnets repelling. On the left, the two repulsion forces are equal. When the magnets are stuck together, so act as a single object, the forces cancel out and have no effect on the pair of magnets.



On the right, the two repulsion forces are not equal. This means that when the magnets are stuck together, the forces do not cancel out and the combined object is able to propel itself using a purely internal force. This just does not happen. If you don't believe me, try picking up a chair you are sitting in while sitting in it with your feet not touching the floor. This in turn means that the pair of forces acting between two objects must be equal in strength and opposite in direction.

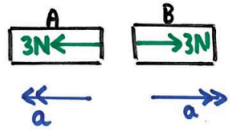
Newton assumed the same thing too, when he analyzed motion, and it is known as his third law of motion:

> **To any action there is always an opposite and equal reaction; in other words, the actions of two bodies upon each other are always equal and always opposite in direction.**
> Whatever presses or draws something else is pressed or drawn just as much by it. If anyone presses a stone with a finger, the finger is also pressed by the stone.

Even in space, where there is nothing to push on, we use Newton's third law to make things move. Given a rocket can not push on anything else, it ejects some material (usually hot gas) and pushes that away at high speed. The rocket pushes back on the gas, and in return, the gas pushes forward on the rocket.

Paradox – *Newton's third law says that if there are two objects pushing each other, the forces are equal and opposite. So they must cancel each other out. Newton's first law says that if the forces on an object cancel out, its motion won't change. So, if you put the two ideas together, the forces al-*

ways cancel out and so nothing's motion can change. But you ***can*** *change motion, so Newton's laws must be wrong, mustn't they?*



The two forces are equal and opposite (third law) but act on different objects. Each block has one force on it, and accelerates in the direction of that force (first law).

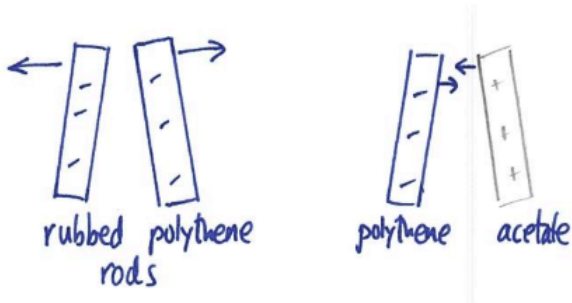
Electricity

20 Charges

The ancient Greeks noticed that lamps made of amber (fossilized tree resin) attracted dust. When they rubbed the lamps, they attracted even more dust.

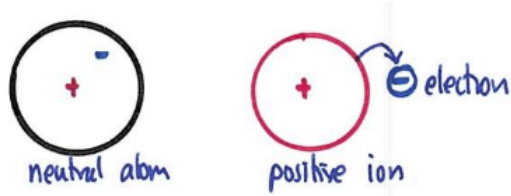
The Greek word for amber was electron, and this gives its name to everything electric, and the name of the particle responsible for most of it. When we rub many materials they can also attract dust. We say that they are charged.

If you rub two identical plastic rods with the same cloth, hang one up, and hold the other close to it, you notice them push apart or repel. However if you rub two different plastics, and bring them near to each other, sometimes the rods pull together or attract.



It follows that there must be at least two different kinds of electric charge. If there were only one type of charge, then you would always get the same effect when you brought two charged objects together. Given that one of these types can cancel out the other, we say that there are only two types of charge and we call one positive and the other negative.

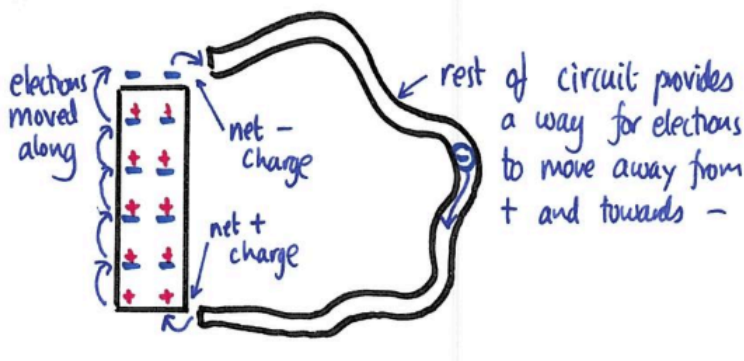
When two objects are charged with the same kind of charge (both positive or both negative) they push apart. However when they are charged with different charges (one $+$ and one $-$) then they pull together. These forces are at the root of all that happens electrically.



Within the last 100 years, we saw evidence that everything is made of atoms. Atoms have positive and negative parts, and the negative parts can sometimes be pulled off and made to move about separately. As these negative parts are responsible for most of what we see as electricity, they were named electrons. Anything which is negatively charged has too many electrons, and anything positively charged has too few.

Electric circuits - Separation

When we make an electric circuit, and put it to use, there are two things going on. In one part of the circuit there is charge separation. Something which would otherwise have been neutral has its positive and negative parts separated so that you have an excess of electrons at one end and a shortage of them at the other. Given that pulling a negative charge away from a positive one takes some force and effort, separating charge is a way of storing energy, and accordingly, this can't be done without an input of energy from some other store.



In a battery or cell, a chemical reaction separates the charge, and the store of chemical energy is reduced as the charge is separated. In a generator,

magnets allow energy involved in the motion to be used to separate the charge; and in a solar cell light itself is used to separate the charge. So the charge separator is the electrical supply.

We are now ready to make an electric circuit by connecting something to our battery.

Completing the circuit

The simplest circuit is made when one end of a metal wire is fixed to the positively charged end of a battery, and the other end of the wire is fixed to the negatively charged end of the battery. Very rapidly, the wire heats up. If the wire is thin enough, it even gets hot enough to glow.

If you repeat our experiment with a piece of string instead of a wire, you find that it does not heat up. Both string and wire, being made of atoms, contain positively charged parts and negatively charged parts (electrons). In string, all of these charged parts are held in place by strong forces. In a metal, all of the positively charged parts, and most of the electrons are also held in place. However some of the electrons are free to move within the wire. When the wire is connected to the battery, these electrons move. This makes an electric current, and as we shall see later, this causes the wire to get hotter.

If it were not for the chemical reaction in the battery, the electrons' motion would neutralise the charge on the terminals of the battery. The excess of electrons at the $-$ end would have moved round to the $+$ where they would have remedied the deficit. However, as this charge moves round the circuit, the chemical reaction in the battery separates more charge, ensuring that there is always an excess of electrons at the $-$ end and a shortage of them at the $+$ end. This keeps the charge flowing through our circuit until the chemical reaction in the battery stops for lack of reactant.

Energy transfer

We might have explained why a piece of string wire won't complete an electric circuit, but we haven't explained why a metal wire might hot. To do this,

we have to look inside the wire and see what the electrons are up to. Frequently they hit something else within the wire and bounce off it. When the electron bounces off, it gives some of its motion and some of its energy away. Consequently the electron loses some of its energy – it bounces off with a lower speed than before the collision. At the same time, the parts of the wire which have been hit gain energy, and so the wire itself gains energy and heats up.

So far, so good. The battery makes the electrons move, and the electrons give that energy to the metal of the wire when they slow down as a result of hitting things. This is perfectly correct, however there is a further complication.

If the electrons slow down as they pass along the wire from the battery, then surely they must emerge from the other end of the wire more slowly than they went in. If this is true, then there must be two consequences. Firstly, the current emerging from the end of the wire must be less than the current entering it. Secondly, as more electrons are entering the wire than leaving it, there must be a build-up of electrons within the wire. This must give the wire a negative charge which increases as time passes.

The problem is that neither of these two consequences is observed. The current entering the wire (measured by an ammeter) is identical to that leaving it at the other end. And the wire remains stubbornly neutral.

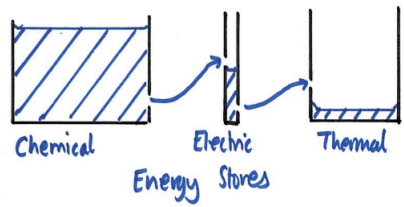
The problem is solved straight away when we realise that although the slowing down is happening throughout the wire, so is the speeding up. The battery sets up an electric field which exists along the whole wire. All of the electrons in the wire experience a force pulling them towards the $+$ end. All of the electrons, let me stress, not just the ones nearest the battery. This force causes all of the electrons to speed up. As they speed up, they hit things more often, and lose energy at a faster rate. Each time an electron hits something, it slows down a bit, but its speed is rapidly restored by the electric force provided by the battery's electric field, and it carries on speeding up until it hits something else.

Very quickly indeed an equilibrium is reached where electrons, on average, have just attained their previous speed when they undergo their next collision. At this time electrons gain just as much energy from the field in between

successive collisions as they lose during one collision. If they try going too fast, they collide more frequently than this, and slow down; if they go too slowly, they collide less frequently, and speed up – either way the equilibrium is soon restored.

Accordingly, we see that the transfer of energy from the electric field to the warmth of the wire is occurring directly at all points along the wire, and the average speed of the electrons is the same at all points in the wire.

If we wish to use a picture of energy stores to understand what is going on, there is a small store of electrical potential energy in the separation of charge. If we think of this store as a tank, the chemical energy store is used to keep this tank topped up, and it is this store which is able to provide what we need to keep the charge moving and indeed to increase the thermal energy store as electrons collide with ions.



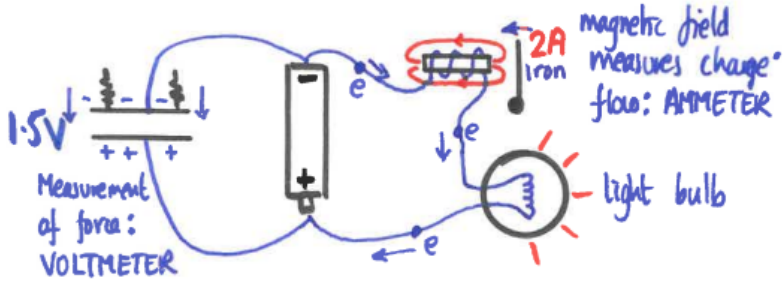
21 Measuring electricity

All of electricity concerns charges. Charge is electrical stuff, and we get our symbol Q for electric charge from the French who simply call it quantity. Perhaps unsurprisingly given the French connection, we measure charge in units called coulombs C, named after a French scientist. When you rub a balloon so it can stay on the ceiling, the charge is very small (about a billionth of a coulomb), so a coulomb is a very large amount of electrical stuff.

It is very difficult to measure charge itself. However we can make measurements of two things charge does, as shown in figure 21. One is to measure it flowing around the circuit. As we will see, a moving charge creates a magnetic field, and this enables the charge flow to be monitored. This flow is called a current. Current is given the symbol I because the French called

it intensity. It is measured in units called amps A, named after a different French scientist. The device which measures current is called an ammeter.

The other way to measure it is to measure the force causing the charge to flow around the circuit. The device for doing this is called a voltmeter. The measurement itself is known by several names, including electromotive force (e.m.f.), potential difference (p.d.) and voltage. It is given the symbol V and measured in units called volts, named after an Italian.



You can measure the current at a point in a circuit (the charge passing that point each second). However you can only measure the potential difference between two points (you need two numbers before you can work out the difference between them).

Sometimes we choose a reference point (usually the negative terminal of the battery) and call this a zero volt point. We can then use a voltmeter to measure the potential difference between our location on the circuit and this reference point. That reading is called the potential of our location.

Contrasting current and potential difference (voltage)

The difference between current and potential difference can cause confusion to students, and they might wonder why we need two measurements (V and I) rather than just one. When we studied energy transfers and work done, we learned that the energy transferred depended on two things — the force applied and also the distance moved. The power, that is the energy transferred each second, then depended on the force applied and the speed. Sometimes one can be traded off against the other for practical reasons, such as choosing a low gear when cycling uphill. This is because in this

situation, you are willing to push the pedals very quickly if it means you don't have to push them as hard. However when cycling at speed on a flat smooth road, you choose a high gear so that by putting a larger force on the pedals, you don't have to move your legs up and down as frequently.

There is a similar situation with electricity, where the force is related to the potential difference (voltage) and the speed is related to the current (charge flow rate). Electronic components as found in most gadgets are damaged if the electrical forces are too great. This limits the potential difference (voltage) which can be used. To get the job done, in compensation, more charge will have to move, so we will need a larger current. On the other hand, when connecting your town or village to the nearest power station we don't want the charge to move too fast — if we do, the wires get really hot and energy is wasted. In this case we choose to use a high potential difference (voltage).

Force is the most visual way of picturing potential difference. However the volt does not quantify force directly, but rather the energy transfer involved with charge as it passes from one side of a component to the other. If the potential difference across a light bulb is 3 V then during the time when one coulomb of charge has been pushed through it, there will be 3 J of energy transfer within it. Thinking back to the tank of electric potential energy on page 57, the potential difference is the height to which we keep the electric tank filled.

Power

The rate of transferring energy mechanically is called the power and can be worked out by multiplying velocity and force. In a similar way, we can calculate electrical power to measure the energy transferred by an electric circuit each second.

The formula is

$$\text{Power} = \text{Current} \times \text{Potential difference}, \quad \text{in symbols:} \quad P = IV.$$

Given that that the current is related to the speed the charges move and the potential difference (also known as the electromotive force) is related to the force on the charges, this formula makes sense.

We can justify the formula further. The potential difference measures the energy transferred for each unit of charge (coulomb) flowing. So the total energy transfer will be given by

$$\text{Energy transfer} = \text{Potential difference} \times \text{Charge}.$$

The energy transfer each second is given by

$$\text{Power} = \frac{\text{Energy transfer}}{\text{Time}} = \frac{\text{Potential difference} \times \text{Charge}}{\text{Time}}.$$

Now current measures the charge flow each second, so

$$\text{Power} = \text{Potential difference} \times \frac{\text{Charge}}{\text{Time}},$$

so

$$\text{Power} = \text{Potential difference} \times \text{Current}.$$

Power is measured in watts (symbol: W) in an electrical circuit, just as it is in a mechanical system, in the same way that we always measure energy transfers in joules regardless of the process going on.

Students often find it easy to see the sense in this equation by making an analogy. For example, the number of tins of soup you bring into your home each year depends partly on how often you shop and partly on how many tins you bring home each time.

A high power could be achieved with a small potential difference providing the current is high enough. A factory canteen could buy its soup one tin at a time, but someone will have to go to the shop many times each day. It would make more sense to have a small number of deliveries with vans or even trucks carrying many tins at a time. This would be equivalent to a circuit of high power with low current but a high potential difference. The potential difference used on the National Grid, which can be $440\,000$ V is equivalent to a delivery using an extremely large truck, or perhaps even a train load of soup!

Direction of current and potential difference

To keep the distinction between current and potential difference (voltage) in mind, always think about the potential difference across the component (the force externally applied to it to make the electrons move) contrasting with the current through.

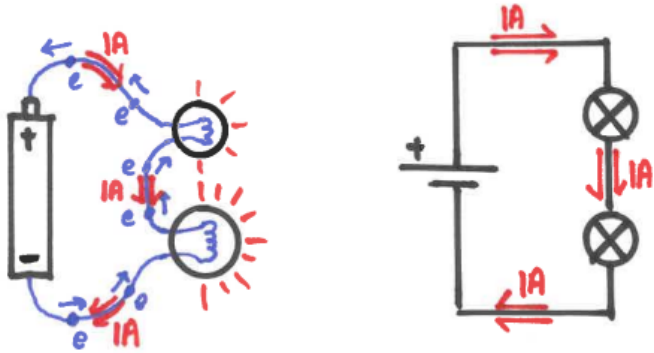
This diagram shows the conventions we use for directions when we study electric circuits. Firstly potential differences are measured as shown by the blue arrow.

We now know that in wires electrons move, and these are negatively charged, so they move from $-$ to $+$ outside of the battery. However, some positive particles can also move in electric circuits (for example, sodium ions in a salt solution). These would move the other way, from $+$ to $-$ when outside of the battery. Given we want to label a direction without having to know what is carrying the charge, we pretend when labelling diagrams that it is only $+$ charge which moves. Accordingly, our diagrams will show the current direction as flowing from $+$ to $-$, even though any moving electrons are actually going the other way.

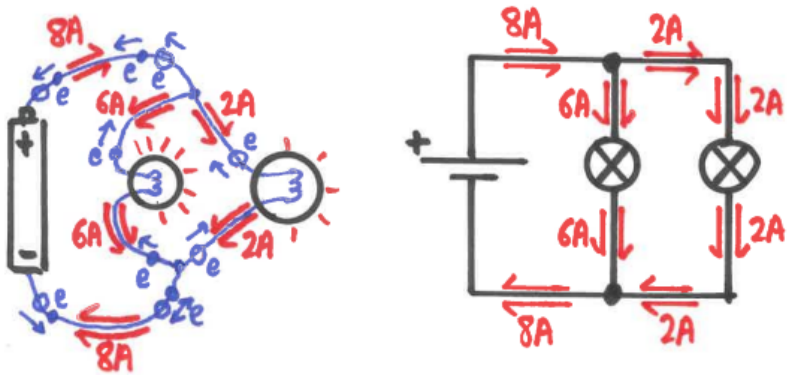
22 Circuit rules

We can explain what is going on in a circuit, for example predicting which light bulb is going to be brightest, if we know the rules which relate currents and potential differences. For each case, we will have two diagrams. The ones on the left give an idea what the circuit will actually look like. However, standard circuit diagrams are clarified versions which show the connections correctly, but use tidier shapes. These are shown on the right in each case.

Rules about current



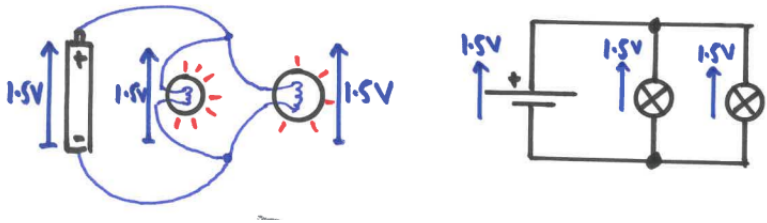
For every electron which goes into a wire at one end, another electron comes out the other end, albeit with some energy transfers having happened along the way. This is because the wire can not make (or destroy) electrons. This means that the if you measure the current in amps on both sides of the same component at the same time, you will get the same answer. Current is not used up as it travels around a circuit.



If a circuit has junctions, then this rule still applies, and so for each electron arriving at the junction, another one leaves. This is known as the conservation of charge. If 8 C in total enters a junction each second we say the current entering is 8 A. It follows that 8 A must also leave it. If there are two routes out of the junction, then the current will be shared between them. If the two routes are identical, then it will be a fair sharing and there will be 4 A flowing each way. However if the routes are not the same, then the current will be split unequally, perhaps 6 A along one route and 2 A along the other. But the total must still be 8 A.

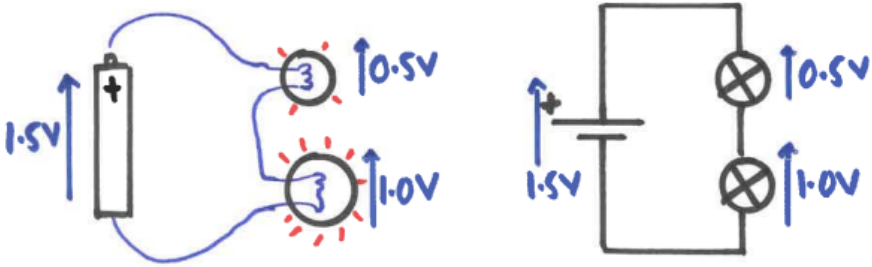
Potential difference in parallel circuits

Charge is not the only quantity to be conserved in an electric circuit. Energy is also conserved. Suppose that we have a 1.5 V battery powering a light bulb. This means that for each coulomb of charge which flows through the circuit, 1.5 J is transferred from the battery to the bulb.



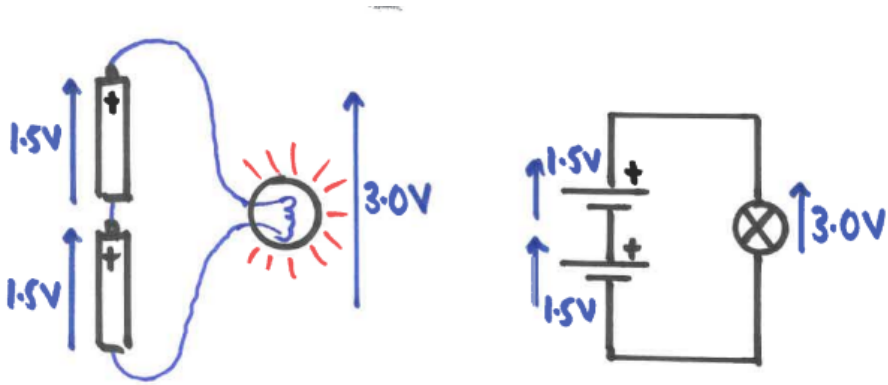
Now let us imagine that the battery is connected to two light bulbs in parallel. Each charge coming from the battery must go through the one light bulb or the other. However either way, the energy it gained by being separated in the battery will be passed to the bulb it goes through. The reduction in energy stored by the battery must be the same as the energy transferred to that light bulb. Now, potential difference measures the energy transfer associated with one coulomb of charge. So a 1.5 V battery transfers 1.5 J to a light bulb for each coulomb which flows through that bulb. Accordingly, the potential difference across each lamp is 1.5 V. So although the charge is shared (you could equally well say that the current is divided), the potential difference is not. Put more bluntly voltage does not split at junctions.

Potential difference in series circuits



On the other hand, if the two light bulbs are connected to the 9 V battery in series, then all of the charge which flows must go through both bulbs. It will go through one and [then] the other. During the time taken for one coulomb to flow, there will be 1.5 J of energy transferred from the battery to the bulbs. However as the total energy can not change, this energy must be shared between the bulbs. So if one bulb received 0.5 J, the other must have received 1.0 J. Given that potential difference (or voltage) is energy transfer for each coulomb flowing through, this means that there will be a potential difference of 0.5 V across one bulb and 1.0 V across the other.

More than one battery

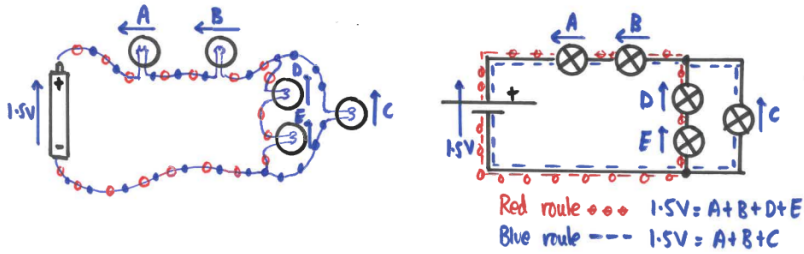


If two batteries are connected in series to a light bulb, then all of the charge which will flow through the bulb will have gone through both batteries. During the time taken for one coulomb to flow, each battery will have contributed 1.5 J to the energy transfer, and accordingly the bulb will receive 3.0 J. The potential difference across the bulb in this case will therefore be $1.5\text{ V} + 1.5\text{ V} = 3.0\text{ V}$.

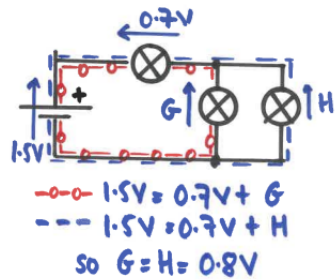
If one of the batteries had been connected the other way round, then its potential difference would be in the other direction. In our diagram, that means its arrow would point the other way. The result is that the force applied to the charges by the batteries would cancel out $1.5\text{ V} - 1.5\text{ V} = 0\text{ V}$, there would be no current flowing and the bulb would not light.

If two opposing batteries have different potential differences, then they will not be able to completely cancel each other out. A 3 V battery opposing a 2 V battery will leave $3\text{ V} - 2\text{ V} = 1\text{ V}$ potential difference to push a small current around the circuit

Rule for potential differences



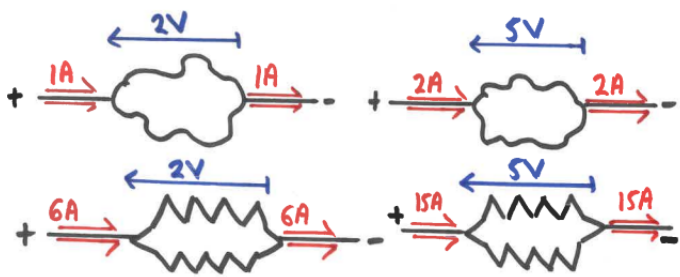
If you choose any route through the circuit from one end of the battery to the other (choosing any direction you like at any junctions you reach), and add up the potential differences across the components you go through on your journey, you will get the potential difference of the battery.



Suppose two components are in parallel with each other. Given that choosing one or the other does not affect the rest of the route (or the potential difference across the battery), it must be that the potential difference across those two components must be the same. It is in general true that components in parallel with each other will have the same potential difference.

23 Resistance

Applying a certain potential difference across an object, say two volts $(2\,\text{V})$, won't always lead to the same current flowing. It depends how well the object conducts electricity. In the diagram below, the spiky sample conducts electricity better than the rounded sample. You can tell this, because for both $2\,\text{V}$ and $5\,\text{V}$ of potential difference, it has a larger current than the rounded specimen. We say that the better conductor (here the spiky sample) has a lower resistance than the rounded sample.



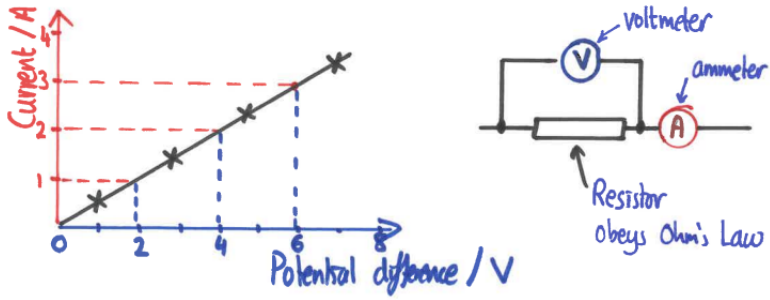
When you apply a larger potential difference across something in a circuit, you put a larger force on the electrons to push them around the circuit. You would therefore expect them to move faster, leading to a larger current through the circuit. This is true for both the rounded and spiky examples in the diagram.

For many materials, we find that doubling the potential difference (or voltage), and so doubling the force on the electrons, leads to a doubling in the speed and therefore the current. These materials are said to obey Ohm's law.

In our diagram, the rounded sample does not obey Ohm's law: to make the current get twice as large we needed to more than double the potential difference. The spiky sample, however does obey Ohm's law: when the potential difference was made $2.5\times$ larger, the current also got $2.5\times$ larger.

Characteristic graph

The graph we might plot to see if something obeys Ohm's law is called its characteristic. The current through it is plotted against the potential difference across it. The component drawn in the diagram, a resistor, does obey Ohm's law. The circuit diagram shows how you would connect a resistor to a voltmeter and ammeter so that the potential difference across it and the current through it can be measured. This little circuit would be attached to batteries of different potential differences to obtain the data in the graph.



Notice in this case that

- For each point on the graph, the potential difference in volts is always twice as big as the current in amps: Potential difference in volts $=$ Current in amps $\times 2$.
- For each point on the graph, the current in amps is always half as big as the potential difference in volts: Current in amps $=$ Potential difference in volts $\div 2$.
- To make the current increase by 1 A, you need to increase the potential difference by 2 V.
- If you increase the potential difference by 1 V, the current increases by 0.5 A.

We can summarize all of these points by writing an equation for the current through the resistor.

$$\text{Current} = \frac{\text{Potential difference}}{\text{Resistance}} \quad \text{in symbols:} \quad I = \frac{V}{R}$$

The resistance here will be 2 units, indicating that for each extra amp you need to increase the potential difference by 2 V. So the resistance is 2 volts per amp or 2 V/A.

Now the 'volt per amp' has a special name, the Ohm. We can't use O as the symbol for Ohm as this could be confused with a zero, so we use the Greek letter O instead, namely omega Ω.

The better something is at conducting electricity, the smaller its resistance will be.

Resistance and Temperature

A sample of a material obeys Ohm's law if its resistance does not change as the current increases providing the temperature is kept steady. Metals fit this very well. Ordinary wires in circuits containing light bulbs and other components tend not to become noticeably warmer when small currents pass, and so these obey Ohm's law as well.

Resistors obey Ohm's law because it is in their job description and they would be sacked and thrown away if they did not. However if you read a resistor's employment contract carefully (sorry, I mean its specification sheet), it does specify a maximum power. If you exceed that power, two potentially nasty things happen. Firstly, the temperature will rise enough to affect the resistance of the resistor significantly. Secondly, it may get sufficiently hot as to pose a hazard. That said, high power resistors are designed to cope with dissipating energy (raising the energy stored thermally in the surroundings) - the high temperature won't hurt *them*, but they may still be too hot to touch.

Light bulbs contain thin filaments of wire which are designed to rise to high temperatures when a current passes. This is what enables them to glow and give off light. At these temperatures, their resistance is much higher than it was when they were unlit. Within the metal is a regular grid of metal ions (eg. Cu^+), and just the right number of free electrons (electrons not attached to a particular ion or atom) to keep the sample electrically neutral.

At room temperature, the ions oscillate or wobble. An electron can hit a wobbling ion in such a way that it loses kinetic energy. To get moving again,

it relies on the energy store of the electric field set up by the battery, which sets up a force pulling it towards the $+$ terminal. This is the origin of resistance and dissipation of energy in a wire. As the temperature increases, the ions wobble more, causing more disruption to the electron motion, thereby making a bigger resistance, a smaller current, providing more warmth to the surroundings, and leading the store of energy in the battery to be used up more quickly.

Silicon and other semiconductors behave differently. In these materials there are electrons which are not *quite* free to move, but which do not require much energy to liberate them. As you raise the temperature of silicon, the energy stored in the oscillating atoms has an increased likelihood of freeing those electrons, which can then carry current. Accordingly, for semiconductors, an increase in temperature is usually accompanied by a rise in current and a drop in resistance. This is what is going on in a thermistor.

For safety reasons, never connect a thermistor directly to a battery without some other resistor in series with it (unless you have an active and effective way of cooling it - such as being in water). While things may start off ok, a vicious cycle soon sets in - the resistance of the thermistor causes it to get hot, which in turn lowers its resistance, so it carries more current, enabling it to get even hotter, lowering its resistance still further enabling it to get hotter still until it melts or causes other hazards. A suitable [ordinary] resistor in series with it effectively limits the current which can flow, protecting the thermistor, and also preventing your students getting hurt.

Light emitting diodes (LEDs) also contain very small pieces of semiconductor material, which are easily damaged by high currents. They also get better at conducting as they carry more current, so they also need to be put in series with a suitable resistor for their protection.

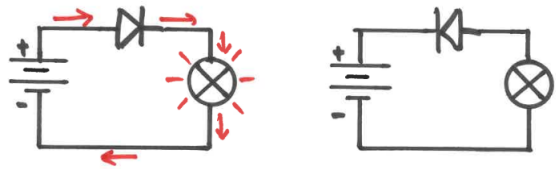
You can estimate the resistance of the safety resistor using

$$\text{resistance} = \frac{\text{battery potential difference}}{\text{maximum safe current}}$$

For LEDs, you would want to limit the current to about 20 mA unless it is a high brightness LED. When working with small thermistors, you would want smaller currents such as 5 mA to prevent them warming up too much.

Diodes

A diode is a clever component which only lets current flow one way through it, and even then only if the voltage is higher than a threshold. The threshold voltage for most diodes is about 0.6 V, but is considerably higher for light emitting diodes (LEDs).

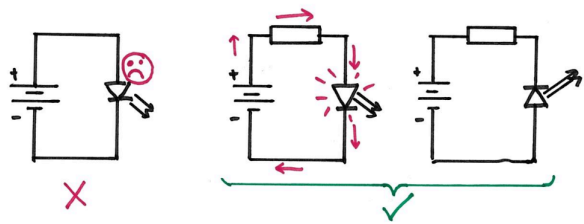


The symbol for a diode, shown in the circuit diagram above, is an arrow. The arrow indicates the direction of the allowed current (on its way from $+$ to $-$). A diode connected the other way round acts almost like a break in the circuit. Even a diode connected the right way round will let negligible current flow through it if the potential difference across it is lower than the threshold voltage.

An explanation of how diodes work is given on page 78.

More Diodes and LEDs

Some diodes are designed to light up when a current passes through them. These are called light emitting diodes (LEDs).



Remember, when connecting a circuit with a battery and an LED, to include a current-limiting resistor as explained on page 69. If you don't do this, and the diode is connected the correct way for charge to flow through it, the current becomes large enough to damage the very small crystal of semiconductor, which then breaks. The LED will not work again, ever.

Assuming a current-limiting resistor has been included in the circuit, a nice demonstration can be made in which if the battery is one way round, the LED lights, and if it is the other way round, the LED does not light.

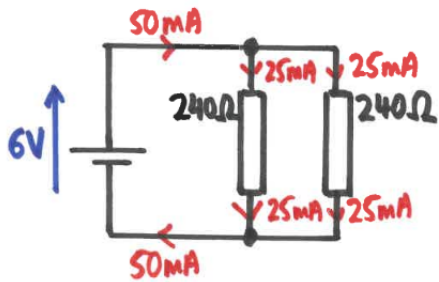
An explanation of how an LED works is given on page 80.

Putting resistors together in parallel

Suppose we have a 6 V battery and two $240\ \Omega$ resistors. If we just connect one of the resistors, the current would be

$$\text{current} = \frac{\text{potential difference}}{\text{resistance}} = \frac{6\text{ V}}{240\ \Omega} = 0.025\text{ A} = 25\text{ mA}.$$

Now suppose we connect both resistors so that each one is connected across the whole battery. In this way, they are connected in parallel. Charge from the battery in effect chooses to go through one or the other, but can't go through both.



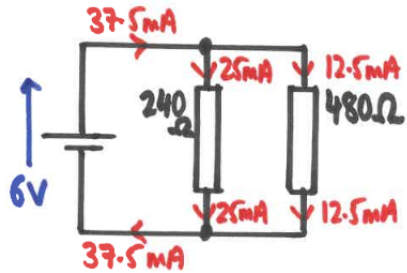
This will not affect the voltage of the battery, so each resistor carries 25 mA as before. The total current will be 50 mA. What this means is that the resistance of the circuit as seen from the battery is

$$\text{resistance} = \frac{\text{potential difference}}{\text{current}} = \frac{6\text{ V}}{0.05\text{ A}} = 120\ \Omega.$$

The effect of putting two identical resistors in parallel is that the combined resistance is half the resistance of the individual resistor.

Now let's connect two different resistors in parallel to our 6 V battery: a $240\,\Omega$ resistor and a $480\,\Omega$ resistor. The current through the $240\,\Omega$ resistor will be 25 mA as before. The current through the other resistor will be

$$\text{current} = \frac{\text{potential difference}}{\text{resistance}} = \frac{6\text{ V}}{480\,\Omega} = 0.0125\text{ A} = 12.5\text{ mA}.$$



The total current is $25\text{ mA} + 12.5\text{ mA} = 37.5\text{ mA}$. Notice that adding a new resistor increased the current as it provided a new route for the charge, even though the new resistor had a larger resistance.

If you wish, you can calculate the resistance of the new combination from this new total current:

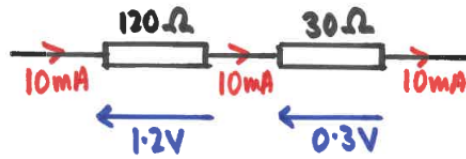
$$\text{resistance} = \frac{\text{potential difference}}{\text{current}} = \frac{6\text{ V}}{0.0375\text{ A}} = 160\,\Omega.$$

This overall resistance is less than the smaller individual resistance.

Putting resistors together in series

The rule when you put resistors in series (so the charge has to flow through one then the other - it can't choose) is much easier: **you just add the resistances**.

So if you put a $120\,\Omega$ resistor in series with a $30\,\Omega$ the circuit has an overall resistance of $120 + 30 = 150\,\Omega$.



Let's see why this works. Suppose the current in the circuit is $10\,\text{mA} = 0.01\,\text{A}$. This current will pass through both resistors (one then the other), and the potential differences across the two resistors will be

$$\text{potential difference} = \text{current} \times \text{resistance} = 0.01\,\text{A} \times 120\,\Omega = 1.2\,\text{V}$$

$$\text{potential difference} = \text{current} \times \text{resistance} = 0.01\,\text{A} \times 30\,\Omega = 0.3\,\text{V}$$

The total potential difference across the circuit will be the sum $1.2 + 0.3 = 1.5\,\text{V}$. In a series circuit like this, the current through each resistor is the same as the current supplied by the battery, but it is the potential difference which is shared between the resistors. The overall resistance of the circuit is

$$\text{resistance} = \frac{\text{potential difference}}{\text{current}} = \frac{1.5\,\text{V}}{0.01\,\text{A}} = 150\,\Omega.$$

We see that the overall resistance is indeed the sum of the original resistances.

Potential division

So, what will happen if we get our series circuit of $120\,\Omega$ and $30\,\Omega$ resistors and connect it to our $6\,\text{V}$ battery?

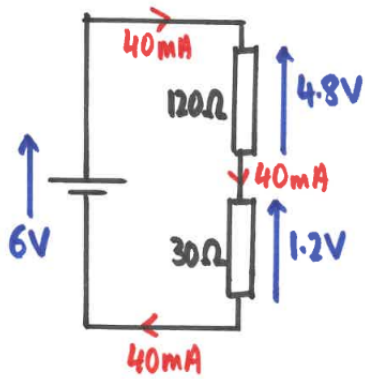
We already know that this combination acts like a $150\,\Omega$ resistor. So the current from the battery will be

$$\text{current} = \frac{\text{potential difference}}{\text{resistance}} = \frac{6\,\text{V}}{150\,\Omega} = 0.04\,\text{A} = 40\,\text{mA}.$$

Now we know this, we can work out the potential difference across each resistor:

$$\text{potential difference} = \text{current} \times \text{resistance} = 0.04\ \text{A} \times 120\ \Omega = 4.8\ \text{V}$$

$$\text{potential difference} = \text{current} \times \text{resistance} = 0.04\ \text{A} \times 30\ \Omega = 1.2\ \text{V}.$$



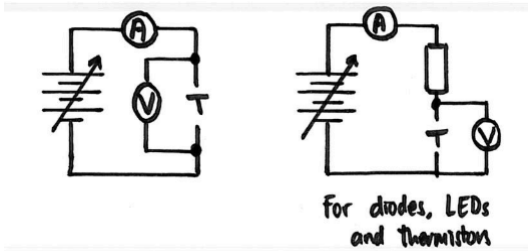
We see that the total potential difference $4.8 + 1.2 = 6$ V which correctly matches the battery potential difference. Notice, however, the way in which this potential difference is shared out between the resistors. One resistor had a resistance four times the other $120\ \Omega = 4 \times 30\ \Omega$. When they were put in series, the potential difference were shared in the same way $4.8\ \text{V} = 4 \times 1.2\ \text{V}$.

One resistor had one fifth of the total resistance $30/150 = \frac{1}{5}$, and it ended up with one fifth of the total potential difference $1.2/6.0 = \frac{1}{5}$. The other resistor, with four fifths of the total resistance ended up with four fifths of the potential difference.

Connecting the potentiometer

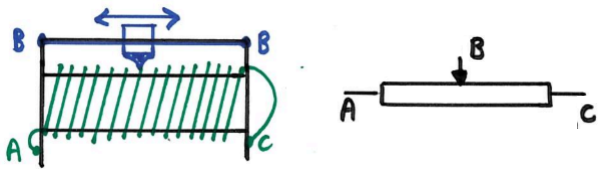
School experiments, including assessed practicals for examinable courses) often require the student to vary the potential difference across a component. There are three ways of doing this in a typical school laboratory.

The most straightforward way is if the school has power supplies with a variable voltage output. The circuits below can then be used to measure the current through and potential difference across a component.



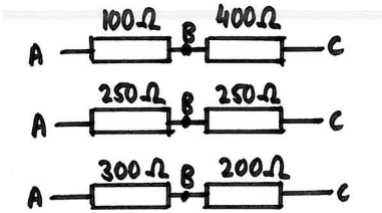
Notice that when investigating a diode, LED or thermistor, we must include a current-limiting resistor. Furthermore, we must set up our voltmeter so that it records the potential difference across the test component only (and not the resistor as well).

Many schools have heavy duty rheostats which often get called into service if variable voltage power supplies are not available.



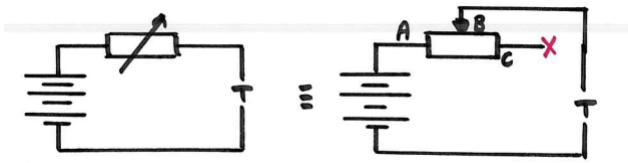
The rheostat is a large potentiometer. In other words, it contains one resistor which is split into two sections by a moving contact. This effectively makes it equivalent to two resistors in series where you can vary the resistance of the two parts subject to the constraint that the total resistance is always the same.

So a $500\ \Omega$ potentiometer (or pot for short) can be set to make any of the possibilities below.



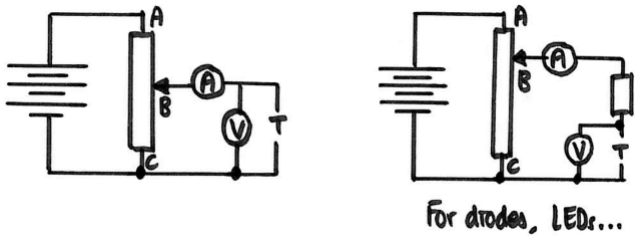
A potentiometer has three connections - the connections to the two ends (A and C in the diagram above), and the connection to the sliding contact (B).

The simplest way of connecting a rheostat (or other potentiometer) is as a variable resistor. In this circuit we use one of the fixed contacts and the sliding contact (eg contacts A and B only).



The disadvantage of this circuit is that while at one extreme it can put the entire potential difference of the battery across the test component, the maximum resistance may not be enough to reduce the potential difference across the test component sufficiently for a good characteristic curve to be drawn. If the test component is a light bulb, you might find that although you can dim the bulb, it never goes dark.

The situation is improved if you use all three of the connections as shown below. In this way, by adjusting the slider, you can deliver any potential difference to the test component from zero up to the battery voltage.



In this circuit, the two resistances of which the potentiometer is made form a potential divider circuit.

If the current taken by the test component is small, you will find that setting the sliding contact half way along the resistance will give half the battery voltage to the test component. However for larger currents drawn, the relationship between position of the slider and potential difference across the test component is not linear.

24 Appendix: Understanding at a deeper level

The content of this section is not needed by students or teachers at GCSE or A Level. However a bit more information about what is going on inside wires, diodes and LEDs as they carry current may well help us understand more of the wonderful mystery of the electric circuit.

The electron as a fermion

When a wire is lying on a table, disconnected, the electrons are not stationary, waiting in hope for a battery to liven up their day. They are already moving up and down the wire at high speed – bouncing off the sides, the ends, other electrons and the other parts inside the wire. You don't normally notice this, because just as many of them are moving one way as moving the other.

However, when you connect the wire to the battery, the motion of the electrons becomes measureable. The battery sets up an electric field in the wire because of its separation of charge. The electrons within the wire all experience this field and are attracted to the $+$ and repelled from the $-$ of the battery. The electrons which are already moving the right way (from $-$ to $+$) speed up. The ones moving the other way slow down. The ones which were going very slowly the wrong way are reversed. And now, we have a net motion of electrons – since more are going one way than the other.

But if the electrons are whizzing up and down the wire (equal numbers whizzing up and down) even before the battery is connected, how come they don't collide with things and lose energy until the battery is put in place? And why, exactly, are the electrons moving in the first place?

The answer goes to the root of what it is to be an electron. Electrons belong to the class of particles called fermions. Fermions are all observed to have a fascinating property – at no single point in time will you find two fermions without some distinguishing feature between them. So if you have two electrons in the same place at the same time, they might, for example, be different by virtue of having different velocities.

Not all electrons are free to move up and down the metal wire. Some are fixed in place. These are said to be localized electrons – they are fixed within a particular atom, and their position is marked as such. Accordingly, there is no problem with two apparently identical electrons being fixed within different atoms – they are obviously in different places.

However, the electrons which move, and therefore 'carry' the electric current, don't behave like this. In a sense, they all belong to the wire as a whole and act as a combined object flowing like a wave up and down the wire. In a sense, they are all in the same place – they are part of the same fluid or wave. Therefore they can only distinguish themselves from each other by having different velocities. This is why they can't all be stationary. If they were all stationary, they would all have the same velocity, so you could not tell them apart, and this is not allowed.

Consequently, the electrons usually take the lowest speeds allowed by this constraint. Yes, they do indeed collide with the fixed parts of the wire and each other before the battery is connected, but they can't lose energy during these collisions. For if an electron lost energy, it would also lose speed, which would mean that its velocity would change to some lower value. But it isn't allowed to have that value, because there is another electron in the wire which already has that lower velocity.

However, when the battery is connected, the electrons moving the 'right way' speed up a tiny bit, while the electrons moving the other way are slowed down by the same amount. This means that it is now possible for one of the fastest electrons moving the 'right way' to be bounced to a very slightly slower speed in the opposite direction and still have a unique velocity. Thus collisions involving a loss of electron energy can now occur. As we have explained earlier, the electron is soon sped up to its original speed by the electric field of the battery, so you don't notice a difference in the average velocity of the electrons at different points along the wire.

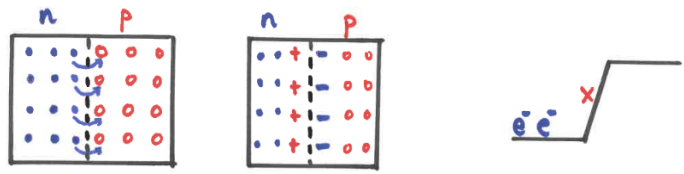
How a diode works

The diode is made of a semiconducting material. This is a material like silicon or germanium which has very few free charge carriers available, and which has four electrons in its outer shell. The material can be doped by adding a

small percentage of another element.

So called n-type semiconductor has a small amount of a material with five outer shell electrons like phosphorus added. This contributes more electrons to the material, and the 'fifth electron' is much freer to move around the material than the regular electrons in the semiconductor.

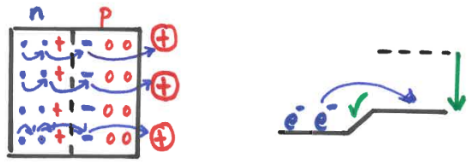
The opposite is when a small amount of boron (or some other material with three outer shell electrons) is added. This makes p-type semiconductor which has fewer electrons than pure silicon. This material can much more easily accept electrons than one without the added boron.



In a diode, thin layers of n-type and p-type semiconductor are joined, as shown in the diagram. Some of the especially mobile 'fifth electrons' from the n-type semiconductor move to join three-electron atoms in the p-type semiconductor.

This causes a charge difference at the barrier. The n-type region has lost electrons, so is now positively charged. The places where the electrons were are called holes. The p-type material has gained electrons so is negatively charged. This region near the junction with no 'fifth electrons' is called a depletion layer. Without mobile charge carriers it acts as an insulator, preventing further charge flow.

There are free electrons ('fifth electrons') in the n-type material further away from the boundary. However, because of the charge separation at the boundary, even if you could get them to cross it, you would have to push them hard to overcome the electrostatic force — you are trying, after all, to push them away from a $+$ charged region to a $-$ charged region. Effectively moving them would be like trying to push something uphill.



When we attach the diode to an electric circuit with the n-type semiconductor connected to the $-$ of the battery, two things happen. Electrons from the battery move through the n-type semiconductor to fill in the holes near the boundary. Meanwhile, the $+$ terminal of the battery attracts the free electrons in the p-type material away from the boundary. It is now possible for electrons to cross the boundary again.

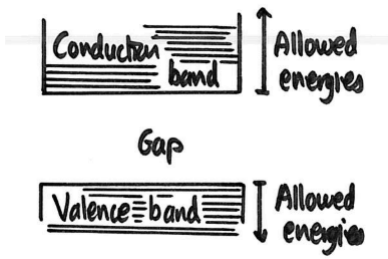
Another way of looking at it is to note that the electric field of the battery has reduced the energy required for an electron to move across the boundary, making the journey more accessible.

For a regular diode, the potential difference set up across the boundary by the depletion zone is about 0.6 V - hence the battery needs to overcome this before charge can flow.

How an LED works

In order to explain how an LED works, we are going to have to explore the energy levels of a semiconductor in more detail.

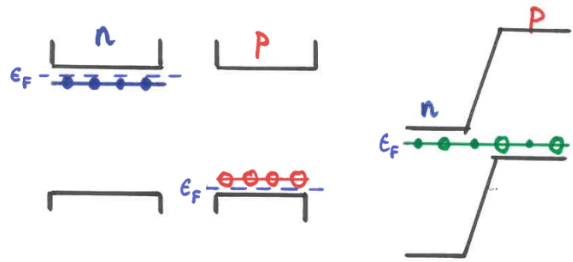
In a semiconductor, the allowed energies of electrons form two allowed regions, with a forbidden region (or gap) between them. The chance of an electron gaining enough energy to cross the gap is very much smaller than the chance of a single lottery ticket winning the jackpot, so the gap limits the ability of the material to conduct.



Usually the lower allowed region (the valence band) is full of electrons. They all move, but because just as many move one way as the other, there is no net current. The upper allowed region (the conduction band) has no electrons.

If an electron from the valence band were to be promoted, it would be free to move into any state of motion in the conduction band — so current could be carried. Furthermore, the vacancy it would leave behind in the valence band (the hole) would break the symmetry of the filled states there allowing the valence band to conduct too.

When we dope a semiconductor, we make extra energy levels. The fifth electrons in n-type semiconductor occupy a energy just below the conduction band. The three-electron atoms in the p-type semiconductor provides an energy level just above the valence band which can accept an electron from it.



So, if we draw the energy level of the most energetic electrons, it will lie between the highest occupied electron energy and the lowest unoccupied electron energy.

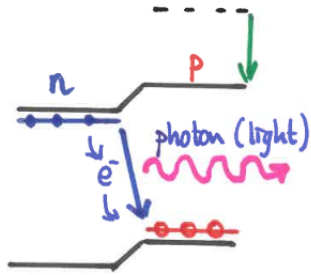
In n-type semiconductor, this will lie between the 'fifth electron' level and the conduction band. In p-type semiconductor, it will lie between the valence band and the 'three electron atom' level.

Now if you use a piece of pipe to connect the bottoms of two buckets, and then pour water in, you notice that once things have settled, the water height in both buckets is the same. If it weren't the pressure difference would push water through the pipe until the levels were equal.

A similar thing happens when you stick a piece of n-type semiconductor next to a piece of p-type semiconductor. Things adjust to ensure the en-

ergy of the most energetic electrons on both sides of the boundary are the same.

This gives rise to the situation shown on the right in the diagram above. This is an alternative explanation of the depletion zone and the potential difference across it.



Now let's make the diode conduct by connecting a battery to counteract this potential. Notice that the effect of doing this is the energies of the most energetic electrons on either side of the boundary are now different as shown in the diagram above.

When an electron now passes from the n-type to p-type region, it will descend in energy. This will reduce its store of electric energy, and so this energy must be transferred somewhere. It is transferred as light. Each time an electron crosses the boundary, one packet (or photon) of light is produced.

Quantum physics has enabled us to find out that the energy of one of these photons is given by hf where h is a fixed number called the Planck Constant with a value of 6.63×10^{-34} Js and f is the frequency of the light in hertz.

The energy removed from the electrical store (and thus the chemical store of the battery) must equal the charge multiplied by the battery, hence qV where $q = 1.60 \times 10^{-19}$ C is the magnitude of charge on an electron.

Putting these two equations together tells you that for light to be emitted $qV > hf$, and so

$$V > \frac{hf}{q}.$$

Using this formula you can see that for a green LED to be lit, where green light has a frequency of about 6×10^{14} Hz, we must have a potential difference of more than 2.5 V.

A red LED requires a lower voltage as the frequency of red light is lower.