Inquiry 2: Collecting and processing data

Interpreting results

When you write an investigation report it will probably contain charts, diagrams or graphs, such as seen in Figure B4.11: a student used an isovolumetric change of a gas to estimate a value for absolute zero, using the apparatus shown in Figure B3.14, Topic B.3.

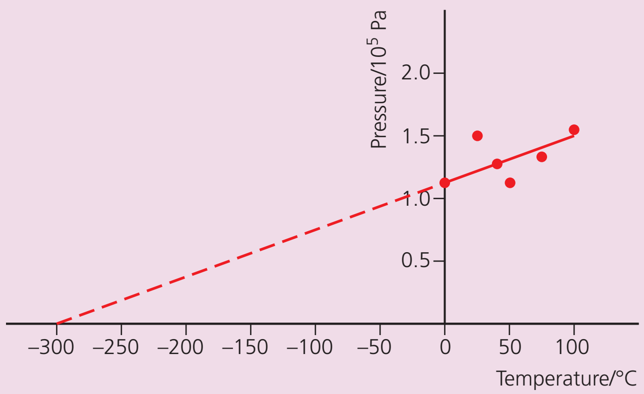


Figure B4.11 Results of an isovolumetric experiment.

Graphical representations summarize the data collected and enable the reader to gain a quick impression of the results of the investigation. Your investigation report should summarize what you have learned from any such representations.

This may involve any, or all, of:

- considering the quantity and spread of measurements taken
- discussing the quality and reliability of graphs
- identifying any relationship that can be identified between the two variables
- describing and explaining any pattern or trend shown by a graph (if no precise relationship is apparent)
- quoting values for intercepts and explaining their significance
- calculating gradients and explaining their significance
- calculating areas under graphs and explaining their significance
- identifying and explaining any maxima or minima (and other turning points).

Anomalous readings (outliers) should not usually be rejected without being checked. The most common reason for an outlier is a simple error in measurement, or recording, which will be evident if it is repeated.

It is unusual that an anomalous reading is confirmed as being correct. But, if it is, you should take further measurements at slightly different values to try to determine the extent of the anomaly.

Ideally, the pattern of the data should be processed quickly at the time of the experiment, often by drawing a rough graph. This prevents the situation whereby you only notice an anomalous reading when processing the data later, when checking the measurement again is not possible. See Figure B4.12 which shows the results of a Boyle's law experiment (Topic B.3).

If not checked, you should include any outliers in your report (and note them as outliers). You need to use your judgement as to whether the outlier should affect your conclusions.

Analyse the graphs shown in the last two figures and discuss the quality of the experimentation that they represent.

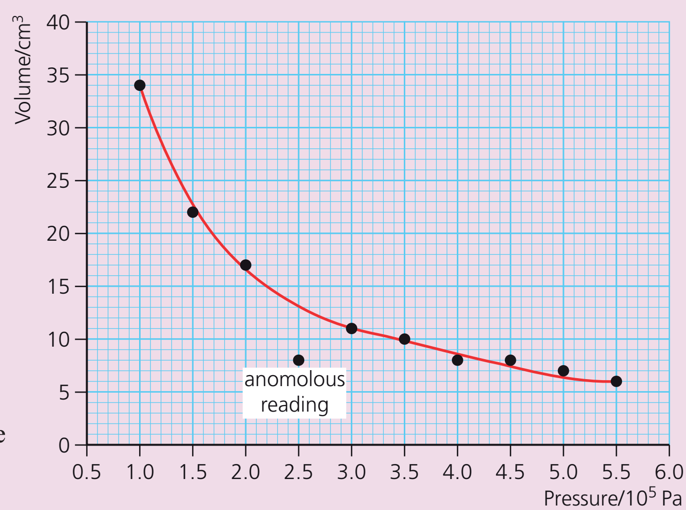


Figure B4.12 An anomalous reading (uncertainty bars not shown)

HL ONLY

WORKED EXAMPLE B4.5

A gas of volume 0.080 m^3 and pressure 1.4×10^5 Pa expands to a volume of 0.11 m^3 at constant pressure when 7.4×10^3 J of thermal energy are supplied.

- **a** Name the thermodynamic process.
- **b** Calculate the work done by the gas.
- **c** Determine the change in the internal energy of the gas.

Answer

- **a** An isobaric process.
- **b** $W = P\Delta V = (1.4 \times 10^5) \times (0.11 - 0.080) = +4.2 \times 10^3$ J
- **c** $Q = \Delta U + W$

  $(+7.4 \times 10^3) = \Delta U + (+4.2 \times 10^3)$

  $\Delta U = +3.2 \times 10^3$ J (The internal energy of the gas increases.)

WORKED EXAMPLE B4.6

The volume of an ideal monatomic gas is reduced in an adiabatic compression by a factor of 8.0. Determine the factor by which the pressure in the gas changes.

Answer

$PV^{\frac{5}{3}} = \text{constant}$

$P_1V_1^{\frac{5}{3}} = P_2V_2^{\frac{5}{3}}$

$\frac{P_2}{P_1} = \left(\frac{V_1}{V_2}\right)^{\frac{5}{3}} = 8^{\frac{5}{3}}$

$\log\left(\frac{P_1}{P_2}\right) = \frac{5}{3} \times \log 8 = 1.505$

$\frac{P_1}{P_2} = 32$

The pressure has increased by a factor of 32.

If this had been an isothermal change the pressure would have increased by the same factor (8.0) as the volume has decreased. In this example the pressure has increased by a bigger factor because the temperature increased in an adiabatic compression.

14 Figure B4.13 represents four successive changes to the state of a gas of constant mass.

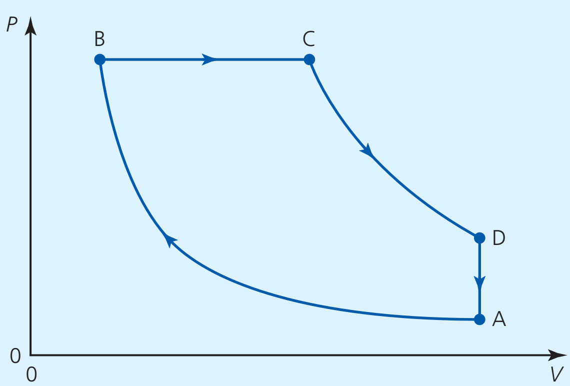


Figure B4.13 Four successive changes to the state of a gas

- **a** Name the processes shown by:
  - **i** BC
  - **ii** DA.
- **b** If CD is an adiabatic process, compare the temperatures at points C and D.
- **c** If AB is an isothermal process, compare the temperatures at points A and B.

15 Copy and complete Table B4.1 by putting 0, +, –, or ± in each box

Table B4.1

		Q	ΔU	W	ΔP	ΔV	ΔT
isothermal	expansion						
	compression						
adiabatic	expansion						
	compression						
isobaric	expansion						
	compression						
isovolumetric	pressure increase						
	pressure decrease						

16 An ideal monatomic gas expands adiabatically from a volume of 3.13×10^{-3} m^3 to 3.97×10^{-3} m^3.

- **a** If the original pressure was 2.60×10^5 Pa, determine the final pressure.
- **b** If the final temperature of the gas was 398 K, what was the starting temperature?

Thermodynamic cycles and *PV* diagrams

SYLLABUS CONTENT

- Cyclic gas processes are used to run heat engines.
- A heat engine can respond to different cycles and is characterized by its efficiency:
  $\eta = \frac{\text{useful work}}{\text{input energy}}$
- The Carnot cycle sets a limit for the efficiency of a heat engine at the temperatures of its heat reservoirs: $\eta_{\text{carnot}} = 1 - \frac{T_c}{T_h}$.

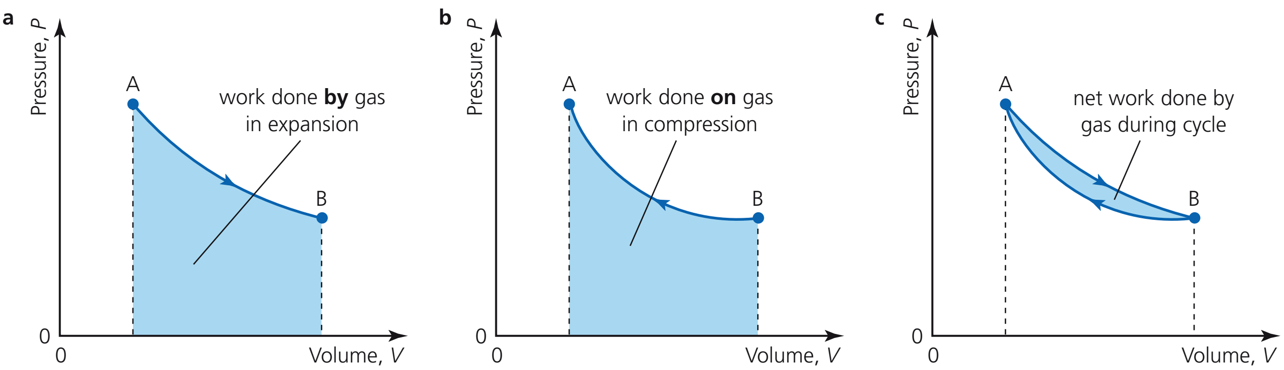
◆ **Working substance** The substance (usually a gas) used in thermodynamic processes to do useful work.

◆ **Cycle (thermodynamic)** A series of thermodynamic processes that return a system to its original state (for example, the Carnot cycle). Usually, the process repeats continuously.

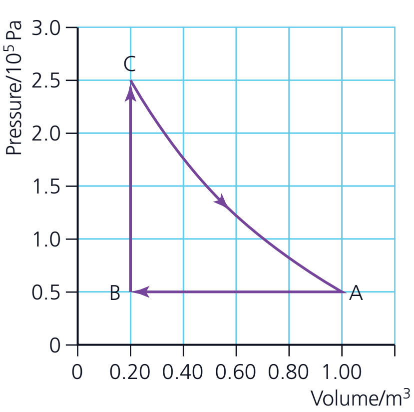
An expanding gas (sometimes called a **working substance**) can do useful work – for example, by making a piston move along a cylinder. However, a gas in a container cannot expand for ever, so any practical device transferring thermal energy to mechanical work must move in **thermodynamic cycles**, involving repeated expansions followed by compressions, followed by expansions and so on. In this section we will discuss some of the physics principles that are fundamental to cyclical processes. It is important to realize that we are not describing details of the actual mechanical processes in any particular type of engine.

The essential process in a *heat engine* is the transfer of thermal energy to produce expansion and do useful mechanical work.

This is represented by path AB in Figure B4.14a. The shaded area under the graph represents the work done *by* the gas during the expansion.



■ **Figure B4.14** Work done in a thermodynamic cycle



■ **Figure B4.15** An idealized complete cycle

In a cyclical process, the gas has to be compressed and returned to its original state. Assume, for the sake of simplicity, that this is represented by the path shown in Figure B4.14b. The area under this graph represents the work done *on* the compressed gas. The difference in areas, shown in Figure B4.14c, is the net useful work done by the gas during one cycle. Of course, if we imagined the impossible situation that, when the gas was compressed, it returned along exactly the same path that it followed during expansion, there would be no net useful work done and no energy wasted.

Figure B4.15 shows a simplified example of an complete cycle. Of course, there are a large number of cycles that could be drawn on a *PV* diagram and, if they are to be considered as the basis of a useful engine, then it is important that they have high efficiencies.

Efficiency of heat engines

We have discussed *efficiency* in Topic A.3, and in the context of this topic it can be restated as:

DB

$$\text{efficiency of a heat engine, } \boldsymbol{\eta} = \frac{\textbf{useful work}}{\textbf{input energy}}$$

Calculating the area under a *PV* diagram is a useful way of determining the efficiency of a thermodynamic cycle.

WORKED EXAMPLE B4.7

Determine the efficiency of the simple cycle shown in Figure B4.14 if the area shown in B4.14a was 130 J and the area in B4.14b was 89 J.

Answer

$$\eta = \frac{\text{useful work done}}{\text{energy input}} = \frac{(130 - 89)}{130} = 0.32\ (32\%)$$

WORKED EXAMPLE B4.8

Figure B4.15 shows one simplified cycle, ABCA, of a gas in a particular heat engine, during which time 1.3×10^5 J of thermal energy flowed into the gas.

- **a** Calculate the work done during the process AB.
- **b** Name the processes AB and BC.
- **c** Estimate the net useful work done by the gas during the cycle.
- **d** What is the approximate efficiency of the engine?

Answer

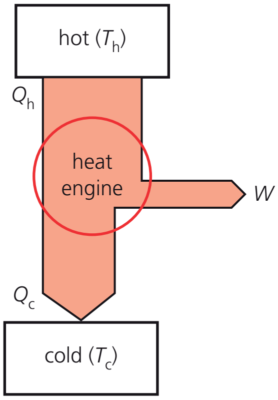
- **a** $W = P\Delta V$ = area under AB = $(0.50 \times 10^5) \times (1.0 - 0.20) = 4.0 \times 10^4$ J (done on the gas)
- **b** AB occurs at constant pressure: isobaric compression. BC occurs at constant volume: isovolumetric temperature increase.
- **c** Net work done by gas = area enclosed in cycle = $(1.0 - 0.20) \times (1.3 - 0.5) \times 10^5$ (estimated from a rectangle having about the same area, as judged by eye)

  $W = 6.4 \times 10^4$ J
- **d** $\text{efficiency} = \dfrac{\text{useful work output}}{\text{total energy input}} = \dfrac{6.4 \times 10^4}{1.3 \times 10^5} \approx 0.50\ (50\%)$

The purpose of a heat engine is to do useful work: to transfer thermal energy into mechanical energy (motion). We know that the opposite process, of converting mechanical energy into thermal energy, occurs around us all the time. For example, rubbing our hands together. Such processes are often 100% 'efficient', and the inevitable dissipation of useful energy in this way is very familiar in the study of dynamics (Topic A.3).

The laws of thermodynamics show us that, in a cyclical process, *it is not possible to convert thermal energy into work with 100% efficiency*. As will see, in practice, 50% is a good output! This is known as the Kelvin form of the second law of thermodynamics (see later).

Heat engines need a flow of thermal energy, which we know always spontaneously flows from a hotter region to a colder region. Some of that energy, but never all of it, is transferred to do useful work. Figure B4.16 represents this in schematic form.



■ **Figure B4.16** Energy flow in a heat engine

◆ **Carnot cycle** The most efficient thermodynamic cycle. An isothermal expansion followed by an adiabatic expansion; the gas then returns to its original state by isothermal and adiabatic compressions.

A temperature difference, $(T_h - T_c)$, is needed between hot and cold reservoirs, so that there is a resulting flow of thermal energy, which operates the engine. Thermal energy Q_h flows out of the hot reservoir and Q_c flows into the cold reservoir. The difference in thermal energy is transferred to doing useful mechanical work, W.

efficiency of a heat engine:

$$\eta = \frac{W}{Q_h}$$

or:

$$\eta = \frac{(Q_h - Q_c)}{Q_h} = 1 - \frac{Q_c}{Q_h}$$

Carnot cycle

The thermodynamic cycle that produces the maximum theoretical efficiency is called the Carnot cycle.

The **Carnot cycle** is an idealized and reversible four-stage process, as shown in Figure B4.17: an isothermal expansion (AB) is followed by an adiabatic expansion (BC); the gas then returns to its original state by isothermal (CD) and adiabatic compressions (DA). Thermal energy is transferred during the two isothermal stages. By definition, thermal energy is not transferred during the adiabatic changes.

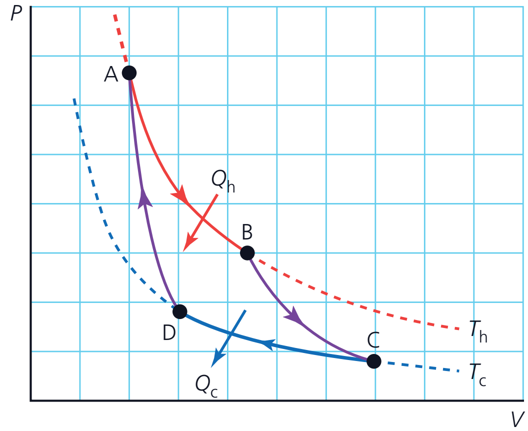
efficiency of a heat engine using the Carnot cycle:

$$\boldsymbol{\eta_{\text{Carnot}} = 1 - \frac{T_c}{T_h}}$$

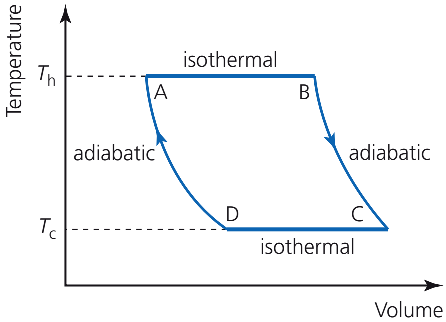
An explanation of the origin of this equation is provided in the *entropy* sub-section.

It is important to realize that this equation represents the most efficient cycle allowed by the laws of physics. This is very different from, for example, processes in which efficiency is limited by energy dissipation due to friction (which also occurs in heat engines).

The Carnot cycle may also be represented on temperature–pressure and temperature–volume graphs (such as shown in Figure B4.18).



■ **Figure B4.17** Carnot cycle



■ **Figure B4.18** Carnot cycle on a volume–temperature graph

WORKED EXAMPLE B4.9

Determine the theoretical thermodynamic efficiencies of Carnot cycles operating between temperatures of:

a 100 °C and 20 °C

b 500 °C and 100 °C

c 150 °C and −150 °C.

Answer

a $\eta_{carnot} = 1 - \frac{T_c}{T_h} = 1 - \left(\frac{293}{373}\right) = 0.21\ (21\%)$

b $\eta_{carnot} = 1 - \frac{T_c}{T_h} = 1 - \left(\frac{373}{773}\right) = 0.52\ (52\%)$

c $\eta_{carnot} = 1 - \frac{T_c}{T_h} = 1 - \left(\frac{123}{423}\right) = 0.71\ (71\%)$

These calculations show us that the efficiency of a Carnot cycle (and others) is limited by the maximum and minimum temperatures that are possible. Extremely high temperatures are technologically difficult to sustain. More importantly, we live in a world which has an average temperature of 288 K.

A waterfall analogy may be helpful. See Figure B4.19. A waterwheel placed in falling water can convert the lost gravitational potential energy into useful kinetic energy of the wheel, but, if for some reason the wheel is placed at a height of $h/2$, a maximum of only 50% of the available energy will be transferred.

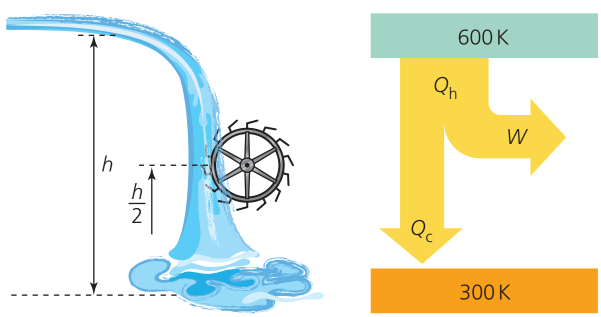


Figure B4.19 Waterfall analogy. Both processes have a maximum efficiency of 50%

If in a Carnot cycle the thermal energy could be transferred between 600 K and 0 K, then the theoretical efficiency would be 100%. But this is not possible, and if we are constrained to an outlet temperature of about 300K, then the maximum possible efficiency is only 50%.

17 a With an outlet temperature of 350 K, calculate the inlet temperature needed to achieve a maximum theoretical efficiency of 45% with a Carnot cycle.

b In practice, why will a higher temperature be needed?

18 Suggest reasons why electrical power stations need cooling towers, such as those seen in Figure B4.20.

Figure B4.20 Cooling towers in Poland

19 Sketch a temperature–pressure diagram for the Carnot cycle.

20 Determine the thermodynamic efficiency of a Carnot cycle in which thermal energy is flowing out of the hot reservoir at a rate of 1.26×10^5 W and into the cold reservoir at a rate of 0.79×10^5 W.

21 Estimate the efficiency of the process shown in Figure B4.16.

22 Figure B4.21 shows the four-stage cycle of a heat engine.

a Which stage is the compression of the gas?

b The temperature at A is 320 K. Calculate the amount of gas in moles.

c Calculate the temperature at point B.

d Estimate the area ABCD. What does it represent?

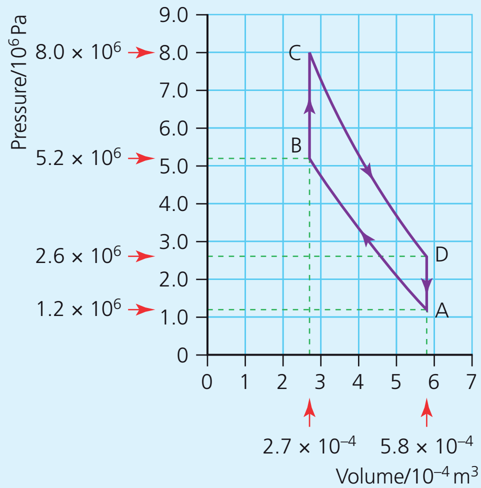


Figure B4.21 The four-stage cycle of a heat engine

23 Using graph paper, make a sketch of the following four consecutive processes in a heat engine. Start your graph at a volume of 20 cm^3 and a pressure of 6.0×10^6 Pa.

a an isobaric expansion increasing the volume by a factor of five

b an adiabatic expansion doubling the volume to a pressure of 1.5×10^6 Pa

c an isovolumetric reduction in pressure to 0.5×10^6 Pa

d an adiabatic return to its original state.

e Mark on your graph where work is done on the gas.

f Estimate the net work done by the gas during the cycle.

◆ **Heat pump** Device which transfers thermal energy from a colder place by doing work.

Heat pumps

A **heat pump** works like a heat engine in reverse, using a work input to enable the transfer of thermal energy from colder to hotter.

Refrigerators and air-conditioners are heat pumps (see Figure B4.22).

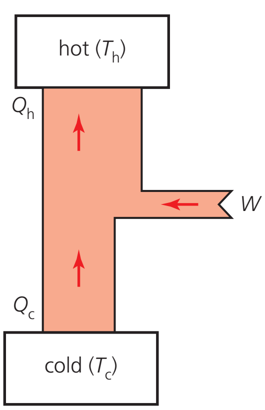


Figure B4.22 Energy transfers in a heat pump

HL ONLY

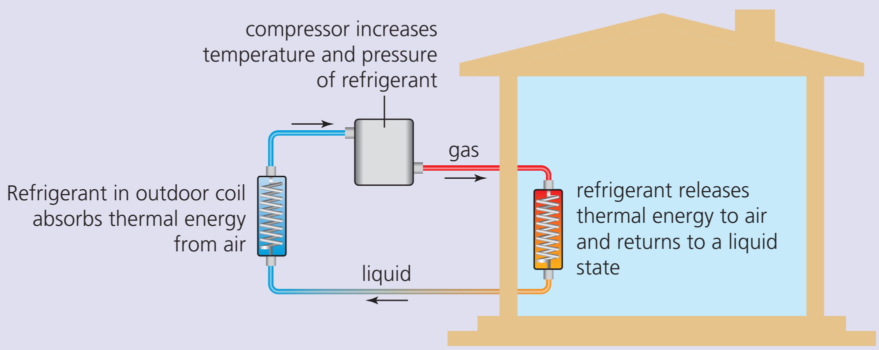
ATL B4A: Research skills, thinking skills

Applying key ideas and facts in new contexts

Using heat pumps for heating homes

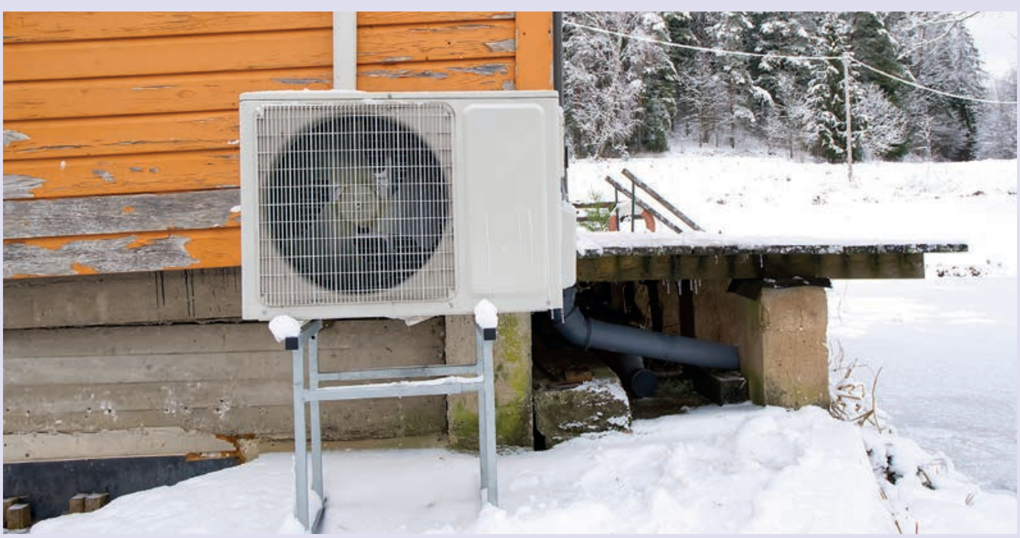
The purpose of refrigerators is to transfer thermal energy from a colder place, inside a refrigerator, to a warmer place (the kitchen). Heat pumps using the same principle can also be used to transfer thermal energy from a colder exterior environment (air or ground) and use it to help to keep the interior of a house warm in winter, although their efficiency may be disappointing in very cold weather, when additional alternative heating may be needed.

Heat pumps use the fact that an evaporating liquid (the 'refrigerant') removes thermal energy (latent heat) from itself and then its surroundings. The thermal energy is released when the gas later is compressed and condenses back to its liquid state. Most heat pumps can be 'reversed' to use as air-conditioners in hot weather. Electrical energy is needed to operate the compressor, as shown in Figure B4.23.



■ **Figure B4.23** Extracting thermal energy from a colder environment

We have seen that the internal energy, *U*, in any substance is proportional to its temperature, (K). It may be surprising to realize that air (for example) at 0 °C has 92% of the internal energy that the same air has at 25 °C.



■ **Figure B4.24** Heat pump outside a home

Why might people choose to install a heat pump to heat their homes, instead of a more conventional fossil-fuel or electric heating system? What could their impact on climate change be?

Using your own research, determine some of the advantages and disadvantages of heat pumps.

Reversible and irreversible changes

SYLLABUS CONTENT

- Processes in real isolated systems are almost always irreversible and consequently the entropy of a real isolated system always increases.

◆ **Irreversible process** A process which cannot be reversed, and in which entropy (see below) always increases. All real macroscopic processes are irreversible.

◆ **Reversible process** A process that can be reversed so that the system and all of its surroundings return to their original states and there is no change in entropy. An impossibility in the macroscopic world.

The required flow of some thermal energy into a cold reservoir in a heat engine is an example of an **irreversible process**, which means that the system *and its surroundings* cannot be returned exactly to their original states.

If the original state of a system *and its surroundings* can be restored exactly, the process is described as **reversible**.

In practice, *all* macroscopic processes can be considered to be irreversible.

Consider watching a swinging pendulum. At first it may seem that the motions keep reversing perfectly. However, if we keep watching, we will notice that the amplitudes decrease, because energy is dissipated out of the system into the surroundings. A video of a pendulum swinging shown in reverse could never be mistaken for the normal behaviour of a pendulum. In fact, almost all events shown in reverse will be obviously just that. Figure B4.25 shows another simple example.

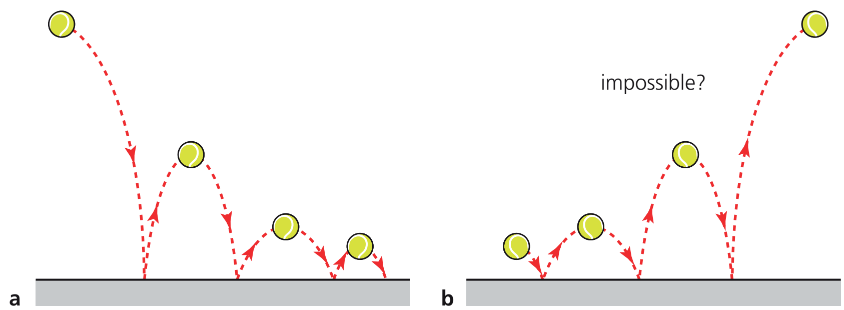


Figure B4.25 Bouncing ball

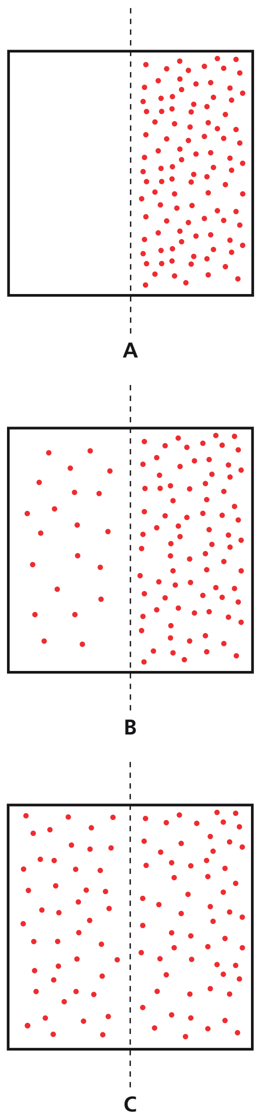
The passage of time seems to be linked to the irreversibility of processes.

So, then we must ask: what is the scientific principle that makes processes irreversible? Why can there be a net energy flow out of a bouncing ball into the ground, but not out of the ground into the ball? The first law of thermodynamics (conservation of energy) does not help us to answer this question.

◆ **Order and disorder (particle)** The way in which particles are arranged, or energy is distributed, can be described in terms of the extent of patterns and similarities.

A bouncing ball (Figure B4.25) has internal energy: its particles have individual *random* potential and kinetic energies. But, in addition, all the particles each have the *same* velocity as the ball as a whole. We need to distinguish between the **ordered** energy of the particles moving in the ball as a whole, and the random **disordered** energies of the same particles. As the ball bounces, more and more ordered energy is transferred to disordered energy. Finally, all the kinetic energies are disordered. The process is irreversible.

Without outside interference, in any process, ordered energy of particles will be *irreversibly* transferred to disordered energy.



■ **Figure B4.26** Gas molecules spreading out in a container

To explain this, we need to consider statistics. Consider an everyday, non-physics example: 20 candy bars are to be distributed among 10 children. The fairest and most ordered way of doing this is to give two bars to each child. However, if the distribution is not controlled, but is entirely random, it is *extremely* unlikely that an even distribution will occur. It is much more likely that, for example, some children will get four and some children will get none. A full statistical analysis can make a reasonably accurate prediction of the overall distribution but cannot predict the number of candy bars given to any particular child.

Continuing the analogy, consider what would happen if we started with an ordered situation in which 10 children each had two candy bars, and then arranged a random re-distribution. The result will be a change to the same overall distribution as described in the previous paragraph.

However, if the distribution was already disordered, further random changes would not affect the overall distribution.

Returning to particle motions, the molecules of a gas move in completely random and uncontrollable ways. What happens to them is simply the most likely outcome. It is theoretically possible for all the randomly moving molecules in a room to go out of an open window at the same time. The only reason that this does not happen is simply that it is statistically *extremely* unlikely.

Consider Figure B4.26, which shows three distributions of the same number of gas molecules in a container. The dotted line represents an imaginary line dividing the container into two equal halves. We can be (almost) sure that A occurred before B, and that B occurred before C. (Probably the gas was released at first into the right-hand side of the container.)

Because it is so unlikely for molecules moving randomly, we simply cannot believe that C occurred before B and A. (In a similar way, we would not believe that if 100 coins were tossed, they could all land ‘heads’ up.) Figure B4.26 only shows about 100 molecules drawn to represent a gas. In even a very small sample of a real gas there will be as many as 10^{19} molecules, turning a highly probable behaviour into a certainty. The simplest way we have of explaining this is that, in the process of going from A to B to C (moving forward in time), the system becomes more disordered.

Similarly, the fact that energy is exchanged randomly between molecules leads to the conclusion that molecular energies will become more and more disordered and spread out as time goes on. Thermal energy will inevitably spread from places where molecules have higher average kinetic energy (hotter) to places with lower average molecular kinetic energies (colder). This is simply random molecular behaviour producing more disorder.

We can be certain that every isolated system of particles cannot spontaneously become more ordered as time progresses. Put simply, this is because everything is made up of particles, and individual atoms and molecules are uncontrollable. Everything that happens occurs because of the random behaviour of individual particles. Of course, we may wish to control and order molecules, for example by turning water into ice, but this would not be an *isolated* system – to impose more order on the water molecules we must remove thermal energy and this will result in even higher molecular disorder in the surroundings.

Two everyday examples may help our understanding: why is it much more likely that a pack of playing cards will be disordered rather than in any particular arrangement? Why is a desk, or a room, much more likely to be untidy rather than tidy? Because, left to the normal course of events, things get disorganized. To produce order from disorder requires intervention and may be difficult, or even impossible. There are a countless number of ways to disorganize a system, but only a relatively few ways to organize it.

Entropy

SYLLABUS CONTENT

- Entropy, S, is a thermodynamic quantity that relates to the degree of disorder of the particles in a system.

◆ **Entropy, S** A measure of the disorder of a thermodynamic system of particles.

◆ **Second law of thermodynamics** The overall entropy of the universe is always increasing. This implies that energy cannot spontaneously transfer from a place at low temperature to a place at high temperature. Or, in the Kelvin version: when extracting energy from a heat reservoir, it is impossible to convert it all into work.

The disorder of a system of particles can be calculated. It is known as the **entropy** of the system.

The concept of entropy, S, numerically expresses the degree of disorder in a system.

Molecular disorder and the concept of entropy are profound and very important ideas. They are relevant everywhere – to every process in every system, to everything that happens anywhere and at any time in the Universe. The principle that molecular disorder is always increasing is neatly summarized by the second law of thermodynamics.

Second law of thermodynamics

SYLLABUS CONTENT

- The second law of thermodynamics refers to the change in entropy of an isolated system and sets constraints on possible physical processes and on the overall evolution of the system.
- The entropy of a non-isolated system can decrease locally, but this is compensated by an equal or greater increase of the entropy of the surroundings.

The **second law of thermodynamics** states that in every process, the total entropy of any isolated system, or the Universe as a whole, always increases.

Figure B4.27 A refrigerator transfers thermal energy from the food and reduces entropy, but where does the energy go?

This is sometimes expressed by the statement 'entropy can never decrease'. But it should be stressed that it is certainly possible to reduce the 'local' entropy of part of a system, but in the process another part of the system will gain *even more* entropy. For example, the growth of a plant, animal or human being reduces the entropy of the molecules that come to be inside the growing body, but there will be an even greater increase in the entropy of all the other molecules in the surroundings that were involved in the chemical and biological processes. Water freezing or the action of a refrigerator, such as seen in Figure B4.27, provide other examples. The internal energy of the contents is reduced as thermal energy is transferred away at the back of the refrigerator. The entropy of the contents is reduced because they are colder, but the entropy of the kitchen is increased by a greater amount because it is hotter.

The statistical analysis of the behaviour of enormous numbers of uncontrollable particles leads to the inescapable conclusion that differences in the macroscopic properties of any system, such as energy, temperature and pressure, must even out over time. This is represented quantitatively by a continuously increasing entropy. This suggests that, eventually, all energy will be spread out, all differences in temperature will be eliminated and entropy will reach a final steady, maximum value. This is often described as the *'heat death' of the Universe.*

HL ONLY

LINKING QUESTION

- Why is there an upper limit on the efficiency of any energy source or engine?

This question links to understandings in Topic A.3.

Disorder and entropy in macroscopic systems

A few examples:

- Gas in a large volume is more disordered / has greater entropy than the same gas at the same temperature in a smaller volume.
- Gas at a higher temperature is more disordered / has greater entropy than the same gas in the same volume at a lower temperature.
- A liquid is more disordered / has greater entropy than a solid of the same material at the same temperature.

Alternative ways of expressing the second law

Consider any two objects at different temperatures placed in thermal contact in an isolated system with no external influences, as shown in Figure B4.28.

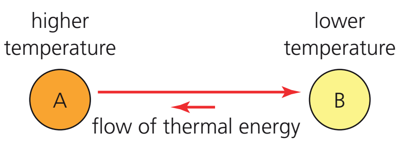


Figure B4.28 Exchanges of energy

Thermal energy can flow from A to B and from B to A, but the net flow of energy is from A to B because the increase in entropy of B will be greater than the decrease in entropy of A (for the same energy transfer). Therefore, the net flow of thermal energy will always be from hotter to colder. This is as we have explained previously in Topic B.1, when discussing particle collisions.. It is an alternative version of the second law of thermodynamics, first expressed by the German physicist *Rudolf Clausius*:

Thermal energy cannot *spontaneously* transfer from a region of lower temperature to a region of higher temperature.

But we *can* use heat engines to transfer thermal energy from colder to hotter by doing external work (heat pumps – see earlier). Thermal energy always flows *spontaneously* from hotter to colder. Insulation can be used to reduce the rate of energy transfer but can never stop it completely. A third version (the *Kelvin* form) of the second law of thermodynamics is expressed in terms of a thermodynamic cycle as:

When extracting energy from a heat reservoir, it is impossible to convert it all into work.

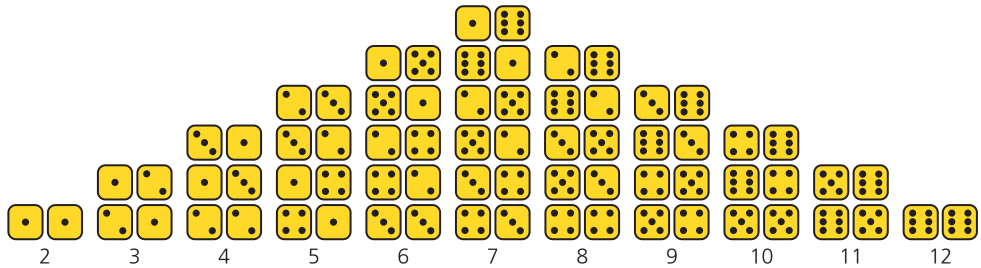
Representing entropy mathematically

SYLLABUS CONTENT

- Entropy can be determined in terms of macroscopic quantities such as thermal energy and temperature, as given by: $\Delta S = \frac{\Delta Q}{\Delta T}$; and also in terms of the properties of individual particles of the system as given by: $S = k_B \ln \Omega$, where k_B is the Boltzmann constant and Ω is the number of possible microstates of the system.

To express the entropy of a system of particles numerically, we need to count the number of ways that the system can be arranged. This is sometimes described as its *multiplicity*. The 'state' of the system can be defined by any property or properties which allows it to be distinguished from other states, for example particle positions or distribution of energies.

First, we will consider an everyday, non-physics example: throwing two six-sided dice and adding the numbers shown to obtain a total. Figure B4.29 shows all the possible combinations. How do we explain that a total of seven is the most likely?



■ **Figure B4.29** Combinations of two dice

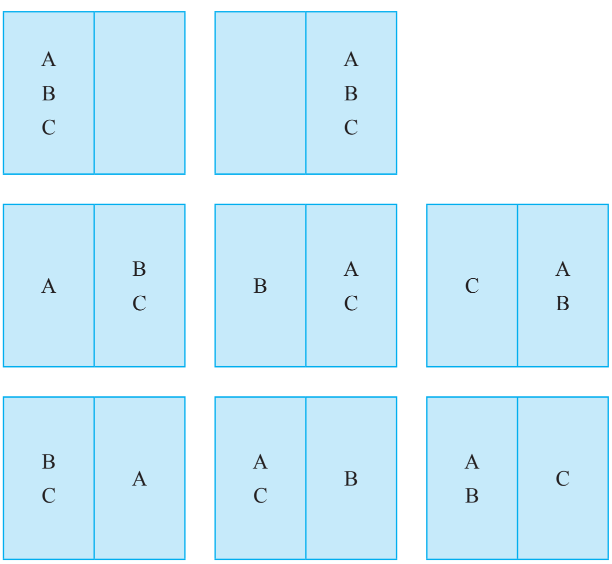
There are 36 possible combinations, with totals ranging from 2 to 12. A total of 2 can only be obtained in one way: 1 + 1 = 2. Similarly, a total of 12 can only be obtained in one way: 6 + 6 = 12. However, a total of 7 can be obtained in six ways: 1 + 6, 2 + 5, 3 + 4, 4 + 3, 5 + 2 and 6 + 1.

Put simply: a total of 7 is most likely because it can be produced in the greatest numbers of ways.

◆ **Microstates** The numerous possible combinations of microscopic properties of a thermodynamic system.

Now consider a very simple system of three freely moving gas particles (A, B and C). We will describe the state of the system in terms of the locations of the particles: to the left, or to the right, of an imaginary line dividing their container in half. We can identify 8 (2^3) possible arrangements, as shown in Figure B4.30. These arrangements may be called **microstates**.

All these eight microstates are equally likely if the particles are moving randomly. However, either of the two 'ordered' microstates, shown in the top line, is less likely than any one of the other six 'disordered' microstates. A system which is ordered will inevitably become disordered.



■ **Figure B4.30** Distributions of three particles

In order to establish the principle, this example has used a small number of particles (3) for simplicity. If larger numbers are used, it becomes clear that returning to an original, ordered arrangement is effectively impossible. Extending the example seen in Figure B4.30: if there were 10 particles, the number of possible disordered microsites would be greater than 1000 (2^{10}). And remember that even small gaseous systems will contain 10^{19} or more particles. (So, the probability of all of the particles being in the left-hand side of a small container is one in two to the power of 10 to the power of 19.).

The greater the number of possible microstates of a system, the greater its disorder and the greater its entropy.

The symbol Ω is used to represent the number of possible microstates of a system (its multiplicity). Clearly, Ω will be a large number. $\ln\Omega$ is more manageable.

Entropy of a system of microscopic particles:

$$S = k_B \ln \Omega \quad \text{(SI Unit: J K}^{-1}\text{)}$$

Nature of science: Science as a shared endeavour

Expressing laws as formulas

The considerable importance of the second law of thermodynamics is undoubted and a broad understanding of the associated concept of entropy is becoming more widespread among the general public. However, entropy is a difficult concept to understand well and, similar to many scientific principles, its true meaning requires the precision of mathematics. Entropy can be determined from the equation $S = k_B \ln \Omega$, but there is no easy application of this equation to everyday life.

The equation was devised by the Austrian physicist Ludwig Boltzmann, who considered the equation to be so important that it was famously carved on his memorial in Vienna. See Figure B4.31. (W was used instead of Ω.)

Figure B4.31 Ludwig Boltzmann

WORKED EXAMPLE B4.10

What is the entropy of a system which has 1×10^{22} microstates?

Answer

$S = k_B \ln (1 \times 10^{22}) = (1.38 \times 10^{-23}) \times 50.7 = 7.0 \times 10^{-22}\,\mathrm{J\,K^{-1}}$

Worked example B4.10 is shown simply to illustrate the principle. In practice, such calculations are often unrealistic. Fortunately, *changes* of entropy, ΔS, in macroscopic situations are much easier to calculate. If thermal energy ΔQ is supplied to a system at a *constant temperature* of T:

DB

change of entropy (using macroscopic quantities): $\boldsymbol{\Delta S = \dfrac{\Delta Q}{T}}$ Unit: $\mathrm{J\,K^{-1}}$

◆ **Entropy change** When an amount of thermal energy, ΔQ, is added to, or removed from, a system at temperature T, the change in entropy, ΔS, can be calculated from the equation $\Delta S = \dfrac{\Delta Q}{T}$.

WORKED EXAMPLE B4.11

Calculate a value for the **entropy change** when 5000 J of thermal energy flows out of a hot cup of coffee at 70 °C into the surrounding room at 25 °C. Assume that the temperatures of the coffee and the room are unchanged. (In practice the coffee will cool by about 5 °C.)

Answer

The entropy of the coffee has decreased by:

$$\Delta S = \frac{\Delta Q}{T} = \frac{-5000}{(273 + 70)} = 14.6\,\mathrm{J\,K^{-1}}$$

The entropy of the room has increased by:

$$\Delta S = \frac{\Delta Q}{T} = \frac{+5000}{(273 + 25)} = 16.8\,\mathrm{J\,K^{-1}}$$

Overall change of entropy in the coffee / room system = 16.8 −14.6 = +2.2 $\mathrm{J\,K^{-1}}$

The fact that the entropy has increased is related to the fact that the thermal energy flowed from a higher temperature to a lower temperature. If the coffee and the room were at the same temperature, there would be no flow of thermal energy and no change of entropy. If we imagined the impossible situation in which thermal energy could flow spontaneously out of the cooler room into the hotter coffee, then entropy of the system would decrease – which never happens.

ATL B4B : Thinking skills

Applying key ideas and facts in new contexts

The true nature of time has always preoccupied scientists. We have seen that the entropy of any system increases with time, and increasing entropy is sometimes described as representing the 'arrow of time'. Is it possible that time is only that: an indication of increasing entropy? Does the second law of thermodynamics imply that time travel is just for science fiction stories, and can never be possible in reality?

LINKING QUESTION

- What are the consequences of the second law of thermodynamics to the Universe as a whole?

HL ONLY

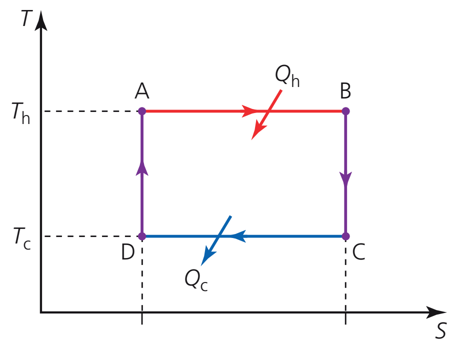
Entropy in the Carnot cycle

The theoretical Carnot cycle is a reversible process, so that there is no overall change of entropy at the end of each cycle. Consider Figure B4.32 and compare it to Figure B4.17.

During the isothermal expansion, AB, thermal energy is supplied, the temperature remains constant but the entropy rises as the volume increases:

$$\Delta S = \frac{\Delta Q_h}{T_h}$$

During the adiabatic expansion, BC, the pressure and temperature decrease, but the entropy remains the same.



■ **Figure B4.32** Temperature–entropy diagram for the Carnot cycle

During the isothermal compression, CD, thermal energy is removed, the temperature remains constant but the entropy falls as the volume decreases:

$$\Delta S = \frac{\Delta Q_c}{T_c}\left(= \frac{\Delta Q_h}{T_h}\right)$$

During the adiabatic compression, DA, the pressure and temperature increase, but the entropy remains the same.

Since:

$$\frac{\Delta Q_c}{T_c} = \frac{\Delta Q_h}{T_h}$$

we can now explain the origin of the equation for the efficiency of a Carnot cycle:

$$\eta = 1 - \frac{\Delta Q_c}{Q_h}$$

But:

$$\frac{Q_c}{Q_h} = \frac{T_c}{T_h}$$

So that:

$$\eta_{carnot} = 1 - \frac{T_C}{T_h}$$

(As highlighted previously.)

24 By discussing what happens to the molecules of the gas, explain the entropy change when a balloon bursts.

25 Coffee, sugar and milk are put in hot water to make a drink. Why it is difficult to reverse the process?

26 Imagine a large container of water, separated into two halves by a removable barrier. Half of the water is at 90 °C and the other half is at 20 °C.
- **a** Explain why, in theory, some of the energy in the water is available to do useful work.
- **b** If the barrier is removed and water from the two halves mixes, what will be the final temperature? Assume that no thermal energy is transferred to the surroundings.
- **c** Why can the particles in the system be described as more disordered after the mixing?
- **d** What has happened to:
  - **i** the total energy of the system
  - **ii** the total entropy of the system?
- **e** Explain why it is now impossible for the system to do any useful work.

27 Consider Figure B4.30.
- **a** How many microstates were there if there were four particles instead of three?
- **b** If the particles were moving randomly, what was the probability that all four were in the right-hand half of the container?

28 Use the equation $S = k_B \ln \Omega$ and Figure B4.30 to determine a value for the change in entropy of a system of three particles when the volume in which they can move is doubled.

29 Calculate the entropy of a system which has 10^{30} microstates.

30 Calculate the total entropy change when 900 J of thermal energy is transferred from a hot reservoir at a constant 550 K to a cold reservoir at a constant 275 K.

31 **a** 3.34×10^4 J are needed to melt some ice at 0 °C. Determine the total change in entropy when the ice melts from thermal energy transferred from the air at a temperature of 25 °C.

b What assumptions did you make?

32 Figure B4.33 represents a Carnot cycle.

a How much thermal energy is transferred between states P and Q?

b How much thermal energy is transferred between R and S?

c How much useful work is done in each cycle?

d Use your answers for **a**, **b** and **c** to determine the efficiency of the process.

e Confirm that the same answer can be calculated from the temperatures involved.

f What physical quantity can be calculated from the area enclosed by the cycle?

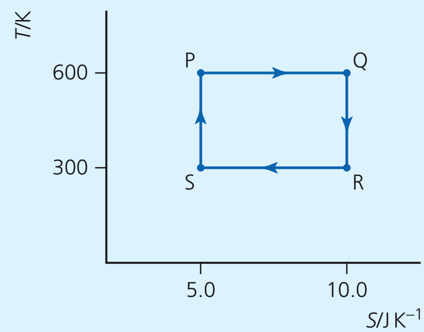


Figure B4.33 A Carnot cycle

33 There are four laws of thermodynamics, but only the first and second are included in this course. They can be summarized in the following humorous form:

- Zeroth: You must play the game.
- *First: You can't win.*
- *Second: You can't break even.*
- Third: You can't quit the game.

What are these comments on the first and second laws suggesting about energy?

Nature of science: Theories

Three versions of the same very important law

The second law of thermodynamics is considered by many physicists to be one of the most important principles in the whole of science. The following quote from Sir Arthur Stanley Eddington (*The Nature of the Physical World*, 1927) may help to convey the importance of this law:

'The law that entropy always increases holds, I think, the supreme position among the laws of nature. If someone points out to you that your pet theory of the Universe is in disagreement with Maxwell's equations – then so much the worse for Maxwell's equations. If it is found to be contradicted by observation – well, these experimentalists do bungle things sometimes. But if your theory is found to be against the second law of thermodynamics, I can give you no hope; there is nothing for it but to collapse in deepest humiliation.'

The second law of thermodynamics can be expressed in different ways depending on the context, and the three versions presented above are slightly different perspectives on the consequences of molecular disorder. Therefore, it is not surprising that, it in the nineteenth century when it was first formulated, the law was the subject of much attention and discussion between prominent scientists in different countries.

B.5 Current and circuits

◆ **Electric charge** Fundamental property of some subatomic particles that makes them experience electric forces when they interact with other charges. Charges can be **positive** or **negative** (SI unit: coulomb, C).

◆ **Opposite charge** Positive and negative charges are described as opposite charges.

◆ **Coulomb, C** The derived SI unit of measurement of electric charge.

◆ **Proton** Subatomic particle with a positive charge ($+1.6 \times 10^{-19}$ C).

◆ **Neutron** Neutral subatomic particle.

◆ **Nucleus** The central part of an atom containing protons and neutrons.

◆ **Electron** Elementary subatomic particle with a negative charge (-1.6×10^{-19} C) present in all atoms and located outside the nucleus.

◆ **Elementary charge, *e*,** 1.6×10^{-19} C

Guiding questions

- How do charged particles flow through materials?
- How are electrical properties of materials quantified?
- What are the consequences of resistance in conductors?

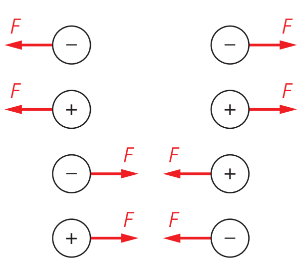
Electric charge

Electric charge is a fundamental property of some subatomic particles, responsible for the forces between them. (Details in Topic D.2.)

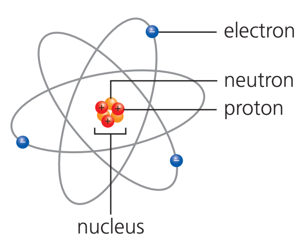
Because there are two kinds of force (attractive and repulsive) as seen in Figure B5.1, we need two kinds of charge to explain the different forces. We call these two kinds of charge, **positive charge** and **negative charge**. The description of charges as 'positive' or 'negative' has no particular significance, other than to suggest that they are two different types of the same thing. Positive and negative charges are often described as **opposite charges**.

Charges of opposite sign attract each other. Charges of same sign repel each other.

Charge is measured in **coulombs**, C. One coulomb is a relatively large amount of charge and we often use microcoulombs ($1\,\mu C = 10^{-6}$ C) and nanocoulombs ($1\,nC = 10^{-9}$ C).



■ **Figure B5.1** Electric forces between similar and opposite charges



■ **Figure B5.2** Simple model of an atom

Figure B5.2 shows a simple visualization of an atom, with three types of subatomic particle. The structure of atoms will be discussed in more detail in Topic E.1.

Protons and **neutrons** are to be found in the small central **nucleus** of the atom. **Electrons** are located in the space around the nucleus.

All protons have a positive charge of $+1.60 \times 10^{-19}$ C and all electrons have a negative charge of -1.60×10^{-19} C.

DB

A charge of magnitude $\mathbf{1.60 \times 10^{-19}\,C}$ is called the **elementary charge**. It is given the symbol *e*.

Since 2019, the elementary charge has been defined to be exactly $-1.602\,176\,634 \times 10^{-19}$ C.

◆ **Neutral** Uncharged, or zero net charge.

◆ **Variable** Quantity that can change during the course of an investigation. Variables can be **continuous** or **discrete**. A variable can be measurable (*quantitative*) or just observable (*qualitative*). A quantity being deliberately changed is called the *independent variable* and the measured, or observed, result of those changes occurs in a *dependent variable*. Usually, all other variables will be kept constant (as far as possible); they are called the *controlled variables*.

◆ **Quantized** Can only exist in certain definite (discrete) numerical values.

There are electric forces between the electrons and the protons in the nucleus. Neutrons do not have any charge; they are **neutral**.

Tool 3: Mathematics

Distinguish between continuous and discrete variables

A **continuous variable** can have, in theory, any value (within the available limits), but a **discrete variable** can only have certain values. An everyday example might be buying eggs: you can buy 1, 2, 6, 10 and so on, but not 1.5 or 3.7 eggs. Values of bank notes are another example. A physical quantity which can only have discrete values is described as being **quantized**.

Figure B5.3 a discrete number of eggs

Any quantity of charge consists of a whole number of charged particles, each $\pm 1.60 \times 10^{-19}$ C. Intermediate values are not possible (with the exception of the sub-nuclear particles *quarks*, but they are not included in this course). For example, it is not possible to have charge with a value of 4.00×10^{-19} C. We describe this by saying that charge is quantized.

Charge is generally given the symbol q. (Q is also sometimes used, but the same symbol is used for thermal energy.)

One coulomb of negative charge is the total charge of 6.24×10^{18} electrons: $\dfrac{1}{1.602 \times 10^{-19}}$

Law of conservation of charge

This is one of the few conservation rules in physics (rules which are always true):

The total charge in an isolated system remains constant.

◆ **Ionization** The process by which an atom or molecule becomes an ion. The required energy is called the ionization energy.

For example, if one or more negatively charged electrons are removed from a neutral atom, this law shows us that the remaining atom must have an equal positive charge. The charged atom is then called an ion and the process is called **ionization**.

TOK

The natural sciences

- Should scientific research be subject to ethical constraints or is the pursuit of all scientific knowledge intrinsically worthwhile?

Increasing knowledge could be life-threatening.

Benjamin Franklin (b 1706) was a famous and influential personality in the USA in the eighteenth century. His experiments with static electricity (see Figure B5.4) certainly endangered lives, but his experiments expanded our scientific knowledge. Travelling into orbit, or to the Moon, are among scientific investigations which could be described similarly.

It can be argued that, if individuals are fully informed and prepared to risk their lives for scientific advancement, then that is entirely their own personal choice. Experiments with animals is another matter, and there is a wide range of well-considered opinions on this matter. Opinions may vary with the type of animal involved and the possible benefits to human society. One notorious demonstration of the effects of poisoning and electricity involved a 'rogue' elephant named Topsy in 1903.

Figure B5.4 Benjamin Franklin famously flew a kite in a lightning storm as part of his investigations into electricity

◆ **Current (electric), I** A flow of electric charge. Equal to the amount of charge passing a point in unit time: $I = \frac{\Delta q}{\Delta t}$.

◆ **Charge carrier** A charged particle which is free to move (mobile).

◆ **Delocalized electrons** Electrons which are not bound to any particular atom or molecule. Sometimes called **'free' electrons**.

Electric currents

SYLLABUS CONTENT

- Direct current (dc), I, as a flow of charged carriers given by: $I = \frac{\Delta q}{\Delta t}$.

Whenever charges flow from place to place we describe it as an **electric current**. We have identified electrons, protons and ions as charged particles, but in an electric current the term **charge carrier** is often used to describe any moving charge. In order for a charged particle to flow as part of an electric current it has to be relatively 'free' to move. We often refer to *mobile* charge carriers.

Some of the electrons in the atoms of metals have enough energy that they are no longer attracted to a particular metal ion. We say that they are **delocalized electrons**, or **free electrons**. This means that electric currents, carried by *free* electrons, can flow through metals better than through most other materials.

In this topic we will only be considering currents in which the charge carriers are electrons. Figure B5.5 represents the electric current of a flow of free electrons through a metal wire. The negatively charged electrons are attracted to the positive terminal of a battery.

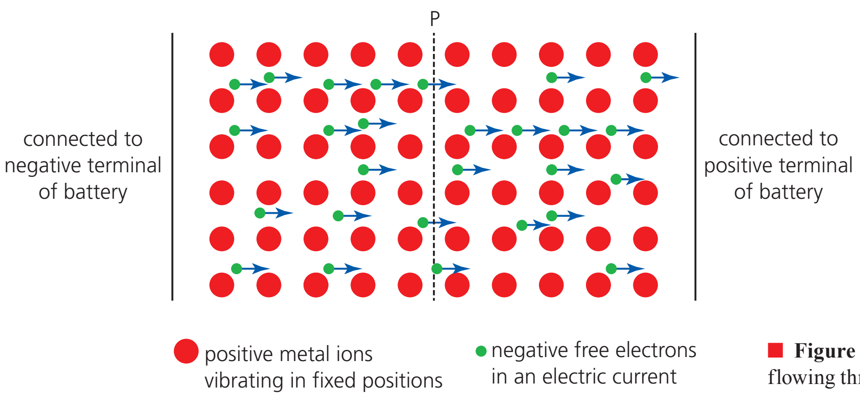


Figure B5.5 Electric current flowing through a metal

When there is no electric current, free electrons normally move around randomly in metals at high speeds, somewhat like molecules in a gas. But, when they form an electric current, a much slower 'drift' speed in the direct of the current is added to the electrons' random movements. This movement cannot be represented in a single diagram such as Figure B5.5.

We define the magnitude of an electric current (given the symbol I) as the amount of charge that passes a point (such as P in Figure B5.5) in unit time:

$$\text{electric current, } \boldsymbol{I} = \frac{\boldsymbol{\Delta q}}{\boldsymbol{\Delta t}}$$

The SI unit for electric current is the **ampere** (amp), A. Milliamps (mA) and microamps (µA) are also in common use.

When a current only flows in one direction it is called a **direct current** (dc). **Alternating currents** (ac) continuously change direction. They are discussed in Topic D.4.

The ampere (amp) is one of the seven base units of the **SI system**. It is defined to be the current in which 1 C of charge (6.24×10^{18} electrons) passes a point in one second. (Before 2019 the amp was defined differently and more obscurely: as the current in two straight parallel wires of infinite length exactly one metre apart in a vacuum, which results in a magnetic force between them of exactly $2 \times 10^{-7}\,\mathrm{N\,m^{-1}}$. This is explained in Topic D.3 and need not be understood here.)

◆ **Ampere, A** SI (fundamental) unit of electric current. $1\,\mathrm{A} = 1\,\mathrm{C\,s^{-1}}$.

◆ **Direct current (dc)** A flow of electric charge that is always in the same direction.

◆ **Alternating current (ac)** A flow of electric charge that changes direction periodically.

◆ **SI system of units** International system of standard units of measurement (from the French 'Système International') which is widely used around the world. It is based on seven fundamental units and the decimal system.

Top tip!

Direct currents are generally more useful than alternating currents, but ac is used for transmitting electrical energy around the world as it is more easily transformed to the high voltages needed to reduce energy dissipation in the wires (discussed briefly in Topic D.3).

◆ **Circuit (electrical)** A complete conducting path that enables an electric current to continuously transfer energy from a voltage source to various **electrical components**.

◆ **Cell (electric)** Device that transfers chemical energy to the energy carried by an electric current.

◆ **Battery** One or more electric cells.

◆ **Terminals (electrical)** Points at which connecting wires are joined to electrical components.

◆ **Conventional current** The direction of flow of a direct current is always shown from the positive terminal of the power source, around the circuit, to the negative terminal. Conventional current is opposite in direction from electron flow.

◆ **Ammeter** Instrument that measures electric current.

◆ **Ideal meters** Meters with no effect on the electrical circuits in which they are used. An **ideal ammeter** has zero resistance, and an **ideal voltmeter** has infinite resistance.

WORKED EXAMPLE B5.1

The current through an LED desk lamp is 50 mA.

a Calculate the amount of charge which flows through the lamp in 1.0 minute.

b How many electrons flow through the lamp every minute?

Answer

a $I = \dfrac{\Delta q}{\Delta t} = 50 \times 10^{-3} = \dfrac{\Delta q}{60}$

$\Delta q = 3.0\,\text{C}$

b $\dfrac{3.0}{1.60 \times 10^{-19}} = 1.9 \times 10^{19}$

Electrical circuits

Circuit diagrams represent the arrangement of components in a circuit.

In order for there to be a continuous flow of current, **electrical components** and wires need to form a complete (closed) loop, called an **electrical circuit**. However, a current cannot flow unless there is a battery (or other electrical power source) included in the circuit, as shown in Figure B5.6, which is drawn in the conventional style.

A single battery is better described as an **electric cell**. When more than one cell is used, the combination is called a **battery**, although in everyday language, one cell is commonly called a battery.

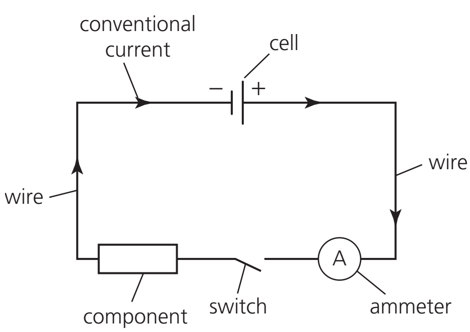


Figure B5.6 Simple electric circuit

The two **terminals** of any battery are labelled positive and negative, and it may be considered that the positive terminal attracts free electrons from the circuit, and the negative terminal repels free electrons. In this way electrons will move around the circuit shown in Figure B5.6 in an anticlockwise direction. However, for historical reasons:

Electric current is *always* shown flowing from positive to negative around any circuit.

This is shown by the arrows in Figure B5.6. This is known as the direction of **conventional current** flow. It was chosen a long time before electrons had been discovered.

Tool 2: Technology

Applying technology to collect data

The magnitudes of electric currents are determined by instruments called **ammeters**, which are connected so that all of the current to be measured flows through them, as shown in Figure B5.6. (This is called being connected *in series*.) Connecting an ammeter in a circuit should not reduce the magnitude of the current it is measuring. Therefore, an **ideal ammeter** will have zero *resistance* to the flow of a current through it. An electronic current sensor responds to the magnetic fields which exist around all currents.

When measuring dc, the current must flow through the ammeter in the correct direction. This is shown by marking the two terminals as positive and negative. Moving through the circuit from the positive terminal on the ammeter, you should arrive at the positive terminal of the battery (or other voltage supply).

Top tip!

In order to understand electrical circuits, it is usually better to consider that the current does *not* begin or end anywhere in particular (at the battery, for example). It is better to consider that the current flows at the same time throughout the circuit, which should be considered as a whole.

(In reality, an electric field moves around the circuit [setting electrons into motion] at a speed close to the speed of light.)

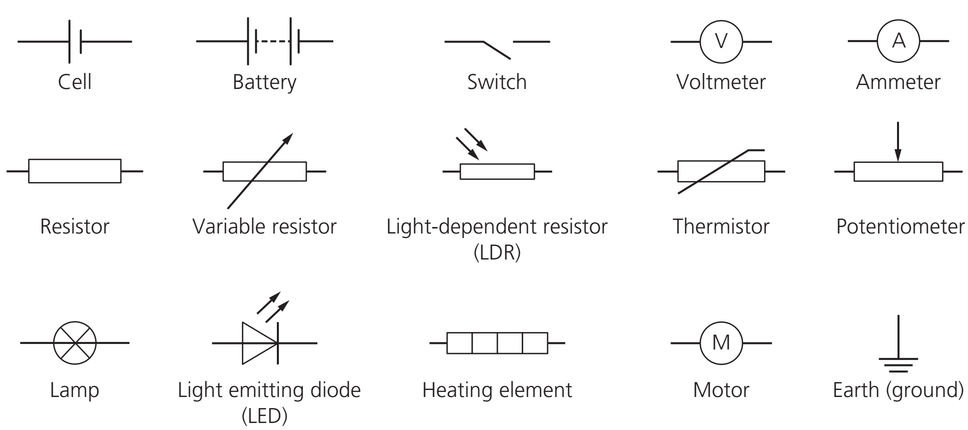


Figure B5.7 Complete list of circuit symbols shown in the IB Physics Data Booklet

Nature of science: Science as a shared endeavour

The use of common symbols and units

The communication of scientific information and ideas between different countries and cultures can be affected by language problems, but this is greatly helped by the use of standard symbols for physical quantities (and units) and for electrical components. Increasingly, English is being used as the international language of science but, naturally, there are many individuals, organizations and countries who prefer to use their own language. Imagine the confusion and risks that could be caused by countries using totally different symbols and languages to represent the circuitry on, for example, a modern international aircraft.

A famous incident occurred in a commercial aircraft flight over Canada in 1983, when the aircraft ran out of fuel because of confusion over the units of volume used for the fuel measurements. Fortunately, there were no serious injuries.

LINKING QUESTION

- In what ways can an electrical circuit be described as a system like the Earth's atmosphere or a heat engine?

This question links to understandings in Topics B.2 and B.4.

1 A carbon atom (carbon-12) contains six protons, six neutrons and six electrons.
  a Sketch this atom in a diagram similar to that seen in Figure B5.2.
  b Calculate the total positive charge in the atom.
  c What is the total negative charge in the atom?
  d We can describe the atom as *neutral*. Explain what this means.
  e If the atom is ionized by the removal of one electron, what is the charge on the ion?

2 2.5×10^{20} electrons flow through a television in one minute.
  a Calculate the total charge which flows in that time.
  b Determine the electric current.

3 Most people find the shape of electricity pylons (that transmit electricity around countries) ugly and would say they spoil the natural beauty of the landscape, for example Figure E5.8.

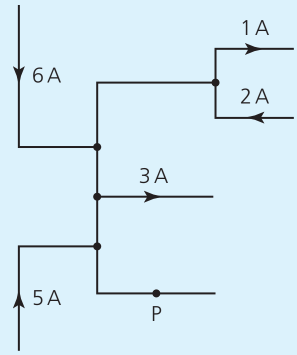
■ **Figure B5.8** Electricity power lines cross some of the most remote countryside in the world

a If the charge flowing through a point on an overhead power line every hour is three million coulombs, what is the current?

b Some architects have suggested that pylons could be designed with interesting and attractive structures that are more sympathetic to the environment, but such pylon designs are more expensive than the usual designs and many people will not be happy to pay more for their electricity. Sketch a pylon design for your country that is attractive, practicable and probably not too expensive.

4 Explain what is meant by a *free* or *delocalized* electron.

5 Determine the current at point P in Figure B5.9. State its direction.



■ **Figure B5.9**

Potential difference / voltage

SYLLABUS CONTENT

▶ The electric potential difference, V, is the work done per unit charge on moving a positive charge between two points along the path of the current: $V = \frac{W}{q}$.

◆ **Voltage** See *potential difference*.

◆ **Volt** Derived unit of measurement of potential difference. $1\,V = 1\,J\,C^{-1}$.

◆ **Potential difference, V** The energy transfered by unit positive charge (1 C) moving between two points. Commonly referred to as voltage.

The term **voltage** is familiar to everyone. Electricity is usually provided to our homes at 110 V or 230 V, and batteries of various lower voltages are used to provide energy to electronic devices.

The voltage of a battery, or other source of electrical energy, is a measure of how much *energy* it can supply to the charge carriers flowing through it. One **volt** means that one joule of energy is transferred by each coulomb of charge moving *between* two specified points.

$1\,V = 1$ joule/coulomb ($J\,C^{-1}$)

Voltage has become the widely used term for the physical quantity that is measured by volts. However, the correct term is **potential difference**, commonly shortened to p.d. The symbol V is used for potential difference, the same letter as used for its unit, V.

The electric potential difference, V, is the work done per unit charge on moving a positive charge between two points along the path of the current: $\boldsymbol{V = \frac{W}{q}}$.

In this topic we are only concerned with free electrons moving in electric circuits, but in Topic D.2, for HL students, we will discuss the movements of both positive and negative charges more generally. When we consider potential differences then, we will need to consider the nature of the charge and the direction of movement more carefully.

We may refer to the potential difference *across* a battery (or other electrical energy source) that is *supplying* energy to a circuit, or to the potential difference across any component(s) that is *using* energy in the circuit.

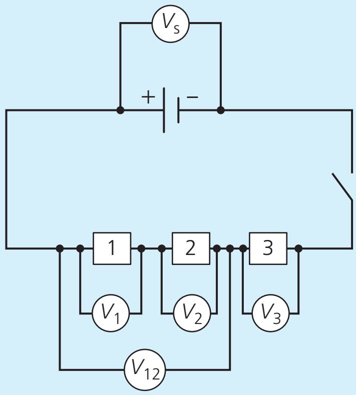
Tool 2: Technology

Applying technology to collect data

Potential differences are measured using voltage sensors (various designs) or **voltmeters**, which are connected as shown by the many voltmeters in Figure B5.10. A voltmeter is always connected *across* (in parallel with) the component(s) it is checking. An **ideal voltmeter** has infinite resistance, so that no current flows through it and it does not affect the p.d. it is measuring. As with ammeters, voltmeters must be connected correctly when using dc circuits: moving through the circuit from the positive terminal on the voltmeter, you should arrive at the positive terminal of the battery (or other voltage supply).

Considering Figure B5.10, if the p.d. supplied to the circuit, V_s is 12 joules to every coulomb (12 V), then when the switch is closed, $V_1+V_2+V_3$ must also equal 12 V (J/C), because the energy transferred into the circuit must be equal to the energy 'used' by the components in the circuit. It is assumed that no energy is transferred in the connecting wires or battery.

Suppose $V_1 = 3$ V and $V_2 = 4$ V, then V_3 must equal $12 - 3 - 4 = 5$ V. V_{12} is measuring the same as $V_1 + V_2$, so it should read 7 V.



■ **Figure B5.10** Connecting voltmeters

◆ **Voltmeter** An instrument used to measure potential difference (voltage).

◆ **Observer effect** When the act of observation, or measurement, changes the phenomenon being observed.

6 Outline what would happen in the circuit shown in Figure B5.6 if the ammeter was replaced by a voltmeter.

7 Explain why replacing the voltmeter, V_s (Figure B5.10) with an ammeter would be a bad idea.

8 Consider the circuit shown in Figure B5.10. If the battery supplied 12 V, the reading on V_1 was 2 V and V_{12} showed a voltage of 5 V, state the readings on the other three voltmeters.

9 400 C of electric charge flow through a lamp in one hour in a country where the electricity mains are supplied at 230 V.
   a Calculate the current in the lamp.
   b How much energy is supplied from the mains to the lamp in this time?

Nature of science: Measurements

The observer effect

We have discussed the use of ammeters and voltmeters to make electrical measurements and referred to the use of 'ideal' meters which will not affect the values of the currents and voltages that they are measuring. However, when taking *any* scientific measurement, we need to consider the possibility that the act of taking the measurement will change what is being measured.

When measuring the pressure in a car tyre, as in Figure B5.11, some of the air in the tyre must flow into the pressure gauge. This will result in a reduction of pressure in the tyre, although it will probably be a very small change.

Many types of thermometer need to absorb or emit thermal energy until they reach thermal equilibrium with their surroundings. This may affect the temperature of the locations that they are measuring. (An infrared thermometer does not have this problem.)

As a non-physics example, doctors are well aware that a patient's blood pressure may well rise when it is being measured because of psychological effects.

■ **Figure B5.11** Measuring the air pressure in a car tyre

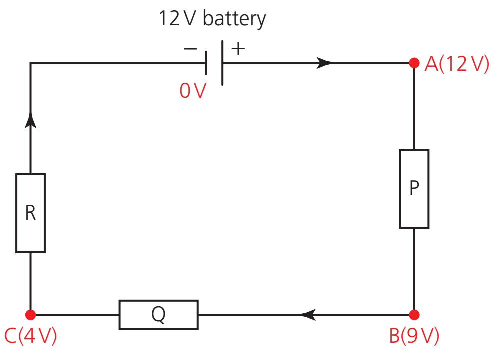
HL ONLY

Using a gravitational analogy may help to explain potential difference

If you are taking the Standard Level examination, you do not need to worry about the deeper meaning of the term *potential difference*; just think of it as *voltage*, and you do not need to remember the following explanation.

A mass will fall towards Earth if it is free to do so. We have explained this by stating that there is a gravitational force acting downwards on it. It then moves from a position of higher gravitational potential energy to a position of lower gravitational potential energy. Alternatively, we can say that any mass may move because of a *difference in gravitational potential* (*energy*).

(The concept of *potential,* defined as potential energy/mass, is introduced for HL students in Topics D.1 and D.2.)



■ **Figure B5.12** Potentials at points around a series circuit

The analogy between gravitational fields and electric fields (introduced in Topics D.1 and D.2) is useful. We will use it now:

A charge will move because of a *difference in electrical potential (energy)* – potential difference, shortened to p.d. – if it is free to do so. It will move from higher electrical potential energy to lower electrical potential energy. A battery provides this p.d. In a similar way, a pump can raise water to a greater height, increasing its gravitational potential energy, but if the water is free to move, it will then fall back down.

Consider the circuit shown in Figure B5.12. We can label the negative side of the battery with a voltage (potential) of 0 V. Since there is a p.d. of 12 V across the battery, the voltage at point A is 12 V. As we move around the circuit, the voltage decreases. For example, it could be 9 V at B and 4 V at C.

A voltmeter connected across P will record a p.d. of (12 – 9) = 3 V.

A voltmeter connected across Q will record a p.d. of (9 – 4) = 5 V.

A voltmeter connected across R will record a p.d. of (4 – 0) = 4 V.

3 + 5 + 4 = 12 V.

Tool 1: Experimental techniques

Recognize and address relevant safety, ethical and environmental issues

Using electrical circuits is an important part of any physics course. For obvious safety reasons, you should work with low voltages. This is done with batteries providing, typically, 9 V or less, but most commonly 1.5 V. However, there is an environmental impact here, as the large number of batteries eventually need to be disposed of. Low voltage (LT) adjustable supplies typically provide 12–15 V. They are connected to the mains supply and need to be checked regularly, including ‘earthing’, for safety.

Electrical resistance

SYLLABUS CONTENT

- Electric resistance and its origin.
- Electrical resistance given by: $R = \frac{V}{I}$.
- Ohm’s law.
- Ohmic and non-ohmic behaviour of electrical conductors, including the heating effect of resistors.
- Properties of electrical conductors and insulators in terms of mobility of charge carriers.

◆ **Conductor (electrical)** A material through which an electric current can flow because it contains significant numbers of mobile charges (usually free electrons).

◆ **Resistance (electrical)** Ratio of potential difference across a conductor to the current flowing through it. $R = \frac{V}{I}$ (SI unit: ohm, Ω).

◆ **Insulator (electrical)** A non-conductor. A material through which a (significant) electric current cannot flow, because it does not contain many charge carriers.

When the same potential difference (voltage) is connected across different electrical components, the currents produced will vary. In general, if the currents are relatively large, the components are described as good **electrical conductors** with low **electrical resistance**. If the currents are small, or negligible, the material is described as a good **electrical insulator**, with a high electrical resistance.

A few substances, most notably silicon and germanium, are described as **semiconductors** because their ability to conduct electricity falls between the obvious conductors and insulators. The electrical behaviour of these materials provides the basis of the electronics industry.

Good electrical conductors, metals, are usually also good thermal conductors. This is because free electrons are important in both processes.

To discuss the origin of electrical resistance we can refer back to Figure B5.5. The greater the number of mobile change carriers (free electrons) in a given volume of the material, the lower we would expect the resistance to be. When the free electrons move through the conductor they will collide / interact with the vibrating metal ions and this is the cause of electrical resistance. We know that particle vibrations in a solid decrease at lower temperatures, so resistance can be reduced by cooling a metal. Conversely, the resistance of a metal will increase if it gets hotter.

Metals are good conductors because they have a large number of mobile charge carriers (free electrons) in unit volume. The vibration of metal ions creates resistance to the flow of electrons.

Electrical resistance, R, is defined quantitatively as follows:

DB

$$\text{electrical resistance} = \frac{\text{p.d.}}{\text{current}} \qquad \boldsymbol{R = \frac{V}{I}}$$

SI unit: **ohm, Ω** (1 Ω = 1 V A^{-1})

◆ **Semiconductor** Material (such as silicon) with a resistivity (explained later in this section) between that of conductors and insulators. Such materials are essential to modern electronics.

◆ **Ohm, Ω** The derived SI unit of electric resistance. 1 Ω = 1 V/1 A.

◆ **Fundamental units** Units of measurement that are not defined as combinations of other units.

◆ **Derived units** Units of measurement that are defined in terms of other units.

Tool 3: Mathematics

Apply and use SI units

There are seven **fundamental (basic) unit**s in the SI system: kilogram, metre, second, ampere, mole, kelvin (and candela, which is not part of this course). The quantities, names and symbols for these fundamental SI units are given in Table B5.1.

They are called ‘fundamental’ because their definitions are not combinations of other units (unlike metres per second, or Newtons, for example). You are not expected to learn the definitions of these units.

Table B5.1 Fundamental SI units used in this course

Quantity	Name	Symbol
length	metre	m
mass	kilogram	kg
time	second	s
electric current	ampere	A
temperature	kelvin	K
amount of substance	mole	mol

Derived units of measurement

All other units in science are combinations of the fundamental units. For example, the unit for volume is m^3 and the unit for speed is m s^{-1}. Combinations of fundamental units are known as **derived units**.

Sometimes derived units are also given their own name (Table B5.2). For example, the unit of force is $kg\,m\,s^{-2}$, but it is usually called the newton, N. All derived units will be introduced and defined when they are needed during the course.

Table B5.2 Some named derived units

Derived unit	Quantity	Combined fundamental units
newton (N)	force	$kg\,m\,s^{-2}$
pascal (Pa)	pressure	$kg\,m^{-1}\,s^{-2}$
hertz (Hz)	frequency	s^{-1}
joule (J)	energy	$kg\,m^2\,s^{-2}$
watt (W)	power	$kg\,m^2\,s^{-3}$
coulomb (C)	charge	A s
volt (V)	potential difference	$kg\,m^2\,s^{-3}\,A^{-1}$
ohm (Ω)	resistance	$kg\,m^2\,s^{-3}\,A^{-2}$

Note that you are expected to write and recognize units using superscript format, such as $m\,s^{-1}$ rather than m/s. The unit for acceleration, for example, should be written $m\,s^{-2}$, not m/s^2.

Express a derived unit in terms of fundamental units

The ohm is a derived unit, but all derived units can be reduced to their fundamental components:

$$V = J\,C^{-1} = Nm \times C^{-1} = kg\,m\,s^{-2} \times m \times A^{-1} \times s^{-1} = kg\,m^2\,s^{-3}A^{-1}$$

$$\Omega = \frac{V}{A} = \frac{kg\,m^2\,s^{-3}\,A^{-1}}{A} = kg\,m^2\,s^{-3}A^{-2}$$

LINKING QUESTIONS

- How does a particle model allow electrical resistance to be explained? (NOS)
- What are the parallels in the models for thermal and electrical conductivity? (NOS)

These questions link to understandings in Topic B.1.

Determining resistance values experimentally

The resistance of a component can be determined as shown in Figure B5.13, recording a pair of values for p.d., V, and current, I.

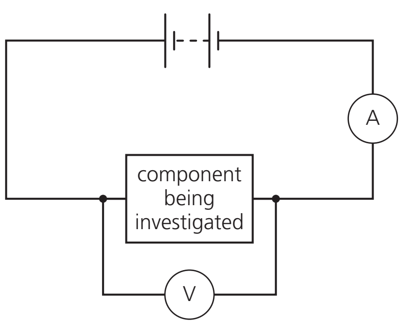


Figure B5.13 Determining the resistance of a component

WORKED EXAMPLE B5.2

The current through an electrical component was 0.78 A when a p.d. of 4.4 V was applied across it. Calculate its resistance.

Answer

$$R = \frac{V}{I} = \frac{4.4}{0.78} = 5.6\,\Omega$$

The value of the resistance obtained should not be assumed to be constant. It may be, but resistances can also change depending on other factors, as discussed later.

I-V characteristics

The resistive properties of a component can be fully investigated by measuring the values of range of different currents produced by varying the p.d. across it. This can be done by using the circuit shown in Figure B5.13 but replacing the battery with a source of *variable* voltage. Alternatively, if only a fixed voltage supply is available, the circuit used in Figure B5.14 can be used. See the worked example later in this topic – in the *Using variable resistors* section (which also explains how using a variable resistor as *potentiometer* is the best method).

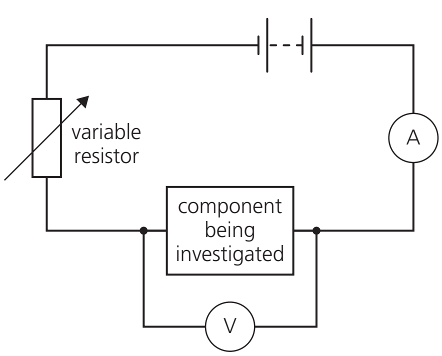


Figure B5.14 Varying current and potential difference using a variable resistor

◆ ***I–V* characteristic:** Graph of current–p.d., representing the basic behaviour of an electrical component.

◆ **Ohm's law** The current in a conductor is proportional to the potential difference across it, provided that the temperature is constant.

◆ **Ohmic (and non-ohmic) behaviour** The electrical behaviour of an ohmic component is described by Ohm's law. A non-ohmic device does not follow Ohm's law.

◆ **Filament lamp** Lamp that emits light from a very hot metal wire. Also called an incandescent lamp.

Common mistake

Resistance cannot be determined from the gradient of a p.d.–current graph, unless the component is ohmic.

The results of these experiments can be shown on current–p.d. graphs. They are called ***I–V* characteristics**.

I–V characteristics are the best way to represent the electrical behaviour of components.

The simplest possible relationship is that the current and p.d. are proportional to each other, as shown in Figure B5.15. This relationship shows that the resistance ($R = V/I$) is constant. The relationship is called **Ohm's law**, and any component that behaves like this is described as being **ohmic**. Metal wires at constant temperatures are ohmic.

Ohm's law: at constant temperature, the current through a metallic conductor is proportional to the p.d. across it: $I \propto V$

Figure B5.15 should be compared to Figure B5.16 which shows the *I–V* characteristic of a metal wire that gets hot. The most common example of this type of **non-ohmic** behaviour is shown by a **filament lamp**. If we took pairs of values for V / I from Figure B5.16, it would show that the resistance ($R = V / I$) increases when the current is greater. This is because, when the current is greater, there are more collisions / interactions between the free electrons and the vibrating metal ions. So that more energy is transferred to the ions, their vibrations increase and the temperature rises.

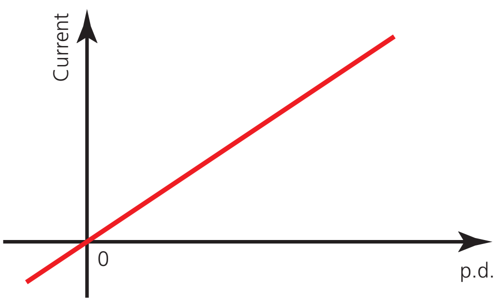


Figure B5.15 Ohm's law for an ohmic resistor

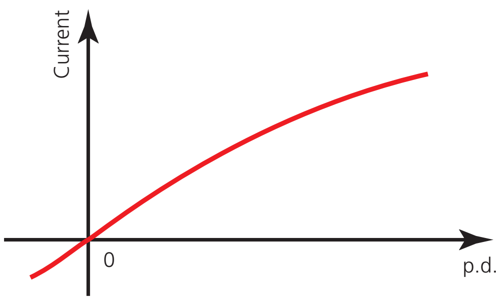


Figure B5.16 A current–p.d. graph for metal wire that gets hot, such as a filament lamp

◆ **Incandescent** Emitting light when very hot.

◆ **Fluorescent lamp** Lamp that produces light by passing electricity through mercury vapour at low pressure.

◆ **Light-emitting diodes (LEDs)** Small semiconducting diodes that emit light of various colours at low voltage and power.

◆ **Diode** An electrical component that only allows current to flow in one direction.

ATL B5A: Research skills, thinking skills

Use search engines and libraries effectively; provide a reasoned argument to support a conclusion

Types of lighting

Figure B5.17 Incandescent filament lamp

Incandescent electrical lamps like that shown in Figure B5.17 were the most popular means of lighting throughout the world for more than one hundred years. Because they need to get very hot to emit light, incandescent lamps are very *inefficient.* They have been replaced by more efficient **fluorescent** lighting and, especially, LED lighting. LED stands for **light emitting diode**.

A **diode** is an electrical component that allows an electric current to pass through it in only one direction. Modern diodes are made from semiconductors, and some of these have the very useful property of emitting light when a current is passing through them. Figure B5.18 shows the *I–V* characteristic of a typical diode.

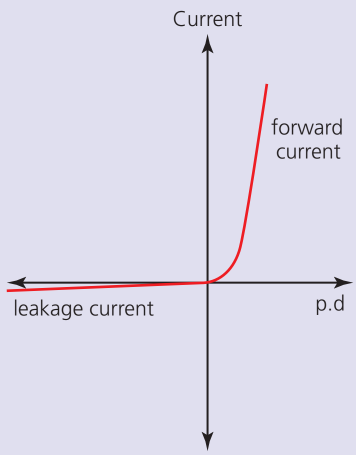
Top tip!

Negative values of current and p.d. represent values in the reverse directions, obtained by turning the battery (or other energy source) the other way around. Simple resistors will behave the same way for currents in either direction, but many other components, diodes for example, need to be connected the 'right' way around.

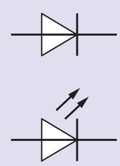
◆ **Peer review** Evaluation of scientific results and reports by other scientists with expertise in the same field of study.

When connected in one way, a 'forward' current is produced, and the diode has very little resistance, as long as the p.d. is greater than a certain minimum value. When connected the other way around, the diode has a large resistance, although a small 'leakage' current is possible. The arrowhead on the circuit symbol shows the 'forward' direction for the current (Figure B5.19).

Individual LEDS are low voltage components and a number of them must be connected *in series* to produce lighting bright enough for a whole room. See Figure B5.20.



■ **Figure B5.18** I–V characteristic for a diode



■ **Figure B5.19** The circuit symbols for a diode and a LED

■ **Figure B5.20** Rings of small LEDs in a ceiling light

Search online to find the relative efficiencies of incandescent, fluorescent and LED lighting. What kinds of lighting do you use at home?

Nature of science: Science as a shared endeavour

Peer review or competition between scientists?

Georg Ohm's famous law was published in Germany in 1827 in the form that the current in a wire, I, is proportional to $(A/L)V$ where A is the cross-sectional area of a uniform metal wire of length L. Two years earlier, in England, Peter Barlow had incorrectly proposed 'Barlow's law' in the form I was proportional to $\sqrt{(A/L)}$, but with no reference to the key concept of voltage, V. It is not unusual for two or more different scientists, or groups of scientists, to be investigating similar areas of science at the same time, often in different countries.

In the worldwide, modern scientific community, with its quick and easy mass communication, new experimental results and theories are quickly subjected to close scrutiny. New ideas are reviewed carefully by other scientists and experts working in the same field, in a process called **peer review**. But 200 years ago, when Ohm was carrying out his research, things were very different. At that time, social factors and the reputation, power and influence of the scientist were sometimes as important in judging new ideas as the value of the work itself. The story of Barlow and Ohm is particularly interesting because in the early stages, the incorrect theory proposed by Barlow was more widely believed.

■ **Figure B5.21** Georg Ohm

◆ **Negligible** Too small to be significant.

◆ **Resistor** A resistance made to have a specific value or range of values.

◆ **Variable resistor** A resistor (usually with three terminals) that can be used to control currents and/or potential differences in a circuit.

◆ **Potentiometer** Variable resistor (with three terminals) used as a potential divider. (see later)

◆ **Light-dependent resistor (LDR)** A resistor, the resistance of which depends on the light intensity incident upon it.

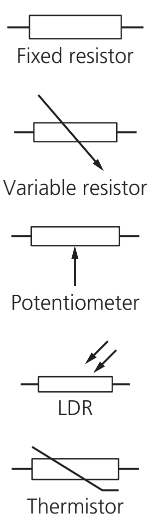
◆ **Thermistor** (negative temperature coefficient) A resistor that has less resistance when its temperature increases. Also called a temperature-dependent resistor

Resistors

All electrical components have resistance, although the resistance of some things, for example connecting wires and ammeters, usually have **negligible** resistance. A component manufactured for its specific resistive properties is called a **resistor**. Resistors are important components in all electrical circuits. Figure B5.22 shows the appearance of a few typical fixed resistors.

■ **Figure B5.22** Fixed value resistors

Apart from resistors of fixed value, later in this topic we will discuss the use of **variable resistors**, **potentiometers**, **light-dependent resistors (LDRs)** and **thermistors**. Their circuit symbols are shown in Figure B5.23.



■ **Figure B5.23** Circuit symbols for resistors: fixed, variable, potentiometer, LDR, thermistor

10 What voltage is needed to make a current of 56 mA pass through a 675 Ω ohmic resistor?

11 a Calculate the operating resistance of a 230 V domestic water heater if the current through it is 8.4 A.

b Explain why you would expect that the resistance would be less when it is first turned on.

12 What current flows through a 3.7 kΩ resistor when there is a p.d. of 4.5 V across it?

13 Calculate the p.d. across a 68.0 Ω resistor if 120 C of charge flows through it in 60 s.

14 Explain what it means if a component is described as non-ohmic.

15 State one reason why the resistance of a component may

a increase as it gets hotter

b decrease as it gets hotter.

16 Sketch an I–V characteristic for a component that has a resistance which decreases as the current through it gets larger.

Electrical resistivity

SYLLABUS CONTENT

▶ resistivity as given by $\rho = \dfrac{RA}{L}$

Investigations into how the resistances of metal wires depends on their dimensions can be undertaken using a circuit similar to that shown in Figure B5.13 or Figure B5.14. The currents should be kept low (or turned on for only short times) to avoid any significant temperature changes in the wires. Three important conclusions can be reached:

1 The resistance of a uniform wire is proportional to its length, *l*

(Assuming the wire has constant thickness and does not change temperature.) As shown by the example in Figure B5.24.

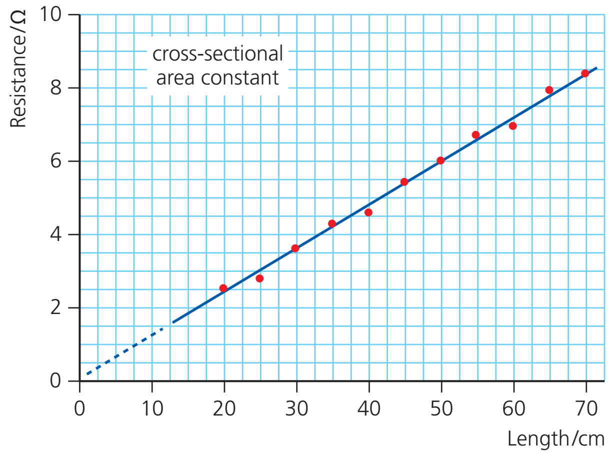


Figure B5.24 Variation of resistance with length of a metal wire

2 The resistance of a uniform wire is inversely proportional to its cross-sectional area, *A*

(Assuming the wire has constant length and does not change temperature.) As shown by the example in Figure B5.25. If the same data is used to plot a resistance–1/area graph, it will produce a straight line through the origin.

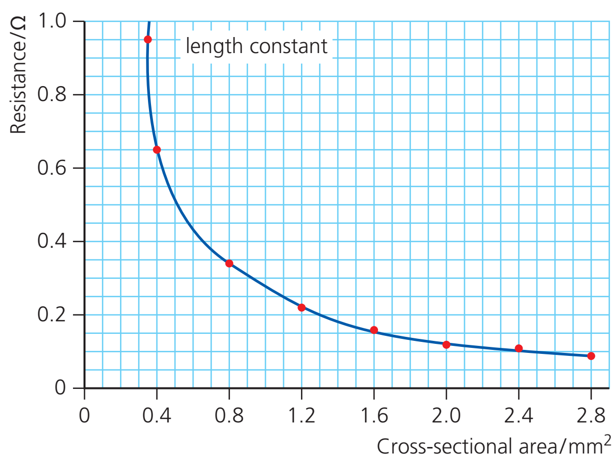


Figure B5.25 Variation of resistance with cross-sectional area of a metal wire

3 The resistance of a uniform wire depends on the metal from which it is made

Combining the last two results, we get:

$$R \propto \frac{l}{A}$$

or:

$$R = \text{constant} \times \frac{l}{A}$$

where the value of the constant depends on the resistive properties of the particular metal.

The constant is called the **resistivity** of the metal, and it is given the symbol ρ.

◆ **Resistivity, ρ** Resistance of a specimen of a material that has a length of 1 m and cross-sectional area of 1 m^2.

resistivity, $\rho = \frac{RA}{L}$ SI unit: Ωm

Common mistake

Note that the SI unit for resistivity is Ω m. Many students think (wrongly) that the unit is ohms per metre, $\Omega\,m^{-1}$.

Clearly, the resistance of a material, like a wire, depends on its shape. For this reason, we cannot refer to the resistance of, for example, aluminium, because we have not specified its shape. The resistivity of a material can be considered as the resistance of a length of one meter, with a cross-sectional area of $1\,m^2$. In other words, the resistance of a cube of the material with sides of 1 m. This is a very large piece of a material, so it is not surprising that the resistivities of good conductors have very low values in SI units. See Table B5.3, which also includes the very high resistivities of some good insulators.

Table B5.3 Resistivities of various substances at 20 °C

Material	**Resistivity/Ω m**
silver	1.6×10^{-8}
copper	1.7×10^{-8}
aluminium	2.8×10^{-8}
iron	1.0×10^{-7}
nichrome (used for electric heaters)	1.1×10^{-6}
carbon (graphite)	3.5×10^{-5}
germanium	4.6×10^{-1}
sea water	$\approx 2 \times 10^{-1}$
silicon	6.4×10^{2}
glass	$\approx 10^{12}$
quartz	$\approx 10^{17}$
Teflon (PTFE)	$\approx 10^{23}$

Variation of resistivity with temperature

As already explained, the resistivity of metals will increase with temperature because of the increased vibrations of the metal ions. The number of free electrons (charge carriers) in metals will not increase significantly unless temperatures are extreme. However, it can be very different with semiconductors and insulators.

The number of charge carriers (per cubic metre) in non-metals can increase significantly with rising temperatures, so that their resistance can decrease considerably as they get hotter. For example, glass is usually described as an insulator, but at 500 °C many types of glass can become good conductors.

WORKED EXAMPLE B5.3

a Determine the resistance of a nichrome wire at 20 °C, if it has a length of 1.96 m and a radius of 0.21 mm.

b Explain why the answer would be different at 100 °C.

Answer

a $\rho = \dfrac{RA}{L}$

Using data from Table B5.3:

$$1.1 \times 10^{-6} = \frac{R \times \left(\pi \times (0.21 \times 10^{-3})^2\right)}{1.96}$$

$R = 16\,\Omega$

b Increased vibrations of metal ions would cause an increase in resistance of the wire.

17 If the wire used to produce the results shown in Figure B5.24 had a resistivity of $4.9 \times 10^{-7}\,\Omega$m, calculate the cross-sectional area of the wire.

18 **a** Use values taken from the graph in Figure B5.25 to show that the resistance was inversely proportional to the area.

b If the wire had a length of 0.56 m, determine its resistivity.

19 Calculate the length of aluminium wire which will have the same resistance as a 1.0 m length of copper wire of the same thickness.

20 The central cable of a high voltage power cable, as seen in Figure B5.26, is made from aluminium and has an effective cross-sectional area of $3.4\,\text{cm}^2$.

a Predict what length of this cable will have a resistance of $1.0\,\Omega$.

b Suggest a reason why the cable is made of thinner strands of aluminium, rather than a single, thicker wire.

Figure B5.26 Power cable

21 A glass rod of length 10 cm and diameter 0.50 cm was heated until its resistance became $10\,\Omega$. Estimate the resistivity of the glass at this high temperature.

22 **a** Calculate the ratio of highest / lowest resistivity as seen in Table B5.3.

b Explain the difference.

Connecting two or more components in the same circuit

SYLLABUS CONTENT

- Combinations of resistors in series and parallel circuits:

  Series circuits

  $I = I_1 = I_2 = \ldots$

  $V = V_1 + V_2 + \ldots$

  $R_s = R_1 + R_2 + \ldots$

  Parallel circuits

  $I = I_1 + I_2 + \ldots$

  $V = V_1 = V_2 = \ldots$

  $\frac{1}{R_p} = \frac{1}{R_1} + \frac{1}{R_2} + \ldots$

◆ **Series connection**
Two or more electrical components connected such that there is only one path for the electrical current, which is the same through all the components.

Components can be connected in series, in parallel, or as a combination of the two. We will use resistors to illustrate the possibilities.

Figure B5.27 shows three different resistors in a **series connection**. All the current follows the same path. Because of the law of conservation of charge, the charge per second (current) flowing into each resistor must be the same as the current flowing out of it and into the next resistor.

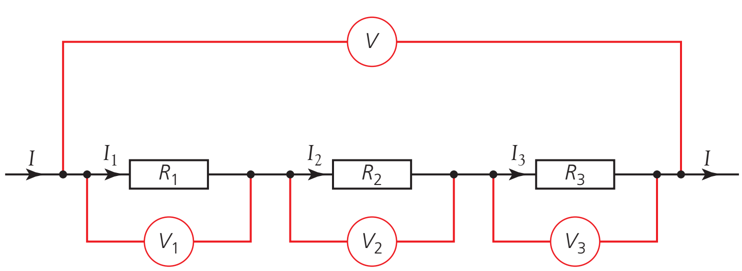


Figure B5.27 Three resistors in series

Currents in series: $I = I_1 = I_2 = \ldots$

The sum of the separate potential differences must equal the potential difference across them all, V, so that:

Potential differences in series: $V = V_1 + V_2 + \ldots$

Using $V = IR$ for the individual resistors, we get $IR_s = IR_1 + IR_2 + IR_3$, so that we can derive an equation for the single resistor, R_s, which has the same resistance as the combination.

Total resistance of resistors in series: $R_s = R_1 + R_2 + \ldots$

◆ **Resistors in series** Resistors connected one after another so that the same current passes through them all. $R_s = R_1 + R_2 + \ldots$

◆ **Parallel connection** Two or more electrical components connected between the same two points, so that they have the same potential difference across them.

Figure B5.28 shows three resistors connected in a **parallel connection**. The current splits into three and they follow different paths between the same two points. Because the resistors are all connected between the same two points, they must all have the same potential difference, V, across them.

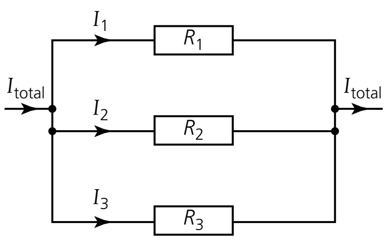


Figure B5.28 Three resistors in parallel

Potential differences in parallel: $V = V_1 = V_2 \ldots$

The law of conservation of charge means that:

Currents in parallel: $I = I_1 + I_2 + \ldots$

Applying $I = \dfrac{V}{R}$ throughout gives:

$$\frac{V}{R_p} = \frac{V}{R_1} + \frac{V}{R_2} + \frac{V}{R_3}$$

Cancelling the V gives us an equation for the single resistor, R_p, which has the same resistance as the combination:

◆ **Resistors in parallel** Resistors connected between the same two points so that they all have the same potential difference across them. $\frac{1}{R_p} = \frac{1}{R_1} + \frac{1}{R_2} + \ldots$

Total resistance of resistors in parallel can be determined from: $\dfrac{1}{R_p} = \dfrac{1}{R_1} + \dfrac{1}{R_2} + \ldots$

All the electrical equipment in our homes is wired in parallel because, in that way, each device is connected to the full supply voltage and can be controlled with a separate switch.

Tool 3: Mathematics

Understand the significance of uncertainties in raw and processed data

How many significant figures are there in 5000 and 8000 (as discussed in Worked example B5.4)?

This example highlights a common problem. Without knowing the context in which this data is presented, we cannot be sure how many significant figures they have. If these are mathematical quantities, then their value is precisely defined and all figures are significant. Equally, if these are measurements of some kind, we would need to know the uncertainty in the measurement to know which were the significant figures. It would be better if the question presented the data in scientific notation!

For numerical data provided in this book, as a rule and for simplicity, we will generally assume that all digits are significant.

WORKED EXAMPLE B5.4

A 5000 Ω resistor and 8000 Ω resistor are connected in series.

a Calculate their combined resistance.

b What is the current through each of them if they are connected to a 4.5 V battery?

c What is the potential difference across the 5000 Ω resistor?

d Repeat these three calculations for the same resistors in parallel with each other.

Answer

a $5000 + 8000 = 13\,000\,\Omega$

b $I = \frac{V}{R} = \frac{4.5}{13\,000} = 3.5 \times 10^{-4}$ A through both resistors (seen on calculator as $3.46... \times 10^{-4}$)

c $V = IR = (3.46 \times 10^{-4}) \times 5000 = 1.7$ V

d $\frac{1}{R} = \frac{1}{5000} + \frac{1}{8000} = \frac{13}{40\,000}$

$R = \frac{40\,000}{13} = 3100\,\Omega$

Both resistors have a p.d. of 4.5 V across them.

Current through 5000 Ω:

$I = \frac{V}{R} = \frac{4.5}{5000} = 9.0 \times 10^{-4}$ A

Current through 8000 Ω:

$I = \frac{V}{R} = \frac{4.5}{8000} = 5.6 \times 10^{-4}$ A

WORKED EXAMPLE B5.5

The lamps shown in Figure B5.29 are all the same.

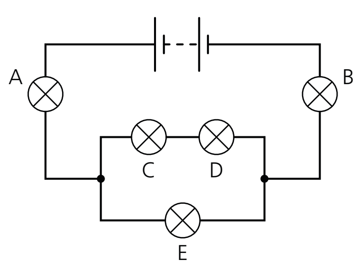


Figure B5.29 Five lamps in circuit

a Compare the brightness of all the lamps (assuming that they are all alight).

b If all the lamps have the same constant resistance of 2.0 Ω, what is the total resistance of the circuit?

Answer

a Lamps A and B will have the same brightness because the same current flows through them both. That same current will be split between lamp E and lamps C and D, so that these three must all be dimmer than lamps A and B.

Lamps C and D will have the same brightness because they are in series with each other.

Lamp E will be brighter than lamps C or D because a higher current will flow through it.

b C and D together will have a resistance of 2.0 + 2.0 = 4.0 Ω.

E in parallel with C / D will have a combined resistance of 1.3 Ω ($1/R = 1/2 + 1/4$).

Total resistance = 2.0 + 2.0 + 1.3 = 5.3 Ω

23 Draw a circuit diagram to represent the following arrangement: two lamps, A and B, are connected to a 12 V battery with a switch such that it can control lamp A only (lamp B is always on). An ammeter is connected so that it can measure the total current in both lamps and a voltmeter measures the p.d. across the battery.

24 Calculate the four possible total resistances that can be made by combining three 10 Ω resistors.

25 Figure B5.30 shows a simple circuit in which the ammeter and voltmeter have been connected in the wrong positions.

a Predict the readings that you would expect to see on the meters. Explain your answer.

b When the positions of the meters were swapped to their correct positions, what readings would you expect to see on the meters?

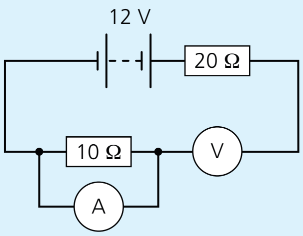


Figure B5.30 Simple circuit with ammeter and voltmeter in wrong positions

26 a Calculate the current that flows through a 12.0 Ω resistor connected to a p.d. of 9.10 V.

b An ideal ammeter would display this current accurately, but what value will an ammeter of resistance 0.31 Ω record?

c Determine the percentage error when using the ammeter in this way.

27 a Determine the currents I_1, I_2 and I_3 in Figure B5.31.

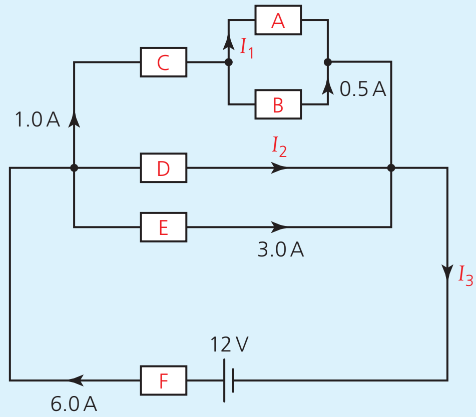


Figure B5.31 Circuit diagram

b If resistor C has a value of 5.0 Ω and resistor E has a value of 2.0 Ω, determine values for the four unknown resistances.

28 Consider the circuit shown in Figure B5.32.

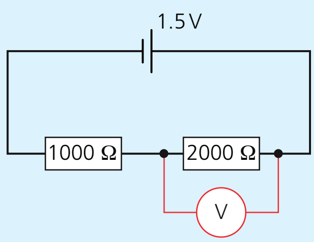


Figure B5.32

a Calculate the current in the circuit before the voltmeter is connected.

b What is the voltage across the 2000 Ω resistor (before the voltmeter is connected)?

c Determine the voltages that will be measured if voltmeters with the following resistances are connected in turn across the 2000 Ω resistor:

i 5000 Ω

ii 50 000 Ω.

Emf and internal resistance

SYLLABUS CONTENT

- Cells provide a source of emf.
- Chemical cells and solar cells as the energy sources in circuits.
- Electric cells are characterized by their emf, ε, and internal resistance, r, as given by: $\varepsilon = I(R + r)$.

◆ **Mains electricity** Electrical energy supplied to homes and businesses by cables from power stations. Also called *utility power.*

◆ **Generator (electrical)** Device that converts kinetic energy into electrical energy.

◆ **Solar cell** Device which converts light and infrared directly into electrical energy. Also called **photovoltaic cell**. A collection of solar cells connected together electrically is commonly called a solar panel.

◆ **Wind generator:** Device that transfers the kinetic energy of wind into electrical energy.

◆ **Dynamo:** A type of electricity generator that produces direct current.

Sources of electrical energy

As individuals, if we wish to use electrical and electronic devices, we need to transfer energy to them using electrical currents. There are several possibilities, including:

- Most homes (but not all) are provided with a p.d. from the '**mains electricity**' generated at electric power stations by various means (most commonly from fossil fuels).
- A home-based **electrical generator** can be used to generate a p.d. from burning a fuel.
- Batteries (also called chemical cells or electric cells) use chemical reactions to provide a p.d. They can be single-use or rechargeable.
- **Solar cells** (also called **photovoltaic cells**) use radiant energy from the Sun to produce a p.d. (see Figure B5.33).
- **Wind generators** use the kinetic energy of moving air to produce a p.d.
- A **dynamo** on a bicycle (for example) can transfer kinetic energy to electrical energy for the lamp.

All of these energy sources have their advantages and disadvantages. These may be assessed by considering:

- convenience of use
- power available
- potential difference available
- whether they contribute to pollution and/or global warming
- whether the energy source is renewable
- whether the power supply is continuous
- whether they supply ac or dc (and the implications of that)
- whether the source is mobile, or fixed to a particular location
- internal resistance of supply (see below)
- cost.

LINKING QUESTION

- What are the advantages of cells as a source of electrical energy?

This question links to understandings in Topic A.3.

ATL B5B: Thinking skills

Asking questions based upon sensible scientific rationale

Imagine that your family have bought a remote house in the countryside for a holiday home, but it has no mains electricity supply. The electricity company can provide a new cable to the house, but the cost would be high. What information would you need to consider in order to decide how to provide energy to the home? You may prefer a renewable energy source, but are they always the best choice?

Emf

The **electromotive force (emf)** of a battery, or any other source of electrical energy, is defined as the total energy transferred in the source per unit charge passing through it.

Electromotive force is given the symbol ε and its unit is the volt, V ($\mathrm{J\,C^{-1}}$). The name electromotive force can cause confusion because it is not a force. For this reason, it is commonly just called 'emf'.

For example, a battery with an emf of 12 V can transfer a total of 12 J to every coulomb of charge that flows through it. However, some of that energy will be transferred within the source itself, as explained in the next section.

■ **Figure B5.33** Solar cells collecting energy in the day to power lamps at night

Internal resistance

◆ **Electromotive force (emf), ε** The total energy transferred in a source of electrical energy per unit charge passing through it.

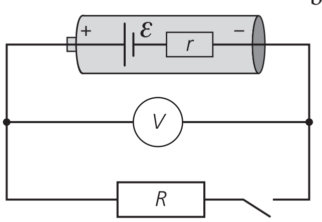
◆ **Internal resistance, r** Sources of electrical energy, for example batteries, are not perfect conductors. The materials inside them have resistance in themselves, which we call internal resistance.

Cells, batteries and other sources of electrical energy are not perfect conductors of electricity. The materials from which they are made all have resistance, called the **internal resistance** of the source, and given the symbol r.

Batteries have resistance, called *internal resistance*, typically less than 1 Ω.

If the internal resistance of a battery is much less than the resistance of the rest of the circuit, its effect can usually be ignored and, as a result, many examination questions refer to batteries or cells of 'negligible internal resistance'. But in other examples, in circuits with high currents and/or low resistance, the internal resistance of an energy source can have a significant effect on the circuit. Internal resistances can vary with temperature and the age of the battery, but in this course, we will assume that they are constant.

Figure B5.34 shows a battery connected to an external resistance, R. In this example, the circuit symbols for the battery and the internal resistance are shown separately to aid understanding, but in practice they are combined and there is no accessible point between them. The voltmeter is assumed to be 'ideal'. To analyse the circuit, we will need to add the internal and external resistances together.



■ **Figure B5.34** A cell in a simple circuit

When the switch is open and there is no current flowing in the circuit ($I = 0$), there will be no voltage across the internal resistance (since $V_r = Ir$) and an ideal voltmeter will display the true value of the emf, ε, of the battery.

When the switch is closed and a current, I, flows, there will be a p.d. of $V_r = Ir$ across the internal resistance, so that the reading on the voltmeter will fall.

Consider a numerical example of Figure B5.34: The battery has an emf of 9.0 V and an internal resistance 0.4 Ω. When the switch is open, the voltmeter will read 9.0 V, but if a current of 2.5 A flows, $V_r = 2.5 \times 0.4 = 1.0$ V. This is commonly called '**lost volts**'. The voltmeter reading will fall from 9.0 V to (9.0 – 1.0) = 8.0 V.

◆ **Lost volts** Term sometimes used to describe the voltage drop (becoming less than the emf) that occurs when a source of electrical energy delivers a current to a circuit. Lost volts (Ir) increase with larger currents.

◆ **Terminal potential difference** The potential difference across the terminals of a battery (or other voltage supply) when it is supplying a current to a circuit (less than the emf).

It should be clear that an ideal voltmeter connected across a battery will only show the emf of the battery if there is no current flowing. At all other times the p.d. will be less, by an amount which depends on the magnitude of the current at that moment.

The p.d. across the terminals of a battery is equal to the p.d. applied to the circuit and is known as the **terminal p.d.**, V_t.

total energy transferred by the cell (per coulomb) = energy transferred to the circuit (per coulomb) + energy transferred inside the cell (per coulomb)

emf of cell = terminal p.d. across circuit + 'lost volts' due to internal resistance of battery

$$\varepsilon = V_t + V_r$$

The same current flows through both resistances, so that using $V = IR$ gives:

$$\varepsilon = IR + Ir$$

or:

$$\boldsymbol{\varepsilon = I(R + r)}$$

WORKED EXAMPLE B5.6

A battery with an emf of 1.5 V and internal resistance 0.82 Ω is connected in a circuit with a 5.6 Ω fixed resistor.

a Calculate the current in the circuit.

b If a (ideal) voltmeter is connected across the terminals of the battery when the current is flowing, what reading will it show?

Answer

a $I = \dfrac{\varepsilon}{R + r} = \dfrac{1.5}{(5.6 + 0.80)} = 0.23\ \text{A}$

b $V = IR = 0.23 \times 5.6 = 1.3\ \text{V}$

Choosing batteries

The emf and the internal resistance of any source of electrical energy are its defining features, although physical size and mass are also important, especially since they will affect the amount of energy that can be stored. 1.5 V is the most common emf produced by an electric cell. They are often connected in series to make a battery which has a greater overall emf, but this will also increase the overall internal resistance. 1.5 V cells connected in parallel will still produce an overall emf of 1.5 V, but the total internal resistance will be reduced. Table B5.4 shows the most common types of batteries. The use of mAh to represent energy storage is explained towards the end of this topic.

Table B5.4 Typical properties of some common batteries

				Energy capacity	
Type of battery		**emf / V**	**Internal resistance/Ω**	**mAh**	**kJ (approx.)**
mobile phone		3.7	0.1	1400	20
AA or AAA		1.5	0.2	1000	5
button / coin battery		3.0	10	100	1
car battery		12	0.01	50 000	2000

Tool 3: Mathematics

Use of units whenever appropriate

The energy stored in a battery *of known voltage* is usually given in terms of the current it could supply for one hour at that voltage. Table B5.4 includes the example of a mobile phone battery, quoted at 1400 mAh. With a voltage of 3.7 V, it can supply a current of 1400 mA (1.4 A) for one hour.

Energy = VIt (as explained below in *electrical power section*) = $3.7 \times 1.4 \times 3600 = 1.9 \times 10^4$ J (≈ 20 kJ)

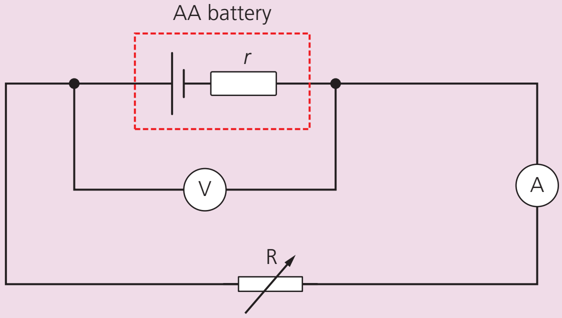
■ **Figure B5.35** E-bike for hire

E-bikes have become very popular. A typical battery might have a mass of about 3 kg and supplies 50 Ah (9 MJ) with an emf of 50 V and internal resistance of 0.1 Ω. This should be enough to travel approximately 30 km on mostly level ground.

Inquiry 2: Collecting and processing data

Designing

A student set up the following circuit to investigate the emf and internal resistance of an AA battery.



■ **Figure B5.36** Investigating internal resistance

The following results (Table B5.5) were obtained when the value of the variable resistance was changed:

■ **Table B5.5** Current and voltage measurements

I/A	V/V
1.0	1.0
0.6	1.2
0.4	1.3
0.2	1.4

The student concluded, by making calculations from these results, that the battery had an emf of 1.5 V and an internal resistance of 0.5 Ω. (Explain how she came to these conclusions.)

The student's teacher said that the conclusions were acceptable, but the experiment was not accurate enough. She suggested that there were ways in which the investigation could be improved, including the drawing of an appropriate graph.

Design an improved investigation using the same arrangement and explain how the data obtained can be used to accurately determine the emf and internal resistance of the battery.

29 Three cells, each of emf 1.5 V and internal resistance 0.20 Ω were connected together. Determine their combined emf and internal resistance if they were connected:

a in series **b** in parallel

30 A very high-resistance voltmeter shows a voltage of 12.5 V when it is connected across the terminals of a battery that is not supplying a current to a circuit. When the battery is connected to a lamp, a current of 2.5 A flows and the reading on the voltmeter falls to 11.8 V.

a State the emf of the battery.

b Calculate the internal resistance of the battery.

c What is the resistance of the lamp?

31 When a battery of emf 4.5 V and internal resistance 1.1 Ω was connected to a resistor, the current was 0.68 A.

a What was the value of the resistor?

b If the resistor was replaced with another of twice the value, predict the new current.

c State an assumption that you made when answering part **b**.

32 If a connecting wire is connected by mistake across a battery or power supply, it is an example of a **short circuit**.

a Calculate the current that flows through a battery of emf 12.0 V and internal resistance 0.25 Ω if it is accidentally 'shorted'.

b Suggest what will happen to the battery.

33 A 1.50 V cell with an internal resistance of 0.100 Ω is connected to a resistance of 500 Ω. Outline why it would be reasonable to assume that the cell has 'negligible internal resistance'. Include a calculation.

◆ **Short circuit** An unwanted (usually) electrical connection that provides a low resistance path for an electric current. It can result in damage to the circuit, unless the circuit is protected by a fuse or circuit breaker.

ATL B5C: Research skills

Use search engines and libraries effectively

Use the internet to research into the latest developments into the design of batteries for electric vehicles, including their weight, charging possibilities and the range of the cars on a fully charged battery. Consider how you will verify the reliability of your sources.

Nature of science: Global impact of science

Scientific responsibility

Battery storage is seen as useful to society despite the well-known environmental issues surrounding their manufacture and disposal. Should scientists be held morally responsible for the long-term consequences of their inventions and discoveries?

Most, if not all, scientific and technological developments have some unwanted, and/or unexpected, side-effects. Most commonly these may involve pollution, the threat of the misuse of new technologies and the implications for an overcrowded world. Or maybe a new technology will result in dramatic changes to how societies function; changes that will have both benefits and disadvantages, many of which will be a matter of opinion.

Should more effort be made to anticipate the possible negative aspects of scientific research and development? Perhaps that is unrealistic, because predicting the future of anything, especially the consequences of as-yet unfinished research, is rarely successful. Of course, there are some extreme areas of research that most people would agree should never be allowed; nuclear or biological weapons, for example. It is important to appreciate that a key feature of much scientific research is that it involves the investigation of the unknown.

If 'society' decides that it wishes to control some area of scientific and technological research because the possible negative consequences are considered to be greater than the possible benefits, who makes those decisions and who monitors and controls the research (especially in this modern international world)? Is it reasonable to expect scientists to be responsible for their own discoveries and inventions? Or will human imagination, the motivation to explore the unknown and the desire (of some people) for fame, power or wealth, inevitably result in every possible new scientific and technological discovery being developed?

◆ **Rheostat** Variable resistance used to control current.

Using variable resistors

SYLLABUS CONTENT

- Resistors can have variable resistance.

Variable resistors come in a wide variety of shapes and sizes. Two examples can be seen in Figure B5.37. **a** is small in size but has a high resistance. Part **b** is larger in size because it is designed to carry much bigger currents, although it has a lower resistance. It is often called a **rheostat**.

Figure B5.37 Two different variable resistors

Many variable resistances have three terminals, one at each end of the resistance and a third movable / sliding contact between them. One possible use has already been seen in Figure B5.14. Only two of the three terminals are being used in that experiment. If the magnitude of the variable resistance is increased, the overall resistance of the circuit increases and the current decreases. At the same time, the p.d. across the variable resistance rises, while the p.d. across the component decreases.

WORKED EXAMPLE B5.7

Consider Figure B5.14. Suppose that the battery supplies a constant 12 V (and has negligible internal resistance), the component has a fixed resistance of 24 Ω and the variable resistor can vary from 0 to 48 Ω.

a Calculate the current in the circuit when the variable resistance is set to:

i 0 **ii** 24 Ω **iii** 48 Ω.

b Determine the p.d. across both components with the same three settings.

Answer

a i $I = \frac{V}{R} = \frac{12}{(24 + 0)} = 0.50\text{ A}$

ii $\frac{12}{(24 + 24)} = 0.25\text{ A}$

ii $\frac{12}{(24 + 48)} = 0.17\text{ A}$

b p.d. across variable resistor = $IR = 0.50 \times 0 = 0$ V. p.d. across component = 12 V

p.d. across variable resistor = $IR = 0.25 \times 24 = 6.0$ V. p.d. across component = 6.0 V

p.d. across variable resistor = $IR = 0.167 \times 48 = 8.0$ V. p.d. across component = 4.0 V

Note that it is not possible to reduce the p.d. across the component any lower than 4.0 V in this arrangement. However, using the same apparatus connected in a different way, it is possible to vary the p.d. across the component from 0 to 12 V. See next section.

Potentiometers

A potentiometer is the name we give to a three terminal variable resistor when its sliding contact is being used to produce a varying p.d.

Used as shown in Figure B5.38, the variable resistor (potentiometer) can provide a p.d., V_{out}, to another part of the circuit, which varies continuously from zero to the full p.d. of the battery, V_{in}. The maximum voltage will be obtained with the sliding contact at the top of the variable resistor (as shown) and the voltage will be zero with the contact at the bottom (when both connections to the other circuit come from the same point).

A potentiometer provides the best way of varying the p.d. across a component in order to investigate its *I*–*V* characteristics (see Figure B5.39).

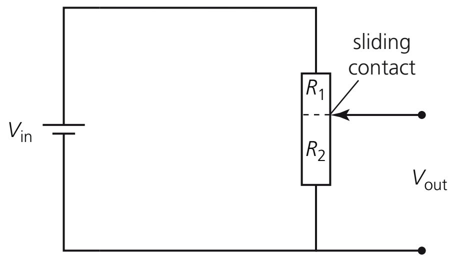


Figure B5.38 A variable resistor used as a potentiometer

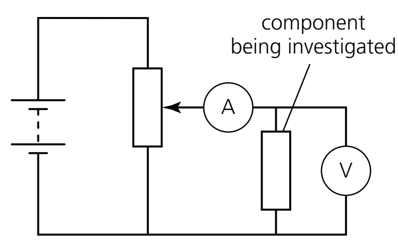


Figure B5.39 A circuit for investigating *I*–*V* characteristics of electrical components

When a potentiometer is connected as the input into a circuit, the value of V_{out} cannot be confirmed without considering the effect of the resistance of the rest of that circuit. Generally, the resistance of the circuit should be much higher than the resistance of the potentiometer, unlike in Worked example B5.8.

WORKED EXAMPLE B5.8

Consider Figure B5.39, using components of the same value as in the last worked example: supply voltage is a constant 12 V, the component being investigated has a resistance of 24 Ω and the variable resistor (potentiometer) can vary from 0 to 48 Ω.

Determine the p.d. across the component when the sliding contact on the potentiometer is:

- **a** at the top
- **b** at the bottom
- **c** at the mid-point.

Answer

- **a** Connected across the full supply p.d.: 12 V
- **b** Not connected across the supply p.d.: 0 V
- **c** Half of the potentiometer (24 Ω) is in parallel with the 24 Ω component, so together they have a resistance of 12 Ω. This means that the other half of the potentiometer (24 Ω) is in series with 12 Ω.

  There will be (24/36) × 12 = 8 V across the top half of the potentiometer and only 4 V across the bottom half, and the component.

Potential-dividing circuits

Figure B5.40 shows an LDR and a thermistor, which are both semi-conducting components. These variable resistors are used as electrical sensors in circuits which control lighting and heating.

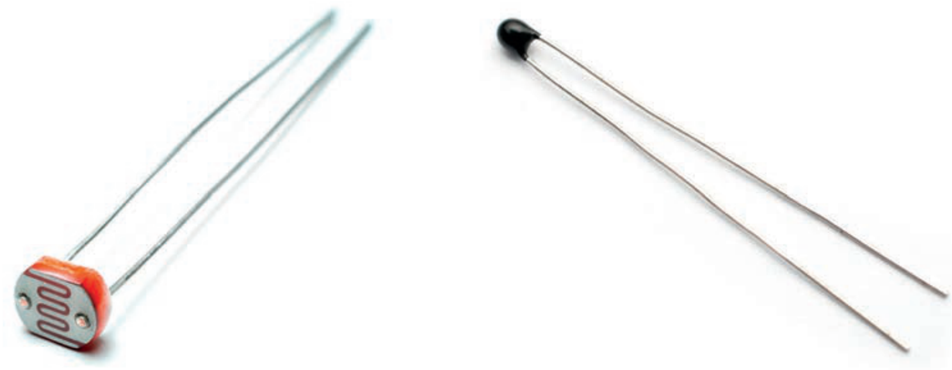


Figure B5.40 LDR (left) and thermistor (right) for sensing changes in light intensity and temperature

◆ **Potential-dividing circuit** Two resistors used in series with a constant potential difference across them. When one resistance is changed, the potential difference across each resistor will change, and this can be used for controlling another part of the circuit.

Figure B5.41 shows how the resistance of an LDR changes with light intensity. (Note that the resistance scale is not linear, it is logarithmic.) In bright sunlight the LDR has a resistance of about $10^2\,\Omega$, but in the dark its resistance rises to about $10^6\,\Omega$ (10 000 × greater). More light energy releases more charge carriers in the LDR.

Sensors are connected in **potential-dividing circuits** as shown (using an LDR) in Figure B5.42.

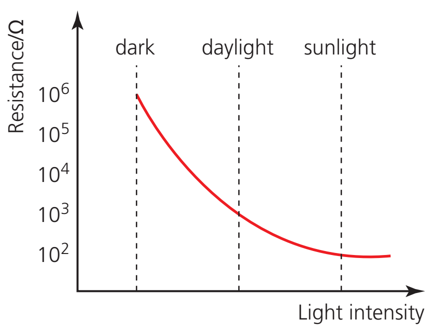


Figure B5.41 Variation of resistance of an LDR with light intensity

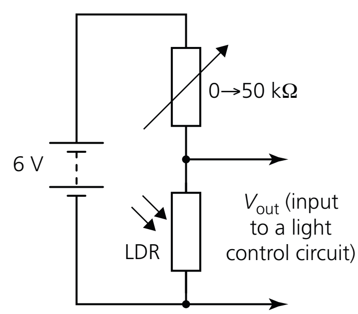


Figure B5.42 An LDR in a potential-dividing circuit

A potential-dividing circuit produces an output voltage which is a fraction of the supply voltage, dependent on the ratio of the values of the resistors it contains.

In this circuit, the p.d. of the supply, V_s, (6.0 V) is always shared / divided between the variable resistance and the LDR in a ratio depending on their resistances, as shown in the following worked example. The p.d. across one of the resistors (in this example, the LDR) is used as the input to another circuit.

Tool 2: Technology

Use sensors

An LDR in a potential-dividing circuit can produce outputs that can be used to roughly *compare* light intensities at different times or locations. However, if more accuracy is required, the output would need to be calibrated to measure the intensity of the light falling on it. This would require comparison with another, reliable light meter, such as shown in Figure B5.43, which usually displays light intensity in the SI unit *lux* (not needed in this course).

Figure B5.43 Light meter

WORKED EXAMPLE B5.9

Consider the circuit shown in Figure B5.42. The light comes on if V_{out} is greater than 5.0 V
If the variable resistance, R_v, is set to 50 kΩ, calculate a value for V_{out}, – which can be the input to the light control circuit – when the room is

a in bright sunlight

b in a dark room. Take values from Figure B5.41.

Answer

a Assuming that the current in both resistors is the same:

$$I = \frac{V_s}{R_{total}} = \frac{V_{LDR}}{R_{LDR}}\left(= \frac{V_v}{R_v}\right)$$

So that:

$$V_{LDR} = V_s \times \left(\frac{R_{LDR}}{R_{total}}\right) = 6.0 \times \frac{100}{100 + (50 \times 10^3)} = 0.12\,\text{V}$$

This p.d. will not be enough to turn the lighting on, as required in a bright environment.

b $$V_{LDR} = V_s \times \left(\frac{R_{LDR}}{R_{total}}\right) = 6.0 \times \frac{1.0 \times 10^6}{(1.0 \times 10^6) + (50 \times 10^3)} = 5.7\,\text{V}$$

This p.d. will turn the lighting on automatically, as required in a dark environment.

The light intensity at which the lighting is turned on, or off, can be adjusted by the choice of value for the variable resistor. In practice the input resistance of the lighting circuit will also have to be considered.

◆ **Thermostat** Component that is used with a heater or cooler to maintain a constant temperature.

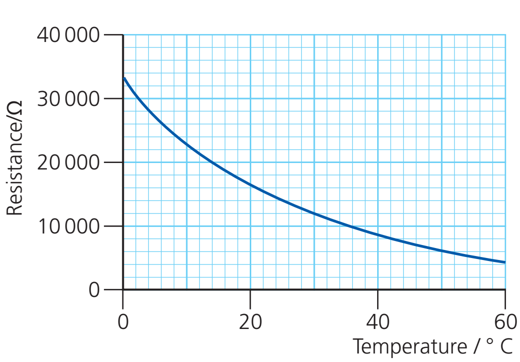


Figure B5.44 Variation of resistance with temperature for a thermistor

Figure B5.44 shows how the resistance of a semi-conducting thermistor changes with temperature. Thermal energy releases more charge carriers in the thermistor, so that its resistance decreases as it gets hotter. This can be used as a sensor (called a **thermostat**) to control temperature. (There is another kind of thermistor which has *greater* resistance if its temperature rises.)

34 a Calculate the potential difference between points A and B in Figure B5.45.

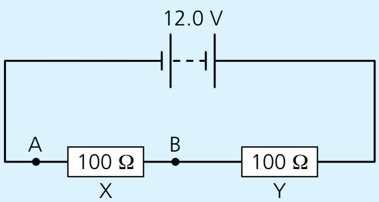


Figure B5.45

b A student wants to connect a lamp which is rated at 2.5 A, 6.0 V. Calculate the working resistance of the lamp.

c The student thinks that the lamp will work normally if he connects it between A and B, in parallel with X. Discuss if the lamp will work as he hopes.

d Another student thinks that the lamp will work if resistor X is removed (while the lamp is still connected between A and B). Calculate the new voltage across the lamp. Will the lamp work normally now?

e Suggest how the lamp could work normally using a 12 V battery.

35 The value of the variable resistor in Figure B5.46 can be changed continuously from 1 kΩ to 10 kΩ.

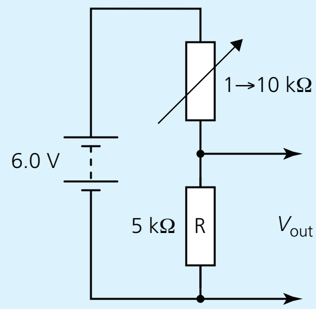


Figure B5.46

a Calculate the maximum and minimum potential differences, V_{out}, that can be obtained across R.

b State (without calculations) how your answer will change if V_{out} is connected across another 5 kΩ.

c Estimate the percentage change in the resistance of the thermistor represented in Figure B5.44, as the temperature changes from 0 °C to 60 °C.

36 a Draw a potential-dividing circuit that could be used to control the temperature of, for example, a refrigerator.

b Make a list of household electrical devices that have thermostats inside them.

37 Describe a laboratory experiment that could be used to obtain results similar to those seen in Figure B5.44. Include a fully annotated diagram.

◆ **Power (electrical)** The rate of dissipation of energy in a resistance.

Electrical power

SYLLABUS CONTENT

▶ Electrical power, P, dissipated by a resistor given by: $P = IV \equiv I^2R = \frac{V^2}{R}$.

If the current through a resistor is, for example, 3 A, then 3 C of charge is passing through it every second. If there is a potential difference across the resistor of 6 V, then 6 J of energy is being transferred by every coulomb of charge (to internal energy). The rate of transfer of energy is 3 × 6 = 18 joules every second (watts).

More generally, we can derive an expression for the **electrical power** dissipated to internal energy in a resistor by considering the definitions of p.d. and current, as follows:

$$\frac{\text{energy transferred}}{\text{time}} = \frac{\text{charge flowing through resistor}}{\text{time}} \times \frac{\text{energy transferred in resistor}}{\text{charge flowing through resistor}}$$

$$\frac{W}{\Delta t} = \frac{\Delta q}{\Delta t} \times \frac{W}{\Delta q}$$

or:

power = current × potential difference $\quad \boldsymbol{P = IV}$

Because $V = IR$, this can be rewritten in two other useful ways:

$$\boldsymbol{P = I^2R = \frac{V^2}{R}}$$

To calculate the total energy transferred in a given time, we know that energy = power × time, so that:

electrical energy = VIt

The heating effect of a current passing through a resistor

Whenever any current passes through any resistance, energy will be transferred to internal energy and then transferred as thermal energy.

This has always been one of the most widespread applications of electricity, including heating water and heating air.

WORKED EXAMPLE B5.10

An electric iron is labelled as 230 V, 1100 W.

a Explain exactly what the label means.

b Calculate the resistance of the heating coil of the iron.

c Explain with a calculation what would happen if the iron was used in a country where the mains voltage was 110 V.

Answer

a The label means that the iron is designed to be used with 230 V and, when correctly connected, it will transfer energy at a rate of 1100 joules every second.

b $P = \frac{V^2}{R}$

$1100 = \frac{230^2}{R}$

$R = 48.1\,\Omega$

c $P = \frac{V^2}{R} = \frac{110^2}{48.1} = 251\text{ W}$

The iron would transfer energy at 0.25 times the intended rate and would not get hot enough to work properly. If an iron designed to work with 110 V was connected to 230 V it would begin to transfer energy at about four times the rate it was designed for; it would overheat and be permanently damaged.

Recall from Topic A.3 that efficiency, η, in terms of energy transfer or power is given by:

$$\eta = \frac{\text{useful work}}{\text{input energy}}$$

So:

$$\eta = \frac{E_{\text{output}}}{E_{\text{input}}} = \frac{P_{\text{output}}}{P_{\text{input}}}$$

◆ **Kilowatt hour, kWh**
The amount of electrical energy transferred by a 1 kW device in 1 hour.

LINKING QUESTION

- How can the heating of an electrical resistor be explained using other areas of physics?

This question links to understandings in Topic B.1.

Tool 3: Mathematics

Use of units whenever appropriate

When we buy a battery or pay for the mains electricity connected to our homes, we are really buying the energy that is transferred by the electric current. In most countries mains electrical energy is sold by the **kilowatt hour**:

1 kWh is the amount of energy transferred by a 1 kW device in one hour, which is the equivalent of 1000 J per second for 3600 s, or **3.6×10^6 J (3.6 MJ)**.

WORKED EXAMPLE B5.11

In a city where the cost of electrical energy is \$ 0.14 for each kWh, predict how much it will cost to operate an air conditioner with an average power of 1500 W for four hours a day for a week.

Answer

Total cost = energy supplied in kWh × 0.14

$= 1.5 \times 4 \times 7 \times 0.14 = \$\ 5.90$

38 A 12 V potential difference is applied across a 240 Ω resistor.

a Calculate:

i the current

ii the power

iii the total energy transferred in 2 minutes.

b What value resistor would have twice the power with the same voltage?

c What p.d. will double the power with the original resistor?

39 A 2.00 kW household water heater has a resistance of 24.3 Ω.

a Calculate the current that flows through it.

b What is the mains voltage?

c Show that the overall efficiency of the heater is approximately 85% if 1.0×10^5 J are transferred to the water every minute.

40 a Determine the rate of production of thermal energy if a current of 100 A flowed through an overhead cable of length 20 km and resistance of 0.001 ohm per metre.

b Comment on your answer.

41 a Calculate the power of a heater that will raise the temperature of a metal block of mass 2.3 kg from 23 °C to 47 °C in 4 minutes (specific heat capacity = 670 J kg^{-1} °C^{-1}).

b Draw a circuit diagram to show how the heater should be connected to a 12 V supply and suitable electrical meters so that the power can be checked.

42 a An electric motor is used to raise a 50 kg mass to a height of 2.5 m in 24 s. The voltage supplied to the motor was 230 V but it was only 8.0% efficient. Determine the current in the motor.

b State two reasons why this process has a low efficiency.

43 a Calculate the value of resistance that would be needed to make a 1.25 kW water heater in a country where the mains voltage is 110 V.

b What current flows through the heater during normal use?

44 If a kWh of electrical energy costs 6.2 rupees, predict how long a 150 W television can be on for a total cost of 100 rupees?

45 In 2022, the best mobile phone batteries were rated at 3.7 V, 5000 mAh.

a Calculate how much energy (J) they store.

b If such a battery is used with an average current of 100 mA, predict how many hours before the battery would be completely discharged.

c Estimate how long completely recharging the battery will take at an average rate of 5 W.

d Suggest why phone manufacturers do not install batteries which store more energy.

ATL B5D: Thinking skills

Applying key ideas and facts in new contexts

Solar panels connected to an outdoor night light are widely available for sale on the internet. Some advertisements make claims which are vague or unrealistic, or only apply to ideal conditions. Suppose you wanted to buy a solar panel with a 25 W LED spotlight which was on for 12 hours every night. Research and compare the products available. Which would you choose, and why?

ATL B5E: Self-management skills

Setting learning goals and adjusting them in response to experience

Reaching the end of Theme B means that you will soon be halfway through the content of the IB Physics syllabus (if you followed the order of this book). Ask yourself some questions and give honest replies.

- Are you doing as well as you hoped at the beginning of the course?
- Could you realistically be doing much better?
- Do you spend enough time studying physics to achieve your goals?
- Are your study methods effective, or are you too easily distracted?
- Are you finding the material interesting? If not, why not?
- Is any of the course content difficult to understand? If so, why?
- Are you using the school's resources effectively?
- Do you use the help of your fellow students and teachers?
- Do you care enough about your physics grade to want to work harder?

Honest answers to these, or similar questions, should lead to setting achievable goals for the rest of the course.

C.1 Simple harmonic motion (SHM)

Guiding questions

- What makes the harmonic oscillator model applicable to a wide range of physical phenomena?
- Why must the defining equation of simple harmonic motion take the form it does?
- How can the energy and motion of an oscillation be analysed both graphically and algebraically?

Oscillations

◆ **Oscillation** Repetitive motion about a fixed point.

An **oscillation** is a regularly repeated backwards-and-forwards movement about the same central point, and along the same path.

Figure C1.1 Oscillations of a humming bird's wings

The importance of the study of oscillations should be apparent from the very wide range of examples, both in physics and more generally. A few scientific examples:

- oscillations of the planets around the sun
- oscillations of the human heart
- oscillations of clocks (mechanical and electronic)
- oscillations of engines and motors
- oscillations of atoms within molecules
- oscillations within musical instruments
- oscillations producing human speech and within the eardrum
- oscillations of waves on water
- oscillations of light and other electromagnetic waves
- oscillations of tides on the ocean
- oscillations of electric currents.

◆ **Simple harmonic motion (SHM)** An idealized oscillation that maintains a constant amplitude and frequency.

◆ **Isochronous** Describing events that take equal times.

Simple harmonic motion (SHM) is a simplified theoretical model representing oscillations. It is the starting point for the study of all oscillations.

In perfect SHM the oscillations always take the same time and there are no resistive forces, so that they continue oscillating indefinitely with no loss of energy and constant amplitude.

This time-keeping property is described as being **isochronous**.

Nature of science: Models

A 'model' in science means a simplified representation of a more complex situation. A model may take many forms, for example: a description in words, a drawing, a theory, an equation, a 3-D construction, a computer program or simulation, and so on.

SHM is a simple model (of oscillations) in a complex world. As the list above suggests, we are surrounded by oscillations, but few, if any, are perfect simple harmonic oscillators. Real oscillators are complex and various. To understand them, we need to first make sure we understand simplified versions.

Such simplifying models are found throughout physics and they are powerful and very useful. They should not be dismissed because of their basic assumptions.

Terms used to describe oscillations and SHM

SYLLABUS CONTENT

- A particle undergoing simple harmonic motion can be described using time period, T, frequency, f, angular frequency, ω, amplitude, equilibrium position, and displacement.
- The time period in terms of frequency of oscillation and angular frequency as given by:

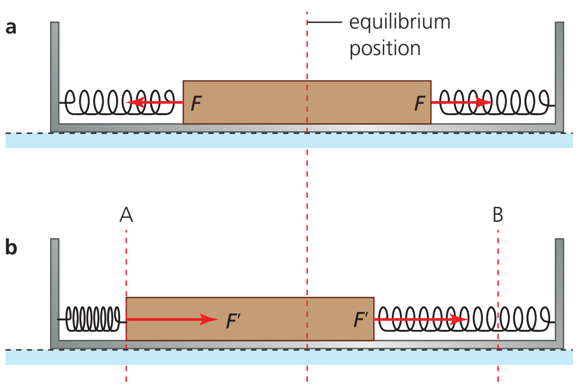
$$T = \frac{1}{f} = \frac{2\pi}{\omega}$$

Although we commonly talk about oscillating masses or oscillating objects, a discussion of perfect SHM often refers to a point mass, or a 'particle'.

◆ Equilibrium position Position in which there is no resultant force acting on an object.

Until a mass is displaced by a resultant force, it will remain in its **equilibrium position**, where the resultant force is zero.

If a mass is then displaced, oscillations may occur if there is always a resultant *restoring force* pulling, and/or pushing, it back towards its equilibrium position. The mass will gain kinetic energy as it moves back where it came from. It will then pass through the equilibrium position (because of its inertia), so that the displacement is then in the opposite direction. Kinetic energy is transferred to some form of potential energy in the system, and when the kinetic energy has reduced to zero, the mass will stop and then reverse its motion. And so on. The motion of a mass between two identical springs on a friction-free surface is a good visualization of this. See Figure C1.2.



■ **Figure C1.2** Oscillations of a mass between two springs

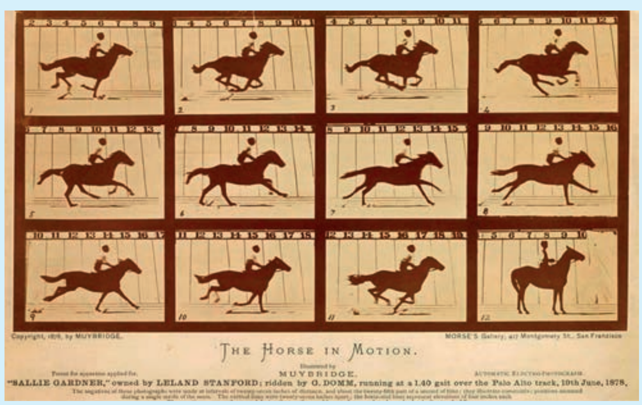
At all times the springs are stretched, but in part **a** of the diagram, the forces from the springs on the mass are equal and opposite. In part **b**, there is a resultant force to the right on the mass from the springs. When released, the mass will accelerate, reach its greatest speed in the centre, and then decelerate until it stops at B. And so on.

Tool 2: Technology

Carry out image analysis and video analysis of motion

Making observations and measurements on fast-moving objects provides many technical difficulties. In the 1870s and 1880s, Eadweard Muybridge was the first to try to analyse motion (most notably of horses) by using quickly taken photographs. See Figure C1.3.

The times between each photograph were known and the position of the horse could be judged from the lines in the background. From this information, the average speed of the horse between each picture could be calculated (and, if relevant, the horse's acceleration). The pictures also revealed previously unconfirmed information about how a horse's legs moved.



■ **Figure C1.3** Famous photographs of a horse in motion (Eadweard Muybridge)

The same principles apply to *video analysis* using a modern video camera, or smart phone app. The motion to be analysed should be just in front of a suitable measurement grid. The video can be replayed in slow motion, or frame-by-frame.

Alternatively, a software program can be used which enables the position of a videoed object to be tracked and measured.

Oscillations similar to that seen in Figure C1.2 can be analysed using a suitable video camera or smart phone. A scale calibrated in millimetres can be placed behind the oscillating mass and a few seconds of action recorded while the mass is oscillating. Replaying the motion in slow motion, or frame-by-frame, can provide information on times and displacements.

Common mistake

The amplitude of an oscillation is *not* the distance between its extreme positions (which equals twice the amplitude). A single movement between the extremes is *not* a complete oscillation (it is half an oscillation).

The displacement, x, of an oscillator is its distance from the equilibrium position in a specified direction. Displacement is a vector quantity. (This term should be familiar from Topic A.1.) The displacement varies continuously, both in size and direction, during an oscillation.

The **amplitude**, x_0, of an oscillation is its maximum displacement while oscillating.

In the idealized example (SHM), where there is no energy dissipation, the amplitude will remain constant. More realistically, each oscillation will have an amplitude which is less than the one before (assuming there is no driving force).

One *oscillation* is completed every time that an oscillating mass returns to a certain position, moving in the same direction.

A compete oscillation is sometimes called a **cycle**. An object which oscillates is called an **oscillator**.

The **time period, T,** of an oscillation is the time taken for one complete oscillation. Unit: s.

The **frequency, f,** of the motion is the number of oscillations in unit time (per second). Unit: hertz, Hz. A frequency of 1 Hz means that there is one oscillation per second.

DB

frequency, $T = \dfrac{1}{f}$

The time period and frequency of an oscillation provide exactly the same information. Typically, we prefer to use the one which is greater than one.

The time period and frequency of most practical oscillations remain constant (they are *isochronous*), even when the amplitude reduces because of energy dissipation.

In physics we often deal with high frequencies, so the following units are in common use: kHz (10^3 Hz), MHz (10^6 Hz), GHz (10^9 Hz).

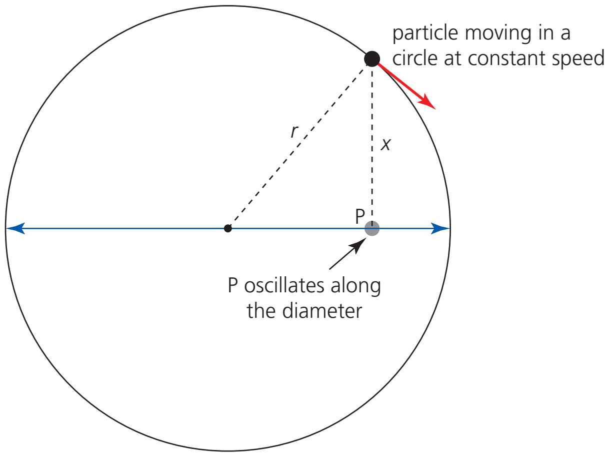
- ◆ **Amplitude** Maximum displacement of an oscillation (or wave).
- ◆ **Cycle (oscillation)** One complete oscillation.
- ◆ **Oscillator** Something which oscillates.
- ◆ **Time period, T** Time taken for one complete oscillation.
- ◆ **Frequency, f** The number of oscillations per unit time, (usually per second). $f = \frac{1}{T}$ (SI unit: hertz, Hz).

Connection between SHM and circular motion

Sometimes we may describe circular motion as an oscillation, and the last two terms described above (frequency and time period) should already be familiar because they were introduced in the circular motion sub-topic in A.2.

There is a close connection between oscillations and circular motion. Indeed, viewed from the side, motion in a circle has exactly the same pattern of movement as a simple oscillation.

Figure C1.4 shows a particle moving in a circle at constant speed. Point P is the projection of the particle's position onto the diameter of the circle.



■ **Figure C1.4** Comparing SHM to motion in a circle

As the particle moves in a circle, point P oscillates backwards and forwards along the diameter with the same frequency as the particle's circular motion, and with an amplitude equal to the radius, r, of the circle.

One complete oscillation of point P can be considered as equivalent to the particle moving through an angle of 2π radians.

The concept of angular velocity, ω, was introduced in Topic A.2 as a key quantity to describe the physical reality of motion in a circle. Because of the close **analogy** between circular motion and SHM (described above), ω is also important in the mathematical description of SHM. However, in the study of SHM, ω is known as **angular frequency**. Unit: rad s^{-1}

◆ **Analogy** Applying knowledge of one subject to another because of some similarities.

◆ **Angular frequency, ω** Similar to angular velocity, but used to represent the frequency of an oscillation in rad s^{-1} (because of the mathematical similarities between uniform circular motion and simple harmonic oscillations.) $\omega = 2\pi f$.

angular frequency, $\boldsymbol{\omega = \frac{2\pi}{T} = 2\pi f}$

Remember that, although the terms period, frequency and angular frequency are all used to describe oscillations, they are just different ways of representing exactly the same information.

TOK

Knowledge and the knower

- How do we acquire knowledge?
- How do our expectations and assumptions have an impact on how we perceive things?

Analogies, correlations and causal relationships

An *analogy* is made when we compare an understanding, process or phenomenon in one area of knowledge to another seemingly unrelated area, and we see similarities. This might enable us to understand some deeper, underlying process that causes both the phenomena – or maybe an analogy just makes it easier to understand what is going on.

If an analogy proves to be useful, but has little, or no other validity, does that justify its use, and does it increase our knowledge of the system to which it applied? What is the difference between a correlation, a causal link and an analogy?

Consider, for example, applying the mathematics of oscillations to variations in animal populations, or to economic cycles.

WORKED EXAMPLE C1.1

Consider the oscillator shown in Figure C1.2. The oscillating mass has a length of 12 cm and the distance between A and B is 18 cm.

a Calculate the amplitude of the oscillation.

b Determine the displacement of the oscillator when its end is in position B.

c State when the mass has its greatest kinetic energy.

d If the period of the oscillator was 1.5 s, calculate its

  i frequency

  ii angular frequency.

Answer

a $\frac{(18-12)}{2} = 3$ cm

b 3 cm to the right

c When it passes through its equilibrium position.

d **i** $f = \frac{1}{T} = \frac{1.0}{1.5} = 0.67$ Hz

  ii $\omega = 2\pi f = 2 \times \pi \times 0.67 = 4.2$ rad s^{-1}

1 The central processing unit of a lap-top computer operates at 3.2×10^9 cycles per second.

a Express this frequency in

i megahertz ii gigahertz.

b Calculate the time period of each cycle.

2 When a guitar string was plucked (once) it oscillated with a frequency of 196 Hz.

a Determine the *angular* frequency of this oscillation.

b Suggest how you would expect the frequency and amplitude to change in the next few seconds. Explain your answer

3 a Show that the angular frequency of the Earth spinning on its axis is approximately 7×10^{-5} rad s^{-1}.

b Determine the total angle (rad) through which it will rotate in 100 hours.

4 A car engine was measured to have 3755 rpm (revolutions per minute). Calculate its angular frequency.

5 Using slow-motion video replay, the angular frequency of the oscillations of a humming bird's wings was found to be 272 rad s^{-1}.
Determine how many times it beat its wings in one minute.

Two commonly investigated oscillators

SYLLABUS CONTENT

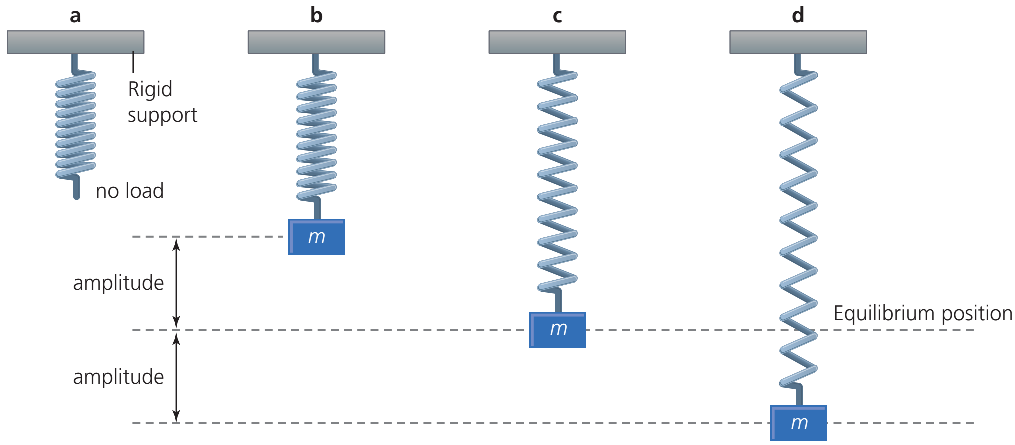
- The time period of a mass–spring system as given by:
$$T = 2\pi\sqrt{\frac{m}{k}}$$
- The time period of a simple pendulum as given by:
$$T = 2\pi\sqrt{\frac{l}{g}}$$

A mass–spring system and a simple pendulum are important for a number of good reasons. These include:

- The proportional relationships between force and displacement are easily understood.
- Their periods of oscillation have a convenient time for measurement in a school laboratory.
- They are good approximations to simple harmonic motion.
- Energy is dissipated slowly, so that the oscillations continue for a long enough time that allows for accurate measurements to be completed.
- They act as starting points for understanding many similar, but more complex, oscillators.

Mass-spring system

We have already briefly discussed a mass oscillating horizontally between two springs (Figure C1.2). A more common arrangement is shown in Figure C1.5. Of course, in this arrangement, the force of gravity (the weight of the mass) also acts vertically, but it does not affect the results because it is constant. The system would have the same time period if it was moved to a location where the gravitational field strength was different.

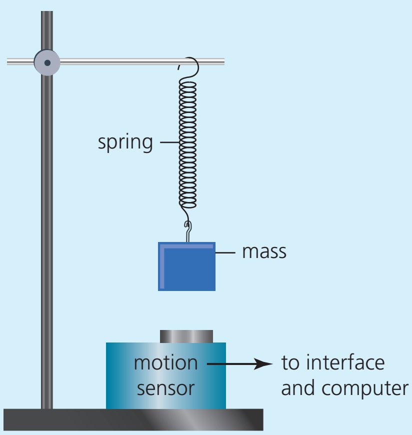


■ **Figure C1.5** Different positions of a mass, *m*, oscillating on a spring

Tool 2: Technology

Use sensors

Position sensors are useful in many aspects of the study of motion, including mechanical oscillations. Figure C1.6 shows how digital data can be collected, which will be processed later by the computer.



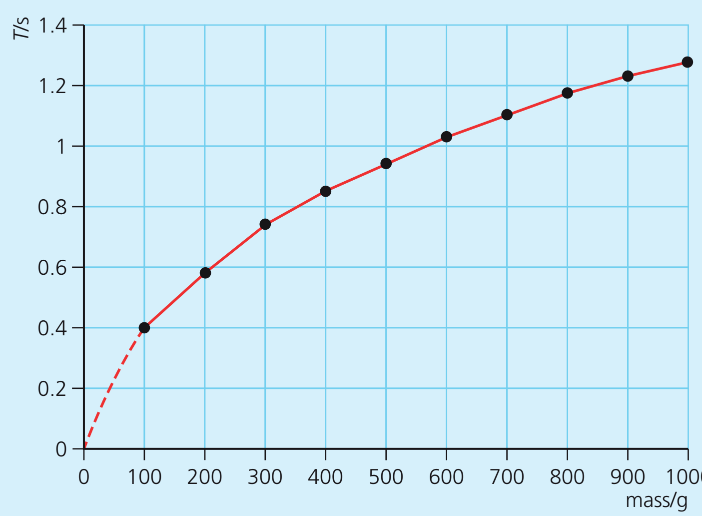
■ **Figure C1.6** Using a sensor to investigate oscillations of a mass on a spring.

We will assume that the spring is never overstretched and that it obeys Hooke's law (see Topic A.2):

$$F_H = -kx$$

Or, in words: the size of the restoring force acting on the mass = spring constant × displacement from its equilibrium position. The spring constant is a measure of the spring's stiffness (= force / deformation).

Laboratory investigations of the time periods of a mass on a spring are straightforward, especially those involving using different masses on the same spring. For typical results see Figure C1.6. Different springs, or combinations of springs, used with the same mass, can be used to investigate the effect of the spring constant, *k*, on the period.



■ **Figure C1.7** Variation of time period, *T*, with mass, *m*, on a spring

Tool 3: Mathematics

Linearize graphs

You may have learnt how to linearize graphs in Topic B.1. The graph seen in Figure C1.6 appears that it might represent the relationship $T^2 \propto m$. Check this by using information from the graph to draw a T^2–m graph (uncertainties in T are ± 0.05s). It should produce a straight line through the origin. Use the graph and the relationship shown below to determine a value for the spring constant, k.

The exact relationship for the SHM of a mass on a spring is as follows:

Time period of a mass-spring system:

$$\boldsymbol{T = 2\pi\sqrt{\frac{m}{k}}}$$

Since $\omega = \frac{2\pi}{T}$, this equation can also be written as: $\omega^2 = \frac{k}{m}$.

This equation applies perfectly only to a point mass acted upon by a separate simple spring system that has a well-defined stiffness (k) and no (significant) mass. There is an extremely large number of other oscillating mechanical systems that have similarities to this simple model but are much more complex. For example, oscillations in buildings and bridges. Figure C1.8 shows a simpler example, in which the mass is spread uniformly along the oscillating system, a ruler.

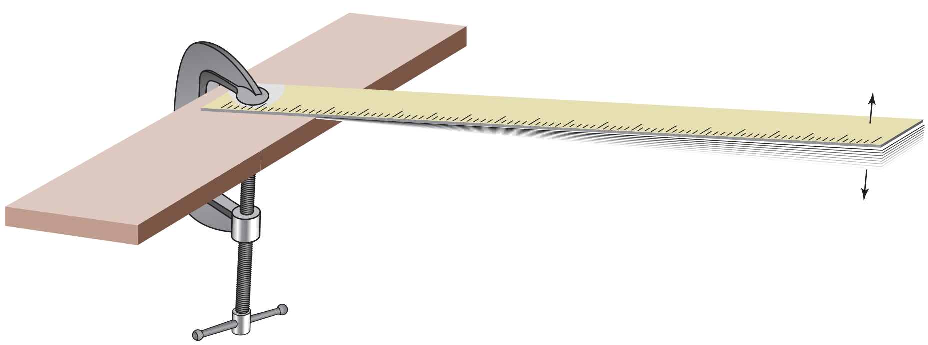


Figure C1.8 Oscillating ruler

Inquiry 3: Concluding and evaluating

Concluding

A student wants to investigate an oscillating ruler, similar to that seen in Figure C1.8, but she varied the mass of the ruler by taping various masses near to its free end. She predicted that the time period of the oscillator could be determined from the same equation as for a mass oscillating vertically on the end of a spring. Table C1.1 summarizes her raw data. Assume that the uncertainties are low.

Table C1.1 Results of vibrating ruler experiment

Mass on end of blade / g	Time period / s
0	too quick to measure
40	0.55
80	0.76
120	0.91
160	1.04
200	1.15

Process the results and reach a conclusion. State whether the student predicted correctly. If not, suggest a possible reason.

WORKED EXAMPLE C1.2

A 200 g mass was placed on the end of a long spring and increased its length by 5.4 cm.

a Determine the spring constant, k, of the spring.

b If the mass is displaced a small distance from its equilibrium position and undergoes SHM, calculate the frequency of oscillations.

c Suggest a possible reason why oscillations with greater amplitude may not be simple harmonic.

Answer

a $F_H = -kx$

$0.200 \times 9.8 = -k \times (5.4 \times 10^{-2})$

$k = -36\,\text{N m}^{-1}$ (the negative sign shows that force and displacement are in opposite directions)

b $T = 2\pi\sqrt{\dfrac{m}{k}} = 2 \times \pi \times \sqrt{\dfrac{0.200}{36}} = 0.47\,\text{s}$

$f = \dfrac{1}{T} = \dfrac{1}{0.47} = 2.1\,\text{Hz}$

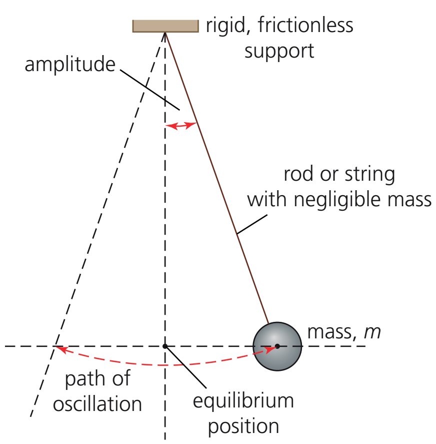
c The spring may become overstretched, so that Hooke's law is no longer applicable.

Simple pendulum

◆ **Pendulum** A weight, which is suspended below a pivot, which is able to swing from side to side. The weight is sometimes called the **pendulum bob**. The concept of a **simple pendulum** is a point mass on the end of an inextensible string.

There are many different designs of **pendulum**, all of which involve a mass swinging from side to side under the effects of gravity. We use the term '**simple pendulum**' to describe the simplest possible model of a pendulum: a spherical mass swinging on a rod, or string (both of which have negligible mass), from a rigid, frictionless support. See Figure C1.9. The mass is often called a **pendulum 'bob'**.

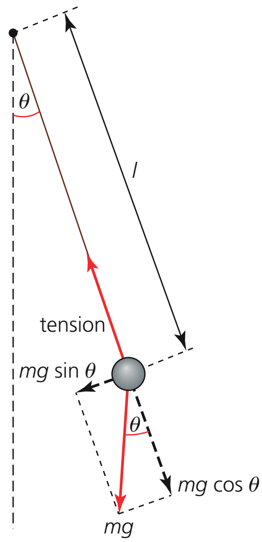
Because the mass is spherical and the rod / string has negligible mass, all of the mass of the pendulum can be assumed to be acting as a point mass at the centre of the sphere. The displacement of the motion can be considered to be the angle between the rod / string and the vertical. The restoring force is provided by gravity: consider Figure C1.10.



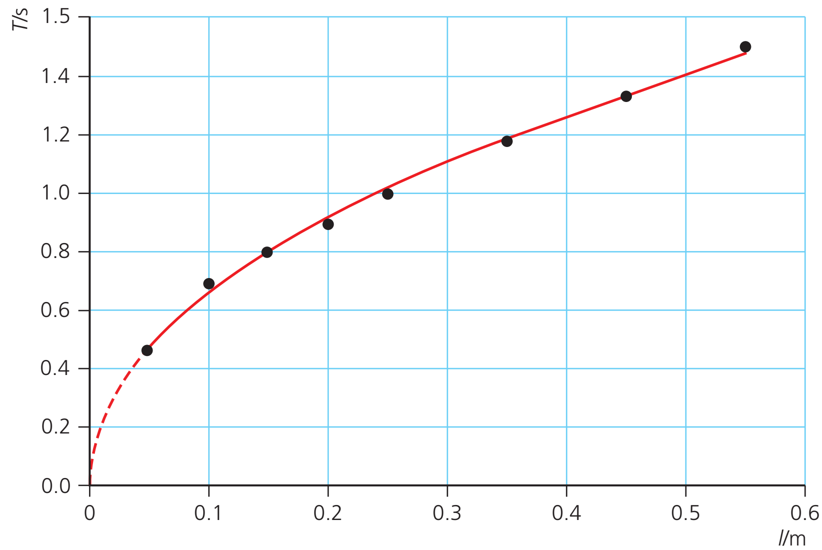
■ **Figure C1.9** Simple pendulum

The weight, mg, of the pendulum is conveniently resolved into two perpendicular components: a force $mg\sin\theta$ provides the restoring force bringing the pendulum back towards its equilibrium position; a force $mg\cos\theta$ keeps the rod / string in tension.

An alternative way of looking at this situation: the pendulum has only two forces acting on it: tension in the rod / string and weight, mg. The resultant of these is the restoring force, $mg\sin\theta$. Practical laboratory investigations of a simple pendulum will confirm that, *for small amplitudes*, the time period, T, depends only its length, l (measured from its point of support to its centre of mass). See Figure C1.11. Period is *not* affected by its mass. This is because doubling the mass, for example, will also result in doubling the restoring force.



■ **Figure C1.10** Components of the weight of the pendulum



■ **Figure C1.11** Results of simple pendulum investigation

The exact relationship for the SHM of a simple pendulum (for small amplitudes) is as follows.

Time period of a simple pendulum:

$$T = 2\pi\sqrt{\frac{l}{g}}$$

Top tip!

The data of Figure C1.11 can be linearized. A graph of T^2–l should be a straight line through the origin. A simple pendulum experiment can be used to determine a value for the gravitational field strength, g. The graph of T^2–l will have a gradient of: $4\pi^2/g$

Tool 3: Mathematics

Use approximation and estimation

For the motion of a simple pendulum to be a good example of SHM, we require that the restoring force ($mg\sin\theta$) is proportional to the displacement (θ). For very small angles, values of $\sin\theta$, $\tan\theta$ and θ (in rad) are almost identical, so that this condition is satisfied. But, as the angle increases, the difference gets greater. Determine the minimum angle for which there is at least a 1% difference between $\sin\theta$, $\tan\theta$ and θ (in rad).

■ **Figure C1.12** Mount Nevado Huascarán in Peru is reported to have the lowest gravitational field strength on Earth; its summit (6768 m) is one of the farthest points on the surface of the Earth from Earth's centre

ATL C1A: Research skills, thinking skills

Use search engines effectively; providing a reasoned argument to support conclusions

The equation highlighted above shows that the period of a simple pendulum would also change if its value was checked in a location where the gravitational field strength had a different value. Using your own research, investigate the variations of the strength of the gravitational field around the Earth. Then state and explain whether experiments in school laboratories would be able to measure those differences.

WORKED EXAMPLE C1.3

Determine a value for the time period of a simple pendulum of length 85.6 cm at a location where the gravitational field strength is

a 9.81 N kg^{-1}

b 1.63 N kg^{-1} (on the Moon's surface).

c This equation predicts that the ratio of the time periods at two different locations is:

$$\sqrt{\frac{g_1}{g_2}}.$$

Do your answers to **a** and **b** confirm that?

Answer

a $T = 2\pi\sqrt{\frac{l}{g}} = 2 \times \pi \times \sqrt{\frac{0.856}{9.81}} = 1.86\,\text{s}$

b $2 \times \pi \times \sqrt{\frac{0.856}{1.63}} = 4.55\,\text{s}$

c $\frac{T_M}{T_E} = \frac{4.55}{1.86} = 2.45$

$\sqrt{\frac{g_E}{g_M}} = \sqrt{\frac{9.81}{1.63}} = 2.45$

Yes, the same ratio is confirmed.

Nature of science: Measurements

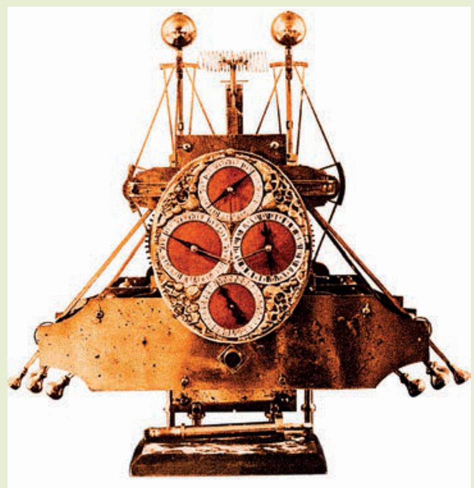
Technological developments in the measurement of time

Famously, Galileo was the first to discover the constant time periods of pendulums, but it was not until about sixty years later, in 1656, that the first pendulum clock was invented by the young Dutch scientist and inventor, Christiaan Huygens. Huygens is widely considered to be one of the greatest scientists and astronomers of all time. By the end of the eighteenth century, the best pendulum clocks could be made with an inaccuracy of less than one second a day. They were used throughout the world as the most accurate timekeepers for more than 250 years.

Accurate timekeepers were also essential for navigation on long journeys by ship. East–west distances (longitude) could be determined from observation of the stars or planets, combined with knowledge of the exact time of observation.

■ **Figure C1.13** Christiaan Huygens

■ **Figure C1.14** Clocks like this were found in many homes



■ **Figure C1.15** James Harrison's first marine timekeeper (1735)

However, pendulum clocks were not designed to cope with the motions of ships on the ocean! The British government offered a valuable reward to anyone who could design a clock that would remain accurate at sea, and hence 'solve the problem of longitude'. It was won by a clockmaker named James Harrison. His design included oscillating spheres on springs, rather than a pendulum.

Today we expect to know the exact time, at any time and place that we want, and we consider such precision to be normal, unworthy of comment. But, before pendulum clocks, most people were unaware of the time to the nearest hour, never mind the nearest minute, or second. (And they probably would not have understood any need for such accuracy!)

It is impossible to over-estimate the importance of the accurate measurement of time in modern life. The physics principle behind the pendulum clock is easily understood, but it took great technological skill (for that era) to manufacture the accurate clocks which had such considerable benefits on everyday life.

6 Determine what mass will have a period of exactly one second when oscillating on the end of a spring which has a spring constant of 84 N m^{-1}.

7 Calculate the angular frequency of a mass–spring system which involves a mass of 1000 g and a spring with $k = 100$ N m^{-1}.

8 A student investigated the effect of using springs of different stiffness (k) on the periods of oscillations, T, of the same mass.
Sketch the T–k graph that should be obtained.

9 Perhaps the world's most famous pendulum was made by Léon Foucault in Paris in 1851. The bob had a mass of 28 kg and its length was 67 m.
   a Calculate the period of this pendulum (the gravitational field strength in Paris is 9.81 N kg^{-1}).
   b Suggest reasons why it could continue to swing for a long time without any continuous energy input.
   c Use the internet to find out why this pendulum was so important.

10 Figure C1.11 shows the results of an investigation into a simple pendulum.
   a Calculate the angular frequency which describes oscillations of a pendulum of length 0.50 m.
   b Determine a value for the strength of the gravitational field from these results.

11 Figure C1.16 shows a girl starting to bounce on a trampoline. To begin with, her feet remain in contact with the rubber sheet and her movement can be considered to be SHM. The girl has a mass of 38 kg and the sheet stretched down by 33 cm when she was standing still in the middle.
   a Calculate a value for the spring constant of the trampoline.
   b Determine the time period of her bounces while she remains in contact with the trampoline's surface.

Figure C1.16 Girl on trampoline

12 Figure C1.17 shows a spring which is part of a car's suspension system.

Figure C1.17 The suspension system of a car

   a Estimate its spring constant by considering how much the wing of a car will depress if you push down hard on it (or sit on it).
   b By considering that the wheel effectively oscillates on the spring, estimate
      i its time period of oscillation
      ii how far a car travelling at 12 m s^{-1} moves during one oscillation.
   c The suspension also incorporates a shock absorber (damper). Discuss why this is necessary.

LINKING QUESTIONS

- How can greenhouse gases be modelled as simple harmonic oscillators?
- What physical explanation leads to the enhanced greenhouse effect? (NOS)

These questions link to understandings in Topic B.2.

Conditions that produce SHM

SYLLABUS CONTENT

- Conditions that lead to simple harmonic motion.
- The defining equation of simple harmonic motion as given by: $a = -\omega^2 x$.

SHM will occur if the restoring force, F, is proportional to the displacement, x, but the force always acts back towards the equilibrium position:

$F \propto -x$

The negative sign is important here. It shows us that the force acts in the opposite direction to the displacement: an increasing force *opposes* increasing displacement.

For an oscillating mass on a spring which obeys Hooke's law ($F_H = -kx$) this condition is obviously satisfied. For a simple pendulum, the restoring force, $mg\sin\theta$, is (almost) proportional to the displacement angle, θ, if the angle is small.

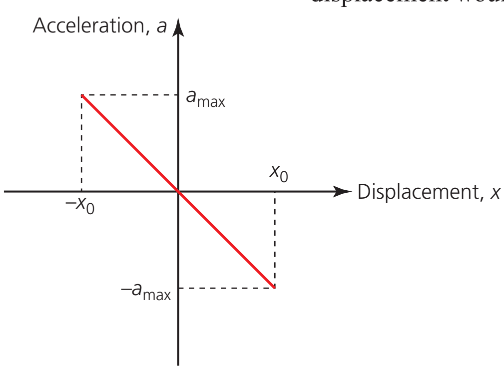
To define SHM we need to refer to the motion, rather than the force. Since $F = ma$, for a constant mass, we can write $a \propto -x$.

◆ **Simple harmonic motion (SHM)** Oscillations in which the acceleration, a, is proportional to the displacement, x, and in the opposite direction, directed back to the equilibrium position. $a \propto -x$.

Simple harmonic motion is defined as an oscillation in which the acceleration is proportional to the displacement, but in the opposite direction (always directed back towards the equilibrium position):

$a \propto -x$

This is represented by the graph shown in Figure C1.18. A graph of the restoring force against displacement would look similar.



■ **Figure C1.18** Acceleration–displacement graph for SHM

If, for example, the maximum displacement (amplitude) of an SHM is doubled, the restoring force and acceleration will also double. This will result in the mass taking the same time for each oscillation, because it is moving twice the distance at twice the average speed. This explains a defining feature of SHM: amplitude does not affect time period.

We can rewrite the equation above as: $a = -\text{constant} \times x$.

The constant must involve frequency because the magnitude of the acceleration is greater if the frequency is higher. The constant can be shown to be equal to ω^2. So that the defining equation of SHM can be written as:

$$\boldsymbol{a = -\omega^2 x}$$

WORKED EXAMPLE C1.4

A mass oscillates horizontally between two springs with a frequency of 1.4 Hz.

a Calculate its angular frequency.

b Determine its acceleration when

- **i** its displacement is 1.0 cm to the right
- **ii** its displacement is 4.0 cm to the left
- **iii** it passes through its equilibrium position.

Answer

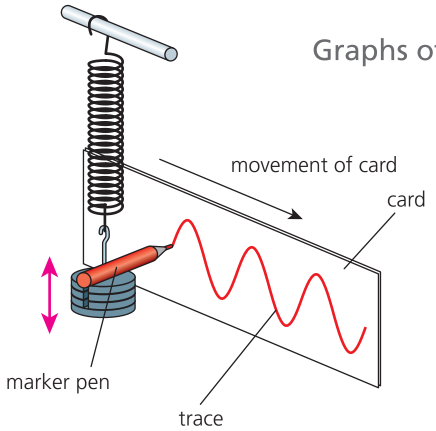
a $\omega = 2\pi f = 2\pi \times 1.4 = 8.8\,\text{rad s}^{-1}$

b **i** $a = -\omega^2 x = -(8.8)^2 \times (+0.010) = -0.77\,\text{m s}^{-2}$ (to the left)

ii $a = -\omega^2 x = -(8.8)^2 \times (-0.040) = +3.1\,\text{m s}^{-2}$ (to the right)

iii $a = -\omega^2 x = -(8.8)^2 \times 0.0 = 0.0\,\text{m s}^{-2}$

Graphs of SHM



■ **Figure C1.19** Recording the motion of an oscillation

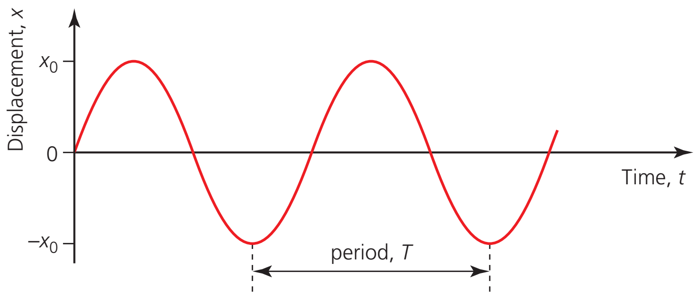
A mass oscillating on a spring could be used with a marker pen to produce a record of the oscillation, as shown in Figure C1.19. If the card moves at a constant speed, the record (trace) produced is effectively a graph of displacement against time. **Data logging** with an appropriate sensor will improve this experiment.

The waveform seen in Figure C1.19 is commonly described as being **sinusoidal** in shape.

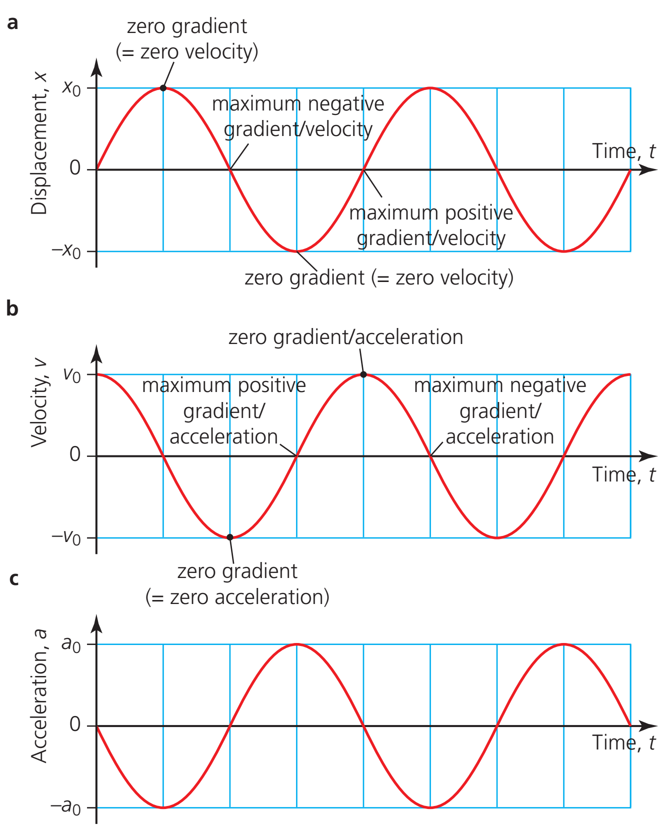
The graph in Figure C1.20 shows the variation in displacement, x, with time, t, for the idealized model of a mass moving with perfect SHM. Here we have *chosen* that the particle has zero displacement at the start of the timing. The motion has an amplitude of x_0. Alternatively, we might have chosen the graph to start with the particle at maximum displacement (or any other displacement).

◆ **Data logging** Connecting sensors to a computer with suitable software to enable physical quantities to be measured and processed digitally.

◆ **Sinusoidal** In the shape of a sine wave (usually equivalent to a cosine wave).



■ **Figure C1.20** Displacement–time graph for simple harmonic motion. Timing was started when the particle had zero displacement



■ **Figure C1.21** Graphs for simple harmonic motion starting at displacement $x = 0$: **a** displacement–time **b** velocity–time **c** acceleration–time

Using knowledge from Topic A.2, we can use this graph to calculate the velocity of the oscillating mass at any particular moment by determining the gradient of the displacement–time graph at that moment:

$$\text{velocity, } v = \frac{\text{change in displacement}}{\text{change in time}} = \frac{\Delta x}{\Delta t}$$

Similarly, the acceleration at any given time can be found from the gradient of the velocity–time graph:

$$\text{acceleration, } a = \frac{\text{change in velocity}}{\text{change in time}} = \frac{\Delta v}{\Delta t}$$

Using this information, three separate but interconnected graphs of motion can be drawn and compared, as shown in Figure C1.21.

The velocity graph has its maximum value, v_0, when the displacement, x, is zero, and the velocity is zero when the displacement is at its maximum, $\pm x_0$.

The acceleration has its maximum value when the velocity is zero and the displacement is greatest. This is to be expected, because when the displacement is greatest, the restoring force, acting in the opposite direction, is greatest.

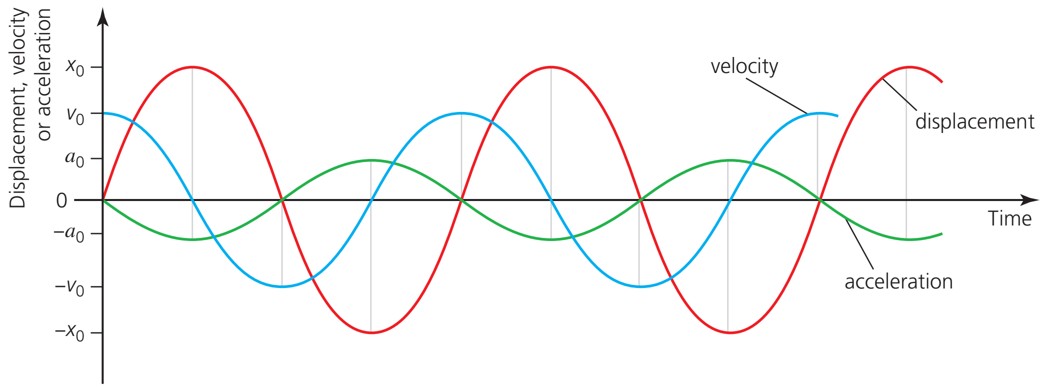
SHM graphs of displacement, velocity and acceleration are all sinusoidal in shape, but their maximum values occur at different times.

Figure C1.22 shows all three SHM graphs drawn on the same axes, so that they can be more easily compared. Note that the amplitudes of the three graphs are arbitrary; they are not interconnected and should not be compared.

LINKING QUESTION

- How can the understanding of simple harmonic motion apply to the wave model? (NOS)

This question links to understandings in Topic C.2.



■ **Figure C1.22** Comparing displacement, velocity and acceleration for SHM, with timing starting at displacement $x = 0$

■ **Figure C1.23** Four oscillations out of phase

Phase difference

This is a convenient point to introduce the important concept of **phase difference**. Figure C1.23 shows an everyday example. The motions of the four children all have the same frequency because all the swings are the same length, but they are each at different points in their oscillations: we say that they are *out of phase* with each other.

A phase difference occurs between two similar oscillations if they have the same frequency, but their maximum values do not occur at the same time.

◆ **Phase difference** When oscillators that have the same frequency are out of phase with each other, the difference between them is defined by the angle (usually in terms of π radians) between the oscillations. Phase differences can be between 0 and 2π radians.

The three graphs shown in Figure C1.21 all have the same frequency and sinusoidal shape, but their peaks occur at different times: there is a phase difference between the waves. This could be quantified by referring to the fraction of a time period, T, that occurs between their peaks. The first peak of the displacement graph occurs $T/4$ after the first peak of the velocity graph. The first peak of the acceleration graph occurs $3T/4$ after the first peak of the velocity graph.

But, remembering that one complete oscillation can be considered as equivalent to moving through an angle of 2π radians, phase differences are more usually quoted in terms of π.

Examples of phase differences

- One quarter of an oscillation, $\frac{T}{4} = \frac{2\pi}{4} = \frac{\pi}{2}$ radians (90°)
- One half of an oscillation, $\frac{T}{2} = \frac{2\pi}{2} = \pi$ radians (180°)
- Three quarters of an oscillation, $\frac{3T}{4} = 2\pi \times \frac{3}{4} = \frac{3\pi}{2}$ radians (270°)

(A phase difference of $\frac{3\pi}{2}$ radians is equal in magnitude to a phase difference of $\frac{\pi}{2}$ radians, but in some circumstances, we might be concerned about which peak came 'first'.)

Referring back to Figure C1.21, can see that there is a phase difference of $\frac{\pi}{2}$ between displacement and velocity, and a phase difference of π between displacement and acceleration.

13 An oscillator moves with SHM and has a time period of 2.34 s.
How far must it be displaced from its equilibrium position in order that its acceleration is $1.00\,\mathrm{m\,s^{-2}}$?

14 During SHM a mass moves with an acceleration of $3.4\,\mathrm{m\,s^{-2}}$ when its displacement is 4.0 cm. Calculate its:

a angular frequency **b** time period.

15 A mass oscillating on a spring performs exactly 20 oscillations in 15.8 s.

a Determine its acceleration when it is displaced 62.3 mm from its equilibrium position.

b State any assumption that you made when answering **a**.

16 Look at the graph in Figure C1.24 which shows the motion of a mass oscillating on a spring. Determine:

a the amplitude

b the time period

c the displacement after 0.15 s

d the displacement after 1.4 s.

17 a Sketch a displacement–time graph showing two complete oscillations for a simple harmonic oscillator which has a time period of 2.0 s and an amplitude of 5.0 cm.

b Add to the same axes the graph of an oscillator which has twice the frequency and the same amplitude.

c Add to the same axes the graph of an oscillator that has an amplitude of 2.5 cm and the same frequency but which is ¼ of an oscillation out of phase with the first oscillator.

18 a Sketch a velocity–time graph showing two complete oscillations for a simple harmonic oscillator which has a frequency of 4 Hz and a maximum speed of $4.0\,\mathrm{c\,ms^{-1}}$.

b On the same axes sketch graphs to show the variation of:

i displacement **ii** acceleration

for the same oscillation.

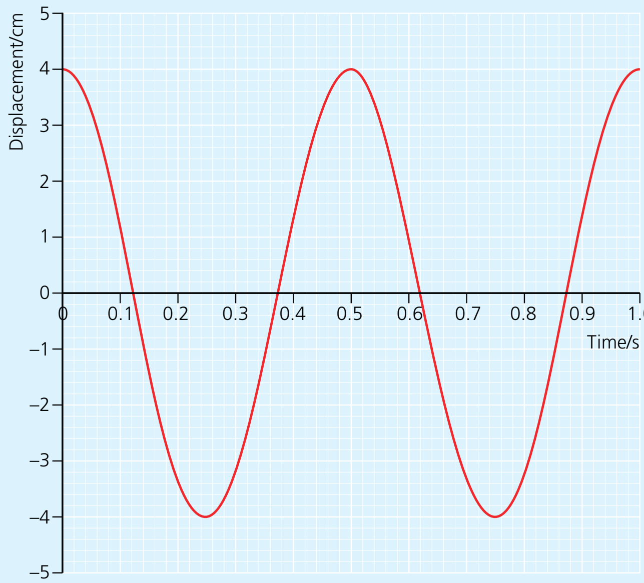


Figure C1.24 Motion of a mass oscillating on a spring

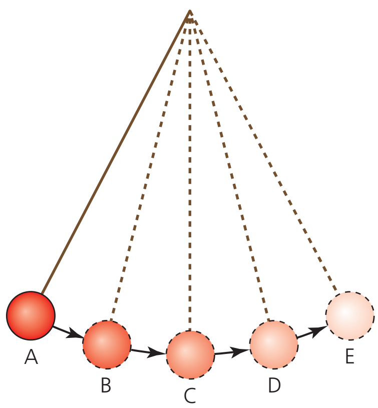
There is more about phase difference, graphs of motion, and the equations which represent them, towards the end of this topic (for HL students).

Energy changes during SHM

SYLLABUS CONTENT

- A qualitative approach to energy changes during one cycle of an oscillation.

When an object which can oscillate is pushed, or pulled, away from its equilibrium position, against the action of a restoring force, work will be done and potential energy will be stored in the oscillator. For example, a spring will store elastic potential energy and a simple pendulum will store gravitational potential energy. When the object is released, it will gain kinetic energy and lose potential energy as the restoring force accelerates it back towards the equilibrium position. Its kinetic energy has a maximum value as it passes through the equilibrium position and, at the same time, its potential energy is minimized. As it moves away from the equilibrium position, kinetic energy decreases as the restoring force opposes its motion and potential energy increases again.



■ **Figure C1.25** A swinging pendulum

As an example, consider Figure C1.25, in which a simple pendulum has been pulled away from its equilibrium position, C, to position A. While it is held in position A, it has zero kinetic energy and its change of gravitational potential energy (compared to position C) is greatest. When it is released, gravity provides the restoring force and the pendulum exchanges potential energy for kinetic energy as it moves through position B to position C. At C it has maximum kinetic energy and the change in potential energy has reduced to zero. The pendulum then transfers its kinetic energy back to potential energy as it moves through position D to position E. At E, like A, it has zero kinetic energy and a maximum change of potential energy. The process then repeats every half time period.

If the pendulum was a perfect simple harmonic oscillator, there would be no energy losses, so that the sum of the potential energy and the kinetic energy would be constant and the pendulum would continue to reach the same maximum vertical height and maximum speed every oscillation. In practice, frictional forces will result in energy dissipation and all the energies of the pendulum will progressively decrease.

All mechanical oscillations involve a continuous exchange of energy between kinetic energy and some form of potential energy.

For a simple pendulum, the potential energy Ep is in the form of gravitational potential energy (see Topic A.3 where ΔE_p was used instead of E_p):

$$E_p = mg\Delta h$$

For a mass on a spring, the potential energy is in the form of elastic potential energy (see Topic A.3, where the symbol E_H was used instead of E_p and Δx was used instead of x):

$$E_p = \frac{1}{2}kx^2$$

Its maximum value is:

$$E_{p_{max}} = \frac{1}{2}kx_0^{\ 2}$$

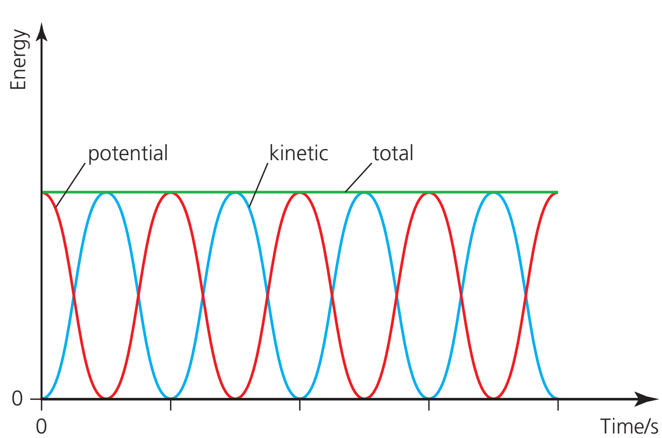
which shows us that generally:

The total energy of an SHM is proportional to its amplitude squared.

Kinetic energy of the mass can be determined from (Topic A.3):

$$E_k = \frac{1}{2}mv^2$$

Figure C1.26 represents these exchanges during several oscillations of perfect SHM.



■ **Figure C1.26** Energy changes during oscillation of perfect SHM

◆ **Damping** Occurs when resistive forces act on an oscillating system, dissipating energy and reducing amplitude.

In more realistic situations, the energies of the system will decrease over time. This is shown in Figure C1.27. Energy dissipation from an oscillating system is called **damping**.

Figure C1.28 shows the variations in energy with displacement during each oscillation of SHM.

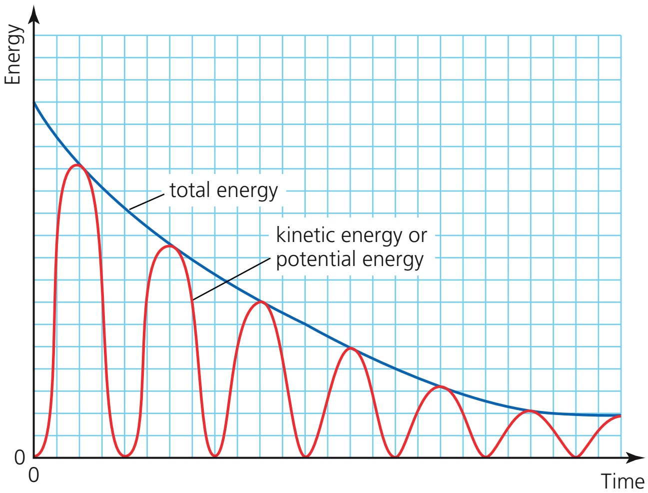


Figure C1.27 Energy dissipation in a practical oscillator

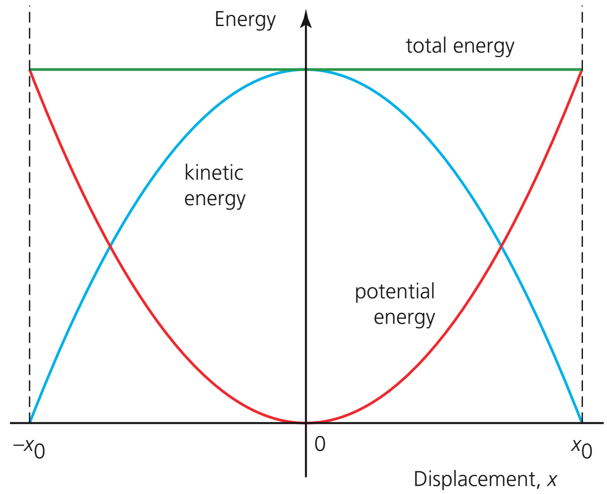


Figure C1.28 Variation of energies of a simple harmonic oscillator with displacement

WORKED EXAMPLE C1.5

A mass oscillating in a mass–spring system has a maximum kinetic energy of 0.047 J.

a If its maximum speed was 85 cm s^{-1}, determine its mass.

b State the maximum value of its potential energy.

c Determine the spring constant if the amplitude of the oscillation was 1.9 cm.

Answer

a $E_k = \frac{1}{2}mv^2$

$0.047 = \frac{1}{2} \times m \times 0.85^2$

$m = 0.13\,\text{kg}$

b 0.047 J

c $E_{p_{max}} = \frac{1}{2}kx_0^{\,2}$

$0.047 = \frac{1}{2} \times k \times (1.9 \times 10^{-2})^2$

$k = 2.6 \times 10^2\,\text{N m}^{-1}$

19 A simple pendulum, of mass 100 g, is released from rest at its maximum displacement, which is 5.0 cm higher than its central position. It then swings with SHM and a frequency of 1.25 Hz.

a Calculate its maximum potential energy.

b Sketch a graph with fully labelled axes to show the variations in potential energy of the pendulum in the first 1.6 s after it was released.

c Determine the maximum speed of the pendulum.

20 The total energy seen in Figure C1.27 shows a nearly 90% decrease in six oscillations.

a Describe where this energy has been transferred to.

b If the oscillating mass was passing through its equilibrium position at time $t = 0$, state what kind of energy is represented by the red line.

c It is suggested that the total energy decreases to the same fraction each oscillation. Analyse data from the graph to check if this is true.

21 a Use the equation:

$$E_p = \frac{1}{2}kx^2 \text{ where } k = 12\,\text{N m}^{-1}$$

to calculate values of elastic potential energy stored in a mass–spring system for SHM displacements, x, of 0, ±0.05 m, ±0.10 m, ±0.15 m, ±0.20 m. Its amplitude was 0.20 m.

b Sketch a graph to display these results.

c On the same axes, add a graph to represent the variations in kinetic energy of the mass.

d If the mass was 50 g, determine its maximum speed.

ATL C1B: Social skills

Resolving conflicts during collaborative work

A group of four students was asked by their teacher to investigate four different oscillators and report back to the rest of their group one week later. One of the four students was chosen to be the team-leader. Three of the students worked well, but the fourth showed no interest and did little. What should the team-leader do about this situation?

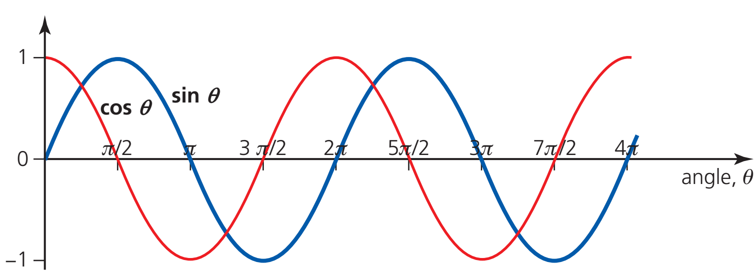
HL ONLY

Calculating displacements and velocities during SHM

SYLLABUS CONTENT

- A particle undergoing simple harmonic motion can be described using phase angle.
- Problems can be solved using the equations for simple harmonic motion as given by:
  $x = x_0 \sin(\omega t + \phi)$
  $v = \omega x_0 \cos(\omega t + \phi)$

Figure C1.29 shows the variations of a perfect sine wave and compares it to a cosine wave. The two waves have identical shapes, but there is a phase difference of $\pi/2$ between them.



■ **Figure C1.29** Comparing sine and cosine waves

The shape of the graph of SHM shown in Figure C1.20 is identical to the sine wave shown in Figure C1.29 and it could be represented by the equation of the form $x = x_0 \sin\theta$.

We know from Topic A.4:

$$\text{angular frequency, } \omega = \frac{\Delta\theta}{\Delta t}\left(\text{or } \frac{\theta}{t}\right)$$

so that: θ (radians) $= \omega t$. And then we then can rewrite: $x = x_0 \sin\theta$ as $x = x_0 \sin\omega t$.

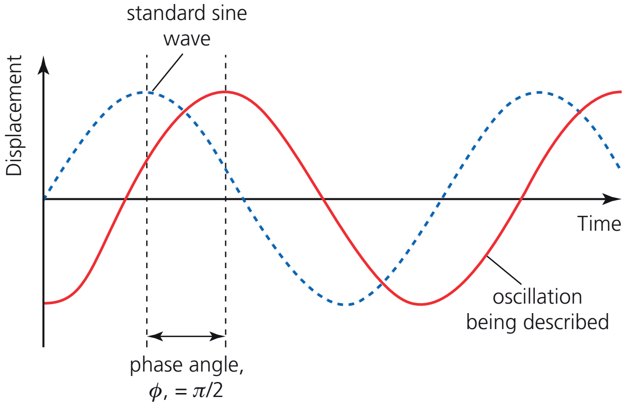
This equation assumes that the initial value of $x = 0$ when $t = 0$. Under those circumstances, it enables us to calculate the displacement of a SHM at any time we choose (if the amplitude and frequency are known).

◆ **Phase angle** The difference in angular displacement of an oscillation compared to an agreed reference point. Expressed in terms of π radians.

HL ONLY

In order to write an equation which allows for all other possibilities, we need to return to the concept of phase difference. We have already discussed the phase difference between two oscillations, but now we will refer to the **phase angle** of a *single* oscillator (sometimes just called *phase*). The oscillations are compared to a theoretical sine wave, which has zero displacement at time $t = 0$.

The phase angle, ϕ, of an oscillation is the fraction of an oscillation (expressed in terms of π) that occurs between when it has zero displacement and a sine wave which has zero displacement at time $t = 0$.



■ **Figure C1.30** Phase angle of an oscillator

Figure C1.30 shows an example.

We can now write a full equation to describe the variation of displacement with time for any SHM:

displacement, $\boldsymbol{x = x_0 \sin(\omega t + \phi)}$

For example, if the lines in Figure C1.30 show the variations in displacement of two oscillators with the same frequency and amplitude, the dotted blue line can be represented by:

$$x = x_0 \sin \omega t$$

while the red line has the equation:

$$x = x_0 \sin\left(\omega t + \frac{3\pi}{2}\right)$$

We have seen (Figure C1.22) that the variations of displacement, x, and velocity, v, for the same SHM oscillator are $\pi/2$ out of phase with each other. So that, the equation representing the velocity of a SHM involves a cosine:

velocity, $\boldsymbol{v = \omega x_0 \cos(\omega t + \phi)}$

Common mistake

The presence of ω at the start of the term on the right-hand side need *not* be explained here, but it is a common mistake to leave it out of calculations.

LINKING QUESTION

- How can circular motion be used to visualize simple harmonic motion?

This question links to understandings in Topic A.2.

The maximum velocity, v_0, will occur when the cosine in the equation has a value of one:

maximum velocity, $v_0 = \omega x_0$

To help to understand these equations, consider first the simplest possible numerical example, with $T = f = 1$ and $x_0 = 1$, for an oscillation beginning at $x = 0$ when $t = 0$. The phase angle is 0, so that the equation for displacement reduces to $x = \sin(2\pi t)$.

Using this equation at, for example, $t = 0.20$ s, leads to $x = +0.95$ m, and at $t = 0.60$ s, $x = -0.59$ m

The negative sign shows that the displacement at $t = 0.60$ s was in the opposite direction to the initial displacement (just after $t = 0$).

Using $v = 2\pi \times \cos(2\pi t)$ with the same data, we get: at $t = 0.20\,\text{s}$, $v = +1.9\,\text{m}\,\text{s}^{-1}$, and at $t = 0.60\,\text{s}$, $v = -5.1\,\text{m}\,\text{s}^{-1}$.

The maximum velocity, $v_0 = \omega x_0 = 2\pi = 6.3\,\text{m}\,\text{s}^{-1}$.

The negative sign shows that the velocity at $t = 0.60\,\text{s}$ was in the opposite direction to the velocity when $t = 0$.

This data is shown in Figure C1.31.

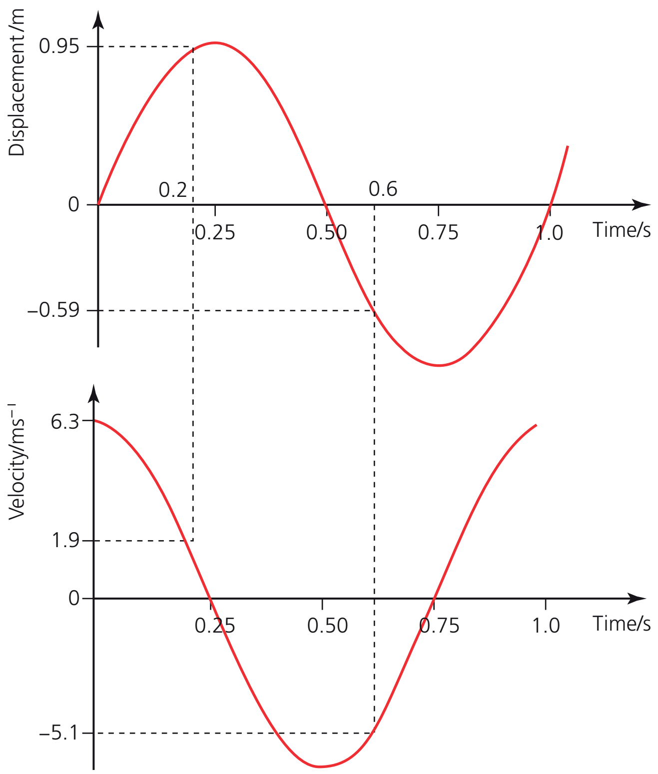


Figure C1.31 Numerical data for displacement and velocity of SHM shown on a graph

WORKED EXAMPLE C1.6

A mass oscillates with SHM of frequency 2.7 Hz and amplitude 1.7 cm. If its phase angle is $\pi/2$, calculate:

a its displacement after 2.0 s

b its velocity after 3.0 s

c its maximum velocity.

Answer

Remembering that $\omega = 2\pi f = 2 \times \pi \times 2.7 = 5.4\pi\,\text{rad}\,\text{s}^{-1}$:

a $x = x_0 \sin(\omega t + \phi) = (1.7 \times 10^{-2}) \times \sin\left((5.4\pi \times 2.0) + \dfrac{\pi}{2}\right) = -1.4 \times 10^{-2}\,\text{m}$

The negative sign shows that the displacement was in the opposite direction to the displacement just after $t = 0$.

b $v = \omega x_0 \cos(\omega t + \phi) = (5.4\pi \times 1.7 \times 10^{-2}) \times \cos\left((5.4\pi \times 3.0) + \dfrac{\pi}{2}\right) = -0.17\,\text{m}\,\text{s}^{-1}$

The negative sign shows that the velocity at $t = 3.0\,\text{s}$ was in the opposite direction to the velocity when $t = 0$.

c $v_0 = \omega x_0 = 5.4\pi \times (1.7 \times 10^{-2}) = 0.29\,\text{m}\,\text{s}^{-1}$

22 A mass is oscillating between two springs with a frequency of 1.5 Hz and amplitude of 3.7 cm. It has a speed of $34\,\text{cm}\,\text{s}^{-1}$ as it passes through its equilibrium position and a stopwatch is started.
Calculate its displacement and velocity 1.8 s later.

23 An object of mass 45 g undergoes SHM with a frequency of 12 Hz and an amplitude of 3.1 mm.

a Determine its maximum speed and kinetic energy.

b What is the object's displacement 120 ms after it is released from its maximum displacement?

24 A mass is oscillating with SHM with an amplitude of 3.8 cm. Its displacement is 2.8 cm at 0.022 s after it is released from its maximum displacement.
Calculate a possible value for its frequency.

25 A simple harmonic oscillator has a time period of 0.84 s and its speed is $0.53\,\text{m}\,\text{s}^{-1}$ as it passes through its mean (equilibrium) position.

a Calculate its speed 2.0 s later.

b If the amplitude of the oscillation is 8.9 cm, what was the displacement after 2.0 s?

26 The water level in a harbour rises and falls with the tides, with a time of 12 h 32 min for a complete cycle. The high tide level is 8.20 m above the low tide level, which occurred at 4.10 am.
If the tides rise and fall with SHM, determine the level of the water at 6.00 am.

27 Discuss what the area under a velocity–time graph of an oscillation represents.

Calculating energy changes during SHM

SYLLABUS CONTENT

- Problems can be solved using the equations for simple harmonic motion as given by:
  $v = \pm\omega\sqrt{x_0^2 - x^2}$
  $E_T = \frac{1}{2}m\omega^2 x_0^2$
  $E_p = \frac{1}{2}m\omega^2 x^2$

We have already discussed the energy exchanges that occur during SHM qualitatively (see Figures C1.26, C.127 and C.128). Now we will interpret those changes in more mathematical detail.

We saw in Topic A.3 that elastic potential energy can be determined from:

$$E_p = \frac{1}{2}kx^2$$

where k is the spring constant.

We also know from earlier in this sub-topic that for a mass–spring oscillator:

$$\omega^2 = \frac{k}{m}$$

which leads to:

potential energy, $\boldsymbol{E_p = \frac{1}{2}m\omega^2 x^2}$

When the mass is at its maximum displacement, x_0, its velocity has reduced to zero. It has zero kinetic energy and it has its maximum potential energy - which is then equal to the total energy, E_T, of the SHM:

total energy, $\boldsymbol{E_T = \frac{1}{2}m\omega^2 x_0^2}$

Since at any point, total energy = potential energy + kinetic energy.

kinetic energy, $E_k = \frac{1}{2}m\omega^2 x_0^2 - \frac{1}{2}m\omega^2 x^2 = \frac{1}{2}m\omega^2 (x_0^2 - x^2)$

The instantaneous velocity, v, at any point in the oscillation can be found by equating:

$$E_k = \frac{1}{2}mv^2$$

with the previous equation. Which leads to:

$\boldsymbol{v = \pm\,\omega\sqrt{(x_0^2 - x^2)}}$

HL ONLY

WORKED EXAMPLE C1.7

An oscillator of mass 750 g oscillates with SHM with an amplitude of 3.47 cm and a period of 1.44 s.

a Calculate its total energy.

b When it has a displacement of 2.15 cm determine:

i its potential energy

ii its kinetic energy.

c When its kinetic energy is 2.00×10^{-3} J, what is its displacement?

d Calculate the velocity of the mass when its displacement is 2.00 cm.

e What is the maximum velocity of the mass?

Answer

a $E_T = \frac{1}{2}m\omega^2 x_0^2$

$= 0.5 \times 0.750 \times \left(\frac{2\pi}{1.44}\right)^2 \times 0.0347^2 = 8.60 \times 10^{-3}$ J

b i $E_p = \frac{1}{2}m\omega^2 x^2$

$= 0.5 \times 0.750 \times \left(\frac{2\pi}{1.44}\right)^2 \times 0.0215^2 = 3.30 \times 10^{-3}$ J

ii $E_k = E_T - E_p = (8.60 \times 10^{-3}) - (3.30 \times 10^{-3}) = 5.30 \times 10^{-3}$ J

c $E_k = \frac{1}{2}m\omega^2(x_0^2 - x^2)$

$2.00 \times 10^{-3} = 0.5 \times 0.750 \times \left(\frac{2\pi}{1.44}\right)^2 \times (x_0^2 - x^2)$

$(x_0^2 - x^2) = 2.80 \times 10^{-4}$

$x^2 = 0.0347^2 - (2.80 \times 10^{-4})$

$x = 9.24 \times 10^{-4}$ m

d $v = \pm\omega\sqrt{(x_0^2 - x^2)} = \left(\frac{2\pi}{1.44}\right) \times \sqrt{(3.47^2 - 2.00^2)} = \pm 12.4$ m s^{-1} (in either direction)

e v_{max} occurs when displacement is zero.

$v_{max} = \pm\omega\sqrt{(x_0^2 - 0^2)} = \pm\omega x_0 = \pm\left(\frac{2\pi}{1.44}\right) \times 3.47 = \pm 15.1$ m s^{-1} (in either direction)

28 A mass of 480 g is suspended on a spring of stiffness 132 N m^{-1}.

a If it undergoes SHM, calculate its time period.

b Calculate its angular frequency.

c If the oscillations have an amplitude of 3.2 cm, determine:

i its maximum kinetic energy

ii its maximum speed.

d Calculate how much potential energy is stored in the system when the displacement is 3.2 cm.

29 An SHM oscillator has a mass of 0.42 kg and a total energy of 1.7 J. If its frequency is 5.7 Hz, determine the amplitude of its motion.

30 A student stretched a vertical spring by placing a mass of 100 g on its end. A second 100 g mass was added and the length of the spring increased by a further 4.7 cm.

a Assuming that it obeyed Hooke's law, determine the spring constant.

b The combined mass of 200 g was then displaced by 5.0 cm so that it oscillated with SHM. What was its period?

c Calculate how much energy was stored in these oscillations.

d Show that the value of the ratio E_k/E_p when the displacement was 2.0 cm was about 5/1.

e Determine by what factor the total energy would increase if the amplitude was increased to 8 cm.

31 A mass was oscillating with SHM at a frequency of 7.6 Hz.

a If its maximum speed was 1.4 m s^{-1}, determine the amplitude of its motion.

b If the mass was 54 g, determine the kinetic energy of it when its displacement was 1.8 cm.

C.2 Wave model

Guiding questions

- What are the similarities and differences between different types of waves?
- How can the wave model describe the transmission of energy as a result of local disturbances in a medium?
- What effect does a change in the frequency of oscillation or medium through which the wave is travelling have on the wavelength of a travelling wave?

What is a wave?

SYLLABUS CONTENT

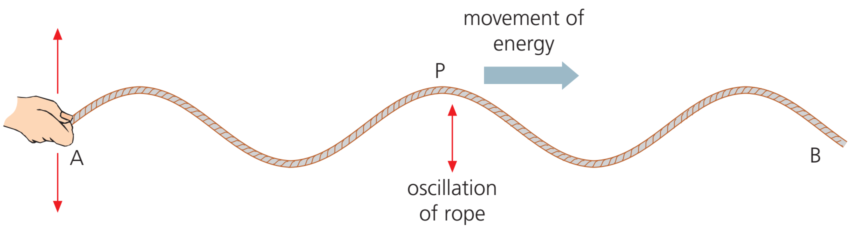
▶ The differences between mechanical waves and electromagnetic waves.

■ **Figure C2.1** Circular waves spreading out on a pond

When we think of waves the first example that comes to mind is probably that of waves on the surface of water, like those seen in Figure C2.1.

When the equilibrium of the surface is disturbed (for example by a falling drop, or touching it with a finger), it results in oscillations of the water surface at that point. Because of the forces between water molecules, the oscillations are transferred to neighbouring molecules a short time later, and then they spread outwards as a two-dimensional wave on the water surface.

A simpler, one-dimensional, example is shown in Figure C2.2: in this example the waves are produced by continuously shaking one end of the rope. Point A is the oscillating source of the wave energy, which travels to the other end, point B. All points on the rope, point P for example, oscillate up and down (as shown)..



■ **Figure C2.2** Creating a wave by shaking the end of a rope

◆ **Wave (travelling)** A wave that transfers energy away from a source. Sometimes called a progressive wave.

◆ **Propagation (of waves)** Transfer of energy by waves.

◆ **Medium (of a wave)** Substance through which a wave is passing.

These two examples are both **travelling waves**. Another kind of wave (a standing wave) is discussed in Topic C.4.

Scientists describe the motion of a wave away from its source as **propagation** of the wave. The matter through which the waves pass is called the **medium** of the wave.

All waves involve oscillations and they can be described as being either 'mechanical' or 'electromagnetic':

◆ **Wave (mechanical)** A wave which involves oscillating masses (including sound).

◆ **Wave (electromagnetic)** A transverse wave composed of perpendicular electric and magnetic oscillating fields travelling at a speed of $3.0 \times 10^8\,\mathrm{m\,s^{-1}}$ in free space.

Mechanical waves involve the oscillations of masses.

Electromagnetic waves, such as light, involve the oscillations of electric and magnetic fields.

The first part of this topic will deal with mechanical waves. Electromagnetic waves are discussed later.

A mechanical travelling wave can be described as an oscillating disturbance that travels away from its source through the surrounding medium (solid, liquid or gas) transferring energy from one place to another. Most importantly, waves transfer energy without transferring the matter itself.

■ **Figure C2.3** Ocean waves transferring a large amount of energy at Brighton, England – there is no continuous net movement of the water itself

For example, ocean waves may 'break' and 'crash' on to a shore or rocks, transferring considerable amounts of energy (that they got from the wind), but there is no net, continuous movement of water from the ocean to the land. A wooden log floating on a lake will simply oscillate up and down as waves pass (unless there is a wind).

Examples of mechanical waves

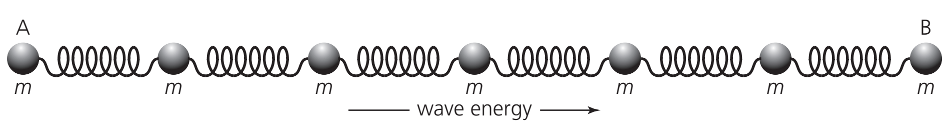
- waves on strings, ropes and springs
- waves on water
- sound (and similar waves in liquids and solids)
- earthquake waves.

Models of mechanical waves

SYLLABUS CONTENT

▶ Transverse and longitudinal travelling waves.

In order to understand more about the propagation of mechanical waves it is convenient to visualize the *continuous* medium in which they are travelling as being composed of *separate* (discrete) *particles* of mass, *m*, separated by springs representing the restoring forces that arise when the medium is disturbed from is equilibrium position. See Figure C2.4. The wave can be produced by shaking the end A, the wave then travels along the system to B.



■ **Figure C2.4** Wave model of masses and springs

Experiments can confirm that the speed of the wave along the system increases if the masses are smaller, or if the springs are stiffer.

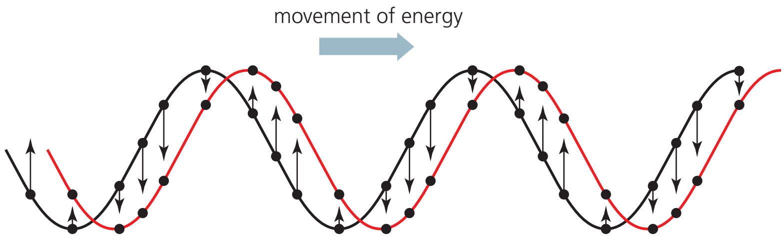
There are two different ways in which A can be shaken: left–right–left–right, or up–down–up–down (as shown). This identifies the two basic kinds of mechanical wave: transverse and longitudinal.

- ◆ **Transverse wave** A wave in which the oscillations are perpendicular to the direction of transfer of energy.
- ◆ **Crest** Highest part of a transverse mechanical wave.
- ◆ **Trough** Lowest point of a transverse mechanical wave.
- ◆ **Longitudinal wave** Waves in which the oscillations are parallel to the direction of transfer of energy.
- ◆ **Compressions (in a longitudinal wave)** Places where there are increases in the density and pressure of a medium as a wave passes through it.
- ◆ **Rarefactions (in a longitudinal wave)** Places where there are reductions in the density and pressure of a gas as a wave passes through it.

Transverse and longitudinal mechanical waves

In a **transverse wave**, each part of the medium oscillates *perpendicularly* to the direction in which the wave is transferring energy.

The waves shown in Figure C2.1, C2.2 and C2.3 are transverse waves. The black line in Figure C2.5 represents the positions of the particles in a continuous medium which is transferring wave energy to the right. The arrows show which way the particles are moving at that moment. The red line represents their positions a short time later. Each particle is oscillating with the same amplitude and frequency, but each particle is slightly out of phase with its neighbour.



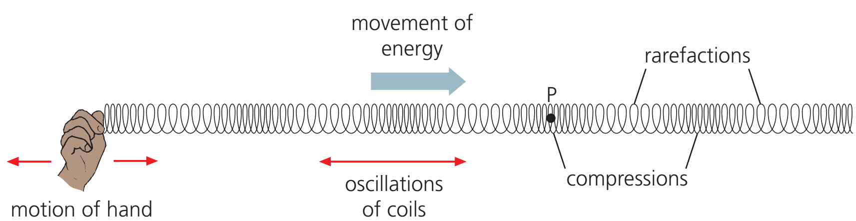
■ **Figure C2.5** Movement of particles as a transverse wave moves to the right

The tops of transverse waves are often called **crests**, while the bottoms of the waves are called **troughs**.

Mechanical waves on strings, ropes and water surfaces are all transverse in nature.

In a **longitudinal wave**, each part of the medium oscillates *parallel* to the direction in which the wave is transferring energy.

Stretched springs are often used to demonstrate waves. They are more massive than strings and this reduces the wave speed, so that the waves can be observed more easily. Stretched 'slinky' springs are particularly useful for demonstrating longitudinal waves. See Figure C2.6, which shows the characteristic **compressions** and **rarefactions** of longitudinal waves on a 'slinky'. Longitudinal waves are sometimes called compression (or pressure) waves.



■ **Figure C2.6** Oscillations of a spring transferring a longitudinal wave

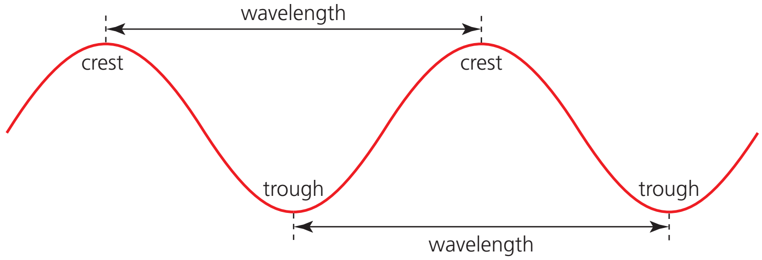
Sound travelling through air is a good example of a longitudinal wave (more details below). Longitudinal compression waves can travel through solids and liquids. Earthquakes are a combination of longitudinal and transverse waves. Transverse mechanical waves cannot travel through gases (or liquids) because of the random nature of molecular movements.

Terms used to describe all types of waves

SYLLABUS CONTENT

- ▶ Wavelength, λ, frequency, f, time period, T, and wave speed, v, applied to wave motion as given by: $v = f\lambda = \dfrac{\lambda}{T}$

The concept of wavelength, λ, is central to the study of waves. See Figure C2.7.



■ **Figure C2.7** One wavelength of a transverse wave

◆ **Wavelength, λ** The distance between two adjacent crests of a wave. More precisely: the shortest distance between two points moving in phase.

◆ **Time period, T** The time taken for one complete wave to pass a point.

◆ **Wave speed, v** The speed at which energy is transferred by a wave.

One **wavelength**, λ, is the shortest distant between two crests, or two troughs. Or the shortest distance between two compressions or rarefactions in a longitudinal wave. More generally, it is defined as the shortest distance between two points moving in phase (SI unit: m).

Displacement, amplitude, time period and frequency have all been discussed before (Topics A.2 and C.1) and are defined in a similar way in the study of waves:

The *amplitude* of a wave is the maximum displacement of the medium from its equilibrium position.

We saw in Topic C.1 that the energy of an oscillation was proportional to its amplitude squared. So, speaking generally, waves with greater amplitude transfer more energy. (We will see in Topic C.3 that the *intensity* of a wave is proportional to its amplitude squared.)

The **time period** of a wave, T, is the time for one oscillation of a particle within the medium, or the time it takes for one complete wave to pass a particular point (unit: second).

The frequency of a wave, f, is the number of oscillations per second of a particle within the medium, or the number of waves to pass a particular point in one second (SI unit: hertz). The following equation is repeated from Topic C.1:

$$f = \frac{1}{T}$$

A wave travels forward one wavelength, λ, every time period, T.

Therefore:

wave speed, $v = \dfrac{\lambda}{T}$

Since $T = 1/f$, we can write:

$$\text{wave speed, } v = f\lambda \left(\text{or } v = \frac{\lambda}{T}\right)$$

WORKED EXAMPLE C2.1

Water waves are passing into a harbour. Five crests are separated by a distance of 9.6 m. An observer notes that 12 waves pass during a time of one minute. Determine:

- **a** the wavelength
- **b** the period
- **c** the frequency
- **d** the speed of the waves.

Answer

a $\lambda = \dfrac{9.6}{4} = 2.4\,\text{m}$

b $T = \dfrac{60}{12} = 5.0\,\text{s}$

c $f = \dfrac{1}{5.0} = 0.20\,\text{Hz}$

d $v = f\lambda = 0.20 \times 2.4 = 0.48\,\text{m}\,\text{s}^{-1}$

1 Consider Figure C2.1. Explain why the amplitude of the waves decreases as they spread away from the central point.

2 Consider Figure C2.2.
   a State the type of wave which is travelling along the rope.
   b If the wave speed is $1.7\,\mathrm{m\,s^{-1}}$, calculate the wavelength produced by shaking the end seven times every 10 seconds.
   c If the rope was replaced by a thinner one, would you predict that the wave speed would increase, or decrease (under the same conditions)? Explain your answer.

3 Describe how the point P on the slinky spring shown in Figure C2.6 moves as the wave passes through it.

4 If you watch waves coming into a beach, you will notice that they get closer to each other.
   a State and explain how their wavelength is changing.
   b Suggest what has caused the waves to change speed.

5 After an earthquake, the first wave to reach a detector 925 km away arrived 149 s later. This type of wave is called a P wave (pressure wave).
   a Suggest whether this is a longitudinal or transverse wave.
   b Calculate the average speed of the wave ($\mathrm{m\,s^{-1}}$).
   c Suggest why your calculation produces an 'average' speed.
   d If the wave had a period of 11.21 s, what was its wavelength?

LINKING QUESTION

- How can the length of a wave be determined using concepts from kinematics?

This question links to understandings in Topic A.1.

Tool 2: Technology

Generate data from models and simulations

Some time after a Primary (longitudinal) wave is received from an earthquake, a different kind of wave will be detected. This is called a Secondary (transverse) wave. If the delay between the detection of the two waves is measured and the speeds of both waves are known, the distance to the original earthquake can be determined.

Set up a spreadsheet that will calculate the distance to the source of an earthquake (dependant variable) for various time delays (independent variable). Assume speeds of waves are $5500\,\mathrm{m\,s^{-1}}$ and $3200\,\mathrm{m\,s^{-1}}$.

Representing waves graphically

Waves can be represented by displacement–position or displacement–time graphs. They both have similar sinusoidal shapes.

Figure C2.8 shows how the displacements of particles (from their mean positions) vary with *distance* from a fixed point (position). x_0 is the amplitude of their oscillations. It may be considered as a 'snapshot' of the wave at one particular moment.

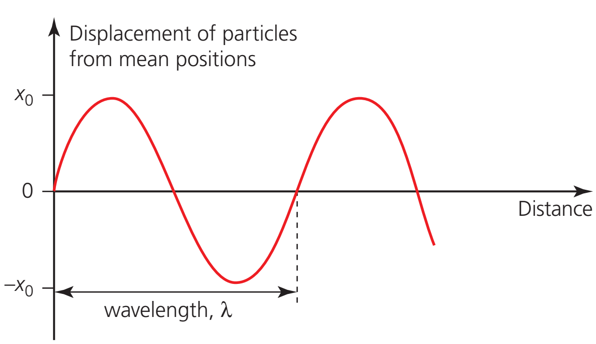
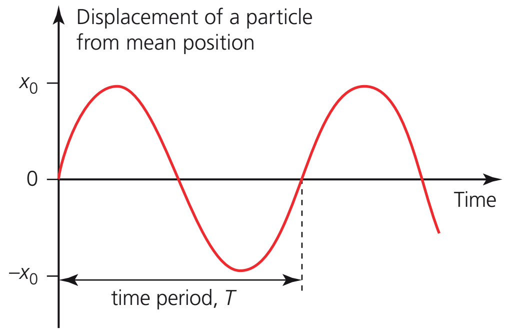


Figure C2.8 Displacement–distance graph for a wave

Figure C2.9 shows how the displacement of a certain particle (from its mean position) varies with time at one precise location. It could be considered as a video of that part of the medium.



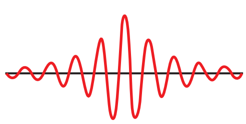
■ **Figure C2.9** Displacement–time graph for a wave

Common mistake

Graphs like these can be used to represent *both* transverse and longitudinal waves. Because of their shape, it is a common mistake to think that they only represent transverse waves. The direction of the displacements shown on the vertical axes of these graphs are not specified, so they could be either

- in the direction of wave travel (longitudinal waves), or
- perpendicular to wave travel (transverse waves).

◆ **Pulse (wave)** A travelling wave of short duration.



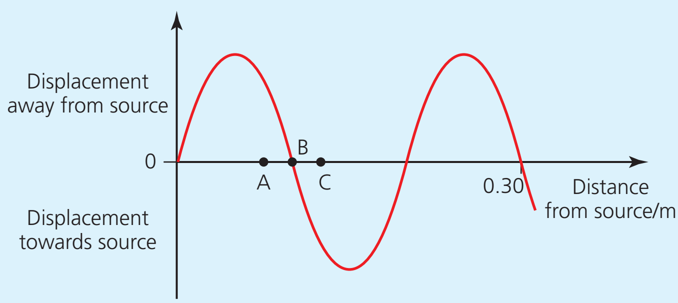
■ **Figure C2.10** Wave pulse

Pulses

A short duration oscillating disturbance passing through a medium may be described as a wave **pulse**. See Figure C2.10 for a simplified representation.

6 Sketch a displacement–time graph for a transverse wave of frequency 4.0 Hz and an amplitude of 2.0 cm. Assume that the wave has its maximum positive displacement at time $t = 0$. Continue the graph for a duration of 0.5 s.

7 Figure C2.11 represents a longitudinal wave.



■ **Figure C2.11** A longitudinal wave

a State its wavelength.

b Describe the instantaneous movement of a particle which is

i at a distance A from the source

ii at a distance C from the source.

c Is there a compression, a rarefaction, or neither, at position B?

8 A wave pulse is made on a water surface by touching it once with a fingertip. Sketch a possible displacement–position graph of the resulting disturbance spreading out on the surface.

Sound waves

SYLLABUS CONTENT

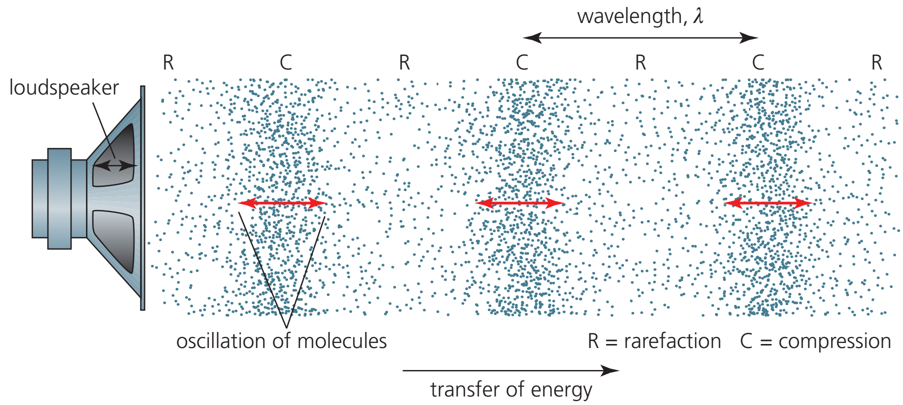
▶ The nature of sound waves.

◆ **Sound** Longitudinal waves in air or other media that are audible to humans.

◆ **Ultrasound** Frequencies of sound above the range that can be heard by humans (approximately 20 kHz).

A vibrating surface will disturb its surroundings and propagate longitudinal waves through the air. The human ear is capable of detecting this type of wave if the frequency falls within a certain range (approximately 20 Hz to 20 kHz). What we hear is called **sound**. Higher frequencies of the same type of wave, which we cannot hear, are called **ultrasound**. (Lower frequencies are called infrasound.)

Figure C2.12 shows how the surface of a loudspeaker can produce longitudinal waves in air. The random arrangement of molecules changes as the wave passes through the air. The compressions and rarefactions result in small periodic changes of air pressure.



■ **Figure C2.12** Arrangement of molecules in air as sound passes through

If the graphs shown in Figure C2.8 and C2.9 represented sound waves, the vertical axes could also be changed to represent variations of air pressure (above and below average air pressure).

Speed of sound

Sound is a mechanical wave involving oscillating particles and, as such, needs a medium to travel through. Sound cannot pass through a vacuum.

Generally, we would expect that sound will travel faster through a medium in which:

- the particles are closer together
- there are stronger forces between the particles.

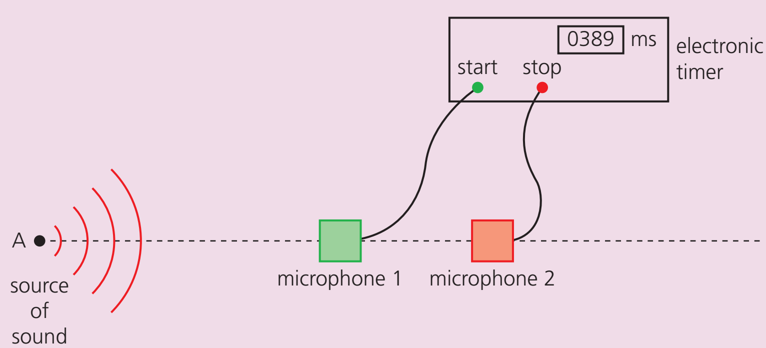
This means that sound usually travels faster in solids than liquids, and slowest in gases, such as air.

The speed of sound in the air around us increases slightly with temperature because then the molecules move faster.

Inquiry 1: Exploring and designing

Designing

Figure C2.13 shows an electronic method for determining the speed of sound.



■ **Figure C2.13** Laboratory experiment to determine the speed of sound

Design an experiment and a valid methodology using this apparatus to determine a value for the speed of sound. Suggest improvements to the design shown in Figure C2.13 so that the speed of sound in air is measured as accurately as possible.

◆ **Pitch** The sensation produced in the human brain by sound of a certain frequency.

◆ **Loudness** A subjective measure of our ears' response to the level of sound received.

◆ **Logarithmic scale (on a graph)** Instead of equal divisions (for example, 1, 2, 3, …), with a logarithmic scale each division increases by a constant multiple (for example, 1, 10, 100, 1000 …).

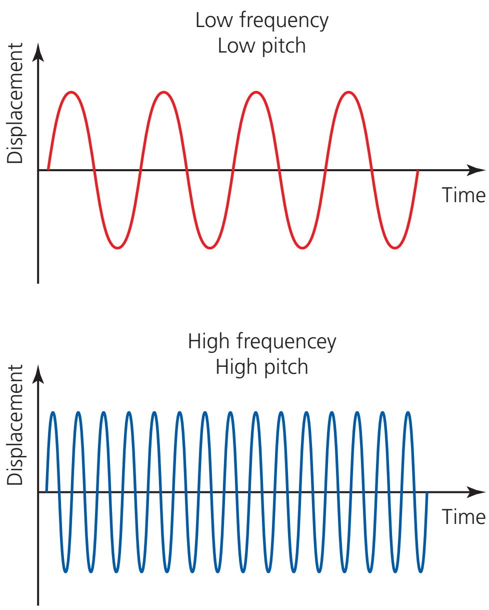
◆ **Decibel** A measure of sound level.

Pitch and loudness of sound

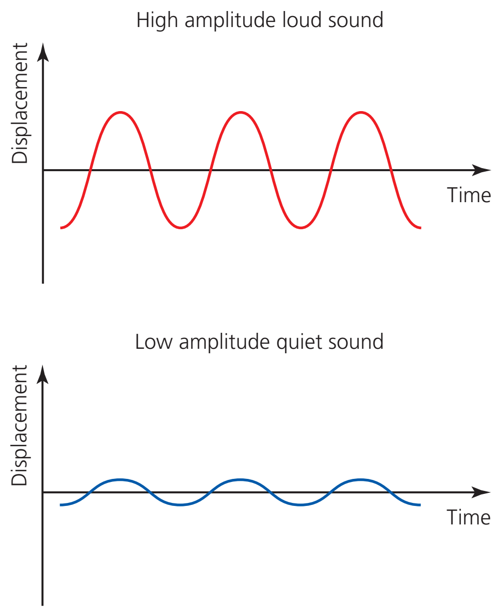
The sounds that arrive at our ears at any one time usually include a range of waves with different frequencies and amplitudes. The oscillations are transferred to our eardrums and our brain interprets them as sounds of different **pitch** and **loudness** (volume).

Figure C2.14 shows pure sound waves of two different frequencies. We describe these sounds arriving at our ears as high pitched and low pitched.

Figure C2.15 shows two waves of the same frequency, but different amplitudes. The wave of larger amplitude transfers more energy and we describe the effects of this as a louder sound.



■ **Figure C2.14** Sounds of different frequency / pitch



■ **Figure C2.15** Sounds of different amplitude / loudness

Tool 3: Mathematics

Logarithmic graphs and power laws

Sound intensity is the power that is carried perpendicularly by sound waves through unit area. It is easily measured by electronic meters, and apps for mobile devices are commonplace.

A normal human ear is capable of detecting sounds with a very wide range of intensities. This makes showing them all on a linear chart impossible. To get over this problem, we use a **logarithmic scale**. On a logarithmic scale (on a chart or a graph) each equal increment represents the fact that the quantity has been multiplied by the same factor (usually 10). As an example, we will consider the **decibel** scale. See Figure C2.16.

A student may wish to investigate the relationship between the intensity of sound (of a constant frequency) and the thickness of material placed between the source and the detector. The student may have no idea what this relationship will be.

Carry out calculations involving logarithmic and exponential functions

Sometimes there is no 'simple' relationship between two variables, or we may have no idea what the relationship may be. So, in general, we can write that the variables x and y are connected by a relationship of the form: $y = kx^p$, where k and p are unknown constants. That is, y is proportional to x to the power p.

Taking logarithms of this equation we get:

$$\log y = (p \times \log x) + \log k$$

Compare this to the equation for a straight line, $y = mx + c$.

If a graph is drawn of $\log y$ against $\log x$, it will have a gradient p and an intercept of $\log k$.

Using this information, a mathematical equation can be written to describe the relationship. Note that logarithms to the base 10 have been used in the above equation, but natural logarithms (ln) could be used instead.

The decibel scale is widely used to compare the intensity of a sound to a reference level. Each additional 10 on the scale represents an increase by a factor of 10 in sound intensity. So, for example, a sound of 50 dB intensity is 10× more intense than a sound of 40 dB. A sound of 60 dB intensity is 100× more intense than a sound of 40 dB, and so on.

Of course, sound intensities decrease with distances from their sources, which are not stated in Figure C2.16, so the numbers should be seen as just a rough guide.

Displaying all parts of the electromagnetic spectrum (later in this topic) is done with a logarithmic scale for the same reason.

Decibels	Example
0	Silence
10	Breathing, ticking watch
20	Rustling leaves, mosquito
30	Whisper
40	Light rain, computer hum
50	Quiet office, refrigerator
60	Normal conversation, air conditioner
70	Shower, toilet flush, dishwasher
80	City traffic, vacuum cleaner
90	Music in headphones, lawnmower
100	Motorcycle, hand drill
110	Rock concert
120	Thunder
130	Stadium crowd noise
140	Aircraft taking off
150	Fighter jet aircraft taking off
160	Gunshot
170	Fireworks
180	Rocket launch

Figure C2.16 An approximate guide to sound levels in decibels

ATL C2A: Research skills

Evaluating information sources for accuracy, bias, credibility and relevance

Find three websites that enable you to check your hearing and follow their instructions. Compare the results and write a short review of your findings.

Were there any differences in the results for each website?

What might account for those differences?

Evaluate the sites in terms of their reliability.

WORKED EXAMPLE C2.2

a Calculate the wavelength of a sound of frequency 196 Hz if the speed of sound in air is $338\,\mathrm{m\,s^{-1}}$.

b If a longitudinal compression wave of the same frequency has a wavelength of 26.1 m in steel, determine the speed of the wave.

c Explain why the wave speed is greater in steel than in air.

Answer

a $\lambda = \frac{v}{f} = \frac{338}{196} = 1.72\,\mathrm{m}$

b $v = f\lambda = 196 \times 26.1 = 5.12 \times 10^3\,\mathrm{m\,s^{-1}}$

c Because the particles are closer together and there are stronger forces between them.

TOK

Knowledge and the knower

- What criteria can we use to distinguish between knowledge, belief and opinion?
- How do we distinguish claims that are contestable from claims that are not?
- How do our interactions with the material world shape our knowledge?

'If a tree falls in a forest and no one is around to hear it, does it make a sound?

■ **Figure C2.17** A fallen tree in a forest

This well-known philosophical question can be answered in different ways, depending on the perspective we take on what is meant by 'sound.'

◆ **Audible range** Range of sound frequencies that can be heard by humans.

If we think of sound only as an effect in the human ear and brain, then the answer is clearly 'no', although there will still be longitudinal waves in the air. If we define sound as a hearable (**audible**) oscillation (regardless of whether anyone is there to hear it), then the answer is 'yes'.

Consider how the knowledge questions above relate to this problem. You may also find the following guiding questions useful:

- Should we believe in things that we have not personally seen / observed / experienced?
- Can we assume that an unobserved event behaves in exactly the same way as an observed event?
- Does observation affect / change the event being observed?
- If the fall of a tree, and any consequential effects, are never observed, is this the same as saying that the tree never fell at all?

9 **a** Sketch a graph to show the air pressure variations (from normal) for a duration of 0.2 s at a certain point through which a sound wave of frequency 100 Hz is passing. Mark one time where there is a compression and one time where there is a rarefaction.
b Determine a value for the period of the wave and show it on the graph.

10 Outline an experiment using hand-held stopwatches to determine a value for the speed of sound in air.

11 Many people know that you can estimate the distance to a storm centre by counting the number of seconds between a flash of lightning and hearing the thunder: about one kilometre for every three seconds. Explain the physics behind this idea.

12 The ultrasound waves used in a medical scanner had a frequency of 9.6 MHz.
a If the wavelength was 0.16 mm, determine the speed of ultrasound waves in the body.
b Suggest three properties of ultrasound that make it useful for obtaining scans from inside the human body.

13 Figure C2.18 shows the use of ultrasound waves (**sonar**) to detect the depth of the ocean below a boat. Waves are produced in a **transducer** and a pulse is directed downwards. The transducer has a diameter of 3 cm. Some wave energy is reflected back from the seabed and then received and detected at the same transducer a short time later. The time delay is used to calculate the depth of the water.

a If the speed of sound waves in sea water is $1520\,\mathrm{m\,s^{-1}}$, calculate the depth of water if the delay between the pulses is 29 ms.

b To limit the spreading of the waves emitted by the transducer it is required that wavelength is much smaller than the size of the transducer. Show that this is true if the waves have a frequency of 214 kHz.

c Suggest why the system uses wave pulses rather than continuous waves.

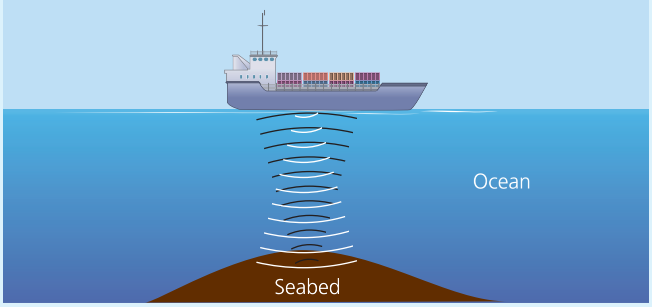


Figure C2.18 Boat using sonar

14 The speed of sound in helium gas is much greater than in air, which is mostly nitrogen (for the same temperature and pressure).
Use knowledge from Topic B.2 to discuss reasons for this difference.

◆ **Sonar** The use of reflected ultrasound waves to locate objects.

◆ **Transducer** Device that converts one form of energy to another. The word is most commonly used with devices that convert to or from changing electrical signals.

◆ **Vacuum** A space without any matter. Also called free space.

◆ **Free space** Place where there is no air (or other matter). Also called a vacuum.

Light waves

In Topic B.1 we described the range of thermal radiations (including light) emitted from various surfaces and 'black bodies' at different temperatures. The true nature of light was not discussed in B.1, but it was a major issue among scientists for hundreds of years.

In the seventeenth century, Isaac Newton believed that a beam of light consisted of particles ('corpuscles'), others thought that light could travel as waves. The wave nature of light was not demonstrated until 1801, when the English physicist Thomas Young showed that light could 'interfere'. This famous experiment and the nature of interference are explained and discussed in Topic C.3.

Light is a transverse electromagnetic wave but, unlike mechanical waves, it does not require a medium to travel through. Light can travel across a **vacuum**, sometimes called '**free space**'. Light travelling from the Sun, through space, to arrive at the Earth is an obvious example.

Visualizing the oscillations of light waves is more difficult than the models of *mechanical* waves that we discussed earlier in this topic. Figure C2.19 shows that light oscillations are high frequency periodic variations in the strength of electric and magnetic fields (which are perpendicular to each other). Electric and magnetic fields are discussed in Theme D.

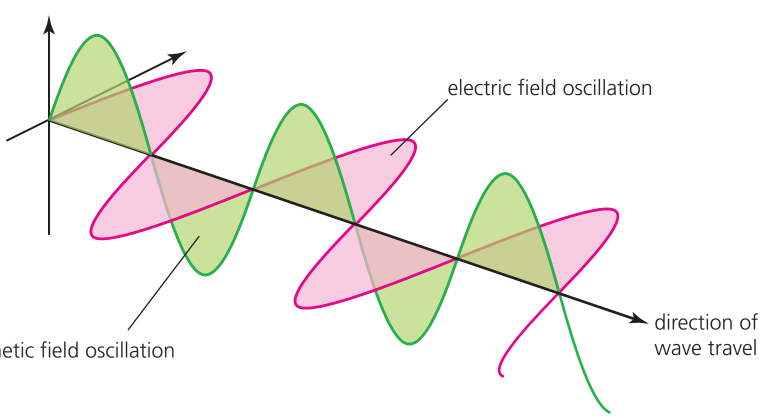


Figure C2.19 Light and other electromagnetic waves are combined electric and magnetic fields

■ **Figure C2.20** Spectrum of visible light

The speed of light in a vacuum is $3.00 \times 10^8\,\mathrm{m\,s^{-1}}$ (more accurately: $299\,792\,458\,\mathrm{m\,s^{-1}}$). It is given the unique symbol '*c*'.

In **transparent** materials light travels at slightly slower speeds. For example, light travels at almost the same speed in air ($299\,970\,500\,\mathrm{m\,s^{-1}}$) as in free space, but at $2.26 \times 10^8\,\mathrm{m\,s^{-1}}$ in water.

The **continuous spectrum** of visible **white light**, from red to violet, is a familiar sight (Figure C2.20). The different colours that we see are created by waves of different frequencies.

Red light has the lowest frequency, violet light has the highest frequency. 'White light' is not a precise scientific term, but it can be assumed to be the same as the light received in the black-body radiation from the Sun on a cloudless day.

◆ **Transparent** Describes a medium that transmits light without scattering or absorption.

◆ **Continuous spectrum** The components of radiation displayed in order of their wavelengths, frequencies or energies (plural: spectra).

◆ **White light** Light which contains all the colours of the visible spectrum with approximate equal intensity.

The fundamental property of a light wave is its *frequency*. If a light wave enters a different medium and then travels more slowly, its frequency cannot change, but its wavelength will decrease ($\lambda = v/f$). However, when we quote data for light waves, it is common to use wavelengths, rather than frequencies. This is because light wavelengths are easier to visualize and measure.

WORKED EXAMPLE C2.3

An orange light has a frequency of $4.96 \times 10^{14}\,\mathrm{Hz}$. Determine its wavelength as it passes through

a air

b a type of glass in which the speed of light has reduced to $1.94 \times 10^8\,\mathrm{m\,s^{-1}}$.

Answer

a $\lambda = \dfrac{c}{f} = \dfrac{3.00 \times 10^8}{4.96 \times 10^{14}} = 6.05 \times 10^{-7}\,\mathrm{m}$

b $\lambda = \dfrac{v}{f} = \dfrac{1.94 \times 10^8}{4.96 \times 10^{14}} = 3.91 \times 10^{-7}\,\mathrm{m}$

Different animals, birds and insects are able to detect different ranges of frequencies. For example, bees are not good at detecting the colour red, but they are able to detect higher frequencies (ultraviolet).

Red light has the longest wavelength in the visible spectrum, approximately $7 \times 10^{-7}\,\mathrm{m}$. Violet has the shortest wavelength, approximately $4 \times 10^{-7}\,\mathrm{m}$.

Use data from the previous paragraphs.

15 a Calculate a typical value for the frequency of red light in air.

b What is the frequency of the same light in glass?

16 Estimate a value for the wavelength of yellow light:

a in air **b** in water.

17 The 'light year' is widely used as a unit of distance in astronomy.
How far does light travel (km) in free space in one year?

18 Briefly outline why light waves are described as electromagnetic waves.

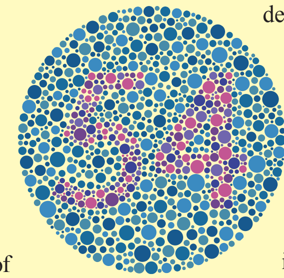
TOK

Knowledge and the knower

- How do our interactions with the material world shape our knowledge?

Perception of colour

We may all agree that light waves have a certain frequency, and whether those waves can be detected in some way by a human eye. There is no ambiguity in that, and most people would agree on the 'colours of the rainbow'. However, how our brains process signals about the light waves detected by our eyes, and how we communicate our impressions of specific colours to other people can be problematic. 'That dress is green' can never be an indisputable scientific fact. 'Colour blindness' (see Figure C2.21) may be an unusual medical condition, but it highlights the fact that human brains can interpret signals in different ways.



■ **Figure C2.21** Test for colour blindness

Added to that, different people, societies and cultures are known to describe colours in different ways. If two people see, or describe, a colour differently, can one be 'right' and the other 'wrong'?

Finally, in terms of physics, it should be pointed out that if you say that a 'dress is green' you probably assume it is being seen under normal lighting conditions, with white light. The colour perceived will change if the lighting is changed. For example, if a red light was used, or it was seen through a yellow filter. Even looking at the dress at night under artificial lighting could change its appearance.

◆ **Ultraviolet** Part of the electromagnetic spectrum which has frequencies just greater than can be detected by human eyes.

◆ **Electromagnetic spectrum** Electromagnetic waves of all possible different frequencies, displayed in order.

◆ **Electromagnetic radiation** Waves which consist of combined oscillating electric and magnetic fields.

Electromagnetic waves

SYLLABUS CONTENT

▶ The nature of electromagnetic waves.

The extent of a visible spectrum such as that seen in Figure C2.20 is limited by:

- the inability of the human eye to detect higher or lower frequencies, and/or
- the ability of any particular source to produce a wider range of frequencies. Light is just a small part of a much wider continuous spectrum.

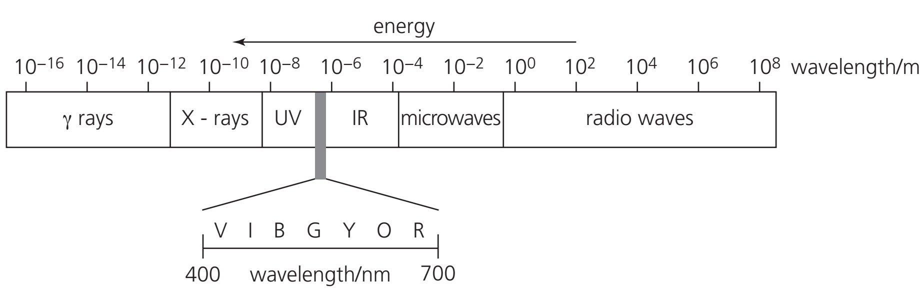
Just beyond the red end of the visible spectrum, there are waves which have longer wavelengths, called infrared. Infrared radiation was discussed in Topic B.1. Just beyond the violet end of the visible spectrum, there are waves of shorter wavelength, called **ultraviolet**.

The complete range (spectrum) of possible electromagnetic wavelengths extends from more than 100 000 km to less than 10^{-16} m. They have different origins and no single source produces all of these waves.

All electromagnetic waves travel at the same speed in vacuum, $c = 3.00 \times 10^8\,\mathrm{m\,s^{-1}}$.

They are all composed of oscillating electric and magnetic fields (Figure C2.19).

Together they are known as the **electromagnetic spectrum** (Figure C2.22). They are also described as **electromagnetic radiation**.



■ **Figure C2.22** Electromagnetic spectrum

'Energy' refers to the energy carried by *photons* of the radiation, as explained later in the course.

Common mistake

Remember that the spectrum is continuous and the boundaries chosen between different named sections are somewhat arbitrary.

LINKING QUESTION

- How can light be modelled as an electromagnetic wave?

This question links to understandings in Topic D.2.

Inquiry 1: Exploring and designing

Exploring

Select sufficient and relevant sources of information

After deciding on a general area of interest, for an investigation you will often need to select and research other sources of information for background knowledge and any physics needed which is beyond the IB course (if appropriate). Your teachers should be an excellent source of advice and information and, obviously, the internet has multiple sources (of various quality). Physics books, science magazines and books from libraries can all be sources of information and inspiration.

Example 1: If you wish to investigate the effect that water vapour in the air has on the rate of evaporation from a water surface, you will need to learn about *humidity*.

Example 2: If you wish to investigate the world-wide use of solar heating of water, you will need to learn about the hours of sunlight in different locations, the variation in altitude of the Sun, comparative costs and so on.

Your intended investigation could be both interesting and unusual, but it needs to be realistic in terms of the apparatus that is available in your school, and the time available. So, it may be wise to check with teachers about whether an intended investigation is sensible under the circumstances.

Any sources of information should be acknowledged in the investigation report, including those which were researched but not used (with a reason given).

Task: Apart from sources on Earth, waves from all parts of the electromagnetic spectrum arrive at Earth from space. Use the internet to gather information about the origins of these waves and to what extent they are able to pass through the atmosphere and reach the Earth's surface.

The list in Table C2.1 shows some origins of electromagnetic waves and a selection of their uses.

Table C2.1 The different sections of the electromagnetic spectrum

Name	Typical wavelength / m	Origins (all are received from Outer Space)	Some common uses
radio waves	10^{2}	electronic circuits / aerials	communications, radio, television
microwaves	10^{-2}	electronic circuits / aerials	communications, mobile phones, ovens, radar
infrared (IR)	10^{-5}	everything emits IR but hotter objects emit much more than cooler objects	lasers, heating, cooking, medical treatments, remote controls
visible light	5×10^{-7}	very hot objects, light bulbs, the Sun	vision, lighting, lasers
ultraviolet (UV)	10^{-8}	the Sun, UV lamps	fluorescence
X-rays	10^{-11}	X-ray tubes	medical diagnosis and treatment, investigating the structure of matter
gamma rays	10^{-13}	radioactive materials	medical diagnosis and treatment, sterilization of medical equipment

LINKING QUESTION

- How are waves used in technology to improve society? (NOS)

This question links to understandings in Topics C.3, C.4, C.5, D.2, D.3 and D.4.

◆ **Quantum physics** Study of matter and energy at the subatomic scale. At this level quantities are quantized.

Top tip!

The fact that electromagnetic waves have some properties that could not be explained satisfactorily by their wave nature had very important consequences. A new 'particle' model for light, introduced at the start of the twentieth century, was the beginning of **quantum physics**. This is introduced in Topic E.2.

LINKING QUESTIONS

- How are electromagnetic waves able to travel through a vacuum?
- Can the wave model inform the understanding of quantum mechanics? (NOS)

These questions link to understandings in Topic E.2.

Nature of science: Experiments

Pure research

The first artificial electromagnetic (radio) waves

Heinrich Hertz (Figure C2.23) was the first to produce and detect artificial electromagnetic waves (1887 in Karlsruhe in Germany). He used high voltage electrical sparks. The electrical currents in sparks involve the necessary high frequency oscillating electric and magnetic fields. Although the distance involved was very small, it was the start of modern wireless communication. It was left to others (such as Guglielmo Marconi) to develop the technology for transmission over longer and longer distances – and then to design techniques to modify the amplitude, frequency or phase of the radio waves to transfer information, such as speech.

Figure C2.23 Heinrich Hertz

Tragically, Hertz died at the age of 36 in 1894. This was long before the far-reaching consequences of his discovery had been exploited.

Hertz had been trying to provide evidence for the electromagnetism theories of James Clerk Maxwell, and he has been widely quoted as saying that his discovery was 'of no use whatsoever'. He was not alone in that opinion at the time.

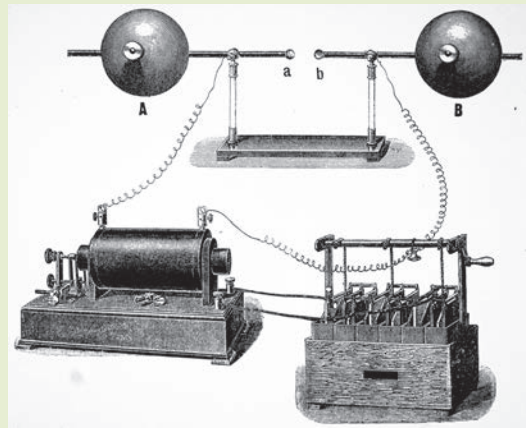


Figure C2.24 Hertz's apparatus for the first artificial production of electromagnetic waves

'Pure research' is about extending knowledge and confirming theories, it is not about solving practical problems. But there are many historical examples of pure research leading to unexpected benefits of major significance – such as radio communication.

Of course, a large number of examples of pure research have *not* produced any worthwhile gains for society. An often-asked question is 'should governments spend large amounts of money on open-ended research which has no obvious benefits (at that time)?'

In terms of laboratory investigations that you might carry out as a student: the common expectation is that they should have an 'aim', which may be answering a specific question. But maybe that is too restrictive?

19 Determine the frequency (in MHz) of a gamma ray which has a wavelength of 4.1×10^{-12} m.

20 A mobile phone network uses electromagnetic waves of frequency 1200 MHz.

- **a** Calculate their wavelength.
- **b** State which part of the electromagnetic spectrum contains these waves.
- **c** Use the internet to find out the frequency used in microwave ovens.
- **d** Suggest why our bodies are not warmed up by using mobile phones.

21 As you are reading this, which types of electromagnetic radiation are there in the room?

22 a State which types of electromagnetic radiation are considered to be dangerous.

- **b** What do they have in common?

23 Outline what properties of X-rays make them so useful in hospitals.

24 a Calculate how long it takes for a Bluetooth signal to travel from a mobile phone to a speaker which is 4.7 m away.

- **b** How much time (to the nearest minute) does it take light to reach the Earth from the Sun?
- **c i** How much time does it take a radio signal to travel to Mars from Earth?
  - **ii** Explain why your answer is uncertain. (Use the internet to obtain relevant data.)

C.3 Wave phenomena

Guiding questions

- How are observations of wave behaviours at a boundary between different media explained?
- How is the behaviour of waves passing through apertures represented?
- What happens when two waves meet at a point in space?

What are the basic behaviours of all waves?

- Reflection
- Refraction
- Diffraction
- Interference

◆ **Ripple tank** A tank of shallow water used for investigating wave properties.

◆ **Wavefront** A line connecting adjacent points moving in phase (for example, crests). Wavefronts are one wavelength apart and perpendicular to the rays that represent them.

Each of these properties will be discussed in this topic. But first we need to consider how we can represent travelling waves in two dimensions on paper, or screens.

Wavefronts and rays

SYLLABUS CONTENT

▶ Waves travelling in two and three dimensions can be described through the concepts of wavefronts and rays.

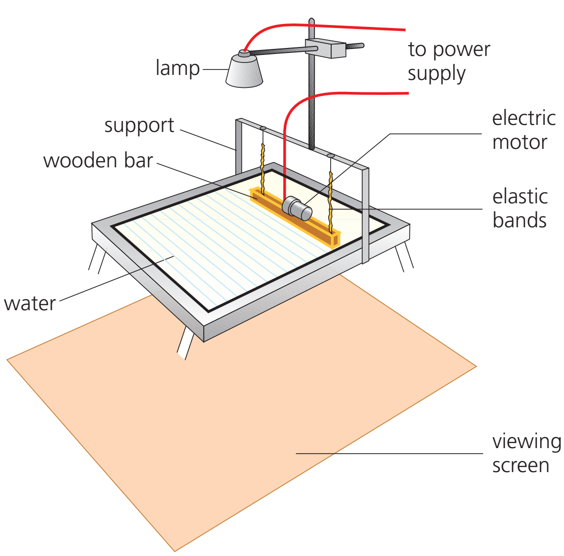


Figure C3.1 A ripple tank is used to investigate wave behaviour

Waves in two dimensions

Figure C3.1 shows a **ripple tank**: a common arrangement used to observe the behaviour of waves. Small waves (ripples) can be made by touching the surface of shallow water at a point, or with a wooden bar. Usually, a motor is attached to the bar to make it vibrate and produce continuous parallel waves at various frequencies. The light above enables the moving waves to be seen on the screen below the tank. (A *stroboscope* is often used to make the waves appear stationary.)

The 'waves' seen on the screen show the positions of wave *crests*. These lines are called **wavefronts**. They are one wavelength apart. More precisely:

A wavefront is a line joining neighbouring points moving in phase with each other.

The blue lines seen in Figure C3.2 represent the pattern produced by regular disturbances of the water at point P. The waves are spreading with equal speed in all directions (in two dimensions), so that the pattern is circular. Wave speed depends on the depth of the water, which is usually constant if the tank is horizontal.

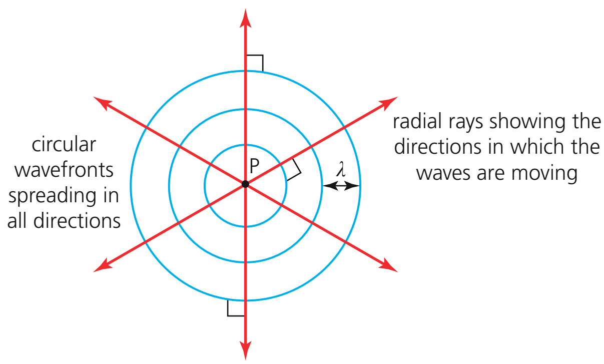
◆ **Ray** A line showing the direction in which a wave is transferring energy.

◆ **Radial** Diverging in straight lines from a point.

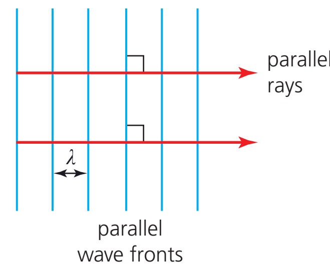
The red lines with arrows in Figure C3.2 are called **rays**. These rays can be described as **radial**, meaning that they are spreading out in straight lines from a point. Radial rays represent circular (or spherical) wavefronts.

Rays are lines showing the direction in which wavefronts are moving and energy is transferred. Rays and wavefronts are always perpendicular to each other.

The wavefronts that we will be considering in this course are either circular (or spherical), see Figure C3.2, or the other common possibility: parallel wavefronts, as shown in Figure C3.3. Parallel wavefronts can be made on the water surface in a ripple tank by using a wooden bar.



■ **Figure C3.2** Circular wavefronts and radial rays spreading from a point source



■ **Figure C3.3** Parallel wavefronts and parallel rays that are not spreading out

The movement of parallel wavefronts is represented by parallel rays.

Waves in three dimensions

◆ **Plane waves** Waves travelling in three dimensions with parallel wavefronts, which can be represented by parallel rays.

◆ **Visualization** Helping understanding by using images (mental or graphic).

Waves spreading from a point with constant speed in three dimensions can be represented by spherical wavefronts. If the source of waves is a very long way away (compared to the wavelength) the wavefronts will be (almost) straight and parallel. Such waves are described as **plane waves**. A common example: light waves from a distant point. The Sun is an obvious source of plane waves.

TOK

Knowledge and the knower, The natural sciences

- How do our interactions with the material world shape our knowledge?
- What is the role of imagination and intuition in the creation of hypotheses in the natural sciences?

Different ways of describing the same thing

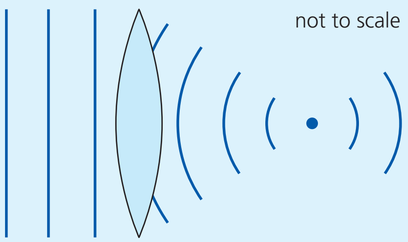
We can use either the concept of wavefronts, or the concept of rays, to describe the movement of the same waves, but they are not normally used at the same time.

The use of wavefronts to describe waves on water surfaces, which are easily *visible*, is easily understood, but the associated concept of 'rays' seems unnecessary. However, when describing another wave phenomenon, the *invisible* passage of light through the air, why do we usually prefer the **visualization** of rays, although they have no physical reality? How might such visualizations extend, affect, or perhaps limit our understanding of the natural world?

1 **a** Describe and explain how parallel wavefronts in a ripple tank appear if they are moving perpendicularly into shallower water.

b Discuss how the wavefronts from a point source would appear if the tank was raised on one side. Explain your answer.

2 Figure C3.4 represents some light wavefronts passing from left to right through a lens.



■ **Figure C3.4** Light wavefronts passing from left to right through a lens

a Explain how you know that the source of light is a long way away.

b Make a sketch of the lens and show the path of five rays to represent the movement of the wavefronts.

c State which word we use to describe the effect of the lens on the light rays.

3 Figure C3.5 shows ocean waves as they approach the coast.
Suggest possible reasons why the separation and direction of the waves change.

■ **Figure C3.5** Ocean waves refracting (and diffracting) as they approach a beach

Transmission, absorption and scattering of waves

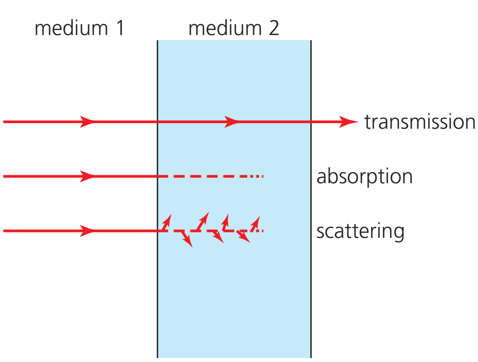
SYLLABUS CONTENT

▶ Wave behaviour at boundaries in terms of transmission.

◆ **Transmission** Passage through a medium without absorption or scattering.

◆ **Opaque** Unable to transmit light (or other forms of energy).

The process in which waves are able to travel through a medium is described as wave **transmission**. For example, light and sound can be transmitted through air and water. During the transmission of waves, there is often *absorption* of energy: some or all of the wave energy is transferred to internal energy within the medium. Waves may also be randomly misdirected by interactions with irregularities within the medium. This is called scattering. Figure C3.6 illustrates these terms: waves are transmitted by medium 1, then enter medium 2, where they are each either transmitted, absorbed or scattered. In reality, all three processes can occur with the same waves.



■ **Figure C3.6** Transmission, absorption and scattering

A medium through which light can be transmitted, and through which we can see clearly, is described as being *transparent*. A medium through which light cannot be transmitted is described as **opaque**.

Wave power, intensity and amplitude

As waves spread out, and/or their energy is dissipated, the power that they transfer is reduced. We usually describe this as a reduction of wave intensity, a concept that was introduced in Topic B.1, and is defined again here:

$$\text{intensity, } I = \frac{P}{A} \quad \text{SI Unit: W m}^{-2}$$

Common mistake

Some books use the symbol A to represent amplitude, but this can cause confusion with the symbol for area. We will use the symbol A to represent area and the symbol x_0 to represent amplitude.

Speaking generally, we know that the energy transferred by a wave is proportional to its amplitude squared (Topic C.2). More precisely:

$$\text{intensity} \propto \text{amplitude}^2$$

Waves spreading from a point without absorption

There are reasons, discussed above, why waves may lose energy during transmission through a medium, but, if the waves were *spreading* out from a point source (that is, they are not plane waves), their intensity will decrease for that reason alone, without any absorption.

In two dimensions (surface waves)

As surface waves spread away from a point source the wavefronts will extend over greater and greater lengths. For example, if a circular wavefront increases its distance from its centre from r to $2r$, then its circumference will increase from $2\pi r$ to $4\pi r$. See Figure C3.7.

In each spreading wavefront the same amount of energy is spread over a greater length, so that the wave amplitude will decrease. If surface waves are a great distance from their source, the wavefronts will become almost parallel to each other, so that no loss of energy / power may be noticeable.

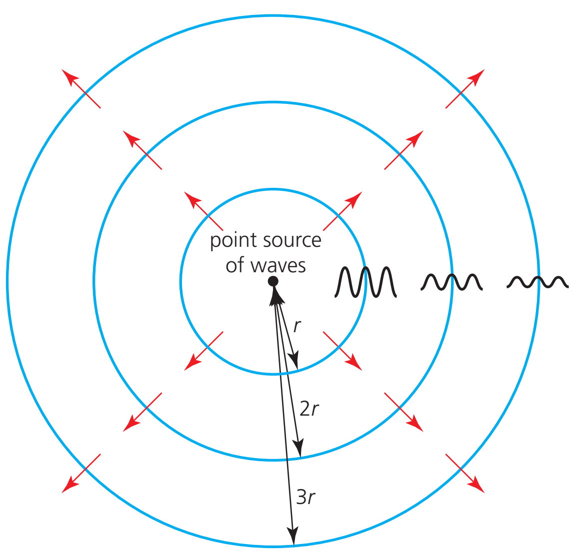
In three dimensions (such as light and sound)

We discussed the spreading of light and infrared waves from the Sun in Topic B.2. Refer back to Figure B2.1 in that chapter. The intensity of any waves spreading radially in three dimensions, without absorption, from a point source follows an inverse square relationship, which is repeated here:

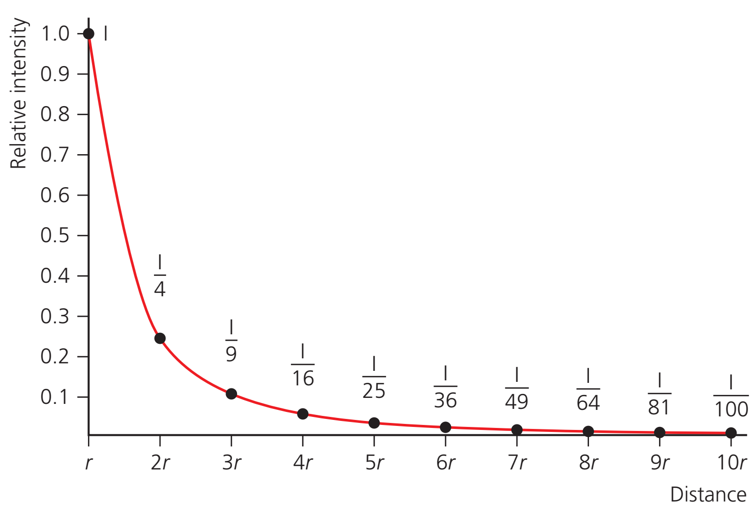
$$\text{intensity, } I \propto \frac{1}{\text{distance}^2}$$

Figure C3.8 represents this kind of relationship in graphical form.

But note that, as before, if the waves are a great distance from their point source, the wavefronts will become almost parallel / plane waves, so that no loss of energy/power due to spreading may be noticeable.



■ **Figure C3.7** Amplitude of circular wavefronts decreases with the distance from source



■ **Figure C3.8** Inverse square relationship

Tool 3: Mathematics

Plot linear and non-linear graphs showing the relationship between two variables

Check the data shown in Table C3.1 to determine if there is an inverse square relationship between x and y:

- numerically
- graphically.

■ **Table C3.1**

x	y
1.34	9
0.96	17
0.81	24
0.70	32
0.64	38
0.59	45
0.55	52
0.52	58
0.49	66

4 Light and infrared radiation arriving perpendicularly on a solar panel have a total intensity of $780\,W\,m^{-2}$.

a If the panel has dimensions of $50 \times 80\,cm$, calculate the incident power.

b Explain why this power will change during the course of the day.

5 A girl is reading a book at night using the light from a single lamp, which may be assumed to be a point source. If the lamp was originally 1.80 m away from the book, show that the intensity of the light on the book doubles if it is moved 0.53 m closer to the lamp.

6 Figure C3.9 shows some typical aerials used for transmitting (and receiving) signals to mobile phones.

■ **Figure C3.9** Typical aerials for transmitting mobile phone signals

a Explain why it desirable that the waves from these aerials are *not* emitted equally in all directions.

b Suggest which feature of the aerials limits the vertical spreading of the waves.

7 The radiation from the Sun which reaches the top of the Earth's atmosphere has an intensity of $1360\,W\,m^{-2}$ (see Topic B.1). It is approximately 40% visible light, 50% infrared and 10% ultraviolet.
The intensity reaching the Earth's surface is approximately $1000\,W\,m^{-2}$.
The approximate proportions reaching the Earth's surface are 44% visible light, 53% infrared and 3% ultraviolet.
Use this data to estimate the percentages of these three radiations which are:

a transmitted by the Earth's atmosphere

b absorbed / scattered by the Earth's atmosphere.

8 Why is the sky blue? (Research on the internet if necessary.)

9 It is considered to be a health risk to expose our ears to sounds of intensity greater than $10\,mW\,m^{-2}$ for more than a few minutes.

a Calculate the total power received on an eardrum of area $0.48\,cm^2$ from this intensity.

b If the sound intensity 2.10 m from a loudspeaker at a rock concert was $0.44\,W\,m^{-2}$, estimate how far away you would need to be in order to reduce the intensity to $100\,mW\,m^{-2}$.

c State two assumptions that you made in answering **b**. Discuss whether these assumptions are reasonable.

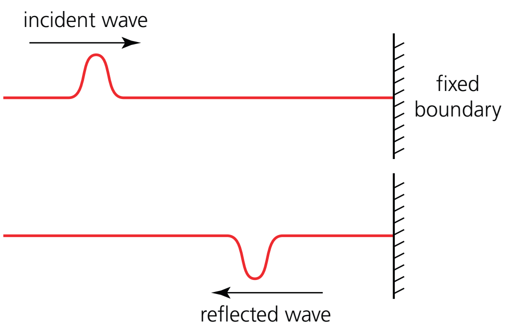
Reflection of waves and rays

◆ **Reflection (waves)** Change of direction that occurs when waves meet a boundary between two media such that the waves return into the medium from which they came.

SYLLABUS CONTENT

▶ Wave behaviour at boundaries in terms of reflection.

When a wave meets a boundary between different media some, or all, of the wave energy will be re-directed back into the first medium. This is called **reflection**.

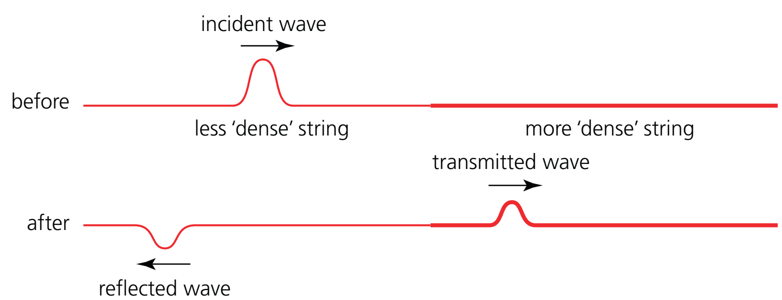


■ **Figure C3.10** Reflection of a pulse off a fixed boundary

To develop understanding, we will first consider the simplest possible example: a single wave pulse on a rope or string, being reflected at a fixed boundary, as shown in Figure C3.10. A wave travelling towards a boundary is called an *incident wave*. The reflected wave is inverted from this type of boundary: there is a phase change of $\pi/2$.

If the incident waves are continuous, they may combine with the reflected waves to produce a standing wave, as discussed in the next topic, C.4.

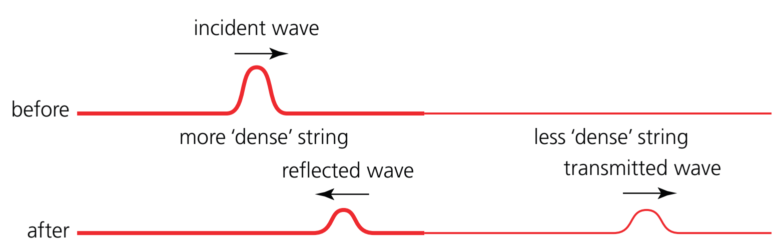
Apart from fixed boundaries, where there is no possibility of transmission, waves may also reflect from a boundary where some transmission occurs. Usually, the wave will have different speeds in the two different media.



■ **Figure C3.11** A pulse travelling into a 'denser' medium

In Figure C3.11 the 'denser' rope has a greater mass per unit length, so that the wave travels more slowly through it. The reflected wave is still inverted. The energy is shared between the reflected wave and the transmitted wave, so that both amplitudes are less than that of the incident wave.

Figure C3.12 shows the situation in which a wave pulse meets a boundary with a medium in which its speed would increase. Note that there is no phase change.



■ **Figure C3.12** Longitudinal waves and pulses behave in a similar way to transverse waves

We will now extend the discussion of reflection to two and three dimensions.

■ **Figure C3.13** Light reflected off and being transmitted by a window

Top tip!

When drawing wavefronts diagrams, make sure that they are *continuous* at the boundaries.

◆ **Ray diagrams** Drawings that represent the directions of different waves or particles as they pass through a system.

◆ **Incidence, angle** The angle between an incident ray and the normal.

◆ **Angle of reflection (rays)** Angle between a reflected ray and the normal.

◆ **Reflection, law of** Angle of incidence = angle of reflection.

Reflected wavefronts

When *parallel* wavefronts meet a plane (flat) boundary, some or all of them will be reflected so that the angle that the incident wavefront makes with the boundary is equal to the angle that the reflected wavefront makes with the boundary. See Figure C3.14.

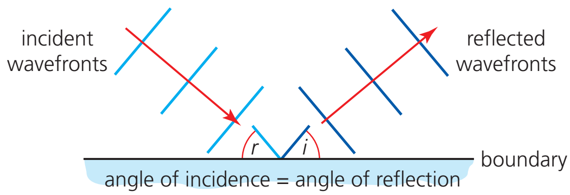


Figure C3.14 Reflection of parallel wavefronts from a straight boundary

When discussing the reflection of light, instead of wavefronts, it is more common to refer to rays, as shown in the **ray diagram** of Figure C3.15. A normal is a line perpendicular to the surface (at the point of incidence). The **angle of incidence** and the **angle of reflection** are measured between the ray and the normal (not the surface).

The **law of reflection**:

> The angle of incidence equals the angle of reflection.

Figure C3.16 shows radial light rays spreading from a point source (an 'object'). When they strike the plane mirror, the law of reflection can be used to determine the directions of the reflected rays. An eye looking into the mirror will see an image located as far behind the mirror as the object is in front.

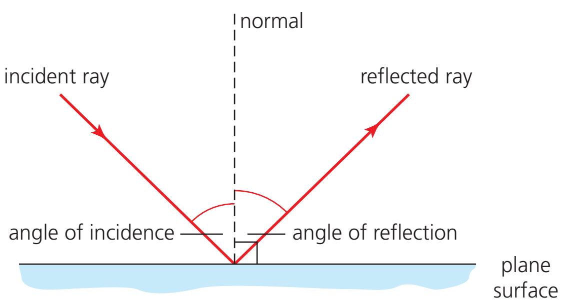


Figure C3.15 Reflection of rays from a plane surface

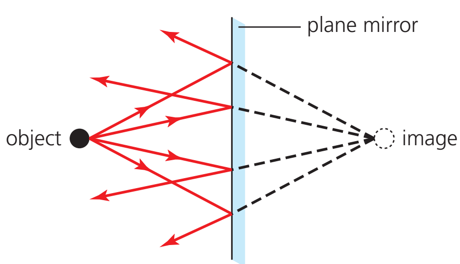


Figure C3.16 Image formed by reflection of rays in a plane mirror

Figure C3.17 represents the same situation using wavefronts instead of rays. The reflected waves appear to come from a point as far behind the reflecting surface as the actual source of the waves (the 'object') is in front.

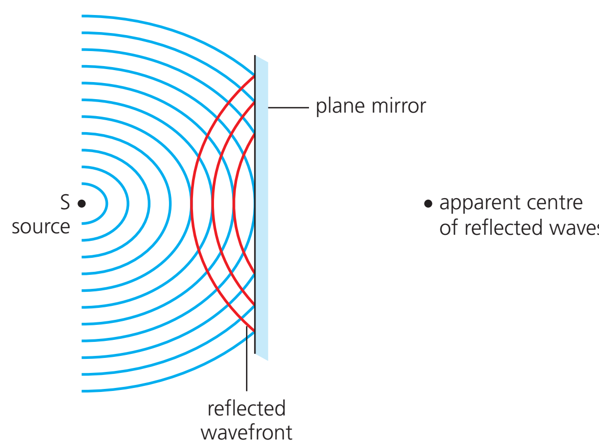


Figure C3.17 Reflection of circular wavefronts by a plane surface.

10 Figure C3.18 shows an idealized pulse on a string approaching a fixed end at a speed of $100\,\text{cm}\,\text{s}^{-1}$. Draw a sketch to show the position of this pulse 0.8 s later.

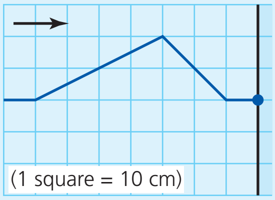


Figure C3.18 An idealized pulse on a string

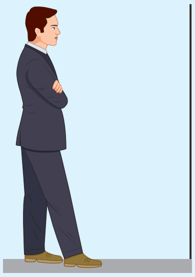


Figure C3.19 A man looking at a wall

11 Figure C3.19 shows a man looking at a wall. Make a rough copy and add to the wall the smallest mirror that would enable the man to see both the top of his head and his feet. Include light rays that explain your positioning of the mirror.

12 Predict if there will be a phase change when light waves reflect off a glass surface. Explain your answer.

13 Figure C3.20 shows light rays from the Sun being reflected to a focus. Describe the shape of the wavefronts

a arriving from the Sun

b being reflected to the focus.

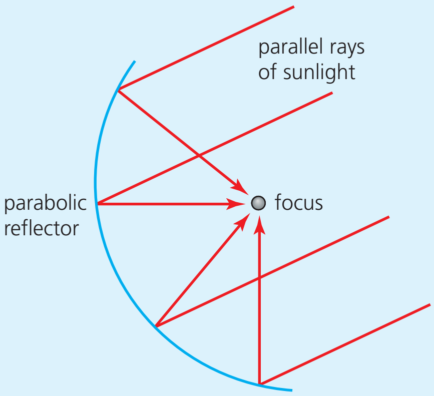


Figure C3.20 Light rays from the Sun being reflected to a focus

Inquiry 1: Exploring and designing

Exploring

Sound reflections in large rooms

Sound reflects well off hard surfaces like walls, whereas soft surfaces, such as curtains, carpets, cushions and clothes, tend to absorb sound. A sound that reaches our ears may be quite different from the sound that was emitted from the source because of the many and various reflections it may have undergone. Because of this, singing in the shower will sound very different from singing outdoors or singing in a furnished room. In a large room designed for listening to music (such as an auditorium, Figure C3.21), sounds travel a long way between reflections.

Figure C3.21 A large auditorium designed for effective transfer of sound to the audience

Since it is the reflections that are responsible for most of the absorption of the sound waves, it will take a longer time for a particular sound to fall to a level that we cannot hear. This effect is called *reverberation*. The longer reverberation times of bigger rooms mean that a listener may still be able to hear reverberation from a previous sound at the same time as a new sound is received. That is, there will be some 'overlapping' of sounds. Reflections of sounds off the walls, floor and ceiling are also an important factor when music is being produced in a recording studio, although some effects can be added or removed electronically after the original sound has been recorded.

Does your school or college have a performance space, such as a hall or a theatre, where you can make sounds and listen to them carefully when they arrive back at your ears? Or is there a performance space nearby you could visit?

1. Research online using search terms such as 'sound reverberation' to find out how professional designers adapt and design performance spaces to change reverberation. Then investigate your chosen performance space, inspecting it for installations which affect sound reverberation.
2. Discuss and suggest what measurements you could make to test the reverberation in the space.
3. Are there any improvements that could be made? State these and explain your reasoning.

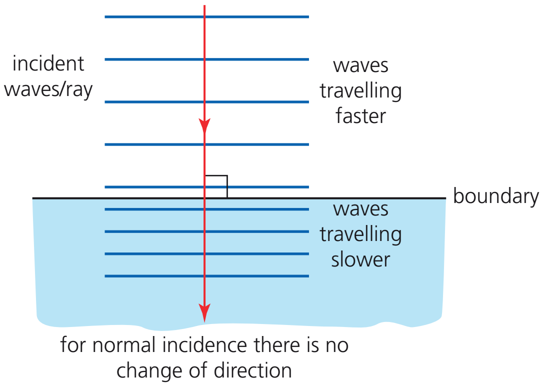


Figure C3.22 Waves slowing down as they enter a different medium

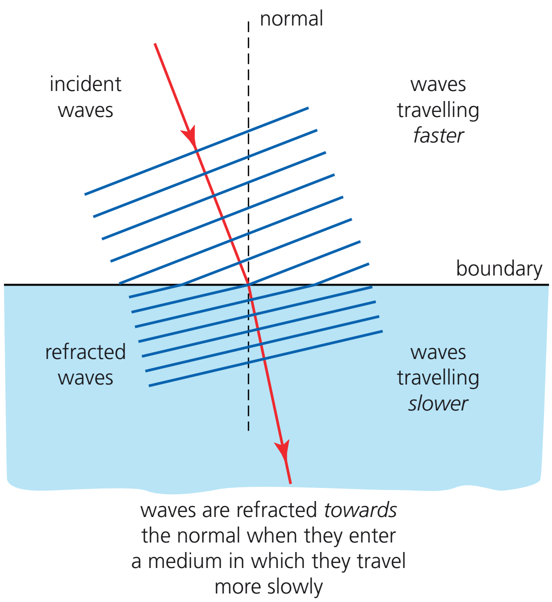


Figure C3.23 Waves refracting as they enter a denser medium

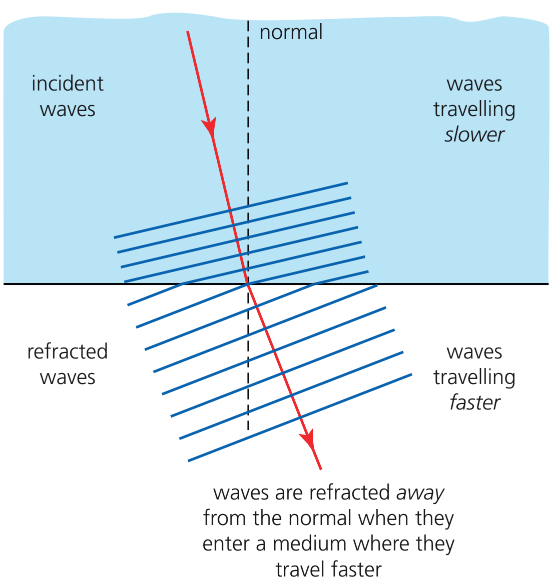


Figure C3.24 Waves refracting as they enter a less dense medium

Refraction of waves

SYLLABUS CONTENT

- Wave behaviour at boundaries in terms of refraction.
- Wavefront-ray diagrams showing refraction.
- Snell's law, critical angle and total internal reflection.
- Snell's law as given by:

$$\frac{n_1}{n_2} = \frac{\sin\theta_2}{\sin\theta_1} = \frac{v_2}{v_1}$$

where n is the refractive index and θ is the angle between the normal and the ray.

Waves will usually change speed when they travel into a different medium. Such changes of speed may result in a change of direction of the wave.

The speed of waves on the surface of water generally decreases as they move into shallower water.

Figure C3.22 shows parallel wavefronts arriving at a medium in which they travel more slowly. The wavefronts are parallel to the boundary and the ray representing the wave motion is perpendicular to the boundary.

In this case there is no change of direction but because the waves are travelling more slowly, their wavelength decreases, although their frequency is unchanged (consider $v = f\lambda$). Now consider what happens if the wavefronts are not parallel to the boundary, as in Figure C3.23.

Different parts of the same wavefront reach the boundary at different times and consequently, they change speed at different times. There is a resulting change of direction which is called **refraction**. The greater the change of speed, the greater the change of direction.

When waves enter a medium in which they travel more slowly, they are refracted towards the normal (assuming that the wavefronts are not parallel to the boundary).

Conversely, when waves enter a medium in which they travel faster, they are refracted away from the normal. This is shown in Figure C3.24; note that this is similar to Figure C3.23, but with the waves travelling in the opposite direction.

The refraction of light is a familiar topic in the study of physics, especially in optics work on lenses and **prisms**, but all waves tend to refract when their speed changes. Often this is a sudden change at a boundary between media, but it can also be a gradual, or irregular change. In Figure C3.25 differences in gas density produce irregular refraction and a blurred image above the fire. Stars twinkle in the night sky because of refraction in a shifting atmosphere.

◆ **Refraction** Change of direction that can occur when a wave changes speed (most commonly when light passes through a boundary between two different media).

◆ **Prism** A regularly shaped piece of transparent material (such as glass) with flat surfaces, which is used to refract and disperse light.

■ **Figure C3.25** Irregular refraction over a fire

Refractive index

The amount of refraction that occurs when a light wave passes from one medium into another depends on the change of speed involved. This is represented numerically by the **refractive index, *n*,** of a medium:

$$\text{refractive index of a medium} = \frac{\text{speed of light in vacuum}}{\text{speed of light in the medium}}$$

$$n = \frac{c}{v}$$

Refractive index is a ratio of speeds, so it does not have a unit.

For example, the speed of light in water is 2.26×10^8, so that:

$$\text{refractive index of water} = \frac{3.00 \times 10^8}{2.26 \times 10^8} = 1.33$$

◆ **Refractive index, *n*** The ratio of the speed of waves in vacuum (or air) to the speed of waves in a given medium.

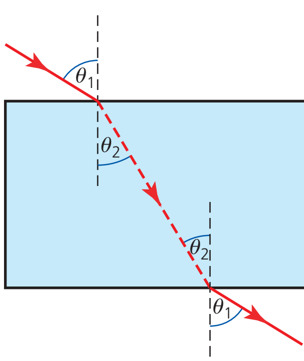
In practice, we will not usually be concerned with light travelling into, or out of a vacuum, but the speed of light in air is almost identical to the speed in a vacuum. This means that we can use the same refractive indices for light passing from air into, or out of, a medium.

■ **Table C3.2** Refractive indices

Medium	Speed of light / $10^8\,\text{m s}^{-1}$	Refractive index
diamond	1.2	2.4
glass	1.8–2.0	1.5–1.7
plastic	1.9–2.3	1.3–1.6
lens in human eye	2.1	1.4
pure water	2.26	1.3
air	2.997	1.0
vacuum	2.998	-

As has been stated in Topic C.2, waves from all parts of the electromagnetic spectrum travel at exactly the same speed ($c = 3.00 \times 10^8\,\text{m s}^{-1}$) in free space (vacuum), but they all travel slower in other media. However, there are also some very small differences in the wave speeds in the same medium, for example, yellow light travels *very slightly* faster than green light in glass. For many applications this is not important, but it can result in the *dispersion* of white light into a spectrum (see later).

Snell's law

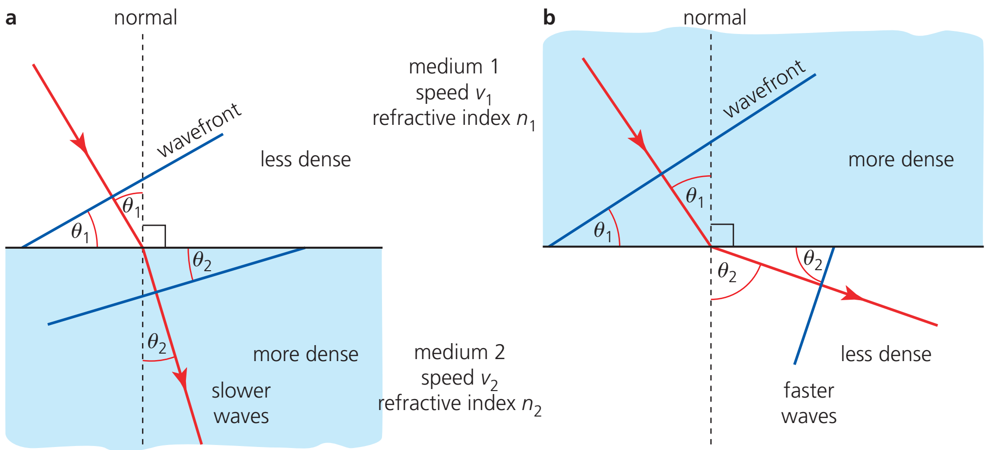


■ **Figure C3.26** Light passing through a parallel-sided transparent block.

When we observe the refraction of light, the only direct measurements that can be made are of the angles involved. Figure C3.26 shows the paths of rays that could be seen in a standard experiment with a parallel-sided glass, or plastic, block. Two measurements are possible: θ_1, the angle that the incident ray makes with the normal and θ_2, the angle that the refracted ray makes with the normal. The rays are refracted *towards* the normal because the speed of light in glass is less than the speed of light in air. Because the block is parallel-sided, the same angles occur again as the ray emerges from the block. A ray of light entering the lower surface (as shown) will follow exactly the same path as a ray entering the upper surface, but in the opposite direction.

Some of the incident light will be reflected off the glass surface, but this has not been shown in the diagram.

Figure C3.27 shows the situation in a little more detail and includes wavefronts.



■ **Figure C3.27** Light rays being refracted **a** towards the normal and **b** away from the normal

The Dutch scientist, Willebrord Snellius, was the first to show how the angles were related to the speeds of the waves and the refractive indices, as light passes from a medium of refractive index n_1, where its speed is v_1, to a medium of refractive index n_2, where the wave speed is v_2 (as seen in Figure C3.26).

Snell's law:

$$\frac{n_1}{n_2} = \frac{\sin\theta_2}{\sin\theta_1} = \frac{v_2}{v_1}$$

◆ **Snell's law (of refraction)** Connects the sines of the angles of incidence and refraction to the refractive indices in the two media (or the wave speeds). $\frac{n_1}{n_2} = \frac{\sin\theta_2}{\sin\theta_1} = \frac{v_2}{v_1}$.

If medium 1 is air, this reduces to:

$$\text{refractive index of medium 2, } n_2 = \frac{\sin\theta_1}{\sin\theta_2} = \frac{c}{v_2}$$

For light entering from air, refractive index, n = sine of the angle of incidence divided by the sine of the angle of refraction.

WORKED EXAMPLE C3.1

Consider Figure C3.26. If the angle of incidence on a plastic block was 48°, and the angle of refraction was 32°:

a determine the refractive index of the plastic, and

b calculate the speed of light in the plastic.

Answer

a $n_{\text{plastic}} = \frac{\sin\theta_1}{\sin\theta_2} = \frac{\sin 48}{\sin 32} = \frac{0.743}{0.530} = 1.40$

b $n_{\text{plastic}} = 1.40 = \frac{c}{v_{\text{plastic}}} = \frac{3.00 \times 10^8}{v_{\text{plastic}}}$

$v_{\text{plastic}} = 2.14 \times 10^8\,\text{m s}^{-1}$

WORKED EXAMPLE C3.2

A light ray travelling in water of refractive index 1.33 is incident upon a plane glass surface at an angle of 27° to the normal.

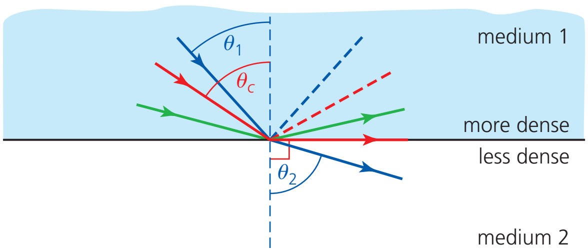
Calculate the angle of refraction if the glass has a refractive index of 1.63.

Answer

$\frac{n_1}{n_2} = \frac{\sin\theta_2}{\sin\theta_1} \Rightarrow \frac{1.33}{1.63} = \frac{\sin\theta_2}{\sin 27°}$

$\sin\theta_2 = 0.3704$

$\theta_2 = 22°$



■ **Figure C3.28** Total internal reflection occurs if the angle of incidence is greater than the critical angle θ_c

Critical angle and total internal reflection

Consider again Figure C3.24, which shows a wave / ray entering an optically less dense medium (a medium in which light travels faster). If the angle of incidence is gradually increased, the refracted ray will get closer and closer to the boundary between the two media. At a certain angle, the refracted ray will be refracted at an angle of exactly 90º along the boundary (see Figure C3.28). This angle is called the **critical angle**, θ_c, shown in red in the diagram.

◆ **Critical angle** Largest angle at which a ray of light can strike a boundary with another medium of lower refractive index, without being totally internally reflected.

◆ **Total internal reflection** All waves are reflected back within the medium. Can only occur when a wave meets a boundary with another medium with a lower refractive index (in which it would travel faster).

◆ **Optically dense** If light travels slower in medium A, compared to medium B, then medium A is described as more optically dense.

For any angle of incidence some light will be reflected at the boundary, but for angles of incidence greater than the critical angle, *all* the light will be reflected back and remain in the denser medium. This is known as **total internal reflection**.

We know that:

$$\frac{n_1}{n_2} = \frac{\sin\theta_2}{\sin\theta_1}$$

but at the critical angle, $\theta_1 = \theta_c$ and $\theta_2 = 90°$, so that $\sin\theta_2 = 1$, and then:

$$\frac{n_1}{n_2} = \frac{1}{\sin\theta_c}$$

Most commonly, the light will be passing from an **optically denser** material (medium 1) like glass, plastic or water, into air (medium 2), so that $n_2 = n_{air} = 1$, and so:

$$n_{medium} = \frac{1}{\sin\theta_c}$$

WORKED EXAMPLE C3.3

a Determine the critical angle for a water / air boundary. ($n_{water} = 1.33$)

b Determine the critical angle for a glass / water boundary. ($n_{glass} = 1.60$)

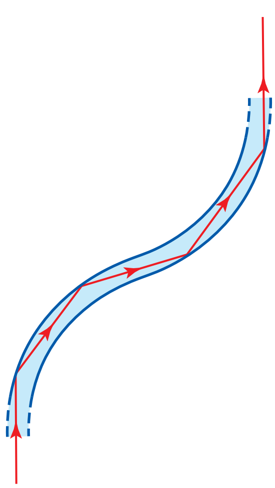
Answer

a $n_{medium} = \frac{1}{\sin\theta_c} \Rightarrow 1.33 = \frac{1}{\sin\theta_c} \Rightarrow \theta_c = 48.8°$

b $\frac{n_{glass}}{n_{water}} = \frac{1}{\sin\theta_c} \Rightarrow \frac{1.60}{1.33} = \frac{1}{\sin\theta_c} \Rightarrow \theta_c = 56.2°$

Applications of total internal reflection

One very important application of total internal reflection is in digital communication. Light passing into a glass fibre can be ‘trapped’ within the fibre because of multiple internal reflections and it will then be able to travel long distances, following the shape of the fibre (see Figure C3.29). The light can be modified to transmit digital information very efficiently.



■ **Figure C3.29** Total internal reflection along a glass fibre

ATL C3A: Communication skills

Clearly communicate complex ideas in response to open-ended questions

Most of the data transferred around the world is done using **optical fibres**. Choose one aspect of this important topic and use a variety of sources to access enough information that you can make an interesting three- to five-minute presentation to the rest of your group.

For example, you could choose one of the following: the choice of wavelength used, the use of binary signals, underwater cables, the *cladding* used in the fibres, the purity of the glass used, possible *bandwidths*, how far the signals can travel without the need for *regeneration*, how optical fibres are connected to copper wires, and so on.

◆ **Optical fibre** Thin, flexible fibre of high-quality glass that uses total internal reflection to transmit light along curved paths and over large distances.

Total internal reflection is also used in *endoscopes* for carrying out medical examinations. Light from a source outside is sent along fibres to illuminate the inside of the body. Other optical fibres, with lenses at each end, are used to bring a focused image outside for viewing directly, or via a camera and monitor (see Figure C3.30).

■ **Figure C3.30** An endoscope can be used to inspect a patient's stomach

14 A ray of light travelling in air was incident at an angle of 47° to the surface of a plastic which had a refractive index of 1.58 Calculate:

- **a** the angle of refraction (to the normal) in the plastic
- **b** the speed of light in the plastic.

15 Parallel water waves travelling at $48\,\text{cm}\,\text{s}^{-1}$ enter a region of shallow water with the incident wavefronts making an angle of 34° with the boundary. If the waves travel with a speed of $39\,\text{cm}\,\text{s}^{-1}$ in the shallower water, predict the direction in which they will move.

16 Light rays travel at $2.23 \times 10^{8}\,\text{m}\,\text{s}^{-1}$ in a liquid.

- **a** Determine the refractive index of the liquid.
- **b** Light rays coming out of the liquid into air meet the surface at an angle of incidence of 25°.
  Calculate the angle of the emerging ray to the normal in air.

17 A certain kind of glass has a refractive index of 1.55. If light passes into the glass from water (refractive index = 1.33) and makes an angle of refraction of 42°, what was the angle of incidence?

18
- **a** Use trigonometry to show that the refractive index between two media is equal to the ratio of wave speeds (v_1/v_2) in the media.
- **b** Show that the refractive index for waves going from medium 1 into medium 2 is given by:
  ${}_1n_2 = n_2/n_1$

19 Explain why it is impossible for any medium to have a refractive index of less than one.

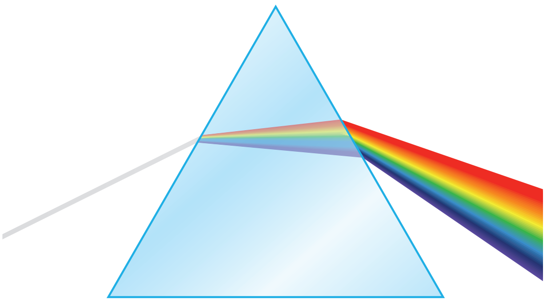
20 The refractive index of red light in a certain type of glass is 1.513.

- **a** If a ray of red light strikes an air / glass boundary at an angle of incidence of 29.0°, determine its angle of refraction.
- **b** A ray of violet light was incident at exactly the same angle (29.0°), but its angle of refraction was slightly less. Explain why.
- **c** If the angle of refraction for violet light was 18.5°, determine values for
  - **i** the refractive index of violet light in this glass
  - **ii** the speed of violet light in the glass.
- **d** We say that the red and violet light rays have been dispersed. Explain what that means. (See also next section.)

21 The speed of light in sea water is $2.21 \times 10^{8}\,\text{m}\,\text{s}^{-1}$. Calculate the critical angle for light striking a boundary between sea water and air.

22 A certain kind of glass has a refractive index of 1.54 and water has a refractive index of 1.33.

- **a** In which medium does light travel faster?
- **b** Describe the circumstances which must occur for light to be totally internally reflected when meeting a boundary between these two substances.
- **c** Calculate the critical angle for light passing between these two media



■ **Figure C3.31** A triangular prism used to produce a continuous spectrum of white light

Dispersion of light into a spectrum

The speeds of different colours (frequencies) of light in a particular medium (glass, for example) are not exactly the same. Red light travels the fastest and violet is the slowest. This means that different colours travelling in the same direction from the same source will not travel along exactly the same paths when they are refracted. When light goes through parallel-sided glass (like a window), the effect is not usually significant or noticeable. However, when white light passes into and out of other shapes of glass (like prisms and lenses), or water droplets, it can be **dispersed** (separated into different colours which spread apart). A triangular prism, as shown in Figure C3.31, is commonly used to disperse white light into a spectrum.

◆ **Dispersion (light)** Separation into different wavelengths / colours (to form a spectrum, for example).

Diffraction of waves

SYLLABUS CONTENT

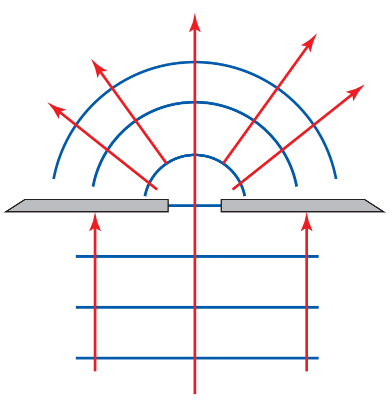
- Wave diffraction around a body and through an aperture.
- Wavefront-ray diagrams showing diffraction.

Waves of all types often encounter obstacles in their path, so it is important to understand how waves pass around (and through) such objects.

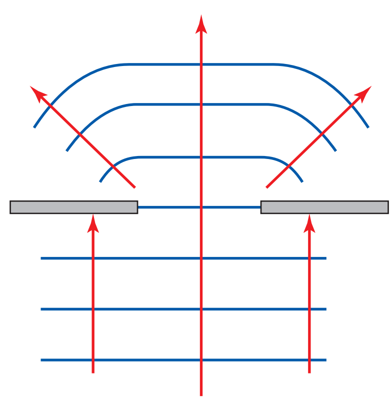
◆ **Aperture** A hole or gap designed to restrict the width of a beam of light (or other radiation).

◆ **Diffraction** The spreading of waves as they pass obstacles, or through apertures.

Waves will tend to spread around corners and as they pass through gaps (**apertures**). This important effect is known as **diffraction**. The simplest and most important example is shown in Figure C3.32. In this diagram the size of the aperture is the same as the wavelength, and the waves spread almost equally in all directions.



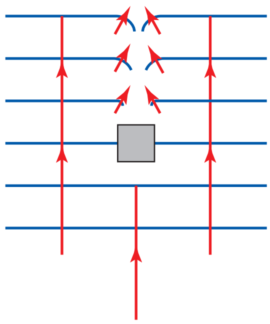
■ **Figure C3.32** Diffraction of plane waves by an aperture of width equal to one wavelength

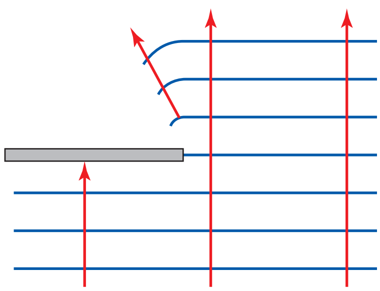


■ **Figure C3.33** aperture > wavelength

For apertures of greater width (compared to the wavelength), the effects of diffraction are less noticeable, as shown in Figure C3.33. Most of the wave energy continues travelling in its original direction. If aperture width is much greater than the wavelength, diffraction usually becomes insignificant.

Diffraction effects are most significant when the size of aperture or object ≈ wavelength.





■ **Figure C3.34** Diffraction around objects

◆ **Laser** Source of intense, coherent, monochromatic light.

◆ **Monochromatic** Containing only one colour / frequency / wavelength (often, more realistically, a narrow range).

◆ **Aerial** A structure that receives or emits electromagnetic signals. Also called an *antenna*.

It is important to realize that diffraction at an aperture also occurs when waves are emitted from a source or aerial, or received by an observer. Consider, for example, sound waves coming from a loudspeaker, or light waves entering an eye.

Figure C3.34 shows how waves can diffract around the edges of objects.

Examples of diffraction

Sound

A typical sound wave may have a wavelength of about 1 m. This is similar to the size of everyday objects and that is why we expect to be able to hear sources that we cannot see (because they are around a corner).

Light

Tool 1: Experimental techniques

Recognize and address relevant safety, ethical or environmental issues in an investigation

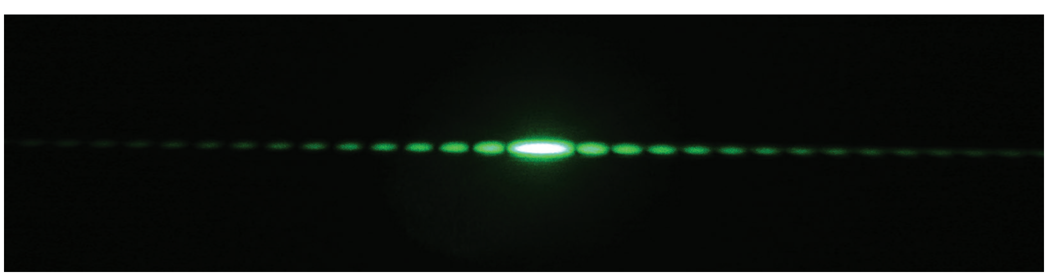
Laser light is very useful in light experiments, especially for observing diffraction. But, because of its intensity, a laser beam can be dangerous if it enters an eye.

Everybody should wear laser googles if they are available. If not, the beam should be kept horizontal and well below eye level. It should also *not* be directed towards a highly reflective surface.

Students should remain in the same places and, as soon as observations have been made, the laser should be quickly turned off.

As we have seen in Topic C.2, light has a very short wavelength (10^{-7}–10^{-6} m). This means that the diffraction of light around everyday objects will not normally be noticed. However, light *is* diffracted, and this can be observed with a very small aperture and a bright light. Figure C3.35 shows the diffraction of **laser** light through an aperture of width less than 0.1 mm. The laser light is **monochromatic**, which means there is only one colour (frequency).

The fact that light diffracts is important evidence of its wave-like nature.



■ **Figure C3.35** Diffraction pattern of light

An explanation of the pattern seen in Figure C3.35 is not important in this section, but is provided in the HL section towards the end of this topic.

Microwaves

A typical wavelength of the microwaves used in a mobile phone network is 0.1 m. Look again at Figure C3.9. The horizontal widths of the apertures emitting the waves encourages waves to be diffracted horizontally, parallel to the ground. The greater height of the aperture reduces the diffraction of waves vertically.

Microwaves are detected and transmitted by an aerial (antenna) inside the phone, which is also similar in size to the wavelength.

Conversely: a microwave beam of similar wavelength from an aircraft detection system at an airport (radar – discussed in Topic C.5) needs to be directional (not spreading out). Aircraft in different directions are located by rotating the aerial. This means that diffraction is not wanted, so the aerial's reflector is designed to be much bigger than the wavelength.

X-rays

X-ray wavelengths are comparable to the sizes of atoms and ions, and their separations, in solids. This means that X-rays are diffracted well by the regular arrangements of atoms / ions in most solids. By analysing X-ray diffraction patterns, we can learn about the structure of matter.

23 Suggest a reason why loudspeakers for producing lower pitched sounds are usually larger in size than speakers for higher pitched sounds.

24 If red light and blue light are passed through the same narrow slit, which will be diffracted more, and why?

25 Bluetooth technology uses a frequency of 2.4 GHz.
- **a** Calculate its wavelength.
- **b** Explain why this makes is suitable for, say, connecting a mobile phone to a Bluetooth speaker.

26 Sketch the wavefronts you would expect to see if parallel water waves of wavelength 1 cm were passing through an aperture of width 5 cm.

27
- **a** State a value for a typical X-ray wavelength.
- **b** Compare your answer to 0.28 nm, the approximate regular spacing of ions in a salt crystal.
- **c** Would you expect a distinct diffraction pattern to occur if X-rays were sent through water? Explain your answer.

Nature of science: Theories

Changing theories about diffraction

The first detailed observations of the diffraction patterns produced in the shadows by light passing through apertures were made more than 350 years ago but, at that time, the phenomenon defied any simple explanation because light itself was not understood, although many believe that a light beam consisted of some type of 'particles'. Many years later, after the wave theory of light became established, a theory of diffraction could be developed that involved the adding together of waveforms arriving at the same point from different places within the aperture. (Depending on the context of the discussion, the addition of waveforms can be variously described as superposition, interference or Fourier synthesis.) The more recent photon theory of light (Theme E) returns, in part, to a 'particle' explanation.

Superposition of waves

SYLLABUS CONTENT

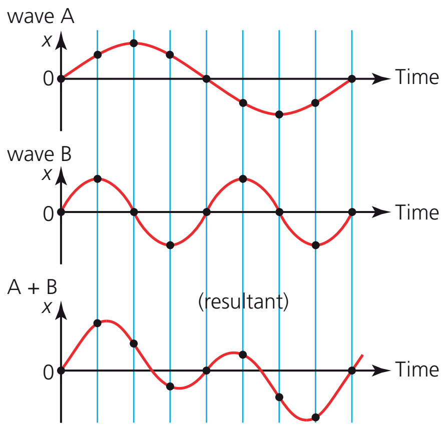
- Superposition of waves and wave pulses.

◆ **Superposition (principle of)** The resultant of two or more waves arriving at the same point can be determined by the vector addition of their individual displacements.

When waves pass through each other at a point, we can add their displacements to determine the overall result at that place at that moment. This is called the **superposition** of waves.

The principle of superposition of waves: the overall displacement is the vector sum of the individual wave displacements.

When similar waves pass through each other they will usually have a wide range of different frequencies and amplitudes. Under these circumstances superposition effects are negligible.



■ **Figure C3.36** Adding wave displacements using the principle of superposition

But when waves have similar frequencies and amplitudes, the effect can be significant. Figure C3.36 shows a simple example of combining wave displacements to find a resultant.

We will use the principle of superposition in the next sub-topic (interference).

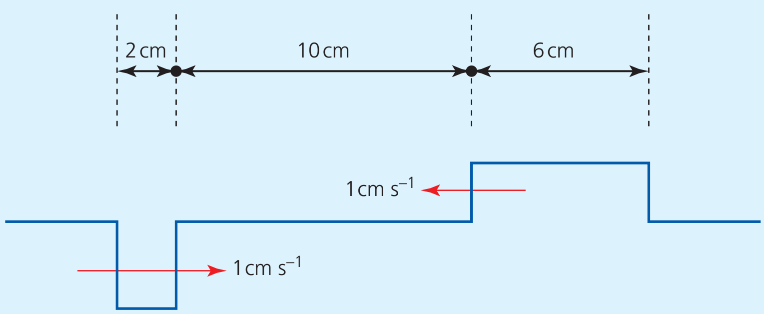
28 a Sketch a displacement–time graph for a sinusoidal oscillation of amplitude 4.0 cm and frequency 2.0 Hz. Start with a displacement of zero and continue for 0.75 s.

b On the same axes draw a graph representing an oscillation of amplitude 2.0 cm and frequency 4.0 Hz.

c Use the principle of superposition to draw a sketch of the resultant of these two waves.

29 Figure C3.37 shows two idealized square pulses moving towards each other. Draw the resultant waveform after

a 6.0 s **b** 7.0 s **c** 8.5 s **d** 12.0 s.



■ **Figure C3.37** Two idealized square pulses moving towards each other

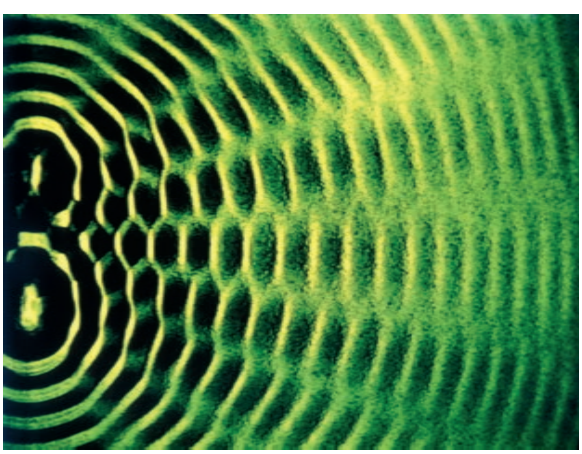
30 Two sinusoidal waves (A and B) from different sources have the same frequency and pass through a certain point, P, with the same amplitude.

a Sketch a displacement–time graph for the resultant waveform if A and B arrive at P in phase.

b Repeat for two waves that arrive at P exactly (π) out of phase.

◆ **Interference pattern (fringes)** Pattern observed when coherent waves interfere.

◆ **Coherent waves** Waves that have the same frequency and a constant phase difference.



■ **Figure C3.38** The interference of water waves on a ripple tank

Interference of waves

SYLLABUS CONTENT

▶ Double source interference requires coherent sources.

When the superposition of wavefronts produces a constant two- or three-dimensional pattern we describe it as an **interference pattern**. Most commonly, this effect occurs between two sources of waves which have the same single frequency and the same wave shape. Such sources (and the waves that they produce) are described as being **coherent**. Figure C3.38 shows an interference pattern produced by two sources of water waves on a ripple tank.

Constructive and destructive interference

SYLLABUS CONTENT

- The condition for constructive interference as given by: path difference = $n\lambda$
- The condition for destructive interference as given by: path difference = $\left(n + \frac{1}{2}\right)\lambda$.

◆ **Path difference** The difference in the distances from a particular point to two sources of waves.

◆ **Interference** Superposition effect that may be produced when similar waves meet. Most important for waves of the same frequency and similar amplitude. Waves arriving in phase will interfere **constructively** because their path difference = $n\lambda$. Waves completely out of phase will interfere **destructively** because their path difference = $(n + \frac{1}{2})\lambda$.

To explain an interference pattern like that seen in Figure C3.38, we need to consider the **path differences** between the waves arriving at various points from the two sources. Consider Figure C3.39, in which the straight lines are representing distances, not rays.

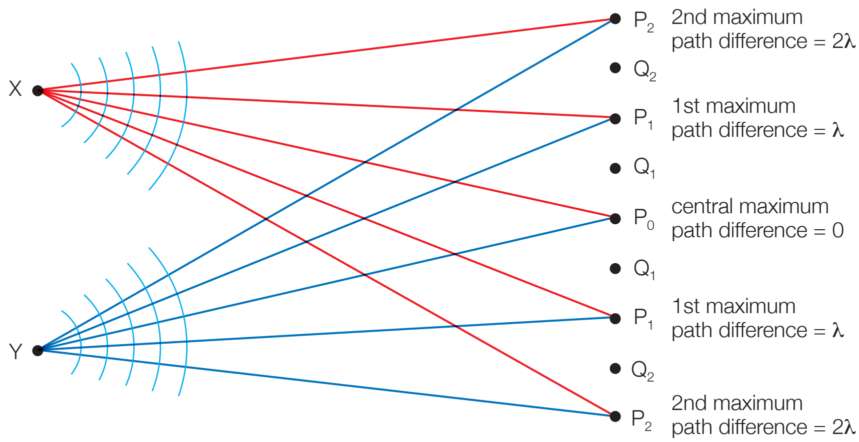


Figure C3.39 Interference and path difference

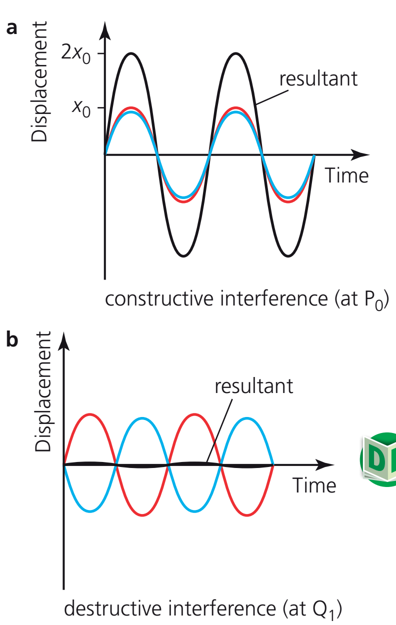


Figure C3.40 a Constructive and **b** destructive interference

If the waves are emitted in phase from points X and Y, and both waves travel the same distance at the same speed to any point such as P_0, the waves will always be in phase at that position. Figure C3.40a shows the result, using the principle of superposition: the resulting wave has double the amplitude of the original waves. This is called **constructive interference**.

The path difference for the waves arriving at P_0, or any other point which is the same distance from X and Y, is zero. Constructive interference will also occur at any place where the waves always arrive in phase, which occurs where the path difference is 1λ, or 2λ, or 3λ … or $n\lambda$ (where n is a whole number), as shown by P_1 and P_2 in Figure C3.39.

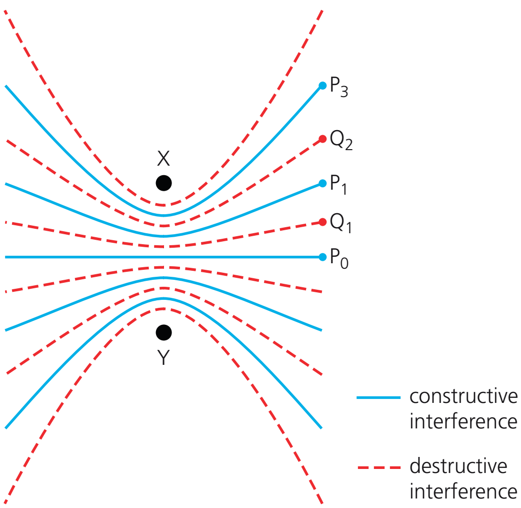
The condition for constructive interference:

path difference = $n\lambda$

Destructive interference is shown in Figure C3.40b. This will occur at all points where the waves arriving from one source have travelled $\frac{1}{2}\lambda$, or $\frac{3}{2}\lambda$, or $\frac{5}{2}\lambda$, or $\left(n + \frac{1}{2}\right)\lambda$ more than the waves from the other source, as shown by the points Q_1 and Q_2 in Figure C3.39.

The condition for destructive interference:

$$\text{path difference} = \left(n + \frac{1}{2}\right)\lambda$$



■ **Figure C3.41** The interference pattern produced by coherent waves from two sources, X and Y

These conditions assume the usual situation: the waves are emitted in phase with each other. If the waves were exactly out of phase, these conditions would be reversed.

Constructive interference and destructive interference describe the extreme possibilities of wave superposition. At other locations the amount of interference varies between these extremes.

From Figure C3.40a, we can see that two sets of waves, each of amplitude x_0, result in a wave of amplitude $2x_0$. Since we know, from earlier in this topic, that wave intensity is proportional to amplitude squared, at places of constructive interference the intensity has quadrupled. This is possible because the intensity at places of destructive interference has been reduced.

Figure C3.41 shows the overall interference pattern produced as described in Figure C3.39. The right-hand half of Figure C3.41 can be compared to Figure C3.38.

Interference is a property *only* of waves, including electromagnetic waves like light. The interference of light cannot be explained, for example, by imagining that light consists of tiny particles. Thomas Young was the first to demonstrate that light could interfere (see below), thus demonstrating for the first time that light had wave properties.

Top tip!

When using two similar wave sources which are in phase, perfect destructive interference, resulting in waves of zero amplitude, is not possible. This is because one wave will always have a reduced amplitude because it has travelled further than the other wave to reach any particular point.

Nature of science: Theories

Competing theories

The nature of light has been widely debated for centuries. Different scientists developed different theories which seemed to partly contradict each other, and no single theory was able explain all the properties of light. This is not unusual in the development of scientific knowledge. There are many modern examples, including the consequences of climate change and the reasons for an ever-expanding Universe.

Clearly, we would prefer a single theory to help explain any particular phenomenon and to make useful predictions. However, if that is not possible at the present time, we can continue to use the best available, but less than perfect, competing theories which have proved to be useful. This is illustrated by a famous quote about the two theories of light, from William Henry Bragg: *'Physicists use the wave theory on Mondays, Wednesdays and Fridays and the particle theory on Tuesdays, Thursdays and Saturdays.'*

Young's double-slit interference experiment

SYLLABUS CONTENT

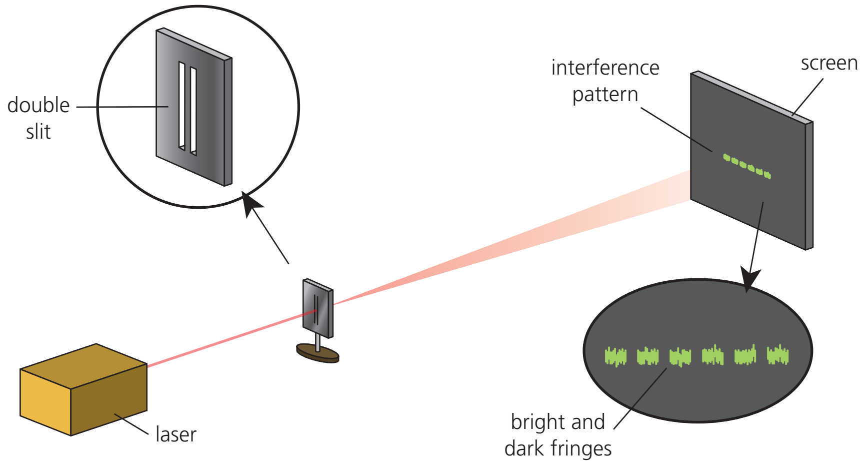
- Young's double-slit interference as given by: $s = \frac{\lambda D}{d}$, where s is the separation of fringes, d is the separation of the slits, and D is the distance from the slits to the screen.

The interference of light waves is not an everyday observation because:

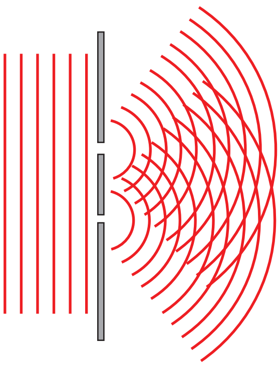
- Separate light sources are not coherent.
- Light has very small wavelengths, so that any interference pattern will be very small and difficult to observe.

◆ **Young's interference experiment** Famous experiment which provided the first evidence that light travelled as waves.

The interference of light can be demonstrated by passing monochromatic laser light through two narrow slits which are very close together (a modern version of **Young's interference experiment** (1801)). The resulting interference pattern can be seen on a distant screen, see Figure C3.42.



■ **Figure C3.42** Interference of light waves

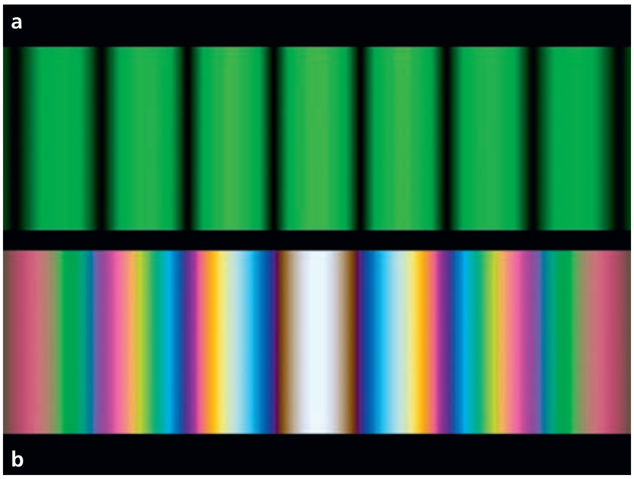


■ **Figure C3.43** Diffracted waves crossing over each other as they emerge from double slits

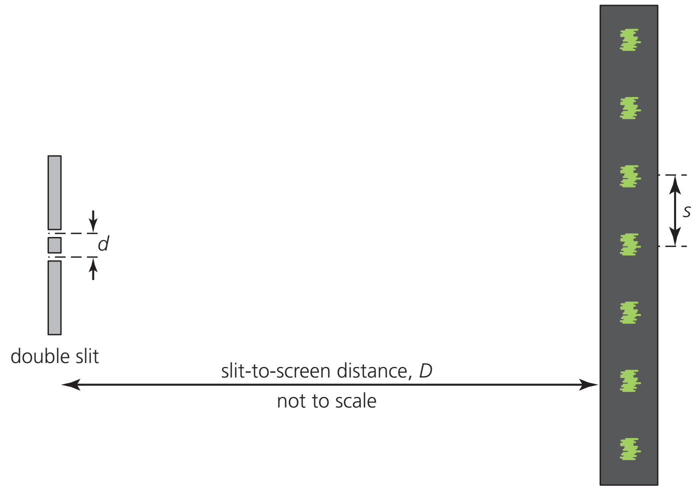
The double slits act as the necessary two coherent sources of waves. Figure C3.43 illustrates an example of the basic principle: each plane wavefront is diffracted into two beams as it emerges from the slits. The waves are coherent because they came from the same original wavefront. The diffracted waves then interfere.

A series of interference fringes is seen on the screen, as shown in the Figure C3.44a. The shape of the pattern depends on the shape of the apertures / slits. The closer the two slits, the wider the spacing of the interference pattern. Figure C3.44b shows the appearance of the fringes if white light is used with the same slits: fringes of different colours / wavelengths occur in slightly different places but overlap.

The wavelength of the light used can be determined from the geometry of the experiment (see Figure C3.45) by using the following equation, which is explained later in this topic for HL students.



■ **Figure C3.44** Interference patterns



■ **Figure C3.45** Geometry of the double-slit experiment

Separation of fringes in Young's experiment:

$$s = \frac{\lambda D}{d}$$

The separation, s, of the centres of the fringes is assumed to be constant, so that it is convenient to measure the total width of a number of fringes. (We are also assuming that $D >> s$ and that the light consists of plane wavefronts arriving in a direction which is perpendicular to the slits.)

WORKED EXAMPLE C3.4

In an experiment similar to that seen in Figure C3.42, the centres of the slits were separated by 0.48 mm and the screen was placed 1.96 m away from the slits. A student measured across nine equally spaced fringes and found that the centre of the first and the centre of the ninth fringe were separated by a distance of 2.25 cm.

a Determine the wavelength of the light used.

b How will the appearance of the fringes change if:

  i the screen is moved to a distance of 4.5 m from the slits

  ii the red light laser is replaced with a green light laser?

Answer

a $s = \dfrac{\lambda D}{d}$

$\Rightarrow \dfrac{2.25 \times 10^{-2}}{8} = \lambda \times \dfrac{1.96}{0.48 \times 10^{-3}}$

$\Rightarrow \lambda = 6.9 \times 10^{-7}\,\text{m}$

b **i** The width and separation of the fringes will more than double, but their intensity (brightness) will be reduced.

  ii Green light has a smaller wavelength than red light, so the fringes will be closer together.

The interference of light as it passes through two or more slits is discussed in more detail for HL students towards the end of this topic.

TOK

The natural sciences

- What kinds of explanations do natural scientists offer?

As we have seen many times, a simple equation (for example, $s = \lambda D/d$) can be used as a starting point to model scientific phenomena. Once that has been thoroughly understood, it can be adapted to more complicated situations. But is this kind of 'modelling' unique to science?

For example, would it be possible in principle to develop a mathematical model to describe the political and/or economic situation in Europe before the start of World War 1? And could such a model be used to predict what happened? Is there something fundamentally different between knowledge in physics and history, or is any historical situation just too complicated, or dependent on human behaviour, for mathematical analysis?

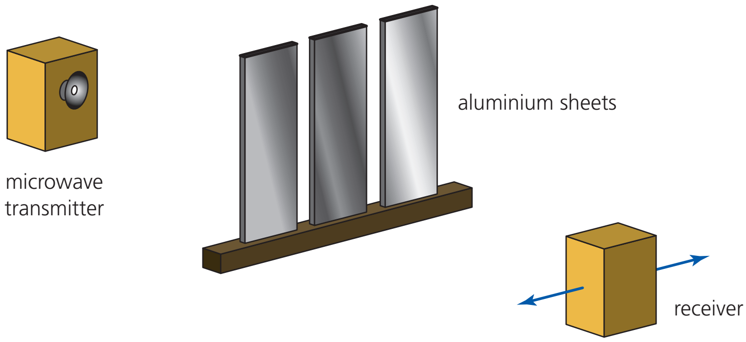
Interference of other types of waves

In theory, all types of waves can produce interference patterns. However, it may be impossible, or very difficult, to produce two *coherent* sources for some types of wave, because the waves are emitted in uncontrolled, random processes. This means that interference will *not* usually be observed with naturally produced waves. That is why the interference of light is not a common phenomenon.

So, to observe interference, we need to turn our attention to artificially produced waves. Apart from the interference of light, two other examples may be demonstrated in school laboratories.

Microwaves

Waves from this section of the electromagnetic spectrum are easily produced by electronic circuits and can have a wavelength of a few centimetres, which is ideal for demonstrations. See Figure C3.46.



■ **Figure C3.46** Interference of microwaves

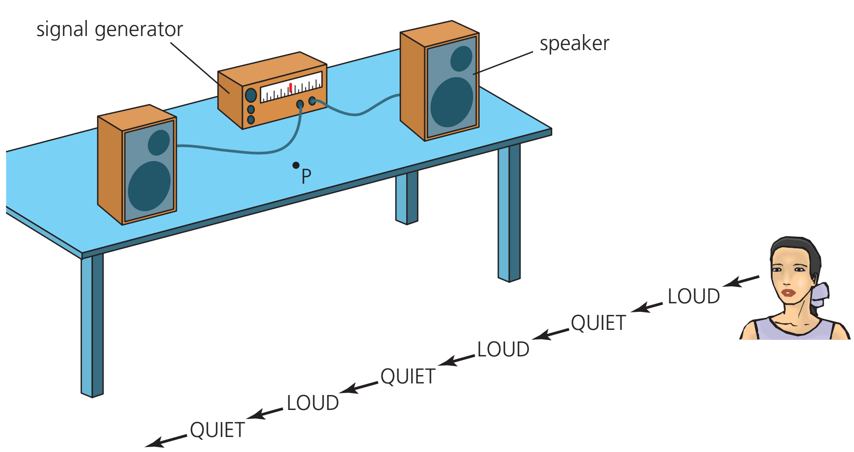
The two gaps between the aluminium sheets each have a width of about one wavelength (≈ 3 cm). They have the same effect as the double slits in Young's experiment: two diffracted, coherent waves emerge from the other side and then interfere. When the microwave detector is moved from side to side, constructive and destructive interference will be detected.

The equation $s \approx \lambda D/d$ *cannot* be used accurately in this situation, or with sound (as described below). This is because the assumptions made about the geometry of the light interference experiment (because of the very small wavelength of light) are not valid in the arrangement shown in Figure C3.46.

Sound

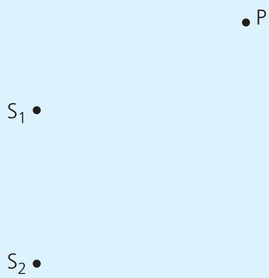
◆ **Signal generator** Electronic equipment used to supply small alternating currents of a wide range of different frequencies.

See Figure C3.47. The **signal generator** provides an oscillating electric current to the two loudspeakers, which then produce coherent longitudinal sound waves of the same frequency. As a listener walks past the speakers, as shown in the figure, the sound intensity rises and falls, because of constructive and destructive interference.



■ **Figure C3.47** Interference of sound waves

31 Figure C3.48 is one-quarter of the real size. It shows two coherent wave sources on a ripple tank and a point P. If the wavelengths are 2.5 cm, take measurements from the diagram to determine what kind of interference occurs at P.



■ **Figure C3.48** Two coherent wave sources on a ripple tank

32 Consider Figure C3.46.

a If constructive interference produces a maximum signal when the receiver is 57 cm from the centre of one slit, and 45 cm from the other, show why 3 cm and 4 cm are both possible values for the wavelength being used.

b Discuss how the actual wavelength can be quickly determined.

33 A teacher wants to demonstrate the interference of light to her class using green laser light of wavelength 5.32×10^{-7} m. She uses double slits of separation 0.50 mm. She would like the fringes to be at least 0.50 cm apart.

a Determine the closest distance she can place the screen to the slits.

b With the slits and screen still in the same places, explain how the teacher can change the experiment to produce fringes which are slightly further apart.

34 Consider Figure C3.47.

The centres of the speakers were 1.20 m apart. And the girl's closest distance to point P was 80 cm, which was where the sound was loudest. The next position where the sound was loud was 50 cm away in the direction shown.

a Are the sound waves from the two speakers emitted in phase, or out of phase?

b Use Pythagoras's theorem to determine the path difference and calculate the wavelength of the sound.

c Describe how the interference pattern would change if

i the sound frequency was increased.

ii the connections to one of the speakers was reversed.

35 The waves used to cook food in a microwave oven reflect repeatedly off the metal walls and can produce interference effects.

a Suggest how this could affect the way in which the food is cooked.

b Research into how microwave ovens are designed to overcome this problem.

36 Outline why the interference of light is not a common observation.

37 A boy stands halfway between two loudspeakers facing each other in a large open space. Both speakers are producing sounds of frequency 180 Hz. In this position he hears a loud sound.

a Explain why the sound level will decrease if he starts to walk towards either speaker.

b Discuss how the sound level will change if he walks from the mid-point along a line perpendicular to a line joining the speakers.

c How far must he walk directly towards one of the speakers before the sound level will rise to a maximum again? Assume that the speed of sound is $342\,\mathrm{m\,s^{-1}}$.

d Explain why this experiment is best carried out in 'a large open space'.

Inquiry 1: Exploring and designing

Designing

Design and carry out an investigation into the properties of the waves emitted by a (television) remote control.

ATL C3B: Self-management skills

Breaking down major tasks into a sequence of stages

A good friend of yours is in the same physics class, but he has not been doing well and lately he seems to have become discouraged because he feels overwhelmed by work and deadlines. There are important physics examinations coming up in three weeks' time. Suggest ways in which you could advise him to prepare for the examinations.

A closer look at single-slit diffraction of light

SYLLABUS CONTENT

▶ Single-slit diffraction, including intensity patterns, as given by: $\theta = \frac{\lambda}{b}$, where b is the slit width.

Common mistake

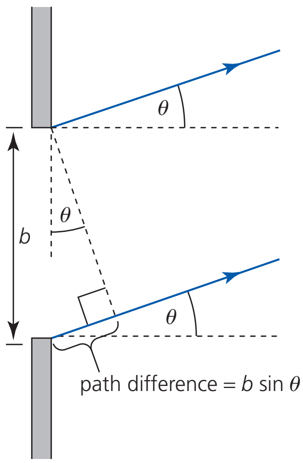
Single-slit diffraction patterns and double-slit interference patterns are similar to each other and easily confused. The most obvious difference is that the central fringe of the single-slit diffraction pattern is brighter and wider than the other fringes.

Figure C3.32 showed a diffraction pattern produced by monochromatic light passing through a narrow slit. It can be produced by an experimental arrangement similar to that seen in Figure C3.42, but with the double slits replaced by a single slit.

In the introduction to diffraction, Figure C3.33 indicated the pattern of diffraction we would expect when waves pass through an aperture greater than the wavelength. Now, we need to explain why the diffraction pattern produced when light passes through a single narrow slit is different: it has fringes similar to that produced by double slits.

Firstly, it is important to understand that even a ‘narrow’ slit of width, for example, $b = 0.1$ mm is much greater than one wavelength, λ, of light. In fact:

$$\frac{b}{\lambda} \approx \frac{0.1 \times 10^{-3}}{5 \times 10^{-7}} = 200$$



■ **Figure C3.49** Path differences and interference

Interference within a single wavefront

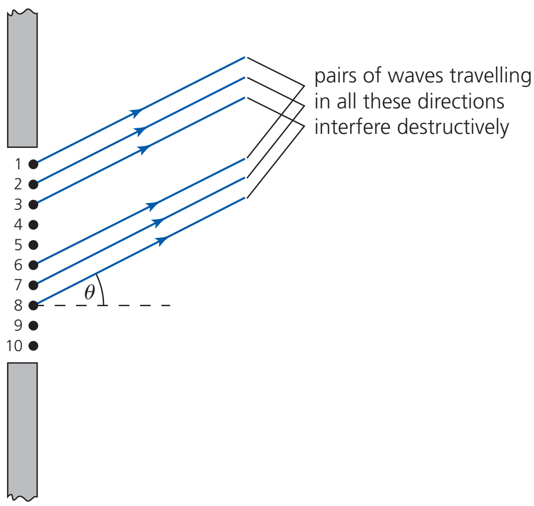
We need to imagine that a plane wavefront emerging from a single slit acts as if it was a series of point sources of ‘secondary wavelets’, maybe one for every wavelength. These ideas were famously first put forward as an explanation for the propagation of all waves by Christian Huygens in 1690.

These wavelets will be coherent and they will interfere.

Firstly, consider how secondary wavelets from the edge of the slit may interfere.

Figure C3.49 shows a typical direction, θ, in which **secondary waves** travel away from a single narrow slit of width, b.

If θ is zero, all the secondary waves will interfere constructively in this direction (straight through the slit) because there is no path difference between them. (Of course, in theory, waves travelling parallel to each other in the same direction cannot meet and interfere, so we will assume that the waves’ directions are very nearly parallel.)



■ **Figure C3.50** Secondary waves that will interfere destructively can be ‘paired off’

Consider what happens for angles increasingly greater than zero. The path difference, as shown in Figure C3.49, equals $b \sin\theta$ and this increases as the angle θ increases. There will be angles at which the waves from the two edges of the slit interfere constructively because the path difference has increased to become equal to a 1λ, 2λ, 3λ … and so on.

But if secondary wavelets from the edges of the slit interfere constructively, what about interference between all the other secondary wavelets? Consider Figure C3.50 in which the slit has been divided into a number of point sources of secondary wavelets. (Ten points have been chosen, but it could be *many* more.)

◆ **Secondary waves** The propagation of waves in two or three dimensions can be explained by considering that each point on a wavefront is a source of secondary waves.

If the angle, θ, is such that secondary wavelets from points 1 and 10 would interfere constructively because the path difference is one wavelength, then secondary waves from 1 and 6 must have a path difference of half a wavelength and interfere destructively. Similarly, waves from points 2 and 7, points 3 and 8, points 4 and 9 and points 5 and 10 must all interfere destructively. In this way waves from all points can be 'paired off' with others, so that the first minimum of the diffraction pattern occurs at such an angle that waves from the edges of the slit would otherwise interfere constructively.

The first minimum of the diffraction pattern occurs when the path difference between secondary waves from the edge of the slit is equal to one wavelength. That is, if $b \sin\theta = \lambda$.

For the diffraction of light, the angle θ is usually small and approximately equal to $\sin\theta$, if the angle is expressed in radians. (This is valid for angles up to approximately 10°, 0.17 rad.). So that:

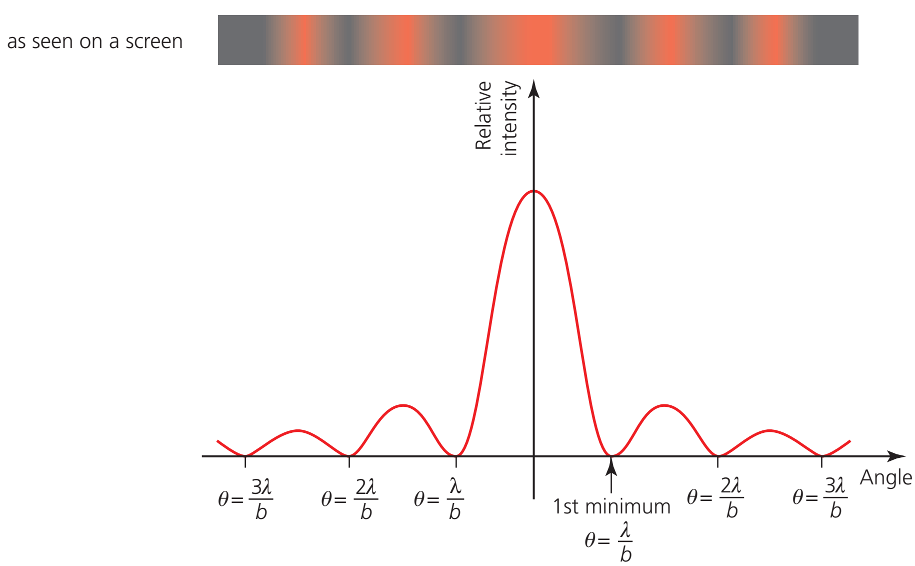
The angle for the first minimum of a **single-slit diffraction** pattern is:

DB

$$\theta = \frac{\lambda}{b}$$

◆ **Single-slit diffraction** The simplest diffraction pattern is that produced by wavefronts interfering after they have passed through a narrow, rectangular slit. The first minimum occurs at an angle such that $\theta = \frac{\lambda}{b}$.

Similar reasoning will show that other diffraction minima will occur at angles, $\theta = 2\lambda/b, 3\lambda/b, 4\lambda/b$... and so on ($\theta_n = n\lambda/b$). These angles are represented on the intensity-angle graph shown in Figure C3.51.



■ **Figure C3.51** Variation of intensity with angle for single-slit diffraction

Top tip!

As we have seen, a full explanation of diffraction involves discussing interference, and this can easily cause confusion, especially if they are thought to be two separate phenomena. To be clear: diffraction is the change of direction that occurs when waves pass gaps and obstacles. Diffracted wavefronts may then undergo superposition effects, which can lead to a pattern being formed which is usually called a 'diffraction pattern', although the principles of interference are used to explain it.

WORKED EXAMPLE C3.5

Monochromatic light of wavelength 663 nm (663×10^{-9} m) is shone through a gap of width 0.0730 mm.

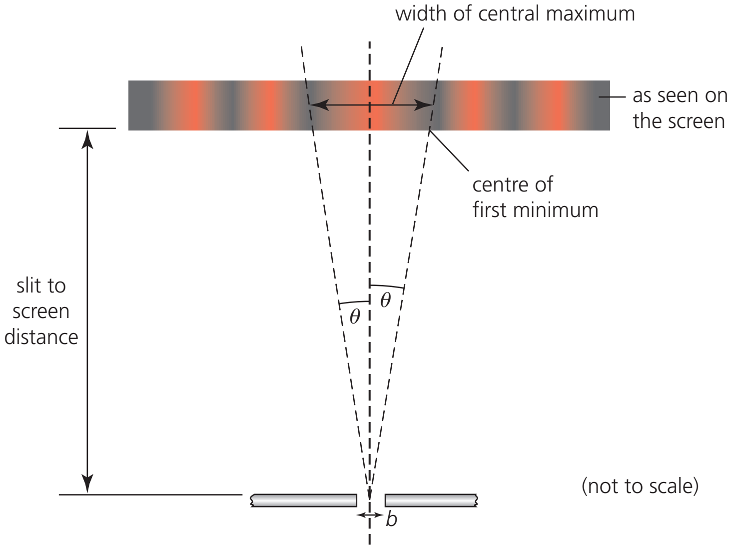


Figure C3.52 Monochromatic light shone through a gap

a Calculate the angle at which the first minimum of the diffraction pattern is formed.

b If the pattern is observed on a screen that is 2.83 m from the slit, determine the width of the central maximum.

Answer

a $\theta = \frac{\lambda}{b} = \frac{663 \times 10^{-9}}{7.30 \times 10^{-5}} = 9.08 \times 10^{-3}$ radians

b $\theta \approx \sin\theta = \frac{\text{half width of central maximum}}{\text{slit to screen distance}}$

half width of central maximum = $(9.08 \times 10^{-3}) \times 2.83 = 0.0257$ m (0.02569... seen on calculator display)

width of central maximum = $0.02569 \times 2 = 0.0514$ m

38 Electromagnetic radiation of wavelength 2.37×10^{-7} m passes through a narrow slit of width 4.70×10^{-5} m.

- **a** State in which part of the electromagnetic spectrum this radiation occurs.
- **b** Suggest how it could be detected.
- **c** Calculate the angle of the first minimum of the diffraction pattern.

39 Determine the wavelength of light that has a first diffraction minimum at an angle of 0.0038 radians when it passes through a slit of width 0.15 mm.

40 When light of wavelength 6.2×10^{-7} m was diffracted through a narrow slit, the central maximum had a width of 2.8 cm on a screen that was 1.92 m from the slit. Calculate the width of the slit.

41 a Sketch and label a relative intensity–angle graph for the diffraction of red light of wavelength 6.4×10^{-7} m through a slit of width 0.082 mm. Include at least five peaks of intensity.

- **b** Add to your graph a sketch to show how monochromatic blue light would be affected by the same slit.

HL ONLY

ATL C3C: Thinking skills

Providing a reasoned argument to support conclusions

Resolution

When light waves enter our eyes, they are diffracted and this affects our ability to see detail. The ability of our eyes (and/or optical instruments) to see objects as separate is called **resolution**. For example, our eyes will not be able to resolve the individual leaves seen on a distant tree.

This is easiest explained by considering two identical sources of light at night, for example car headlights a long way away. Considering the distance involved, we can assume that the light waves are effectively emitted from points sources, but when they are received by the eye, the images on the back of the eye (retina) are not points, but diffraction rings. (Rings are formed because the aperture in the eye is circular.) Consider Figure C3.53.

When the lights are close (Figure C3.53a), we can see two headlights, but when the lights are a long way away (maybe 5 km or more), the diffraction patterns overlap and cannot be distinguished – only one light is seen.

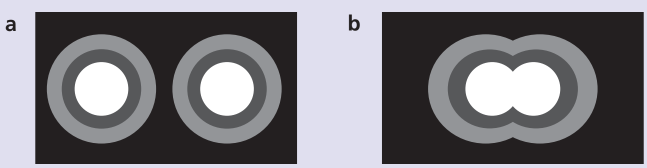


Figure C3.53 Images of two point sources observed through circular apertures that are **a** easily resolvable and **b** just resolvable

Most of the wave intensity will be concentrated near the centre of diffraction rings, so that under most circumstances, diffraction effects are not significant. However, when we want to see fine details under a microscope, or make astronomical observations on distant objects, diffraction is the major factor limiting resolution.

We have seen that the amount of diffraction at an aperture can be represented by the ratio λ/b, so that the resolution of any telescope, or a microscope, can be improved by increasing the width of the aperture receiving the waves, or reducing the wavelength involved (often not possible).

Radio telescopes such as that seen in Figure C3.54 are much larger than optical telescopes. Explain why. Support your explanation with scientific reasoning.

Figure C3.54 The Jodrell Bank radio telescope in England, UK

◆ **Resolution (optical)** The ability of an imaging system to identify objects as separate.

Two slits, multiple slits and diffraction gratings

SYLLABUS CONTENT

▶ Interference patterns from multiple slits and diffraction gratings as given by: $n\lambda = d\sin\theta$.

LINKING QUESTIONS

- What can an understanding of the results of Young's double-slit experiment reveal about the nature of light?
- What evidence is there that particles possess wave-like properties such as wavelength? (NOS)

These questions link to understandings in Topic E.2.

In this section we will discuss

- the advantages of passing light waves through more and more parallel slits
- how the pattern of single-slit diffraction affects interference between two or more slits.

Earlier in this topic we discussed the interference patterns seen when light waves pass through two narrow slits which are close together: for a given wavelength, λ, we saw that the spacing, s, of the pattern seen on a fixed screen depends on the separation of the slits, d, and the slits to screen distance, D $\left(s = \lambda\frac{D}{d}\right)$. See Figure C3.45.

Before we look at this in more detail, we need to obtain a more generalized equation which can be used for predicting where constructive interference occurs with any number of slits. Consider Figure C3.55, which shows two (almost) parallel rays from adjacent slits representing plane waves which interfere constructively when they have an angle θ to the original direction in which the waves were travelling. (For the moment, we will not consider any interference effects between wavelets from the same slit.)

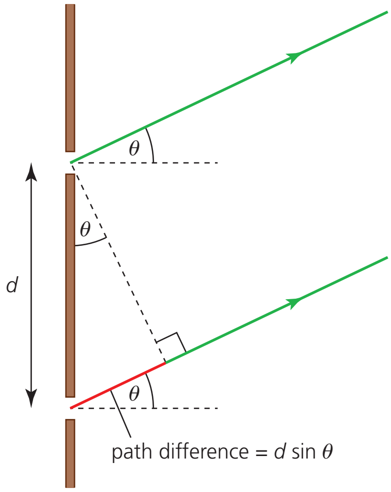


Figure C3.55 Explaining path difference = $d \sin\theta$

The extra distance travelled by the lower ray, the path difference, can be determined from the right-angled triangle: path difference = $d \sin\theta$.

Since the waves interfere constructively, we know that the path difference must also be equal to a whole number, n, of wavelengths. This leads to:

Constructive interference occurs at angles such that:

$$n\lambda = d \sin\theta$$

In other words, constructive interference will occur at angles which have sines equal to λ/d, $2\lambda/d$, $3\lambda/d$, $4\lambda/d$ … and so on.

$$\sin\theta = \frac{n\lambda}{d}$$

For $n = 1$ and small angles:

$$\sin\theta = \frac{s}{D} = \frac{n\lambda}{d}$$

(which can be re-arranged to give:

$$s = \frac{\lambda D}{d}$$

as used previously for two slits).

Similarly, destructive interference occurs at angles such that:

$$\left(n + \frac{1}{2}\right)\lambda = d \sin\theta$$

These equations can be used with any number of slits. Indeed, its most common application is with the very large number of slits of a *diffraction grating*, as discussed below. In such arrangements the angles are typically large enough that *sin* θ must be used, not θ in radians.

Note that the angles at which constructive and destructive interference occur depends on the separation of the slits, but *not* the number of slits.

Figure C3.56 represents these equations in the form of an intensity–$\sin\theta$ graph (for a small number of slits).

Figure C3.57 shows how what is seen on the screen relates to the various angles and values of n.

Common mistake

Students often confuse the equations $\theta = \lambda/b$ and $\sin\theta = n\lambda/d$.

$\theta = \lambda/b$ predicts the small angle for which diffraction at a single slit, of width b, produces an intensity minimum.

$\sin\theta = n\lambda/d$ predicts the angles at which interference of light from two or more slits, of separation d, produces intensity maxima.

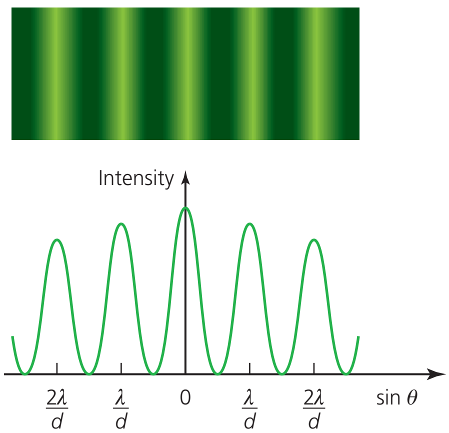


Figure C3.56 Variation of intensity with angle for multiple-slit interference

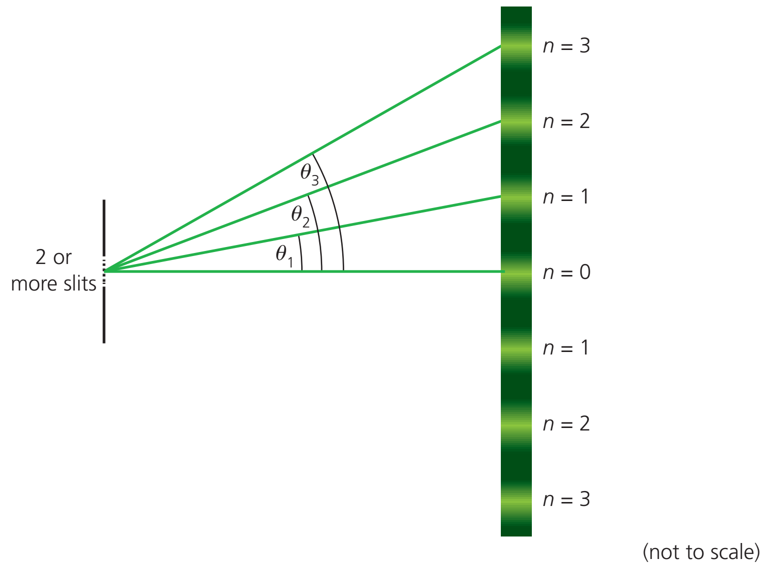


Figure C3.57 Separation and numbering of fringes seen on a screen

HL ONLY

WORKED EXAMPLE C3.6

After passing green laser light through a few parallel slits of spacing 0.059 mm, an interference pattern was seen on a screen placed 3.12 m from the slits. It was similar in appearance to that seen in Figure C3.56. If the distance from the centre of the pattern to the centre of the third fringe was 8.42 cm

a Determine the:

i tan **ii** sin

of the angle, θ_3, between two rays from the slits going to the centre of the pattern and the centre of the third fringe.

b What is this angle in:

i degrees **ii** radians?

c Calculate a value for the wavelength of the laser light.

Answer

a **i** $\tan\theta_3 = \dfrac{8.42}{312} = 0.0270$

ii sine of the same angle = 0.0270 (same as tan)

b **i** 1.55°

ii 0.0270 (same as tan and sin)

c $n\lambda = d\sin\theta$

$3\lambda = (0.059 \times 10^{-3}) \times 0.0270$

$\lambda = 5.3 \times 10^{-7}\,\text{m}$

Because the angles are small the same answer for the wavelength could have been determined from

$s = \dfrac{\lambda D}{d}$

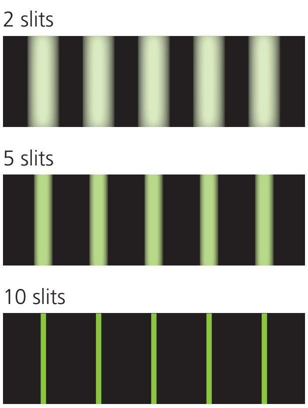


Figure C3.58 How an interference pattern changes as more slits are involved

◆ **Multiple slits** By increasing the number of parallel slits (of the same width) on which a light beam is incident, it is possible to improve the resolution of the fringes / spectra formed.

◆ **Diffraction grating** A large number of parallel slits very close together. Used to disperse and analyse light.

Effect of having more slits (of the same width and separation)

It has already been stressed that the number of slits does not affect the directions in which constructive interference occurs. So, what is the advantage of having a light beam pass through more slits? Certainly, this should make the fringes brighter / more intense, but there is another, more important, reason for using **multiple slits**: the intensity peaks become 'sharper', more precisely located. This is shown in Figure C3.58, which compares the sharpness of the interference peaks obtained with different number of slits.

This effect is used particularly in diffraction gratings.

Diffraction gratings

A **diffraction grating** (see Figure C3.59) is a very large number of slits, very close together. A typical grating has 600 parallel slits every mm. We usually refer to this as 600 lines/mm. The separation of the centres of such lines, d, is 1.67×10^{-6} m, which is equivalent to about (only) three average wavelengths of light. This is much smaller than the width of the slits we have been discussing so far.

This very small separation of the slits, d, results in large values of $\sin\theta$ (consider $n\lambda = d\sin\theta$). If the light incident on the grating is also spread over a large number of lines, the intensity peaks will be well separated, intense and sharp. Figure C3.60 compares the intensity peaks produced by a diffraction grating to those produced by double slits. (The difference in intensity levels will be even greater than that shown.)

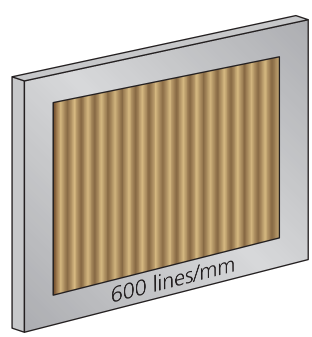
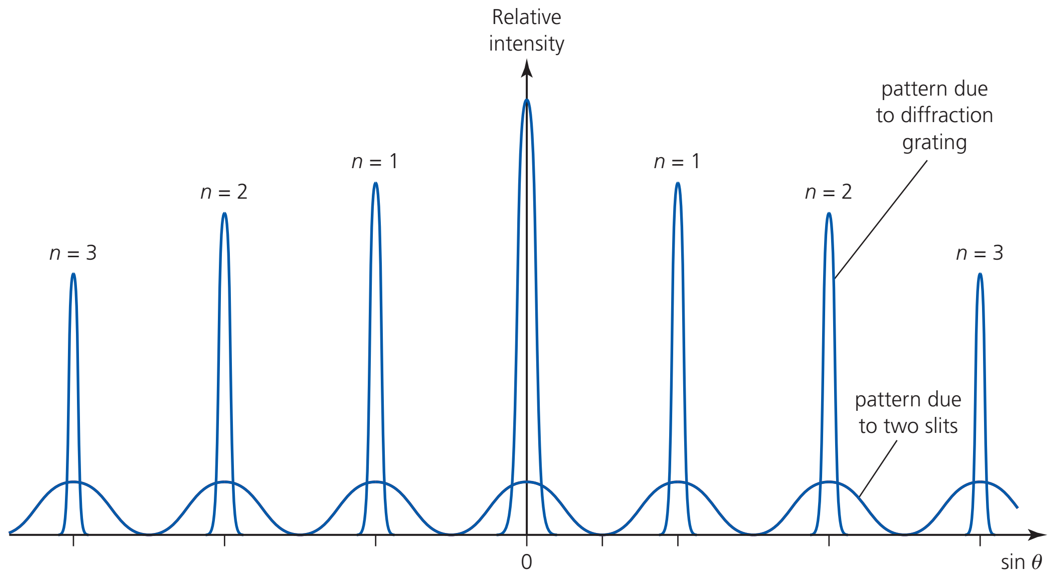


Figure C3.59 Diffraction grating



■ **Figure C3.60** Comparing the maxima produced by double slits and a diffraction grating using monochromatic light

WORKED EXAMPLE C3.7

Use the data in the paragraphs above to calculate the angles (in degrees) for the first, second and third peaks from the centre, for light of wavelength 594 nm (5.94×10^{-7} m).

Answer

$n\lambda = d\sin\theta$

For $n = 1$, $\sin\theta = \dfrac{\lambda}{d} = \dfrac{5.94 \times 10^{-7}}{1.67 \times 10^{-6}} = 0.3557 \Longrightarrow \theta = 20.8°$

For $n = 2$, $\sin\theta = 2 \times 0.3557 = 0.7114 \Longrightarrow \theta = 45.3°$

For $n = 3$, $\sin\theta = 3 \times 0.3557 = 1.067 \Longrightarrow \sin\theta > 1$, which is not possible.

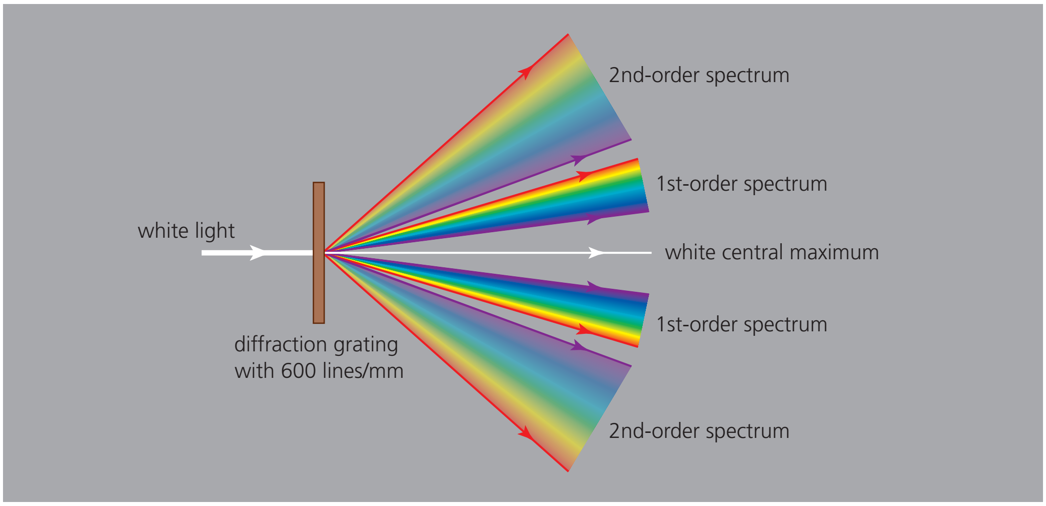
Only two peaks can be seen under these conditions. This is common when using gratings. One or two intense peaks are usually all that are required.

Observing spectra with diffraction gratings

Diffraction gratings are widely used for producing spectra and determining unknown wavelengths of light. (Prisms can also be used.)

Continuous white light spectrum

If white light is sent through a diffraction grating, different wavelengths/colours will be sent in slightly different directions. The light will be *dispersed* into spectra, as seen in Figure C3.61.



■ **Figure C3.61** White light passing through a diffraction grating

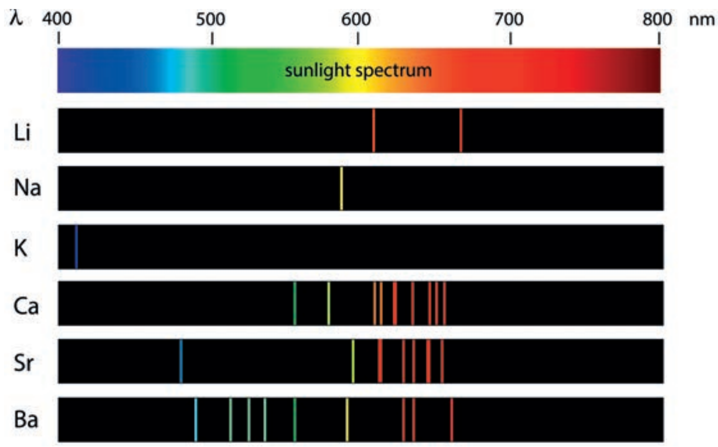
◆ **Spectral orders**
Considering the diffraction grating equation, $n\lambda = d\sin\theta$: different orders correspond to different values of n.

◆ **Spectrum, line**
A spectrum of separate lines (rather than a continuous spectrum), each corresponding to a discrete wavelength and energy.

We commonly refer to the **order of a spectrum**. All the wavelengths / colours with $n = 1$ are called the first-order spectrum, all the wavelengths / colours with $n = 2$ are called the second-order spectrum and so on.

Line spectra

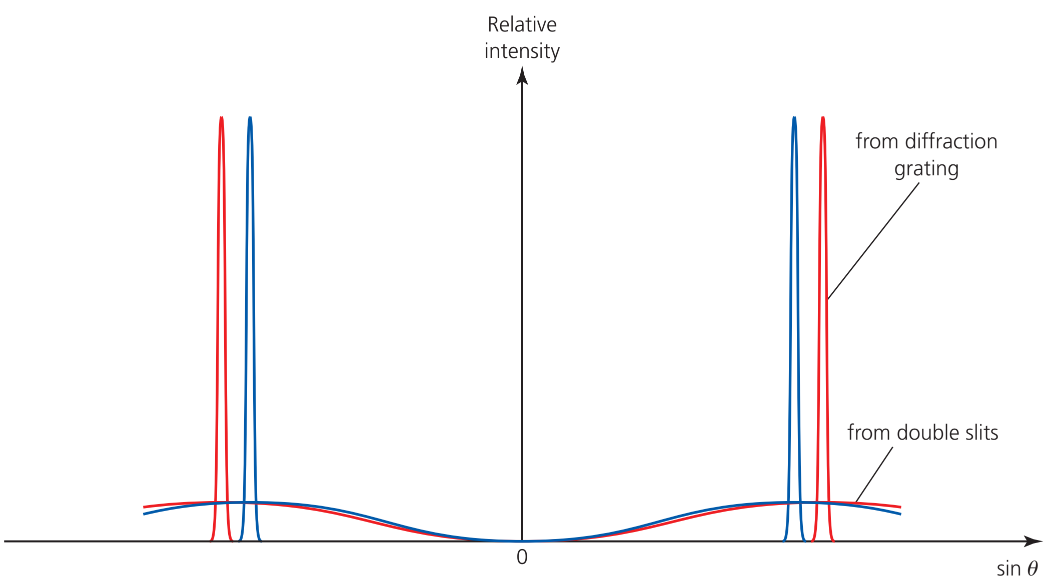
If atoms of a particular element, or some compounds, are given enough energy they will emit light in the form of a **line spectrum** (see Figure C3.62), not a continuous white light spectrum. This is discussed in much more detail in Topic E.1, and a full understanding is not expected here.



■ **Figure C3.62** Line spectra of various elements

Diffraction gratings offer an excellent way of producing and measuring line spectra. This is shown in Figure C3.63, which compares the ability of double slits and a diffraction grating to produce separate lines. The grating produces a much greater resolution.

Diffraction grating have very large numbers of lines very close together. This means that they are excellent at producing intense line spectra with high resolution.



■ **Figure C3.63** High resolution produced by diffraction gratings

Modulation by single-slit diffraction

SYLLABUS CONTENT

▶ The single-slit pattern modulates the double-slit interference pattern.

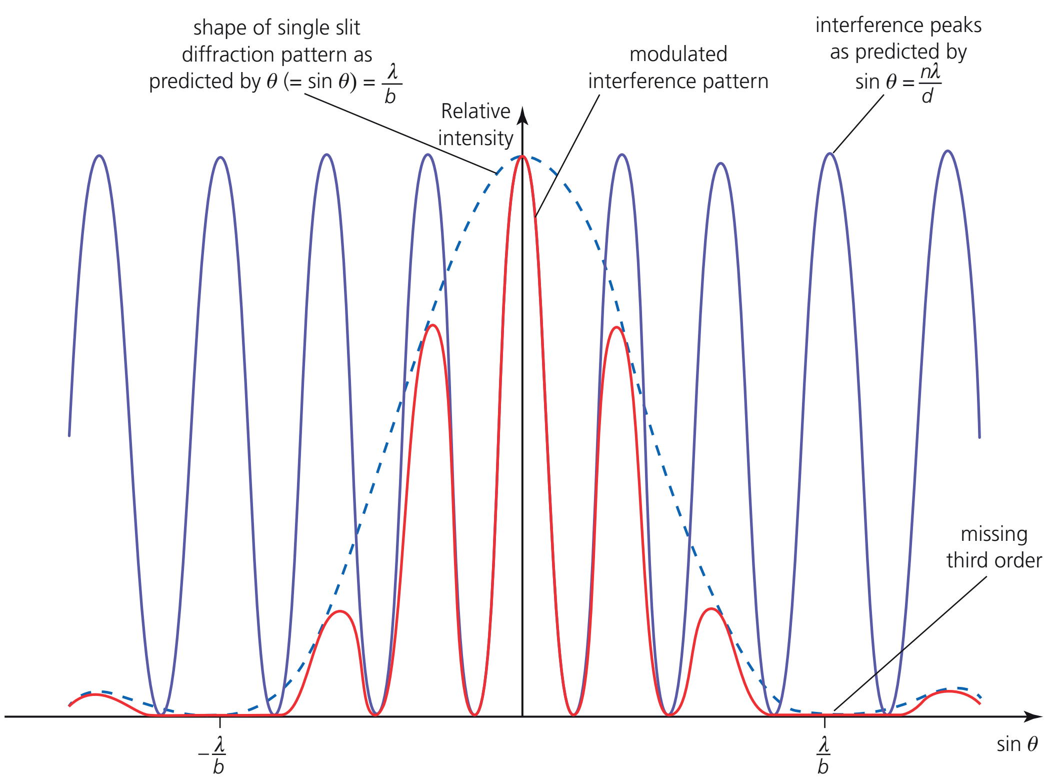
The explanation of the interference effects produced between light waves coming from multiple slits (including diffraction gratings) has so far ignored one very important factor: the effect of the single-slit diffraction pattern produced by the interference of secondary wavefronts within each and every slit. (We will assume that all slits are identical.)

The conclusions from a detailed superposition analysis of all wavefronts can be expressed in the following description:

◆ **Modulation** Changing the amplitude (or frequency) of a wave according to variations in a secondary effect.

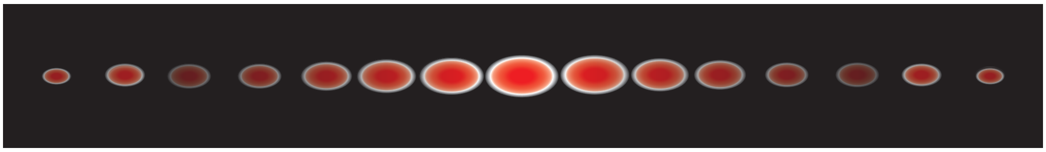
The intensity of the interference peaks produced by double slits, or multiple slits, is **modulated** by the shape of the single-slit diffraction produced by each individual slit.

This is explained by Figure C3.64. Note that *d* must always be larger than *b*, so that the spacing of the interference pattern must always be smaller than the spacing of the diffraction pattern from each slit. The shape and spacing of the single-slit diffraction pattern act as a guiding 'envelope' for the size of the interference peaks.



■ **Figure C3.64** How single-slit diffraction modulates multiple slit interference

This 'modulation' can result in some orders being suppressed or missing. For example, the third order in Figure C3.64 and the fourth and fifth orders in Figure C3.65.



■ **Figure C3.65** Missing orders because of the modulation effect of single-slit diffraction

HL ONLY

HL ONLY

42 Light of wavelength 460 nm is incident normally on a diffraction grating with 200 lines per millimetre. Calculate the angle to the normal of the third-order maximum.

43 A diffraction grating is used with monochromatic light of wavelength 6.3×10^{-7} m and a screen a which is perpendicular distance of 2.75 m away.
Calculate how many lines per millimetre are on the diffraction grating if it produces a second-order maximum 68 cm from the centre of the pattern.

44 Monochromatic light of wavelength 530 nm is incident normally on a grating with 750 lines per millimetre. The interference pattern is seen on a screen that is 1.82 m from the grating.
Calculate the distance between the first and second orders seen on the screen.

45 A prism can also be used to produce a spectrum. Outline why red light is refracted less than blue light in a prism, but red light is diffracted more than blue light by a diffraction grating.

46 When using white light, explain why red light in the second-order spectrum overlaps with blue light in the third-order spectrum.

47 Sketch a relative intensity against $\sin\theta$ graph for monochromatic light of wavelength 680 nm incident normally on a diffraction grating with 400 lines per mm.

48 A diffraction grating produced two first-order maxima for different wavelengths at angles of 7.46° and 7.59° to the normal through the grating. This angular separation was not enough to see the two lines separately.
What is the angular separation of the same lines in the second order?

49 Figure C3.66 shows a centimetre ruler placed next to an interference pattern seen on a screen which was placed 3.40 m away from double slits.
- **a** If the separation of adjacent slits was 0.64 mm, determine the wavelength of the light being used.
- **b** Estimate the width of the individual slits.
- **c** How would the pattern change if the light passed through 10 slits (of the same width and separation)?

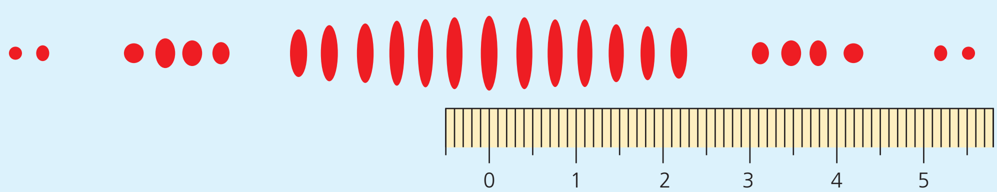


Figure C3.66 A centimetre ruler placed next to an interference pattern

C.4 Standing waves and resonance

Guiding questions

- What distinguishes standing waves from travelling waves?
- How does the form of standing waves depend on the boundary conditions?
- How can the application of force result in resonance within a system?

The nature of standing waves

SYLLABUS CONTENT

- The nature and formation of standing waves in terms of superposition of two identical waves travelling in opposite directions.
- Nodes and antinodes, relative amplitude and phase difference of points along a standing wave.

So far, the discussion of waves has been about travelling waves, which transfer energy progressively away from a source. Now we turn our attention to waves that remain in the same position.

Consider two travelling waves of the same shape, frequency, wavelength and amplitude moving in opposite directions, such as shown in Figure C4.1, which could represent transverse waves on a string or a rope.

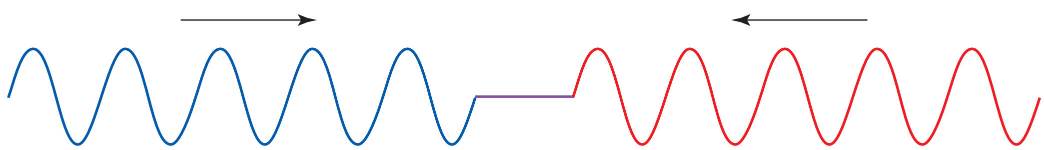


Figure C4.1 Two travelling waves of the same shape, frequency, wavelength and amplitude moving in opposite directions

As these waves pass through each other, they will combine to produce an oscillating wave pattern that does not change its position. Such patterns are called standing waves.

Standing waves usually occur in confined systems in which waves are reflected back upon each other repeatedly.

Typical examples of standing wave patterns (on strings) are shown in Figure C4.2. Note that a camera produces an image over a short period of time (not an instantaneous image) and that is why the fast-moving string appears blurred. This is also true when we view such a string with our eyes.

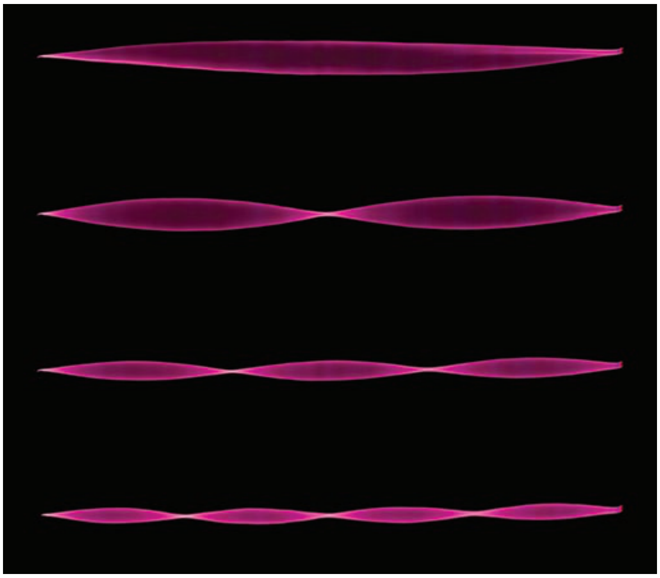


Figure C4.2 Standing waves on a stretched string

Simple standing wave patterns can be produced by shaking one end of a rope, or long stretched spring, at a suitable frequency, while someone holds the other end stationary. Patterns like those shown in Figure C4.2 require high frequencies (because a string is much less massive than a rope), but can be produced in a laboratory by vibrating a stretched string with a mechanical vibrator which is controlled by variable electrical oscillations from a signal generator. This apparatus can be used to investigate the places at which the string appears to be stationary and the frequencies at which they occur.

◆ **Standing wave** The kind of wave that can be formed by two similar travelling waves moving in opposite directions. The most important examples are formed when waves are reflected back upon themselves. The wave pattern does not move and the waves do not transfer energy.

◆ **Nodes** The positions in a standing wave where the amplitude is zero.

◆ **Antinodes** The positions in a standing wave where the amplitude is greatest.

Nodes and antinodes

A standing wave pattern remains in the same place.

Points where the displacement is always a minimum (often zero) are called **nodes**. Points where the amplitude is greatest are called **antinodes**.

Figure C4.3 represents the third wave from the photograph in Figure C4.2 diagrammatically. Note that the distance between alternate nodes (or antinodes) is one wavelength.

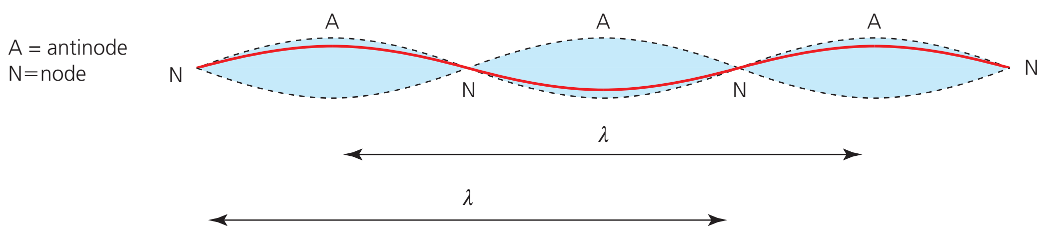


Figure C4.3 Nodes and antinodes in a standing wave

A system in which there is a standing mechanical wave has both kinetic energy and potential energy, but energy is not transferred *away* from the system in the form of a wave along the system. However, there will be energy dissipation within the system because of resistive forces, so that the amplitude of the standing wave will decrease, unless it is sustained by energy transferred from an external driving frequency (see later).

Nodes occur at places where the two waves are *always* exactly out of phase. At other places, the displacements will oscillate between zero and a maximum value which depends on the phase difference between the waves moving in opposite directions. At the antinodes the two waves are always exactly in phase. You should try using a computer simulation to illustrate this time-changing concept.

Between adjacent nodes all parts of the medium oscillate in phase with each other with the same frequency. Each position has a constant amplitude, but the amplitudes vary as shown in Figure C4.4.

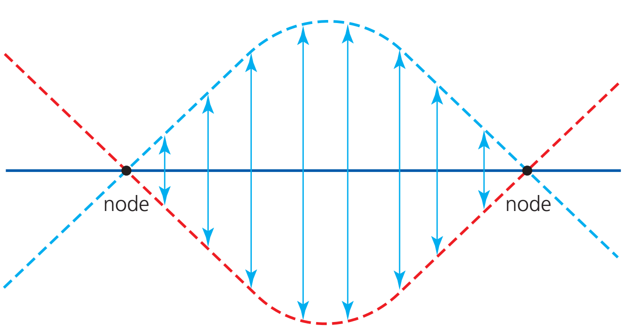


Figure C4.4 Variation of amplitude between nodes

Standing waves are possible with any kind of wave (transverse or longitudinal) moving in one, two or three dimensions. But, for simplicity, discussion will be confined to one-dimensional waves, such as transverse waves on a stretched string, or longitudinal sound waves in pipes.

Boundary conditions

◆ **Boundary conditions** The conditions at the ends of a standing wave system. These conditions affect whether there are nodes or antinodes at the ends.

Standing waves occur most commonly when waves are reflected repeatedly back from boundaries in a confined space. The frequencies of the standing waves will depend on the nature of the ends of the system. These are called the **boundary conditions**. For example, the ends of a string, or rope, may be fixed in one position or free to move; the ends of a pipe could be closed or open. When the ends are free to move, we would assume that standing wave has antinodes there. There will be nodes at fixed ends.

The boundary conditions of a standing wave system describe whether there are nodes or antinodes at the end of the system.

Comparing standing waves with travelling waves

Table C4.1 Comparison of standing waves and travelling waves

	Standing waves	Travelling waves
Wave pattern	standing (stationary)	travelling (progressive)
Energy transfer	no energy is transferred	energy is transferred through the medium
Amplitude (assuming no energy dissipation)	amplitude at any one place is constant but it varies with position between nodes: maximum amplitude at antinodes, zero amplitude at nodes	all oscillations have the same amplitude
Phase	all oscillations between adjacent nodes are in phase	oscillations one wavelength apart are in phase; oscillations between are not in phase
Frequency	all oscillations have the same frequency	all oscillations have the same frequency
Wavelength	twice the distance between adjacent nodes	shortest distance between points in phase

Standing wave patterns in strings

SYLLABUS CONTENT

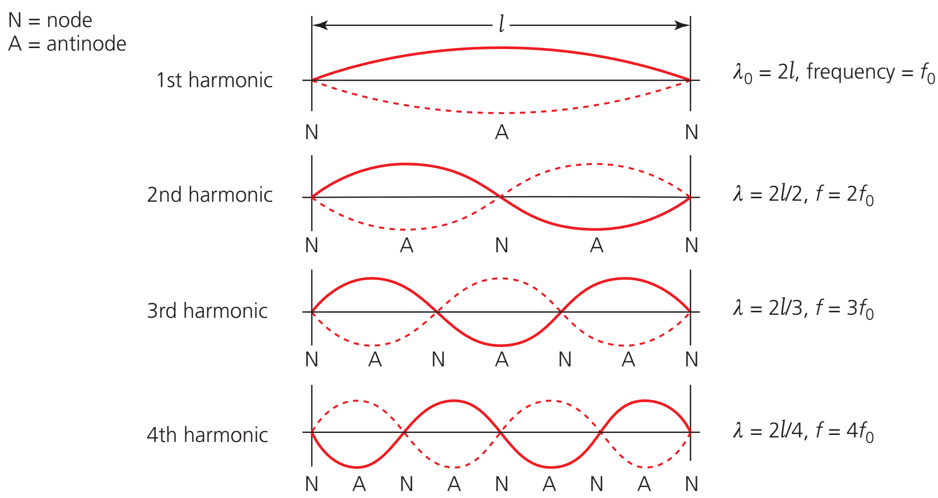
- Standing waves patterns in strings.

◆ **Harmonics** Different frequencies (modes) of standing wave vibrations of a given system. The frequencies are all multiples of the frequency of the **first harmonic**.

◆ **Modes of vibration** The different ways in which a standing wave can arise in a given system.

If a stretched string *fixed at both ends* is struck or plucked, it can only vibrate as a standing wave with nodes at both ends. The amplitude of the standing wave will usually decrease quickly as energy is dissipated.

The simplest way in which it can vibrate is shown at the top of Figure C4.5. This is known as the **first harmonic**. It is usually the most important mode of vibration, but a series of other harmonics (**modes of vibration**) is possible and can occur at the same time as the first harmonic. Some of these harmonics are shown in Figure C4.5.



■ **Figure C4.5** Modes of vibration of a stretched string fixed at both ends

The first harmonic is the standing wave with the lowest possible frequency, f_0 (greatest wavelength). Other harmonic are mathematical multiples of this frequency.

If a string is fixed at both ends, the wavelength, λ_0, of the first harmonic is $2l$, where l is the length of the string. The speed of the wave, v, along the string depends on the tension and the mass per unit length. If the wave speed is known, the frequency of the first harmonic, f_0, can be calculated using $v = f\lambda$ (from Topic C.2):

$$f_0 = \frac{v}{\lambda_0} = \frac{v}{2l}$$

The wavelengths of the harmonics are, starting with the first (longest), $2l$, $\frac{2l}{2}$, $\frac{2l}{3}$, $\frac{2l}{4}$ and so on. The corresponding frequencies, starting with the lowest, are f_0, $2f_0$, $3f_0$, $4f_0$ and so on.

The wavelength of the first harmonic can be found from the length of the system and the boundary conditions.

$v = f\lambda$ can then be used to connect frequency and wave speed.

WORKED EXAMPLE C4.1

A stretched rope with two fixed ends has a length of 0.98 m, and waves travel along it with a speed of $6.7\,\mathrm{m\,s^{-1}}$.

a Calculate:

  i the wavelength of the first harmonic

  ii the frequency of the first harmonic.

b Calculate:

  i the wavelength of the fourth harmonic

  ii the frequency of the fourth harmonic.

c Explain how your answers would change if the tension in the rope was increased.

Answer

a i $\lambda_0 = 2l = 2 \times 0.98 = 2.0\,\mathrm{m}$ (1.96 seen on calculator display)

ii $f_0 = \frac{v}{\lambda_0} = \frac{6.7}{1.96} = 3.4\,\mathrm{Hz}$

b i $\lambda = \frac{2l}{4} = \frac{\lambda_0}{4} = \frac{1.96}{4} = 0.49\,\mathrm{m}$

ii $f = 4f_0 = 4 \times 3.4 = 14\,\mathrm{Hz}$

c The wavelengths would remain the same, but the frequencies would increase because the wave would travel faster if the tension was increased.

Standing waves on strings and ropes are usually between fixed ends, but it is possible that one, or both, boundaries could be ‘free’. If the two ends are free (an unusual event), there will be antinodes at each end / boundary, so that the frequency of the first harmonic will be the same as for fixed boundaries, which has nodes at each end.

◆ **Oscilloscope** An instrument for displaying and measuring potential differences that change with time.

◆ **Waveform** Shape of a wave.

If there is a node at one end and an antinode at the other, the wavelength of the first harmonic will be greater and its frequency will be lower. An example of this situation would be a standing wave produced on a chain hanging vertically. (In that case, the wavelength would not be constant: see Question 4.)

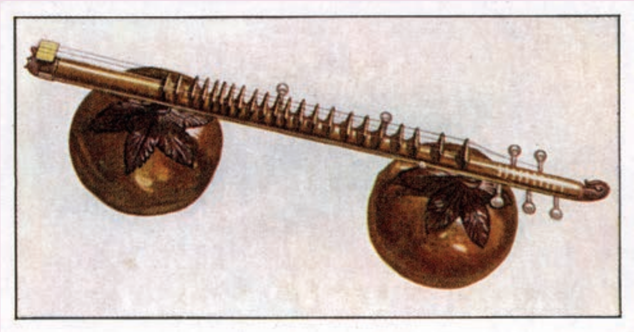
Inquiry 1: Exploring and designing

Designing

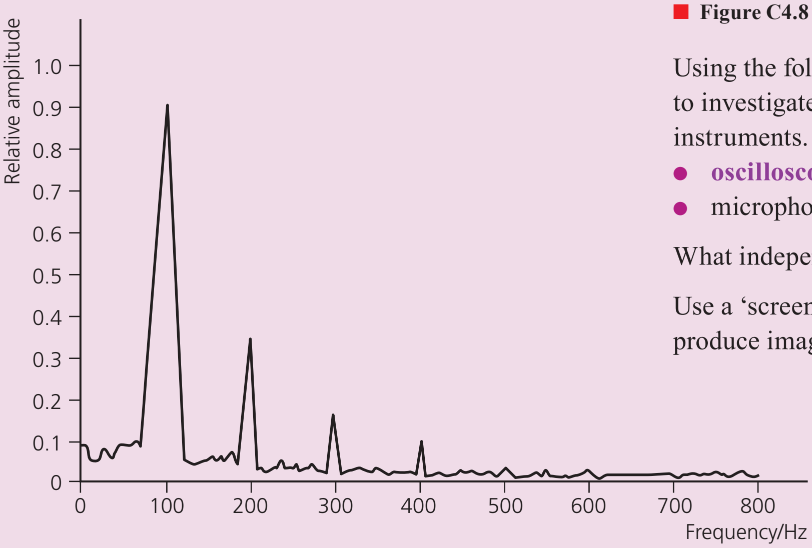
Musical instruments

An amazing variety of musical instruments have been used all around the world for thousands of years (see Figure C4.6). Most involve the creation of standing wave patterns (of different frequencies) on strings, wires, surfaces or in tubes of some kind. The vibrations disturb the air around them and thereby send out sound waves (musical notes).

When musical notes are played on stringed instruments, such as guitars, cellos and pianos, the strings vibrate mainly in their first-harmonic modes, but various other harmonics will also be present. This is one reason why each instrument has its own, unique sound. Figure C4.7 shows a range of frequencies that might be obtained from a guitar string vibrating with a first harmonic of 100 Hz.



■ **Figure C4.6** Vina, an Indian stringed instrument



■ **Figure C4.7** Frequency spectrum from a guitar string

The factors affecting the frequency of the first harmonic are the length of the string, the tension and the mass per unit length. For example, middle C has a frequency of 261.6 Hz. The standing transverse waves of the vibrating strings are used to make the rest of the musical instrument oscillate at the same frequency. When the vibrating surfaces strike the surrounding air, travelling longitudinal sound waves propagate away from the instrument to our ears.

■ **Figure C4.8** Creating standing waves on a cello

Using the following apparatus, design an experiment to investigate the sound waves produced by different instruments.

- **oscilloscope** or app on computer or phone
- microphone.

What independent variables will you change or control?

Use a 'screen capture' technique on the oscilloscope to produce images and compare the **waveforms** produced.

WORKED EXAMPLE C4.2

An experiment similar to that shown in Figure C4.9 can be used to determine the speed of a wave along a stretched string.

a If the length of the vibrating string was 79.4 cm and the first harmonic had a frequency of 140 Hz, calculate the wave speed.

b Determine the wavelength and frequency of the fifth harmonic.

c A student finds the following formula on the internet:

$$\text{wave speed} = \sqrt{\left(\text{tension} \times \frac{\text{length}}{\text{mass}}\right)}$$

If the mass of the vibrating string was 2.6 g, and the tension was 147 N, show that use of this formula confirms the answer to part **a**.

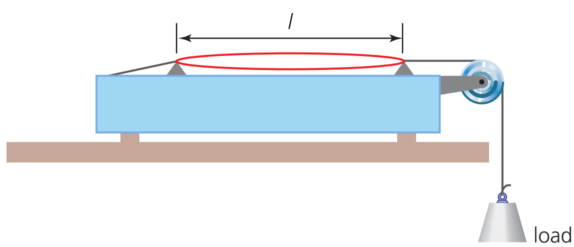


Figure C4.9 Experiment to determine wave speed

Answer

a $v = f_0\lambda_0 = 140 \times (2 \times 0.794) = 222\,\text{m s}^{-1}$

b $5f_0 = 5 \times 140 = 700\,\text{Hz}$

$$\frac{\lambda_0}{5} = \frac{1.588}{5} = 0.318\,\text{m}$$

c $v = \sqrt{\left(147 \times \frac{0.794}{0.0026}\right)}$

$= 2.1 \times 10^2\,\text{m s}^{-1}$

This is within 5% of the answer to part **a**, so the two answers are consistent, within experimental uncertainties.

1 A string on a violin has a length of 32.8 cm and produces a note of 262 Hz (middle C).

a Calculate the speed of the wave on the string.

b State any assumption that you made.

c Suggest what will happen to the speed of the wave on the string and the frequency of the note if sometime later the same string has lost some tension.

2 The velocity of a transverse wave on a string of length 28 cm is 240 m s^{-1}.

a Calculate the frequency of the second harmonic of a standing wave on this string when both ends are fixed.

b Determine the wavelength of the sound that this produces in the surrounding air. Assume the speed of sound in air is 340 m s^{-1}.

3 A teacher wishes to show his class standing waves on a thin string, similar to those seen in Figure C4.2. He uses a vibration generator set at a frequency of 384 Hz.

a Determine the length of string that is needed to produce the fourth harmonic if the speed of the wave is 285 m s^{-1}.

b Using the same apparatus, the teacher increases the length of the string to exactly 2.00 m. Predict the wavelength and frequency of the third harmonic in this new arrangement.

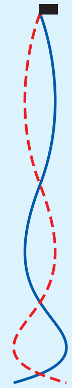
4 The top of a thin chain hanging vertically is shaken with increasing frequency until the standing wave pattern seen in Figure C4.10 is produced.

a Suggest why the wavelength on this system decreases towards the lower end of the chain.

b Describe the boundary conditions of this system.

c State which harmonic is shown in the picture.

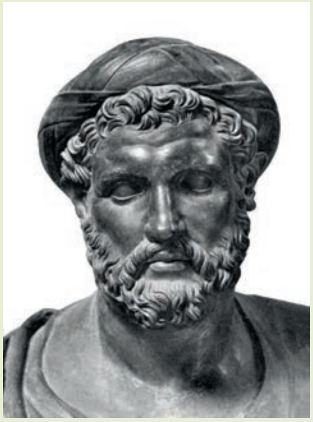
d Explain why it is not possible to produce the second harmonic on this system.



■ **Figure C4.10** A standing wave on a hanging chain.

Nature of science: Patterns and trends

Pythagoras is often credited with being the first to realize that there was a mathematical relationship between musical notes and the dimensions of the instrument that produced them. That was about 2500 years ago. More generally, this may have been one of the first occasions when mathematical / 'scientific' reasoning was used to describe features of everyday life.



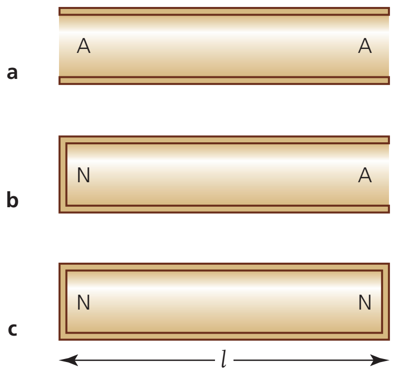
■ **Figure C4.11** Pythagoras

Standing wave patterns of air in pipes

SYLLABUS CONTENT

▶ Standing waves patterns in pipes.

Standing longitudinal waves of sound can be created easily in the air contained by a pipe / tube / column. The air may be set into motion by, for example. the simple action of blowing across the open end of the pipe. The sound produced by blowing across the top of an empty bottle is an everyday example. Musical wind instruments, like trumpets or flutes, use the same principle.



■ **Figure C4.12** Nodes and antinodes at the ends of open and closed pipes

As with strings, in order to understand which wavelengths can be produced, we need to consider the length of the system and the boundary conditions.

The first harmonic is the standing wave with the greatest possible wavelength, λ_0, and lowest possible frequency, f_0. Other harmonics are multiples of this frequency.

After determining possible wavelengths, the equation $v = f\lambda$ can then be used to predict harmonic frequencies if the wave speed is known.

Figure C4.12 show the three possible combinations of boundary conditions.

A pipe open at both ends must have antinodes, A, at the ends, and at least one node, N, in between. A pipe open at only one end has one antinode and one node as its boundary conditions.

A pipe closed at both ends (an unusual situation) must have nodes at the ends and at least one antinode in between.

Figure C4.13 graphically represents the first three harmonics (vibration modes) for a pipe open at both ends.

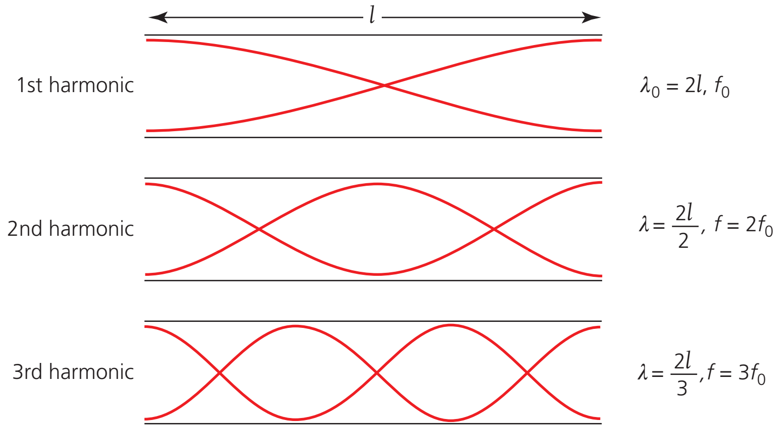


Figure C4.13 The first three harmonics in a pipe open at both ends

The wavelength of the first harmonic (twice the distance between adjacent nodes or antinodes) is $2l$.

Common mistake

Note that the representations of standing longitudinal waves seen in Figure C4.13 and Figure C4.15 may cause confusion: the curved lines in the diagrams are an indication of the maximum *sideways* displacement of vibrating air molecules. They should not be mistaken for transverse waves, like those on a string. Figure C4.14 shows how the second harmonic shown in Figure C4.13 is representing the movements of some molecules.

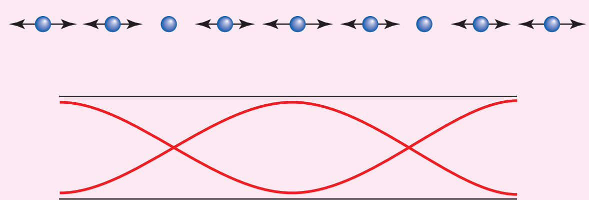


Figure C4.14 How the second harmonic can represent the movements of some molecules

For a pipe which is closed at one end, but open at the other, the possible standing waves are shown in Figure C4.15. The first harmonic has a longer wavelength and lower frequency than for a similar length pipe which is closed, or open, at both ends. Note that only odd-numbered harmonics are possible with these boundary conditions.

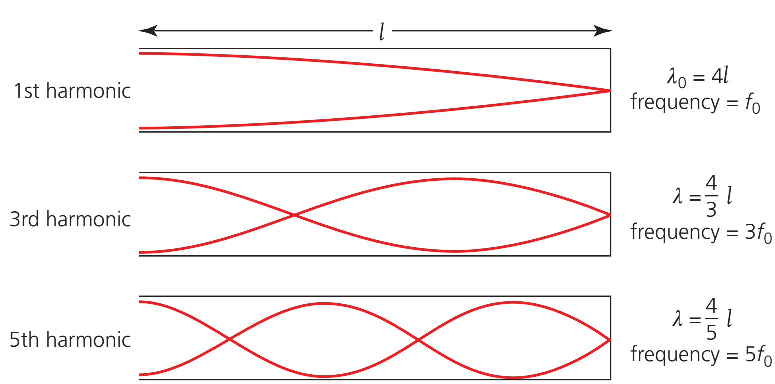


Figure C4.15 Harmonics in a pipe open at one end

One way of demonstrating standing sound waves in air is by using a small loudspeaker attached to the end of a long, clear plastic tube, which is closed at the other end. See Figure C4.16. Electrical signals of different frequencies can be applied to the loudspeaker which then sends sound waves of the same frequencies down the tube. When the waves reflect back off the end of the tube, a standing wave can be set up only *if* the frequency is equal to the frequency of the one of the possible harmonics. Some powder can be scattered all along the pipe and when the loudspeaker is turned on and the frequency carefully adjusted, the powder is seen to move into separate piles. This is because the powder tends to move from places where the vibrations of the air are large (antinodes) to the nodes, where there are no vibrations. The tube may be considered as closed at both ends.

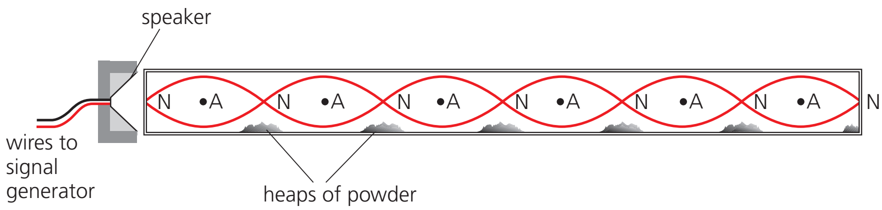


Figure C4.16 Demonstrating a standing wave with a loudspeaker

WORKED EXAMPLE C4.3

Consider Figure C4.16.

a State which harmonic is shown in Figure C4.16.

b Calculate the wavelength of this standing wave if the length of the tube is 73.5 cm.

c Determine the speed of the sound wave in air if the frequency used by the loudspeaker was 1410 Hz.

d Calculate the theoretical frequencies of the first two observable harmonics if the tube was left open at the end on the right-hand side.

Answer

a sixth

b 73.5/3 = 24.5 cm

c $v = f\lambda = 1410 \times 0.245 = 345\,\mathrm{m\,s^{-1}}$ (345.45 seen on calculator display)

d See Figure C4.15. The wavelength of the first harmonic will be 4 × length of tube = 2.94 m.

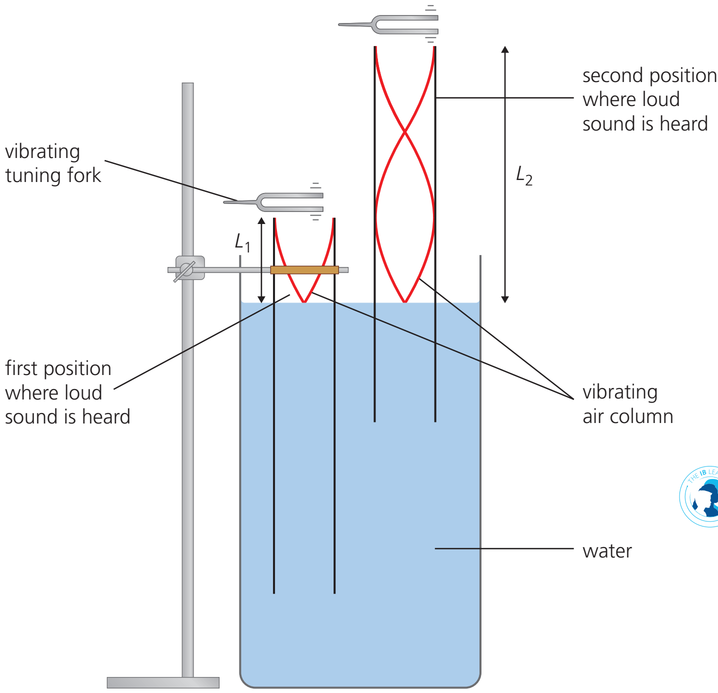
$$f_0 = \frac{v}{\lambda} = \frac{345.45}{2.94} = 1.18 \times 10^2\,\mathrm{Hz} \text{ (117.5 seen on calculator display)}$$

The second harmonic does not occur.

Frequency of third harmonic = $3f_0 = 3 \times 117.5 = 353$ Hz

Figure C4.17 shows another way of investigating standing waves in pipes. The length of the pipe is easily changed by raising or lowering it in the tall container of water. The pipe is always open at one end and closed at the other. Air in the pipe is disturbed by the pure, single frequency emitted from a **tuning fork** (see Figure C4.20) placed just above the open end of the tube, and the length of the pipe is adjusted until the sound of the standing wave becomes audible.

◆ **Tuning fork** Device designed to vibrate at only one precise frequency.



■ **Figure C4.17** Demonstrating standing waves with a tuning fork

The length of the tube seen on the left of the Figure C4.17 has been adjusted to the shortest length for which a standing wave can be heard. This will be the first harmonic for a tube of that length. The position on the right corresponds to the third harmonic for a tube of the new length. (The speed, wavelength and frequency of the wave are unchanged.) There is no second harmonic possible with these boundary conditions (open at one end, closed at the other). Higher harmonics may be heard if the tube and container are long enough.

ATL C4A: Social skills

Working collaboratively

Suggest how physics and music students could work together to produce a presentation called 'The Science of Music' for the rest of their year group.

Consider how you could share your presentations online for future students.

For these questions, where necessary, assume that the speed of sound in air is 342 m s^{-1}.

5 **a** If the tuning fork in Figure C4.17 had a frequency of 384 Hz, show that the length of the pipe needed for the first harmonic to be heard is about 20 cm.

b How far will the pipe need to be raised until the next harmonic is heard?

6 Sketch the first three harmonics for air in a pipe which is closed at both ends.

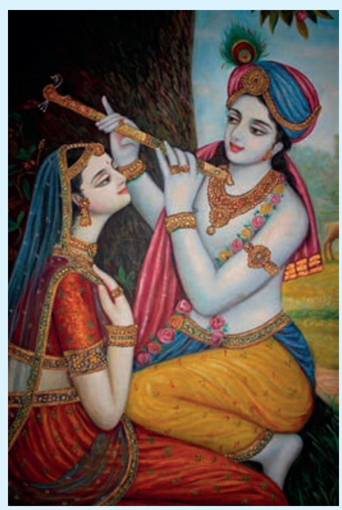
7 A pipe has a length of 1.32 m and is closed at both ends. Determine the wavelength and frequency of its third harmonic.

8 The flute is the oldest of musical instruments. See Figure C4.18. There are a very large number of designs. It can be considered as a pipe which is open at both ends. Sound is produced when air is blown across an opening near the end of the pipe.

a Determine what length (cm) of pipe will produce a first harmonic of frequency 493 Hz.

b Explain the purpose of the holes along the length of the flute.

c A clarinet is a similar type of musical instrument to a flute, but the pipe is closed at one end. Compare the length of the pipes of a flute and clarinet that produce musical notes of the same frequency.



■ **Figure C4.18** Radha listening to Krishna's flute

9 Apparatus similar to that shown in Figure C4.17 was used to investigate the variation of the speed of sound with air temperature. At a temperature of 30 °C, adjacent nodes were found to be separated by a distance of 23.3 cm when using a frequency of 750 Hz.

a Determine the speed of sound at this temperature.

b Give a molecular explanation of why this speed is greater than the 342 m s^{-1} used in other questions.

10 Water being poured into a bottle may produce many sounds, but there will usually be a noticeable increase in frequency of the sound as the bottle fills up. Explain this effect.

TOK

The natural sciences

- What is the role of imagination and intuition in the creation of hypotheses in the natural sciences?

When electrons were discovered in 1897, they were believed to be tiny solid particles. Twenty-seven years later, it was proved that electrons have wave properties (see Theme E). This led on to the theory that electrons could only be confined within atoms in the form of standing waves.

Most people will struggle to visualize electrons as three-dimensional standing waves in unimaginably small atoms. But, after about one hundred years, this still remains the accepted theory. There is no good reason for us to believe that the atomic-scale particles must behave in ways that are the same as masses which we can see with our eyes.

Natural frequencies of vibration

SYLLABUS CONTENT

- The nature of natural frequency.
- The effects of light, critical and heavy damping on the system.

◆ **Frequency, natural** The frequency at which a system oscillates when it is disturbed and then left to oscillate on its own, without influence from outside.

◆ **Vibration** Mechanical oscillation (usually of relatively small amplitude).

When many objects are struck briefly by an external force, they vibrate 'freely', or 'naturally' (although, for most objects, the vibrations may be insignificant and/or very short-lasting because energy is quickly dissipated into the surroundings). Such vibrations often disturb the air around them and send longitudinal waves into the environment, which may be heard as sound, if they have a suitable frequency.

When an object is disturbed and then left to oscillate without further interference, it is said to oscillate at its **natural frequency** (or frequencies).

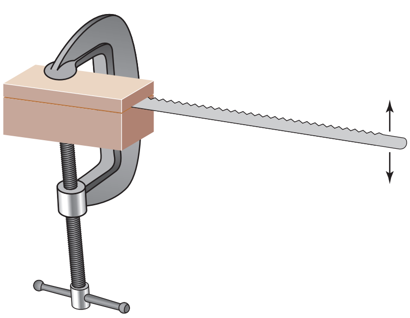


Figure C4.19 Vibrating hacksaw blade

Relatively small amplitude *mechanical* oscillations are commonly called **vibrations**.

The simplest examples of natural frequencies are those of a simple pendulum and mass on a spring, as discussed in Topic C.1. A further example is a clamped ruler, or hacksaw blade, as shown in Figure C4.19.

Vibrating objects will oscillate at a natural frequency(s) which depends on their dimensions and masses.

Tool 3: Mathematics

Linearize a graph

A student has read that the square of the time period of a hacksaw blade oscillator is proportional to the cube of the vibrating length.

What graph, involving frequency, should she draw in order to see if the relationship is correct?

An object made of only one material in a simple shape, such as a tuning fork (see Figure C4.20), may produce a single, 'pure', natural frequency. But most objects will have more complicated structures and a range of natural frequencies, although one frequency may dominate.

The *two-dimensional* standing wave patterns of a horizontal metal plate can be observed by placing small grains, such as fine sand or salt, onto a surface that is disturbed into vibration at the plate's natural frequencies. See Figure C4.21. (Oscillations can be easily maintained using a mechanical oscillator, driven by a signal generator, to vibrate the plate.)

■ **Figure C4.20** This tuning fork produces a frequency of exactly 440 Hz (a musical A)

■ **Figure C4.21** Demonstrating the oscillations of a metal plate

Damping

The motions of all macroscopic objects have resistive forces of one kind or another acting against them. Resistive forces will always act in the opposite direction to the instantaneous motion of an oscillator, and result in a reduction of its speed and kinetic energy.

Therefore, as with all other mechanical systems, useful energy is transferred from an oscillator into the surroundings (dissipated) in the form of thermal energy and maybe some sound. Consequently, an oscillator will move at slower and slower speeds, and its successive amplitudes will decrease in size. This effect is called damping.

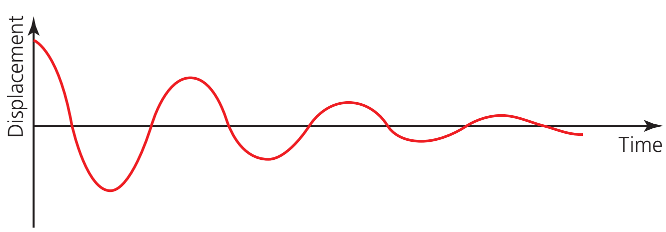
Top tip!

The magnitude of each successive peak of the graph shown in Figure C4.22 can be determined by multiplying the previous value by the same fraction. For example: $1, \frac{1}{2}, \frac{1}{4}, \frac{1}{8} \ldots$

(This is known as an exponential series, or a geometric sequence.)

Damping is the dissipation of energy from an oscillator due to resistive forces.

It is common for the frequency (and time period) of a vibration to remain approximately constant during damping, as shown by the graph in Figure C4.22. This is because, although the displacements are reduced, the speeds and accelerations also decrease.

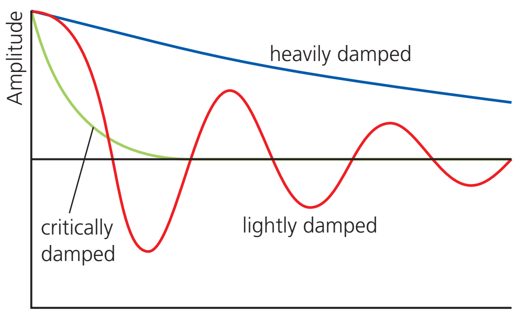


■ **Figure C4.22** Decreasing amplitude (at constant frequency) of a damped oscillation

Tool 3: Mathematics

Propagate uncertainties in processed data

Successive amplitudes (in cm) of the **free vibration** of a hacksaw blade were measured to be 2.7, 2,4, 2.1, 1.8, 1.6, 1.4, 1.3, 1.1. The measurements were made to the nearest millimetre. Considering uncertainties in measurement, is it possible that this data fits an exponential pattern?



■ **Figure C4.23** Light damping, heavy damping and critical damping

Damping can be investigated experimentally using simple apparatus like that shown in Figure C4.19, but with horizontal cards of different areas taped to the blade in order to increase air resistance.

The amount (degree) of damping in oscillating systems can be very different, as shown in Figure C4.23.

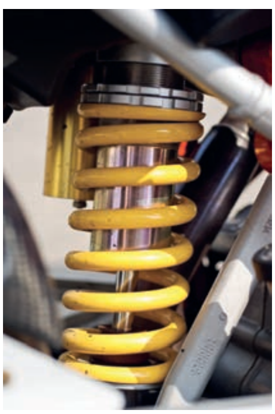
Some oscillations are *heavily damped* because of considerable frictional forces. In effect no oscillations occur because resistive forces are such that the object takes a long time (compared to its natural time period) to return to its equilibrium position.

Conversely, an oscillator may be *lightly damped*, so that it continues to oscillate, taking a relatively long time to dissipate its energy. A pendulum and a mass oscillating on a spring are good examples of lightly damped systems.

If an oscillation is opposed by resistive forces, such that it settles relatively quickly (compared to its natural time period) back into its equilibrium position, without ever passing through it, the process is described as **critical damping**. A car's suspension (see Figure C4.24) is an example of this kind of damping.

◆ **Vibration (free)** Vibration without any external influence.

◆ **Damping (critical)** When an oscillating system returns relatively quickly to its equilibrium position without oscillating.



■ **Figure C4.24** A car's shock absorber

11 Describe how the natural frequency of
- **a** a simple pendulum,
- **b** a mass hanging vertically from a spring

can be increased.

12 A student was investigating the vibrations of a hacksaw blade, as shown in Figure C4.19. She displaced the end of the blade and then left it to vibrate freely, but she found that the vibrations were too quick for her to observe easily.
- **a** Suggest how she could decrease the frequency of the vibrations.
- **b** The blade then vibrated with a frequency of 1.0 Hz, while its amplitude was (almost) constant, at 0.50 cm, for several seconds.
  Sketch a displacement–time graph for the first two seconds of its motion.
- **c** The vibrations were then damped.
  Add a second line to your sketch for part **b**, to represent the damped oscillations.
- **d** Suggest how the student could have damped the vibrations.

13 Outline why the fine sand shown in Figure C4.21 moves into places which demonstrate the standing wave pattern on the plate.

14 Figure C4.25 shows an automatic door closer.
- **a** Describe its intended purpose.
- **b** What type of damping does it exhibit?
- **c** Give an example of where it might be used.

■ **Figure C4.25** Automatic door closer

◆ **Vibration (forced)** Vibration affected by external periodic forces.

Forced vibrations

A **forced vibration** (oscillation) occurs when an external oscillating (periodic) force acts on a system. This may tend to make it oscillate at a frequency which is different from its natural frequency.

■ **Figure C4.26** How can we increase the amplitude of a swing?

We are surrounded by a range of oscillations. It is important to consider how these oscillations affect other things around them. In other words, what happens when an external oscillating force is continuously applied to a separate system?

To understand this, it is helpful to consider a very simple example: what happens when we keep pushing a child on a swing (see Figure C4.26)?

In this case it is fairly obvious: it depends on when we push the swing and in which direction. If we want bigger swings (increasing amplitudes), then we should push once every oscillation in the direction in which the child is moving at that moment. In more scientific terms, we would say that we need to apply an external force that has the same frequency as, and is in phase with, the natural frequency of the swing.

The most important examples of forced oscillations are those in which the frequency of the external force (**driving frequency**) is the same as the natural frequency. The child on the swing described above is a good example of this. When a regular periodic stimulus to a system results in an increasing amplitude the effect is called resonance.

◆ **Frequency, driving** The frequency of an oscillating force (periodic stimulus) acting on a system from outside. Sometimes called forcing frequency.

◆ **Resonance** The increase in amplitude that occurs when an oscillating system is acted on by an external periodic force that has the same frequency as the natural frequency of the system. The driving force must be in phase with the natural oscillations of the system.

◆ **Frequency–response graph** Graph used to show how the amplitude of a system's oscillations responds to different driving frequencies.

◆ **Resonant frequency** The frequency at which resonance occurs.

Resonance

SYLLABUS CONTENT

- The nature of resonance including the amplitude of oscillations based on driver frequency.
- The effect of damping on the maximum amplitude and resonant frequency of oscillation.

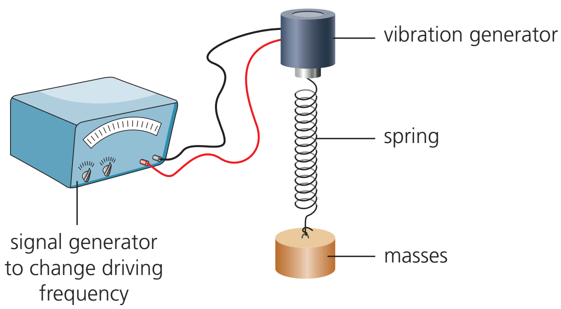
Resonance is the name given to the increase in amplitude and energy of an oscillation that occurs when a periodic external driving force has the same frequency as the natural frequency of a system.

The oscillations of the driving force must be in phase with the natural oscillations of the system.

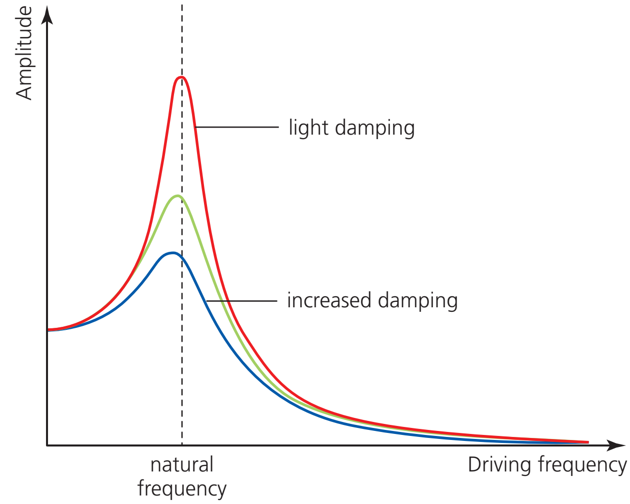
Simple quantitative laboratory experiments into the effects of resonance can be difficult to perform, but they can produce interesting results that show how varying the driving frequency affects the amplitude of an oscillating system. When the force is first applied, the oscillations may be erratic, but the system will settle into a regular pattern of movement with a measurable maximum amplitude.

Figure C4.27 shows a possible arrangement. The resonant frequency of the mass–spring system can be changed by using different springs and/or different masses. The driving frequency is provided by the vibration generator, which can be driven using different frequencies from the signal generator.

A typical **frequency–response graph** drawn from the results of an experiment like that shown in Figure C4.28 rises to a maximum amplitude at the **resonant frequency**. This occurs when the rate of energy dissipation (damping) has risen to a level that is equal to the power supplied from the source of the driving frequency.



■ **Figure C4.27** Investigating the resonance of a mass on a spring



■ **Figure C4.28** Typical frequency–response curves with different degrees of damping

The resonant frequency is at, or very close to, the natural frequency, but the sharpness and height of the peak also depend on the amount of damping in the system. The greater the damping, the greater the dissipation of energy and, therefore, the smaller the amplitude. The value of the resonant frequency reduces slightly with greater damping.

If there is a powerful input, or very little damping, amplitudes of vibration can become large and this may have destructive, or useful, consequences.

The energy of an oscillation is proportional to its amplitude *squared*.

There may be smaller resonance peaks at values of the driving frequency which are equal to the natural / resonant frequency divided by 2, 3, 4 and so on (not shown in the diagram).

LINKING QUESTIONS

- How does the amplitude of vibration at resonance depend on the dissipation of energy in the driven system?
- How can resonance be explained in terms of conservation of energy?

These questions link to understandings in Topic A.3.

Inquiry 2: Collecting and processing data

Processing data

Table C4.2 shows the results obtained in an experiment similar to that seen in Figure C4.27. The uncertainties in this experiment were significant.

Draw a graph of these results, including uncertainty bars, with a curve of best fit to determine a value for the resonant frequency. You should not assume that the largest recorded measurement is the peak of the graph.

■ Table C4.2 Results of resonance experiment

Applied frequency / Hz ± 2 Hz	**Maximum amplitude / cm ± 0.5 cm**
4	1.7
6	2.1
8	2.7
10	4.9
12	5.4
14	3.3
16	2.5
18	1.9

Examples of resonance

Some examples of resonance are useful but many more are unwanted, and we usually try to reduce their destructive effects. Avoiding resonance in all types of structure is a major concern for engineers and it provides an interesting combination of physics theory and practical engineering.

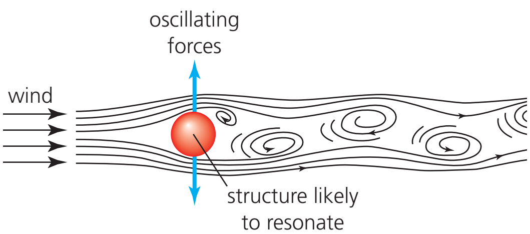
Unwanted resonance

■ **Figure C4.29** Resonance may be one reason why some buildings collapse in an earthquake

Parts of almost all machinery (and their immediate surroundings) may vibrate destructively when their motors are operating at certain frequencies. For example, a washing machine may vibrate strongly when the spinner is running at a certain frequency, and parts of vehicles can vibrate when the engine reaches a certain frequency, or they travel at certain speeds.

Earthquakes may well affect some buildings more than others. The buildings that are most damaged are often those that have natural frequencies closest to the frequencies of the waves produced by the earthquake (Figure C4.29).

Strong but steady winds, or currents, can also cause dangerous resonance in structures such as bridges and towers. This is often due to the effect of eddies and vortices as the wind or water flows around the structures. See Figure C4.30.



■ **Figure C4.30** A steady wind can cause oscillating forces because of the alternate way in which vortices can be formed

If you have ever crossed a small suspension bridge for walkers (such as the one in Figure C4.31), you will probably know how easy it is to set it vibrating with increasing amplitude by shaking it or stamping your feet at a certain frequency. This is because it would be too difficult or expensive to build such a simple bridge with a natural frequency that is very different from a frequency that people can easily reproduce, or to use a design that incorporated damping features.

■ **Figure C4.31** Walker on a suspension bridge in Nepal

The resonance of bridges has been well understood for many years and the flexibility of suspension bridges makes them particularly vulnerable. The famous collapse of the newly built Tacoma Narrows Bridge in the USA in 1940 is widely given as a simple example of resonance caused by the wind, although this is only part of a much more complex explanation. Videos of the collapse are easily found on the internet. In June 2000, the Millennium Bridge across the River Thames in London had to be closed soon after its opening because of excessive lateral (sideways) oscillations due to resonance (Figure C4.32).

In this case positive feedback was important. The slow oscillations of the bridge made people sway with the same frequency, and their motion simply increased the periodic forces on the bridge that were causing resonance. The problem was solved by adding energy-dissipating dampers, but it was about 18 months before the bridge could reopen.

■ **Figure C4.32** The Millennium Bridge in London was affected by resonance

To reduce the risk of damage from resonance engineers can:

- alter the shape of the structure to change the flow of the air or water past it
- change the design so that the natural frequencies are not the same as any possible driving frequencies – this will involve changing the stiffness and mass of the relevant parts of the structure
- ensure that there is enough damping in the structure and that it is not too rigid, so that energy can be dissipated.

Useful resonance

- The molecules of certain gases in the atmosphere oscillate at the same frequency as thermal radiation emitted from the Earth. These gases absorb energy because of resonance; this results in the planet being warmer than it would be without the gases in the atmosphere. This is known as the greenhouse effect, as discussed in Topic B.2.
- Microwave ovens use electromagnetic radiation in the microwave region. The microwave wavelength equals a vibrational frequency of water molecules, so that the molecules absorb the radiation.
- Your legs can be thought of as pendulums with their own natural frequency. If you walk with your legs moving at that frequency, energy will be transferred more efficiently and it will be less tiring (we tend to do this without thinking about it).
- Quartz crystals can be made to resonate using electronics – the resulting oscillations are useful in driving accurate timing devices such as watches and computers.
- The sound from musical instruments can be amplified if the vibrations are passed on to a supporting structure that can resonate at the same frequency. An obvious example would be the strings on a guitar causing resonance in the box on which they are mounted. Because the box has a much larger surface area it produces a much louder sound than the string alone.
- Magnetic resonance imaging (MRI) is a widely used technique for obtaining images of features inside the human body. Electromagnetic waves of the right frequency (radio waves) are used to change the spin of protons (hydrogen nuclei) in water molecules in the patient's body.

LINKING QUESTION

- How can the idea of resonance of gas molecules be used to model the greenhouse effect? (NOS)

This question links to understandings in Topic B.2.

15 Consider the oscillations of a mass on a spring, as shown in Figure C1.5.

- **a** Use the formula from Topic C.1 to determine the natural frequency of a 740 g mass oscillating vertically on a spring which has a spring constant of $17.2\,\mathrm{N\,m^{-1}}$.
- **b** Sketch a frequency–response curve to show how the amplitude of oscillation could change when the applied frequency from the vibration generator increases from 0.5 Hz to 2.0 Hz.
- **c** Discuss how the results of the experiment will change if the mass was placed in water (in a beaker). Illustrate your answer by adding a second curve to your sketch for part **b** of this question.

16 **a** Estimate the natural frequency of your leg when it swings freely like a pendulum.

- **b** If, when walking, your leg moves with the same frequency, predict your approximate speed, in $\mathrm{km\,h^{-1}}$.

17 **a** Use the internet to find out a typical frequency of waves generated by an earthquake.

- **b** Suggest ways in which an architect / civil engineer can ensure that their designs do not resonate dangerously at that frequency.

18 It is claimed that an opera singer can shatter a wine glass using sound resonance. Research the internet for any video evidence of this effect. Quote your conclusions and sources.

19 A wing mirror on a car resonates at multiples of its natural frequency of 20 Hz.

- **a** Sketch a graph to show the frequency response of the mirror as the rpm (revolutions per minute) of the car engine increase from 1000 to 4000.
- **b** Suggest how the vibrations of the mirror could be reduced.
- **c** Add a second curve to your graph to show the new frequency response.

20 Carbon dioxide gas in the Earth's atmosphere, CO_2, is an important cause of the greenhouse effect. In a simplified model, the molecule may be visualized as a simple harmonic oscillator, with the two oxygen atoms oscillating to and from a central carbon atom, as shown in Figure C4.33.

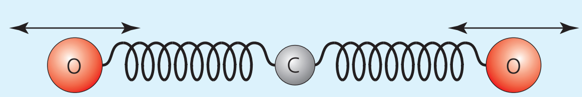


Figure C4.33 Oscillations of a carbon dioxide molecule

- **a** If each oxygen molecule has a mass of 2.7×10^{-27} kg, use the formula for the time period of a mass on a spring, with $k = 530\,\mathrm{N\,m^{-1}}$, to determine a value for a resonant frequency of the carbon dioxide molecule.
- **b** In what part of the electromagnetic spectrum are waves of this frequency to be found?

C.5 Doppler effect

Guiding questions

- How can the Doppler effect be explained both qualitatively and quantitatively?
- What are some practical applications of the Doppler effect?
- Why are there differences when applying the Doppler effect to different types of waves?

What is the Doppler effect?

SYLLABUS CONTENT

- The representation of the Doppler effect in terms of wavefront diagrams when either the source or the observer is moving.

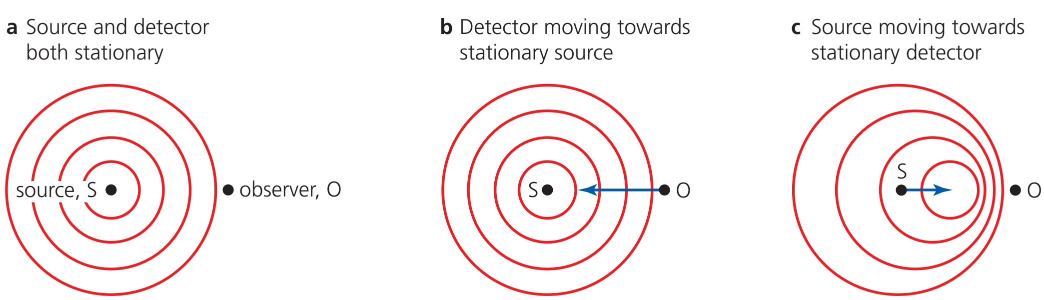
◆ **Doppler effect** When there is relative motion between a source of waves and an observer, the emitted frequency and the received frequency are not the same.

The **Doppler effect** is the name given to a phenomenon that is observed when there is relative motion between a source and an *observer* of waves. The term 'observer' is being used here to represent the person, or the device, receiving waves (of any type). The Austrian physicist Christian Doppler first identified in 1842 the effect that bears his name.

When there is relative motion between a source of waves and an observer, the waves received by the observer will have a different frequency (and wavelength) than the waves emitted by the source. This is called the Doppler effect.

The easiest way to explain the Doppler effect is by drawing wavefronts. Figure C5.1a shows the common situation in which a stationary source, S, emits waves that travel towards a stationary observer, O, with the same speed in all directions. Figure C5.1b shows an observer moving directly towards a stationary source and Figure C5.1c shows a source moving directly towards a stationary observer.

Similar diagrams can be drawn to represent the situations in which the source and detector are moving apart. The Doppler effect may be better understood by observing computer simulations which show *moving* sources or observers.



■ **Figure C5.1** Wavefront diagrams to demonstrate the Doppler effect

The observer in Figure C5.1b will meet more wavefronts in a given time than if it remained in the same place, so that the received frequency, f', is greater than the emitted frequency, f.

Since $\lambda = \frac{v}{f}$ (from Topic C.2) and the wave speed, v, is constant, the received wavelength will be less than the emitted wavelength.

In Figure C5.1c, the distance between the wavefronts (the wavelength, λ) between the source and the observer is reduced, which again means that the received frequency will be greater than the emitted frequency. ($f = \frac{v}{\lambda}$ and the wave speed, v, is constant.)

The most common everyday examples of the Doppler effect are with sound, but the effect is usually only noticeable if the sound is loud and the movement is fast, from a moving vehicle, for example.

Doppler effect for sound waves

SYLLABUS CONTENT

- The nature of the Doppler effect for sound waves.

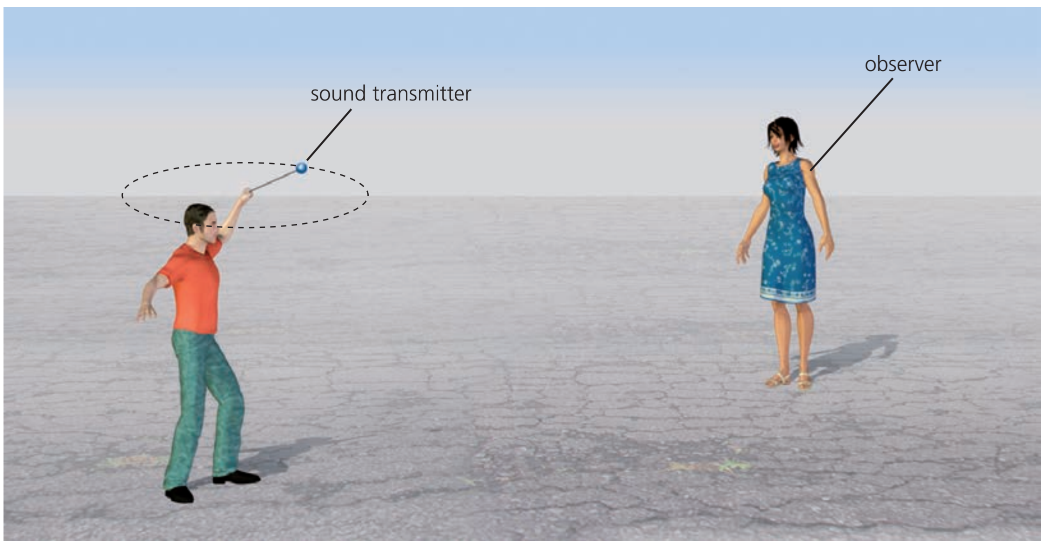


Figure C5.2 Demonstrating the Doppler effect with sound waves

Figure C5.2 shows a way in which the Doppler effect with sound can be demonstrated. A small source of sound (of a single frequency) is spun around in a circle. When the source is moving towards the observer a higher frequency is heard by the observer; when it is moving away, a lower frequency is heard.

The speed of sound through air depends only on the physical properties of the air. It does not vary with the motion of the source, or the observer. However, the speed with which sound *passes* a moving observer depends on their relative speeds. For example, a sound travelling away from a stationary source at $340\,m\,s^{-1}$, will pass an observer who is moving directly away from the source at $200\,m\,s^{-1}$, with a speed of $140\,m\,s^{-1}$.

The most common and easiest understood examples of the Doppler effect for sound include trains or cars which are moving quickly at a constant speed in an approximate straight line towards, or away from, a stationary observer.

Common mistake

Ambulances and police sirens are often given as examples of the Doppler effect, which they are. But be aware that they also emit sounds of varying frequency and loudness which should not be confused with the Doppler effect itself.

■ **Figure C5.3** These bats in Malaysia use the Doppler effect to navigate

Many types of bats use the Doppler effect. See Figure C5.3.

The Doppler effect for sound is covered in more mathematical detail later in this topic at the Higher Level.

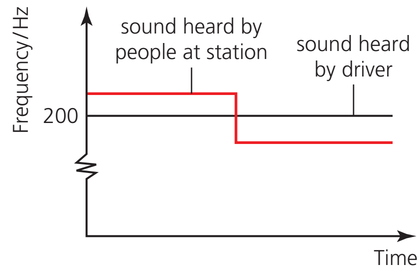
WORKED EXAMPLE C5.1

A train travelling at a constant speed approaches a station where it will not stop. The driver of the train sounds a warning horn, of frequency 200 Hz, as the train approaches and then passes through the station. Sketch graphs (on the same axes) to show how the frequency heard by:

a the driver varies with time

b the people at the station varies with time.

Answer

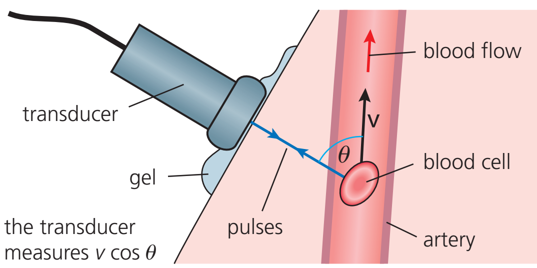


■ **Figure C5.4** There is a sudden frequency change as the train passes the people at the station

Top tip!

When the motion of the source is directly towards, or away from, the observer (or the other way around), there is *sudden* change of frequency when they pass. But the change of frequency is gradual if they do not pass close to each other.

The measurement of the rate of blood flow in an artery is an application of the Doppler effect for sound (ultrasound), which is shown in Figure C5.5.



■ **Figure C5.5** Measuring blood flow rate using the Doppler effect

Pulses of ultrasonic waves are sent into the body from the transducer and are reflected back from blood cells flowing in an artery. (A transducer is a general term used to describe any device which converts another form of energy into, or from, electrical energy.) The received waves have a different frequency because of the Doppler effect, and the measured change of frequency can be used to calculate the speed of the blood flowing in the artery. This information can be used by doctors to help to diagnose many medical problems. Because the waves usually cannot be directed along the line of blood flow, the calculated speed will be the component ($v\cos\theta$).

1 Draw diagrams similar to Figure C5.1b and c to represent the wavefronts when:

a an observer is moving directly away from a stationary source

b a source is moving directly away from a stationary observer.

2 A hospital patient had a Doppler ultrasound scan to check the blood flow in an artery, as shown in Figure C5.5. The transducer used a frequency of 6.0 MHz, and a healthy blood flow rate was expected to be about $10\,\text{cm}\,\text{s}^{-1}$.

a State whether the detected frequency will be higher, or lower, than 6.0 MHz. Explain your answer.

b i Explain why a calculation based only on the frequency change predicts that the blood speed in an artery is lower than its true value.

ii If the patient has a medical problem affecting blood flow rate, predict what effect this will have on the frequency measurements.

c Suggest what properties of ultrasound make it useful for this medical examination.

Tool 3: Mathematics

Sketch graphs, with labelled but unscaled axes, to qualitatively describe trends

A ‘sketch’ graph should be drawn neatly, using a ruler where appropriate. The axes should be clearly labelled, with zeros and negative values indicated (if applicable). Usually there is no requirement to add scales, or numerical values to the axes, but important features should be labelled.

3 Sketch a frequency–time graph to show how the sound heard by the observer in Figure C5.2 varies during one oscillation of the sound transmitter.

4 Figure C5.6 shows the shape of waves spreading from the front of a boat (its bow).

a Describe the overall shape of the wave fronts.

b Explain why bow waves similar to these can cause damage to the surroundings if the boat is close to shore.

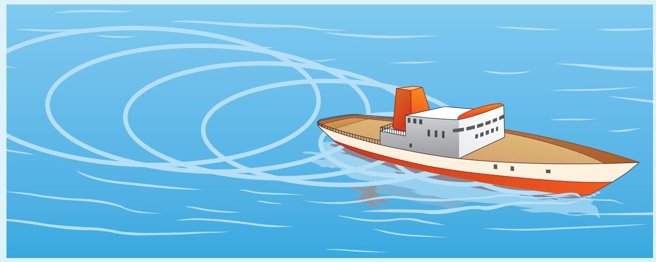


Figure C5.6 The shape of waves spreading from the front of a boat

Nature of science: Global impact of science

Shock waves: breaking the sound barrier

As an object, like an aircraft, flies faster and faster, the sound waves that it makes get closer and closer together in front of it. When an aircraft reaches the speed of sound, at about 1200 km h^{-1}, the waves superpose to create a ‘shock wave’. This is shown in Figure C5.7.

When an aircraft reaches the speed of sound it is said to be travelling at ‘Mach 1’ (named after the Austrian physicist, Ernst Mach). Faster speeds are described as ‘supersonic’ and twice the speed of sound is called Mach 2, and so on. As Figure C5.8 shows, the shock wave travels away from the side of the aircraft and can be heard on the ground as a ‘sonic boom’.

For many years some engineers doubted if the sound barrier could ever be broken. The first confirmed supersonic flight (with a pilot) was in 1947. Now it is common for military aircraft to travel faster than Mach 1. Concorde and Tupolev 144 were the only supersonic passenger aircraft in regular service, but their use has been discontinued. There are plans to introduce a new supersonic passenger aircraft before the year 2030.

Choose search terms to find out about the planned designs for future passenger / commercial supersonic aircraft.

Explore the reasons why previous supersonic airliners were discontinued. Were these reasons technological, scientific, economic, environmental, political? Explain your reasoning.

It is possible to use a whip to break the sound barrier. If the whip gets thinner towards its end, then the speed of a wave along it can increase until the tip is travelling faster than sound (in air). The sound it produces is often described as a whip ‘cracking’.

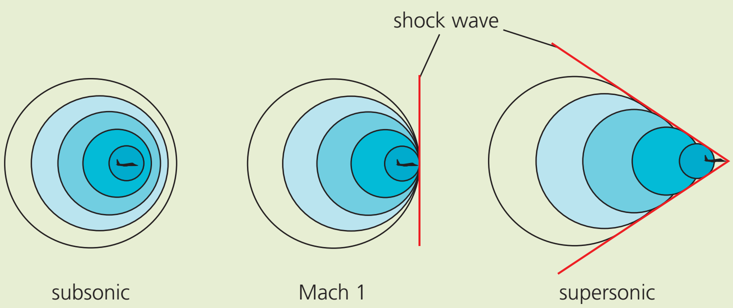


Figure C5.7 Creating a shock wave in air

Figure C5.8 Aircraft breaking the sound barrier

LINKING QUESTION

- What are the similarities and differences between light and sound?

This question links to understandings in Topic C.2.

Doppler effect for light and other electromagnetic waves

SYLLABUS CONTENT

- The nature of the Doppler effect for electromagnetic waves.
- The relative change in frequency, or wavelength, observed for a light wave due to the Doppler effect where the speed of light is much larger than the relative speed between the source and the observer, as given by: $\frac{\Delta f}{f} = \frac{\Delta\lambda}{\lambda} \approx \frac{v}{c}$.

Apart from mechanical waves like sound, the Doppler effect also occurs with electromagnetic waves, but the situation is more complicated because the speed of electromagnetic waves, as measured by any observer, is always the same: it is unaffected by the speed of the source, or the speed of the observer (this is a relativistic effect – see Topic A.5 – however, this theory is not of concern here).

The following equation for the change (shift) in frequency, Δf (= received frequency – emitted frequency), or shift in wavelength, $\Delta\lambda$, can be used if the relative speed between source and observer, v, is very much less than the speed of the electromagnetic waves, c ($v << c$). This is usually a valid assumption because of the very high value of c.

DB

$$\frac{\Delta f}{f} = \frac{\Delta\lambda}{\lambda} \approx \frac{v}{c}$$

Using the Doppler effect to determine speeds

Microwaves of known frequency, f, are easily produced and transmitted as a beam which can be directed at a moving object. A small fraction of the radiation will be arrive back at the source after reflecting off the object.

◆ **Radar** A system which uses microwaves to detect the distance, direction and speed of moving objects.

The change in frequency, Δf, due to the Doppler effect can be used to determine the speed of an object.

■ **Figure C5.9** Radar dish

Examples include:

- Speed 'guns' are used by police for checking the speed of moving vehicles.
- **Radar** is used for monitoring aircraft, or boat movements. (See below for an explanation of radar.)
- Radar is used for tracking the movement of storms.

In all these examples, the microwaves travel from the source to the object and then back to the source. For this reason, a factor of 2 should be added to the right-hand side of the equation above.

Radar (**RA**dio **D**etection **A**nd **R**anging) is a system used for determining the direction, distance and speed of an aircraft (or other object). Pulses of microwaves are sent from a rotating aerial (see Figure C5.9 for an example). After a very small fraction of a second, some microwaves which were reflected off the aircraft are received back at the aerial. The time delay can be used to determine the distance to the aircraft and the orientation of the aerial provides information about the direction to the aircraft.

The use of the Doppler effect with the radar also enables a direct calculation of aircraft speed and has the advantage of being able to ignore reflections from objects which are not moving or are only moving slowly.